Guidelines for Preserving New Forms of Scholarship

Jonathan Greenberg, NYU Libraries; Karen Hanson, Portico, ITHAKA; Deb Verhoff, NYU Libraries

September 2021, Corrected February 2022, September 2024

DOI: https://doi.org/10.33682/221c-b2xj

Background

Scholars are making extensive use of new digital technologies to express their research. Publishers, in turn, are working to support increasingly complex publications that are not easily represented in print. Examples include publications with embedded visualizations, multimedia, data, complex interactive features, maps, annotations, or that depend on third-party platforms or APIs, such as YouTube or Google Maps. These publications present formidable challenges for long-term preservation.

To study this challenge, a group of digital preservation institutions, libraries, and university presses worked together on an Andrew W. Mellon Foundation funded project, Enhancing Services to Preserve New Forms of Scholarship, led by New York University Libraries. Publishing organizations included NYU Press, Michigan Publishing, the University of Minnesota Press, UBC Press and Stanford University Press. Preservation service organizations included CLOCKSS, Portico and the libraries of the University of Michigan and NYU. Together, they examined a variety of enhanced ebooks and identified which features can be preserved at scale using tools currently available. Their findings, combined with the knowledge and research of experts in preservation, publishing, and copyright, resulted in this set of guidelines and best practices.

Intended Audience

These recommendations will guide publishers to create digital publications that are more likely to be preservable. They are meant to be shared with authors, editors, digital production staff, software developers and those who design and maintain publishing platforms. We hope that publishers and platforms will adapt these guidelines to create versions that take into account local workflows, technologies, and cultures. The guidelines were derived from research performed by professional preservation services who work at scale. They will also aid the wider publishing and preservation communities, from the individual content creator to those who steward digital collections and ensure long term access.

These guidelines are neither categorical nor prescriptive in nature. They represent a variety of approaches for improving preservability, out of which only a subset may apply to a particular publication. Publishers, authors, platform developers, and preservationists will sometimes need to weigh technological creativity against preservability, and some pathbreaking digital work may always resist preservation. Nonetheless, in order to make these decisions, those involved with digital publishing should understand the implications of complex digital features for digital preservation. As these guidelines demonstrate, many pitfalls are easy to avoid if planning begins early in the publication process.

Terminology

Publication resources: The digital materials that make up the publication. “Resource” does not necessarily refer to a specific file but to the content and form. Some resources may consist of multiple files or be expressed in different formats. The publication itself is a resource. If separate from the body of the publication, each of the supporting items are also resources - each figure graphic, video clip, piece of software, dataset, 3D visualization etc. The publisher can work with a preservation service to determine which publication resources are required to represent its core intellectual components, and which files are appropriate to represent each resource.

Core intellectual components: The aspects of a publication that are considered integral to the understanding of it. Rather than pertaining to specific digital resources or renditions of a publication, this is a more abstract sense of the facets of the work. For example: the linear text divided by section headings, the media placed at specific locations within the text, the additional digital resources supplied by the author with descriptive information tied to them. This kind of abstraction is helpful in preservation for communicating requirements and then designing a strategy that covers the important aspects of a work.

Keywords

The following keywords were applied to the guidelines to aid with navigation.

- EMBEDDED RESOURCES: Non-text resources such as images, audio, video, visualizations, etc. that appear in a publication.
- EXPORT PACKAGES: A package of content created to represent a publication for the purpose of transferring it to a preservation service.
- EPUB: An electronic publication file format and technical standard published by the International Digital Publishing Forum.
- PLANNING: Bringing a preservation mindset into pre-production and design phases.
- PUBLISHING PLATFORMS: Software used to manage and provide access to one or more publications.
- RIGHTS: Issues pertaining to copyright, terms of use/service, or licensing of publication resources.
- SOFTWARE AND DATA: In general, these refer to raw and executable materials offered in support of the key points of the text.
- THIRD-PARTY DEPENDENCIES: A component of a publication that is dependent on a service or platform that is outside of the control of the publisher.
- WEB-BASED PUBLICATIONS: Publications designed primarily to be presented in the form of a website rather than downloaded as a document (as with EPUB or PDF).

1. Communicate early in a new project with those who will be involved in preserving the content.

Keywords: PLANNING

This document offers a variety of general recommendations for creating publications that are more preservable. If working with specific preservation partners, however, their capabilities, standards used, and services offered may vary greatly. It is helpful to engage with these partners early to discuss new projects. This gives the preservation partners an opportunity to indicate whether they have local practices that differ from the suggestions here. For example, some preservation services may prefer that specific file formats or metadata standards are used. Discussing the project early can improve the preservation outcomes.

These other guidelines can facilitate conversations with preservation partners:

5. Establish formatting rules for common features
6. Keep preservation partners informed of changes
9. Define the version of record for your context
10. Define and document the core intellectual components of a work

2. If selecting a new publishing platform, assess technology options in light of preservation priorities.

Keywords: PLANNING, PUBLISHING PLATFORMS

Other guidelines in this publication lay out many aspects of platform technology that could be considered, but in brief, here are some indicators that a platform may have features that facilitate preservation: The platform utilizes appropriate standards relevant to the publishing community e.g. standardized metadata, exports to common formats, accessibility standards. The platform uses established technologies rather than being dependent on newer more experimental technologies that may not be well supported. The platform itself is well established and broadly adopted. There are existing workflows for preservation. The platform has a comprehensive export option that includes all raw materials, dependencies (e.g. fonts), descriptive metadata, and packaging metadata that describe how it all fits together. The export package supports, through completeness and use of standards, a complete migration to a new platform with equivalent features, rather than being closely tied to the current platform. In the absence of an export, the platform includes a predictable structure or API that could facilitate content discovery, enumeration, or harvesting from an external source. Finally, the platform does not have an over-abundance of built-in features that will not be used - these can add bulk and complexity to workflows including for preservation.

3. Use existing standards to guide design decisions while developing, enhancing, or implementing a publishing platform.

Keywords: PLANNING, PUBLISHING PLATFORMS

If you are developing a new publishing platform, or have control over how publishing platform features are designed or implemented, use existing standards to guide decisions. For example,

there are standards for bibliographic data (e.g. ONIX, Dublin Core), full-text data (e.g. TEI, EPUB), annotations (e.g. W3C's Web Annotation Data Model), persistent identifiers (e.g. DOIs, Handles, ARK IDs), citations (e.g. MLA, BibTeX), metrics (e.g. COUNTER), accessibility (e.g. W3C's Web Content Accessibility Guidelines) and more. Preservation workflows scale best when working with common standards.

4. Use platform features and plugins as intended. If you enhance existing features, be consistent and document how it is different from the default.

Keywords: PLANNING, PUBLISHING PLATFORMS

When using out-of-the-box software solutions for your publishing platform, export and preservation workflows are often designed around the built-in functionality of that software. For this reason, it is helpful to use platform features as intended. If the built-in functionality of the publishing software does not meet local requirements, avoid making undocumented, one-off changes to core code in order to get something working quickly. Instead, attempt to formalize and document any changes to the out-of-the-box software so that the new functionality is reusable in other publications and internally consistent within the platform. If the platform software has a formal process for applying enhancements (e.g. a plugin process), make use of this. Ensure any export processes are modified to align with the local changes and if working with a preservation partner, communicate any local changes to the software. The risk of not following a formal process may be loss of the new features during preservation, updates, or platform migration. An undocumented customization can disrupt the preservation of entire publications.

These other guidelines may be helpful when implementing new features:

3. Use existing standards when implementing features
5. Establish formatting rules for common features
6. Keep preservation partners informed of changes that affect the publications

5. Establish some basic formatting rules for common enhanced features.

Keywords: PLANNING, PUBLISHING PLATFORMS

Consistency is the key to a scalable preservation workflow, and so if a publisher or platform supports multimedia content or other enhanced features, establish basic rules early on and continue to express these features in a consistent way. Limit formats and arrangements as much as possible. For example, if one embedded video is an MP4 with no caption, another is a WEBM and has a caption in a box, and another still is a Vimeo video with a caption but no box, for some approaches these minor inconsistencies can cause problems when performing preservation activities at scale. These potential variations should be clearly defined and constrained.

These other guidelines may be helpful when implementing new features:

3. Use existing standards when implementing features
6. Keep preservation partners informed of changes that affect the publications

6. Keep preservation services informed about changes to the structure of publications and scope of file types that are accepted.

Keywords: PLANNING, PUBLISHING PLATFORMS

For platforms and publishers working with a preservation service, preservation workflows will be designed based on the sample publications provided. If these are not representative of the full range of functionality that the publishing platform supports, then the preservation workflow developed may miss things that the publisher wants to preserve. Keep a record of the scope of variations that might be found in a publication. As formatting rules for a publication change, expand, or new file formats or arrangements can be expected, inform your preservation institution so that they can adapt their workflows accordingly and avoid missing important features.

For more about changes that should be communicated to a preservation service:

4. Document any changes to the default functionality of a platform
5. Establish and document basic formatting rules
10. Define and document the core intellectual components of a work

7. Ensure that Terms of Use for publication platforms that have user-contributed content are appropriately transparent and cover preservation of this content.

Keywords: PLANNING, PUBLISHING PLATFORMS, RIGHTS

If a publication platform enables user contributed content and that content is managed by the platform e.g. annotations or comments, the platform's Terms of Use should clearly define the rights related to that content, especially if they may wish to preserve it or migrate it as part of the context of the publication. If a publication is likely to be archived with this context intact, the implementation of these features and their associated terms should factor in ethical consideration of how a user's information is displayed on the platform, and how they are informed and consent to the use of the content.

See also:

55. Ethical concerns of user-contributed content

8. For Terms of Service with third party applications, include information about preservation and reuse of content created within the application.

Keywords: PLANNING, PUBLISHING PLATFORMS, RIGHTS, THIRD-PARTY DEPENDENCIES

If a publication platform integrates third party applications for features such as annotations or comments, the publisher should ensure that the terms of service for that application provide

appropriate permission for preserving and migrating that content over time. For example, Hypothesis' Terms of Service[1] specify that the copyright of annotation data is CC0.

See also:

14. Avoid being dependent on third party services for core features
15. Plan a strategy for preservation when third party dependencies exist

9. Define "version(s) of record" in the context of your platform and ensure there are consistent rules around this.

Keywords: PLANNING, PUBLISHING PLATFORMS

A preservation service will work with a publisher to determine the version(s) of record. If there may be multiple versions of record, or if draft versions are considered significant, the parameters of these should be clearly defined. In addition, these versions should be identified in a formal way so that automated updates can occur as needed while retaining clarity across the preservation copies.

These guidelines relate to other aspects of versioning:

23. Express versioning in bibliographic metadata
31. Assign new identifiers to significant versions

10. Define and document core intellectual components that need to be preserved.

Keywords: PLANNING

For each work, establish what readers need in order to perceive the authors' intellectual and rhetorical contributions, acknowledging that the current form of the publication may not be available in the future with changing technologies and social frameworks.

11. Use non-proprietary, broadly supported and adopted open file formats.

Keywords: PLANNING, RIGHTS, EXPORT PACKAGES, SOFTWARE AND DATA

The Library of Congress updates their Recommended Formats Statement[2] regularly. This is a helpful quick reference for selecting a format that is stable when there is an opportunity to choose. If converting data from a proprietary format to an open file format results in some data loss, consider saving both. For less established or proprietary formats, consider recording the type, version, and software used to generate and play the file—this can be included in the metadata or documentation.

---

1 Hypothesis Terms of Service. (2021, May 8). Hypothesis. Retrieved August 19, 2021, from https://web.hypothes.is/terms-of-service
2 Recommended Formats Statement. (n.d.). Library of Congress. Retrieved August 19, 2021, from https://www.loc.gov/preservation/resources/rfs/index.html

These guidelines may also be considered during file format selection:

13. Acquire the highest quality version of media to use for preservation
34. For EPUBs, opt for core media types, as defined by the EPUB specification

12. Have a discussion as early as possible about audio, video and other media assets.

Keywords: PLANNING, EMBEDDED RESOURCES, THIRD-PARTY DEPENDENCIES

Thinking through the best ways to present and preserve media assets such as video early in the publication cycle will allow for lead time to implement best practices for preservation, such as procuring and/or licensing media for local hosting or exclusively for preservation, or to choosing remote services better suited to web harvesting.

13. When embedding audio or video, acquire the highest quality version of the media files to use as a preservation copy.

Keywords: EMBEDDED RESOURCES

Publications with embedded media content often require that these files are compressed, streamed or otherwise optimized for delivery and access. When possible, acquire and retain copies of the original, high resolution, media files for the purpose of long-term preservation. Formats for preservation should be open, non-proprietary or widely adopted. Higher quality uncompressed versions of media files are preferred for inclusion in an archival package.

Where there is a choice of which file format to use, consider this guideline:

11. Use non-proprietary, broadly supported and adopted open file formats

14. Avoid depending on externally hosted web services for content that is required for the publication to function or is a core intellectual component of the work.

Keywords: PLANNING, THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES

Move supporting files such as multimedia, fonts, JavaScript, and CSS, local to the publication or inside the application used for publishing. This helps ensure the vital components of the work can be easily packaged together, reduces ongoing maintenance, and helps ensure exports contain all necessary resources.

If this is impractical in the live environment, other guidelines may be relevant:

15. Develop a strategy to capture any external media content
16. Captions for non-text features add meaningful context
20. Ensure all core intellectual components of a work are reflected in the export package
29. Consider a preservation-specific EPUB in your workflow
51. Host media files local to the website

15. For third-party hosted media content that is a core intellectual component of the work, plan a preservation strategy to capture this content.

Keywords: PLANNING, RIGHTS, THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES

Sometimes it is necessary or preferable to reference or embed third-party media content that is outside of the control of the publisher but integral to the understanding of the work. For these features, anticipate that their availability may be temporary and make plans to ensure that they are not only preserved, but sustained in some form within the publication while they are on the publisher platform. In the case of an embedded YouTube video, for example, some options to support preservation might include: retaining or requesting a copy of the video file; getting permission to take a copy of the content using the YouTube-DL tool in order to bring it into the local publication; or archiving and linking to a copy on the Internet Archive. An informative caption can help support future readers if the content is unavailable.

These guidelines may also improve preservability of third party hosted media:

12. Start discussions about multimedia early in the project
14. Avoid externally hosted media
16. Captions for non-text features add meaningful context
20. Ensure all core intellectual components of a work are reflected in the export package

16. Create a meaningful caption for all non-text features embedded in a publication.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, EPUB, WEB-BASED PUBLICATIONS

EPUB web based publications embedded resources third-party dependencies Embedded enhanced features, especially those that link to resources outside of the publication or use an unusual format, are at the highest risk of failing in the future. For this reason, a meaningful caption is vital for providing clues to future readers about what they should expect to find in that location in the text, and preferably some means of finding it and accessing it. Ideally, this caption would include a title, source, unique persistent identifier (e.g. DOI, ARK ID, or Handle), and a link to an archived copy if different from the identifier. Though any link could ultimately fail, this information would at least provide clues to where the user might find an archived copy. When creating captions, apply the standards available within the format you are using to support automated parsing. For example, HTML5 has the <figure> and <figurecaption> elements. “Alt" tags are also widely used to supply context if a feature cannot be viewed. In this respect, a meaningful caption may also meet standards for digital accessibility.

Where non-text features are supplied as separate publication resources, this guideline may also be relevant:

24. Create metadata for each publication resource

17. Use unique persistent identifiers to link to or cite external resources.

Keywords: PLANNING, PUBLISHING PLATFORMS

When referencing an external resource in a publication, see if there is a version of the resource that has a unique persistent identifier and if so use that identifier to reference it. While all "persistent" identifiers can eventually break depending on whether they are properly maintained, they are more likely to last than other links and uniquely identify a resource. Another option for tackling “link rot” - the term for when links stop working - is to use a web archiving snapshot service such as archive.today[3] or Internet Archive’s Save Page Now[4] service to archive the page and reference the resulting snapshot as an alternative link in the document. Robust Links[5] are one way to present this to users.

These guidelines cover other instances that may benefit from use of identifiers:

27. Assign persistent identifiers to publication resources and use them
31. Assign identifiers to significant new versions of the work

18. For linear publications, export packages should include core intellectual components of the work separated from the publishing platform and transformed to an existing full-text standard.

Keywords: PUBLISHING PLATFORMS, EXPORT PACKAGES

If a publication is document-like, exporting and transforming the core intellectual components to an existing standard for full text publications e.g. to EPUB, TEI, or JATS/BITS XML is a robust approach. This includes publications that contain multimedia or remote content since these enhanced features can be managed more easily at scale when the rest of the publication is expressed in a standard form. Existing standards can be validated at scale, support both platform migration and preservation, and may steer enhanced features to be expressed more consistently to work with the document.

These guidelines may also be helpful when considering the export package for a linear publication:

3. Use existing standards for export formats
10. Identify and document the core intellectual components of a work
20. Ensure exports cover all core intellectual components

[3] Archive.today. (n.d.). Archive.Today. Retrieved August 19, 2021, from https://archive.today
[4] Wayback Machine, Save Page Now. (n.d.). Internet Archive. Retrieved August 19, 2021, from https://web.archive.org/save
[5] Robust Links - Motivation. (2020, June 29). Robust Links. Retrieved August 19, 2021, from https://robustlinks.mementoweb.org/about

19. Publication export packages created for preservation should have a simple package structure with consolidated structured metadata and a predictable file and folder structure.

Keywords: PUBLISHING PLATFORMS, EXPORT PACKAGES

Excessive small metadata files or a complex folder hierarchy within an export package adds complexity to the workflow. Ideally, export processes consolidate metadata into one file per publication, and the folder and file structure are mostly flattened, predictable, and use a consistent naming convention. Metadata should be fully expressed within the metadata file, not via file- and folder names, and should include references to the files being described so that they are easily connected. The complexity of a submitted information package has an impact on the ability of a preservation service to efficiently and quickly convert it to an archival information package. Reducing the number of separate metadata files and folders reduces processing time and can improve stability in the long term by simplifying migration either to a preservation system or to another platform. To the extent that the goal is an automated preservation workflow, the export packages should be consistent across publications.

See also:

22. Use an appropriate metadata serialization within the export package

20. Represent all core intellectual components of the work in the export package.

Keywords: PUBLISHING PLATFORMS, EXPORT PACKAGES

In addition to the main text and embedded or supplemental media, other features or content such as annotations, high-quality versions of media, supporting data, and peer reviews may be considered integral to the work in some cases. If so, these resources should be part of the export package so that they can be preserved alongside the publication. Special provisions may need to be made for artifacts that are hosted outside of the platform to include them in the export.

See also:

10. Identify and document the core intellectual components of a work

21. Provide structured bibliographic metadata alongside exported publications.

Keywords: EXPORT PACKAGES

Each publication should have structured bibliographic metadata associated with it. This should be expressed as a separate file stored adjacent to or within the publication package. When possible, this should be expressed in a standard format such as e.g. ONIX, JATS, or Dublin Core. In order to process metadata at scale, the file naming convention, location of the file relative to the publication, and format should all remain consistent.

These guidelines may add context when deciding how to format bibliographic metadata and where to store it:

3. Use existing standards when creating metadata
22. Express metadata in an appropriate structured format
30. Add bibliographic metadata to an EPUB
45. Embed bibliographic metadata in a web page

22. Express metadata in a structured format that is appropriate for the metadata.

Keywords: EXPORT PACKAGES

When exporting metadata, ensure that the data format used to express it is appropriate for the content. For example, a CSV file will work for very simple metadata, but if the fields contain formatting, values that include new lines, or express specific data types, a CSV export could become unreliable or difficult to process. A structured format such as JSON or XML is generally more appropriate and can be validated for errors more easily.

23. Ensure any bibliographic metadata associated with a publication includes an expression of versioning.

Keywords: PUBLISHING PLATFORMS, EXPORT PACKAGES

Current publishing platforms can support frequent updates and new versions. These should be expressed clearly through the metadata so that the preserved copies can be properly distinguished from each other. If something has changed, it should be reflected in the version and date and where necessary, new exports should be provided.

These guidelines also relate to versioning:

9. Determine the version of record in you context
31. Assign new identifiers to significant versions of a work

24. Provide structured descriptive metadata for each publication resource.

Keywords: PUBLISHING PLATFORMS, EXPORT PACKAGES

Many publication resources that are supported by modern publishing platforms warrant their own description to ensure they are properly credited, interpreted, and rendered with context in the future. Where possible, include descriptive metadata for each resource. Use an existing standard for guidance on what to include, e.g. Dublin Core.

These guidelines add additional context to creating metadata for publication resources:

16. Captions for non-text features add meaningful context
22. Express metadata in an appropriate structured format
25. Express the license information in the resource-level metadata
26. Describe connections between resources in the metadata
27. Assign and use unique persistent identifiers for publication resources

25. Ensure that the license for each publication resource covers preservation, and express the license in a structured consistent way as part of the metadata.

Keywords: RIGHTS, EXPORT PACKAGES

When a publisher acquires rights for resources that are part of the publication, these should also include rights pertaining to the preservation of those resources. Express these rights in the metadata in a way that allows a preservation institution to determine what they have permission to preserve and relate them to the relevant material.

These guidelines may also support the creation of license metadata:

8. Clarify the license related to preserving third party web resources
24. Create descriptive metadata for each publication resource
40. Embed license information in the HTML

26. If a publication has many components or has complex integrations, ensure the metadata describes any important relationships between resources.

Keywords: EXPORT PACKAGES

While developing export processes, attention should be given to describing each resource in the package. If the relationships between the resources are also significant, ensure that this is expressed in the metadata as well. For example, if several data files are dependent on each other, or two items are versions of the same thing, or something should interact with the publication in a specific way, these relationships should be expressed so that they can remain connected in the preserved copy. Ask, what information is needed to restitch the seams between the resources in your package?

See also:

27. Assign persistent identifiers to publication resources, they can help perpetuate connections between resources

27. Assign unique persistent identifiers to each publication resource that may be referenced as an independent artifact. Use these to reference the resource within the publication and metadata.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, EPUB, WEB-BASED PUBLICATIONS

Some platforms support assigning each publication resource its own descriptive metadata and landing page making it possible to cite them independently of the text as a whole. In these cases, if the publisher has the capacity to assign unique persistent identifiers such as valid DOIs, ARK IDs, or handles to each publication resource and to provide this as part of the metadata, this can help maintain connections between the components of a publication and sustain citation links. As an example, consider the case where a video is embedded in an EPUB and it has a caption under it that includes a registered DOI. The DOI points to a page dedicated

to the published video. If the publisher no longer has that material, a preservation service may have the option to register the location of its preservation copy with doi.org so that the link would point to a new location. If a resource is local to the publication and is not intended to be cited or described independently, then a meaningful caption provides useful context, but creating persistent identifiers isn't necessary.

These guidelines also relate to the use of identifiers:

17. Use persistent identifiers to link or cite external resources
24. Create descriptive metadata for each publication resource, include identifiers
31. Assign persistent identifiers to significant versions

28. Embed and follow character encoding in publications, metadata, and websites.

Keywords: EPUB, WEB-BASED PUBLICATIONS, PUBLISHING PLATFORMS, EXPORT PACKAGES

Correct handling of character encoding can make an enormous difference to whether a publication is properly rendered. Encoding type should be expressed in the metadata, and/or within the publication as appropriate for the format. For example, websites may include encoding in the metatags and/or the charset property of the HTTP headers.

29. Consider a preservation-specific EPUB in your workflow.

Keywords: EPUB

EPUB3 has the potential to be a solid format for preservation when creators (1) abide by the official standard; (2) keep to core media types; (3) avoid encryption; and (4) encapsulate all required resources within the EPUB file. Including large files, remote resources, or interactive features, however, can make the EPUB large and therefore impractical for general distribution. Where the publishing platform has mechanisms for generating EPUBs, implementing a workflow for a preservation-specific EPUB3 that can be created alongside the public-facing ebook would be a boon to preservation services.

When preserving a website rather than an EPUB, this guideline may be relevant:

58. For a custom web application, consider encapsulation early

30. Store structured bibliographic metadata in or alongside the EPUB.

Keywords: EPUB

Each EPUB should have structured bibliographic metadata associated with it. This can be expressed in the package metadata within the EPUB (the OPF or equivalent), or as a separate file stored adjacent to the EPUB—this may be generated during export e.g. ONIX, JATS, Dublin Core. In order to process metadata at scale, the file naming convention, location of the file relative to the EPUB, and format should all remain consistent.

These guidelines relate to bibliographic metadata for other formats:

21. Provide structured bibliographic metadata with exported publications
45. Embed bibliographic metadata in the <head> of a web-based publication

31. Ensure ISBNs, DOIs, or other persistent identifiers are appropriately assigned to significant versions of the work.

Keywords: PLANNING, PUBLISHING PLATFORMS, EPUB

If a publication contains digital enhancements that are important enough to warrant preservation, the publication inclusive of its enhancements may be substantial enough to warrant a new ISBN, DOI, or other persistent identifier. This practice would ensure that the new version can be easily distinguished from other unenhanced versions of the publication in the preservation system.

These guidelines also relate to management of versions and use of identifiers:

9. Define the version of record in your context
17. Use persistent identifiers to link or cite external resources
23. Include version information in bibliographic metadata

32. Supply an unencrypted version of the EPUB to the preservation institution.

Keywords: EPUB

If EPUBs are encrypted by the publisher for distribution and compatibility with specific EPUB readers, they will need to make the unencrypted version available to the preservation institution. Encryption will make opening or even validating an EPUB difficult, perhaps impossible.

33. Use open, non-obfuscated, non-copyrighted fonts and embed them in the EPUB.

Keywords: EPUB, RIGHTS

Some publishers may use copyrighted fonts and obfuscate them in order to protect the rights when embedded in the EPUB. Because obfuscated fonts create both a technical and copyright challenge for preservation, open fonts should be used.

34. For EPUBs, opt for core media types, as defined by the EPUB specification.

Keywords: EMBEDDED RESOURCES, EPUB

The EPUB specification defines a list of core media types that are supported. Using formats outside of this list introduces an additional risk for preservation since EPUB reader tools may not support these formats. Publishers should therefore consider whether using something outside of these types is justifiable given that doing so may result in the loss of that media.

This more general guideline may also be useful to consider:

11. Use non-proprietary, broadly supported and adopted open file formats

35. Embed video and other media content in publications that are published as EPUBs.

Keywords: EMBEDDED RESOURCES, EPUB

Embedding media resources within an EPUB ensures that a future reader will be able to locate these resources and view them in the original context of the work. In order to keep the overall size on an EPUB manageable for access, it may be advantageous to embed lower quality copies of the media and link to higher resolution versions via persistent links such as DOIs.

These guidelines also refer to where media content is hosted:

14. Avoid depending on externally hosted web services in general
29. Consider a preservation-specific version of the EPUB

36. Where remote resources or non-core media types must be used in an EPUB, define a fallback.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, EPUB

Where there is a strong justification for using remote resources or non-core media, EPUB supports a fallback option that allows something else that is supported to be displayed in its place. This functionality should be used in these instances.

These guidelines may also be relevant when considering use of non-core media types:

29. Consider a preservation-specific version of the EPUB
41. Harvesting the content of iframes may have unpredictable outcomes

37. Validate the EPUB using EPUBCheck and resolve issues.

Keywords: EPUB

Basic errors in the formatting of an EPUB may not have an impact on the presentation in your favorite EPUB reader. Anything that does not follow the EPUB specification, however, may cause problems in other tools and with future playback. W3C's EPUBCheck[6] is a tool that can help identify any formatting issues and provides an opportunity to resolve them before distribution or preservation.

38. If external web content is visually embedded in an EPUB, include URLs in the package metadata.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, EPUB

[6] GitHub - w3c/epubcheck: Validation tool for EPUB. (n.d.). World Wide Web Consortium. Retrieved August 19, 2021, from https://github.com/w3c/epubcheck

If externally linked web content must be visually embedded in an EPUB, recognize that it is at very high risk for loss. If the content cannot be moved inside the EPUB container using supported features, this material should have an informative caption and be described clearly in the structural metadata within the EPUB. Specifically, the package's manifest metadata should have an item that: (a) specifies the resource URL (b) lists "remote-resources" as a property, and (c) defines a fallback item. If the embedded web content is not supplied to the preservation service, but can be successfully harvested, this additional metadata could facilitate a preservation workflow to identify and capture these features using an appropriate harvesting tool. If for example a visually embedded Google Trends chart no longer displays active content in the future, an archived web page with this chart could be accessed instead. This content should be noted consistently and documented as part of the publication that needs to be preserved. In general, any consistency that makes it easy to automatically identify the visually embedded web-based features within the text increases the chance of designing a scalable workflow to manage it.

These guidelines may also be relevant to embedding web content in an EPUB:

16. Captions for non-text features add meaningful context
40. Indicate the license status of resource in the HTML around the object
41. Use HTML iframes with caution
42. Facilitate a local web archive workflow for iframe content

39. Avoid using iframes to embed multimedia in EPUBs.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, EPUB

Iframe, short for "inline frame," is an HTML tag that can be used to embed the content from any URL inside an HTML-based document such as an EPUB or webpage. Some publishers may use an iframe to embed things like YouTube videos, or advanced media players into an EPUB. It is more sustainable to use html <video> or <audio> elements when embedding audio or video. EPUB3 readers are not required to support iframes. If used, the content may not render in all EPUB3 readers and is at a high risk of loss through link rot.

These guidelines are also be relevant to embedding media in EPUBs:

12. Start discussions around multimedia early in the process
14. Avoid external dependencies in general
34. Opt for core media types when embedding multimedia in an EPUB

40. If external web content is embedded in a publication, identify the rights and note intention for collecting this content as part of the publication.

Keywords: RIGHTS, THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, EPUB, WEB-BASED PUBLICATIONS

A preservation service may not collect web content outside of the agreed upon domain names unless copyright for the content being harvested is clear. If third-party pages and features that are visually embedded in an EPUB or a web-based publication are meant to be preserved, it should be possible to identify which content publishers have the right to collect so that a web crawler can be configured to include or exclude it. One way to differentiate could be to

consistently express the rights in the metadata that is supplied to the preservation service. Another option is to apply structured metadata describing the rights status to the HTML. The Creative Commons REL documentation[7] includes examples of this that cover both page- and object-level licenses - this approach could support automated harvesting decisions at either level. Alternatively, a publisher could supply a list of domain names to include for harvest during the initial preservation workflow configuration.

These guidelines may also be useful to consider when embedding external web content:

25. Add license information to resource-level metadata
38. List the URLs for external web content in the metadata
45. Embed metadata that includes a license in the <head> of a web page

41. Use the HTML iframes with caution, and discuss specific cases with preservation specialists.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, EPUB, WEB-BASED PUBLICATIONS

An HTML iframe can contain a wide range of types of content, from a wide range of sources, which makes them a challenge for preservation. The quality of automated website archiving in general can vary greatly. If an iframe is embedded in an EPUB or website, the more inconsistent, complex, and dynamic their content, the more likely they will be lost in an automated process. If these features are important to preserve, consider a manual process to capture and package the intellectual components of the iframe content in another form. For example, a video or screenshot with a caption that links to the website might be a sufficient fallback for conveying the contents of the iframe.

These guidelines may also be relevant to use of iframes:

38. List the URLs for each embedded iframe in the metadata
39. Avoid use of iframes in EPUBs
42. Facilitate a local web archiving workflow to support iframes

42. If external web content is visually embedded in an EPUB via an iframe, reduce risk of total loss of these features by facilitating a local web archiving workflow.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, EPUB

Preservation services might not support a workflow that automatically harvests the content of iframes embedded within an EPUB. Even with such a service, the quality could vary greatly, and the content might change following publication. If fallback options are not sufficient a more stable approach would be for the publisher to create an archived copy of the web page featured in the iframe. While there are tools that can be run locally by the publisher to perform single page archiving, there are also third party archiving services such as archive.today or Internet Archive's Save Page Now service that allow you to archive a single page before publication and

[7] CC REL by Example. (n.d.). Creative Commons. Retrieved August 19, 2021, from https://labs.creativecommons.org/2011/ccrel-guide

generate a persistent link for the embedded web content. This link could be included in a descriptive caption under the embedded feature. Publishers should test the outcome of these single page captures as quality can vary depending on the complexity of the website and the harvest method applied.

These guidelines may also be relevant:

14. Avoid dependence on externally hosted platforms for core features
38. Avoid the use of iframes to embed multimedia

43. Include a sitemap.xml file for web based publications.

Keywords: PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS

Sitemaps containing links to all of the content in a website ensure that website archiving crawlers will be able to locate all of the content. Doing so may also improve search engine optimization. Sitemaps that are intended to facilitate web archiving should include links for all texts, resource landing pages, downloads, and views of the data i.e. API URLs that are called dynamically while the user is interacting with the page and each combination of query parameters that may appear.

This guideline will make creating a sitemap simpler:

46. For websites, give each page state its own URL

44. Keep methods of linking within a web based publication simple.

Keywords: PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS

Successful website archiving is contingent on a harvester visiting each URL that forms the work. If a full list of URLs is not supplied to the harvesting tool via a sitemap or through other configuration, automation may be used to discover the URLs. Automated website crawling tools can easily identify the target of simple HTML <a> or <link> tags with a relative or full URL, and will include them in a crawl. Many websites, however, use JavaScript actions to fetch content. Crawlers may not be able to identify the URLs that are loaded by JavaScript causing the content to be missed during an automated archiving process. Similarly, hyperlinks that are within compiled features e.g. compiled 3D visualizations, can be difficult or impossible for a crawler to discover. When designing web content, consider the value of using simple HTML links so that crawlers can identify the URLs that make up a work. Note that as with <link> tags, the target URLs of <a> tags will likely be crawled even if they do not display text on the page, and so they can be used to guide a crawler to relevant content. Conversely, a crawler cannot determine which of these tags link to content that is not vital to the work, and so using these tags for other purposes or having hidden link tags that are never used can guide the crawler to things that may be out of scope for an archived copy of the publication, such as previous or unused iterations of a page.

This guideline may make changes for efficient crawling less critical:

43. Include a sitemap for all web-based publications

45. Embed structured bibliographic metadata in the <head> of a web-based publication.

Keywords: PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS

This can help facilitate a fully automated web harvest of content in situations where an export is not a feasible approach. Bibliographic metadata is a vital component of a publication preservation package. As with other metadata it's best to use a broadly adopted standard such as Google Scholar, Dublin Core, or PRISM. Cover the core bibliographic information to make the publication findable, and be consistent. An expression of the material's license, for example, through <link rel="license" href=...>, is valuable since this can support an archive's understanding of whether the material can be preserved and how it can be reused. Note that HTTP Link headers can also be used to convey some metadata and can be applied to the HTTP Response of both HTML and non-HTML web resources. An approach to this is described on signposting.org[8].

These guidelines may also be relevant when generating bibliographic metadata:

21. Provide bibliographic metadata with exported publications
30. Bibliographic metadata in the context of EPUBs
40. The license for external resources can be expressed in HTML

46. On a website, give each unique page state its own URL.

Keywords: PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS

Data driven websites can technically display different sets of resources from the server at the same URL. If different views of a page share the same URL, however, this means that retrieving a web page from a web archive could have unpredictable results. It is therefore helpful to ensure that, where reasonable, the URL reflects any filters or properties that change what is loaded into the browser from the server via the path or querystring (the part of the URL following the question mark). This allows the different states of a page to be bookmarked, but also makes it possible to utilize a sitemap to express the full range of resources that make up the website. While a sitemap can include API calls that might be used for dynamically generated views, sitemaps are easier to maintain if these views are also reflected in the browser's address bar.

This is another guideline about making URLs web archive friendly:

49. Parameters should not be added to the URL unnecessarily

47. Plan the structure of URLs so that they are more likely to persist over time; redirect when needed.

Keywords: PLANNING, PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS

Key URLs for a publication, such as a publication's home page, should not change over time. If they must change, redirect the original URL to the new location. Apart from helping to decrease broken links from other websites, using a well planned URL structure can help with website

[8] Signposting the Scholarly Web. (n.d.). Signposting the Scholarly Web. Retrieved August 19, 2021, from https://signposting.org

preservation. Ensuring the publication's URL does not change over time can make it easier to manage and connect different versions of the publication that are preserved and avoid duplication.

These guidelines discuss identifiers, another way to support URL persistence:

27. Persistent identifiers can be used at the publication resource level
31. Persistent identifiers should assigned to new versions of the work

48. Where multiple publications are on a single domain or subdomain, plan the structure of URLs so that the scope of each publication is easily identified via its URL path.

Keywords: PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS, PLANNING

Where there are multiple publications on the same domain or subdomain, and each one spans multiple pages, using a consistent and hierarchical naming convention in the URL path helps web harvesting tools identify its scope. For example, if the publication content is organized in these directories: example.org/book-slug/text, example.org/book-slug/resources, a crawler can be set to generate an archive of the resources within the "book-slug" directory.

49. On a website, parameters added to a URL should reflect the data that is loaded from the server and not be added unnecessarily.

Keywords: PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS

Website crawling and playback of web archives use URLs as unique references—this includes the query parameters (after the "?" and, for some tools, after the "#"). Adding parameters to the URL that do not affect what data is loaded from the server, or simply reflect a default where the page is the same with or without the property, complicates the capture and playback of the web archive and bloats the size of the crawl since every URL is captured as if it is a new page even if the content is identical.

This guideline is also useful for creating web archive friendly URLs:

46. Assign each unique page state one, and only one, URL

50. For highly dynamic websites that use a lot of JavaScript to interact with the server in the background, consider a "progressive enhancement" design approach.

Keywords: PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS

Many modern websites depend on JavaScript to load data from the server as the user interacts with the site creating a dynamic experience. This can make it difficult for a web crawler to automatically create a functional copy of a web page since it may not be able to predict all user behaviors that pull new content from the server. Some web developers design websites using a "progressive enhancement" approach in which a baseline of functionality is supported for a variety of environments, including those with scripts disabled. Where this approach is used, the version of the site presented to the user will change if they choose to disable, or cannot support,

JavaScript in their environment. They will instead see a scriptless version of the site that presents the core intellectual components of the page in a more static form. If this functionality exists or can be easily supported, it can serve as an alternative way to capture pages using web archiving in cases where the full dynamic version cannot be crawled automatically.

This guideline describes an alternative way to manage JavaScript-rich features:

53. For dynamic web page features, favor designs that pre-load data

51. When embedding video and other media in web-based publications, host the media files local to the website using standard HTML tags rather than depending on third party services for streaming.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, WEB-BASED PUBLICATIONS

Linking to media that is hosted on YouTube or Vimeo is a threat to platform and content longevity, especially for media that is owned or managed by third parties. In order to mitigate against future link rot and the general instability of archiving streamed content, where appropriate (technically and legally), host a local copy of any media assets and embed it in the web page using standard HTML5 media tags. In order to keep the overall size of embedded media manageable for access and for the purpose of web archiving, it may be advantageous to embed lower quality copies of the media and link to higher resolution versions via persistent links such as DOIs.

See also:

12. Start discussions about multimedia features early
14. Avoid depending on externally hosted web services

52. If parts of a publication are only accessible via a search feature, implement a corresponding browse feature that would allow a web crawler to discover all content via simple links.

Keywords: PUBLISHING PLATFORMS, WEB-BASED PUBLICATIONS

Platforms with good search engine optimization implement paths to navigate every page via links. This is also useful for web archiving since both search-engine crawlers and web-archiving crawlers use similar mechanisms to discover all pages of content.

These guidelines also help a website crawler discover all content:

43. A sitemap can help website crawlers reach unlinked content
44. Use simple links to help a website crawler find content
46. Ensure each page state has its own unique URL

53. For web-based features that call for a dynamic user experience within a single webpage—e.g. pop-ups, annotations, data visualizations, maps—consider designs that pre-load all data when the page initially loads in the browser.

Keywords: EMBEDDED RESOURCES, WEB-BASED PUBLICATIONS, THIRD-PARTY DEPENDENCIES

In order to improve the likelihood that content published to the web will be able to be captured via web archiving methods, developers could preload any content that would otherwise depend on user interactions. For example, rather than repeatedly making small API calls as the user interacts with a feature, if the dataset that supports the feature is small enough, load the data as a JSON file when the page loads so that further server calls are not necessary.

This guideline describes another approach:

50. Consider a "progressive enhancement" design to support a scriptless environment

54. Avoid embedding social media posts directly in a publication.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, WEB-BASED PUBLICATIONS

Avoid using the "embed" option to insert a social media post into your publication. This can be unstable for preservation and for long-term sustainability since posts or accounts may be deleted. If the social media post is integral to the work, consider first taking a screenshot that can be embedded into the publication as an image. Underneath, a caption should indicate the origin of the post. Finally, use a web archive service such as archive.today or Internet Archive's Save Page Now service to create a copy of the post—be sure to test the results, since archiving social media posts can be unreliable. The two links (live and archived) could be referenced as a citation or footnote depending on local practices.

These guidelines are also relevant to embedding social media posts in a publication:

8. Ensure terms of service cover preservation of data in third-party services
14. Avoid depending on third party services for core intellectual components
55. Consider ethical implications of embedding social media posts

55. Before including social media or user contributed content in the list of things to be preserved, consider whether it is ethical/appropriate to preserve it.

Keywords: THIRD-PARTY DEPENDENCIES, WEB-BASED PUBLICATIONS

Some publications, especially in a web environment, may include social media posts or user contributed content that are automatically included in the archive package, especially with a web harvesting approach. Before implementing these features or including them in publications, consider whether taking copies of them infringes on individual rights or safety. Preservation services may not be able to evaluate specific situations in a scalable way, and so it's important to avoid including these in the preservation scope if there is uncertainty around them. This may involve designing a website in a way that certain content can easily be excluded e.g. keeping

this content at a separate URL that can be skipped during crawling. The Documenting the Now[9] project website includes information about ethical collection of social media content.

These guidelines discuss legal and technical considerations for preservation:

8. Ensure terms of service cover preservation of data in third-party services
54. Avoid embedding social media posts in a publication

56. For maps that display a single location, replace dynamic mapping features with a still image of a map and link to the dynamic version.

Keywords: THIRD-PARTY DEPENDENCIES, SOFTWARE AND DATA, EMBEDDED RESOURCES, WEB-BASED PUBLICATIONS

Dynamic maps such as those generated with Google Maps, consist of many smaller map tiles that are loaded on the fly as users pan and zoom. Web crawlers cannot easily capture this experience, nor can this be exported. If the map is not the focal point of the work and is being used to present a small number of locations, consider using one or more still images. Display the place name and coordinates for the pin in the caption and provide a link to a live map.

These guidelines offer alternative ways to manage dynamic map features:

16. Captions add important context to non-text features
53. Consider web page designs that pre-load all data when the page loads

57. For web-based features that require perpetual or open ended communication with a live server, consider alternative strategies to provide a representation of this feature to a preservation service.

Keywords: THIRD-PARTY DEPENDENCIES, EMBEDDED RESOURCES, WEB-BASED PUBLICATIONS

Some web-based features require communication with a server that is driven by an unpredictable user interaction or utilizes an open-ended number of URLs to retrieve the data to support that feature. These features cannot be exported easily due to their dependence on a live website and cannot be captured well using web archiving, which depends on identifying every unique URL. Examples include: dynamic maps (e.g. Google Maps), full text or faceted search, web forms, data visualizations (e.g. ArcGIS), IIIF image viewers, and streamed content. Some features can be redesigned to remove their dependency on a live server, but if they can't, publishers will need to consider what can be preserved. There are many strategies for this, for example: create a simpler static version of the feature that incorporates the key features for the purpose of preservation; embed a local copy of a server based resource rather than depend on a third party service; supply code or data for the feature with documentation for re-assembling the functionality; record a video of the interaction as it behaves in the published environment for future playback; or, a combination of these.

[9] Documenting the Now. (n.d.). Documenting the Now. DocNow. Retrieved August 19, 2021, from https://www.docnow.io

These guidelines offer alternative ways to manage features that depend on a live server:

16. Captions add important context to non-text features
53. Consider web page designs that pre-load all data when the page loads
63. Supply raw data, documentation for data visualizations

58. For a custom web application built for a single publication, consider encapsulation of features early.

Keywords: PLANNING, WEB-BASED PUBLICATIONS

If publishers are involved early enough in the development process for a custom web application that is being built for a single publication, they should encourage developers and authors to make choices that avoid external dependencies or to have fallback mechanisms when external dependencies fail. For example, if a connection to Google Maps fails, fall back to a still image and the vector coordinates. Developers can test their site by running it in a virtual environment with no internet connection. If it works, it is not only likely to be easier to preserve, but also much more sustainable and easier for the publisher to maintain.

These guidelines may be referred to when considering encapsulation:

14. Avoid depending on externally hosted web services
51. Embed multimedia locally
56. Avoid embedding map visualizations where a static representation would suffice

59. For a custom web application built for a single publication, or where custom software is part of the project, keep the infrastructure and installation process simple, select widely used technologies.

Keywords: PLANNING, WEB-BASED PUBLICATIONS

All websites have to be maintained in order to be sustained on the live web. An over-complicated web application will not only degrade more quickly and be more expensive to maintain, it will likely be even more difficult to preserve as an application. Unless the focus of the project is experimental technology, use technologies and programming languages that will be easily supported by technical staff. Do not unnecessarily overcomplicate the infrastructure and code. A helpful reference to building sustainable projects is the University of Victoria's Endings Principles for Digital Longevity[10].

These guidelines may also be helpful when considering publication software:

2. General considerations for designing or selecting publication platforms
3. Favor existing standards

[10] Team, T. E. P. (2021, August 3). Endings Principles for Digital Longevity Version 2.0. Building Sustainable Digital Humanities Projects. https://endings.uvic.ca/principles.html

60. Request an installation script for a custom web application built for a single publication, or where custom software is part of the project.

Keywords: SOFTWARE AND DATA, WEB-BASED PUBLICATIONS

For custom websites or software, publishers should request an installation script from the authors or developers. This can be used in combination with a clean installation package (one that is unpolluted by extraneous files and data generated in the live environment during deployment and use) to install the software or website in a new environment. In addition to the install script, the authors or developers should provide a document listing the machine requirements and any dependencies that will be installed or used by the script. If a script is not available, at minimum the authors or developers should provide documentation that describes the requirements, dependencies, and detailed installation process with sample commands as appropriate. This information can be placed in a README file placed in the root of the project. While installation scripts may stop working as technology evolves, they provide information about how to get the software working and can be vital context for a preservation service, or when migrating to new infrastructure.

These guidelines also discuss the installation package for a web application:

61. Create installation packages for custom websites that don't require a live server
62. Create installation packages for custom websites that do require a live server
67. Keep the source code and compiled version of the software

61. For a custom web application built for a single publication that does not rely on a live web server, work with the developers and authors to prepare a clean, self-contained, installation package for the purpose of backup and preservation.

Keywords: PLANNING, WEB-BASED PUBLICATIONS

When a custom publication is developed using plain HTML5, CSS, and JavaScript that does not communicate with a live web server, it may be possible to run the entire application from a local machine by opening it in a browser. In this case, a clean application package should be created and retained by the publisher as a backup and for preservation. Work with the developer and author to ensure that this preservation copy: functions fully offline; does not contain any system files, server information, or logs; uses relative links that do not contain a specific domain name; and contains only local stylesheet, font, or JavaScript references. If there are features that depend on a third-party service, e.g. for search or commenting, that are not a core intellectual component of the work, these can be disabled. A README file should be placed in the root of the application folder to describe the project, instructions, dependencies, versions of technologies used, and details of any unique features that might be useful for playback later. The entire package can be stored as a zip file. If updates happen once the application is deployed on the live server, these should be reflected in the clean preservation copy and a version number should be expressed in the package.

62. For a custom web application built for a single publication that requires a web server to run, work with the developers and authors to

prepare a clean self-contained, installation package for the purpose of backup and preservation.

Keywords: PLANNING, WEB-BASED PUBLICATIONS

When other methods of preserving a web publication (export, web crawling) cannot appropriately capture the important properties of a publication because it is dynamic and data-driven, a preservation institution may attempt to preserve the application itself with the goal of running it in an emulated web server environment in the future. In order to do this, the preservation institution would require a clean installation package as well as documentation of the requirements, dependencies, and installation process. A preservation copy could be created during the publication process. Work with the developer and author to ensure this preservation copy: functions fully in a self-contained web server that does not have access to any resources outside of the machine; does not contain any server information or logs; uses relative links that do not contain a specific domain name; and contains only local stylesheet, font, or JavaScript references. Where features require a live third-party site, consider a local functionality that could replace it adequately in this package. Overall, it would be beneficial for the developers of the publication to design any website with sustainability and encapsulation in mind—ensuring files are local to the application where possible and that there is a simple way to fallback to local functionality for integrations such as third-party resources.

These guidelines also discuss the installation package for a web application:

58. Consider encapsulation of custom-built web applications early
60. Request an installation script for custom software and websites
61. Produce packages for software and websites that don't require a live server

63. For data visualizations, retain raw data and documentation about them.

Keywords: THIRD-PARTY DEPENDENCIES, SOFTWARE AND DATA, WEB-BASED PUBLICATIONS

Data visualizations tend to be a particular arrangement of one or more raw datasets. Data visualization formats can obscure parts of the underlying data that they are derived from. They may also be compiled or complex. All of these properties could potentially make the data difficult to open, validate, or comprehend in the future. To preserve a publication in which data visualizations are core intellectual components, request underlying raw data from the author. Request supporting documentation that would enable a future reader to retrace the author's steps from the raw data to the visualization. Images or videos of the visualization may also be helpful for recreating it. For both visualization and raw data formats, as with all supplements, ideally the files will be an open or broadly adopted format. The Library of Congress Recommended Formats Statement can help with selecting formats. In the case of vector data, for example, there is not a broadly adopted open format, but Shapefile, while proprietary, is broadly adopted and openly documented. There are a variety of tools that can read Shapefiles which increases the likelihood that it will continue to be supported in some form.

These guidelines may also be relevant when considering preservation of data visualizations:

11. Use non-proprietary, broadly supported and adopted open file formats
57. Use alternative approaches for features that require communication with a server
64. Use meaningful file names and field names in your data, supply documentation

64. For data, use meaningful file names and field names and supply documentation.

Keywords: SOFTWARE AND DATA

For data that is to be preserved as part of, or as a supplement to a publication, provide metadata and documentation that explains the context of the data, and the meaning and limits of each file or field. In the absence of strong documentation, using file and field names that convey some meaning can be helpful to support data reuse. For example, a field named "otherstuff_12" is less useful than "weight_in_kg."

65. Ensure irrelevant or private administrative data is removed from data exports.

Keywords: EXPORT PACKAGES, SOFTWARE AND DATA

Do not send administrative data to a preservation archive unless it is integral to the work. For example, when exporting a SQL database, you may need to exclude or anonymize the content from user tables, indexes that support a specific UI, non-public communications, or logs. Only archive the essential data that can be made publicly accessible.

These guidelines refer to the creation of the installation package:

61. Create installation packages for custom websites that don't require a live server
62. Create installation packages for custom websites that do require a live server

66. Include a README file in order to provide context needed to understand any data or custom software that is included in the publication package.

Keywords: EXPORT PACKAGES, SOFTWARE AND DATA

For data, software, or any resource that has a complex arrangement of files, if structured metadata cannot be supplied, a common convention is to include a README file from the author. Written using a plain text file format, this should be a note to future users who wish to use the files. It should include information such as, scope, purpose, author(s), relevant dates, license for reuse, dependencies, field names/descriptions, and instructions for use.

See also:

68. Provide documentation for software

67. When software is included in an archive package, provide both the compiled software and the source code if available.

Keywords: SOFTWARE AND DATA

Compiled software may be opaque or impossible to modify, while source code may be impossible to compile if build dependencies become unavailable. Supplying both can enable different preservation pathways. If compiled software can no longer run due to an incompatible

operating system, it may be possible to match it with an appropriate emulator. Source code provides future users with an opportunity to understand what the software is if the documentation is insufficient and may also allow modifications to the software to work in a different environment or context. Ensure that the software license is expressed within the package and appropriate for reuse.

These guidelines refer to the creation of the installation package:

60. Request an installation script for custom software
61. Create installation packages for custom websites that don't require a live server
62. Create installation packages for custom websites that do require a live server

68. When software is included in an archive package, provide documentation.

Keywords: EXPORT PACKAGES, SOFTWARE AND DATA

Consider what a future user of the software might need to know to run the software and understand how it should work. Ensure this is covered by the documentation. For example, what is the software for? What are the supported operating systems and versions? Are there any dependencies or requirements? How do you install it? How do you use it? What should it do if it is working? What is its license? In the case where software is not possible to preserve, visual and narrative documentation of the user experience can provide vital context.

This guideline refers to another common method for documenting software:

66. Use a README file to document data or software