Article-level enhancements in the humanities and social sciences

Hosting supplementary material

Archiving supplemental materials

DOIS for journals: linking and beyond
HOSTING SUPPLEMENTARY MATERIAL: TECHNICAL CHALLENGES & SUGGESTED BEST PRACTICES

ANDREA LAUE
Debates about the definition of supplementary material point to a larger transition in scholarly and technical publishing from print to electronic articles as versions of record. In the past, a piece of evidence was deemed “supplemental” because it couldn’t be presented in print. The medium defined what constituted rhetorically valid evidence. As online versions overtake print, the impulse is to imagine that medium is no longer a constraint. Any and all supplemental information can be “included” online.

In her now famous editorial in *Cell*, Emilie Marcus writes of setting “limits” in defining supplementary materials. Marcus defines supplemental materials via their proximity to the main argument. The point is to admit only evidence that is integral to the paper and appropriate for the medium, whether that medium be electronic or print. What’s appropriate for the electronic article may currently exceed and continue to expand faster than that of print, but it’s not without bounds.

Commenting on online display conventions for supplemental materials, Ian Brown of HighWire Press issued the caveat that “‘supplemental materials’ are truly that—they should supplement the primary information presented in the article, but not themselves be central to it. Technical limitations of various media make integration of data more difficult (e.g., it may be hard to effectively incorporate audio datasets into print), but those limitations ideally would never affect whether core information was treated as ‘supplemental’ or not.” Brown gets at the issue of medium from a slightly different tack here, arguing that materials considered “supplemental” in print might best be modeled as part of the core article in an electronic environment.

The challenge, then, is twofold. First, the rhetorical challenge of presenting the most appropriate evidence given the medium of publication. Second, the technical challenges associated with any medium of publication, whether it be print or online.

The task of describing, displaying, and searching supplemental materials online remains underexplored territory. Simply dumping those materials on the web and providing links is only marginally better than citing a video in a print article. Hosting of those materials poses technical and business challenges, and issues surrounding preservation are just now getting the attention they deserve.

**Technical Challenges**

Supplemental materials present several challenges to online publishers of scholarly and technical content. Five technical challenges will be discussed here—the first second and fourth in some depth and the remaining two in a more cursory manner:

- Display
- Search
- Hosting
- Article markup
- Preservation

CONTINUED »
1. DISPLAY

Audience and device are two key variables to consider when designing display of supplemental materials. Different audiences are likely to have different levels of interest in supplemental materials, and different devices are likely to have different capacity for displaying them.

Ian Brown, HighWire, suggests that “a ‘complete’ presentation of an article should include descriptions of, references to, and (where possible) ready access to all material used to create the article. This so that researchers are able both to understand the authors’ conclusions (by reading the article itself) and evaluate the methodologies used and quality of the conclusions (by looking at supporting data).” Descriptions should characterize the content of the supplementary materials as well as the technical requirements for viewing them. An article should contain explicit links to all supplementary items as well, connecting that extra material to a specific argument made within the article.

Some publishers now include supplementary materials sections in their (electronic) articles. “Publishers opting for such presentation might consider grouping materials into sets, one that supports the conclusions of the article and another that prompts additional questions and future research,” suggests Brown. Evidence that complicates findings might also be addressed and presented in a separate section.

Different audiences are likely to have different levels of interest in supplementary materials. John Sack, Director of HighWire, has noted that clinicians are less likely to consult supplementary materials than researchers. Clinicians—readers generally—are also increasingly likely to read research articles on a mobile device, which poses access and data management issues when it comes to supplementary materials.

Ironically, mobile devices and printed PDFs pose similar challenges when it comes to supplementary materials. Both make simple descriptions of the supplementary materials desirable, as access to the materials themselves cannot be assumed. Sack advised that the PDFs have supplementary materials copied into the PDF so as to avoid taking readers by surprise when they discover that the article they printed isn’t complete. This might also be addressed by offering two versions of the article PDF, one with supplementary material included and one without it.

Brown suggests a “package” download, whereby a reader could retrieve the article and all associated supplements. This “one stop shopping” presented alongside the article PDF would address the needs of clinicians and researchers both, while presenting each with the other option.

Concern over length and complexity of PDFs reminds me of earlier anxieties about connection speeds and file sizes of PDFs. Perhaps we’ll soon see “article-only PDFs” and “full PDFs,” the latter containing all supplementary materials suitable for print, much like we once had low-resolution “screen” PDFs for reading and high-resolution “print” PDFs for downloading.

2. SEARCHING

Searching locally-hosted, non-XML supplementary resources and remotely-hosted resources of any format presents particular challenges. Some of these challenges are technical (e.g., tagging of video so that transcripts can be searched and hits accurately located within the video’s binary object) and others are conceptual (e.g., does a text search operate on the article and dataset in the same manner). Adding the complexity of federated search (local and remote resources) to that presents a daunting challenge.

Todd McGee, Assistant Director of Application Systems at HighWire Press, argues for integrated searching of article and supplemental material content. “By definition supplementary data does not stand on its own,” observes McGee, and “that being the case the supplemental data should be indexed along with the original article and should, by default, be included in a simple fulltext search.” When it comes to search, supplemental materials should be treated like other components of an article, McGee feels. Search options should “follow the pattern that a publisher follows for other sections of an article/chapter.”

Search results lists that include “hits” within supplemental materials should clearly indicate the parent/child relationship of the article and supplemental materials. A results list from a general search should take pains to label supplemental materials and clearly associate the “hit” with the argument that invoked that material as evidence. If a given piece of supplemental material is hosted remotely, the search result should include a link directly to the supplemental object.

A targeted search of supplemental data should follow the same conventions as other section- or object-specific searches, suggests McGee. “Typically this would be to return a link to the parent item; optionally that link might be to a positional anchor to the section of the article where the supplementary data can be viewed.”

A publication of the National Information Standards Organization (NISO)
McGee advises publishers to request transcripts for audio/video supplements to make this material searchable and accessible to all users, regardless of their abilities. Supplements supplied as Text, HTML, PDF, etc. should be indexable by most search engines. Compressed files or very large datasets present special cases and in most cases would not be directly indexable by many search engines.

Including descriptive metadata of supplements not susceptible to standard indexing would offer some means for improving searchability. This metadata might be indexed by default, or it might be called on only in those cases where transcripts or other surrogates were unavailable. The NLM Journal Publishing DTD (see #4 below) isn’t optimized for this purpose, however, so pursuing this strategy will likely require some creative interpretations of or extensions to the standard content model.

3. HOSTING

Hosting supplementary materials presents a variety of challenges and requires constant attention to an always shifting set of formats and requirements. From a vendor’s perspective, the key decision is whether to host in-house or contract with a third party for the technical expertise and transmission bandwidth to serve supplementary materials.

Digital video is a good example here. Competing formats (Flash, QuickTime, WindowsMedia) with different server and player requirements are constantly moving targets. Investing in and maintaining the infrastructure to host and securing the bandwidth to serve the videos involves potentially significant cost. Alternatively, there are third-party vendors willing to partner with publishers and take on this portion of the risk. Then the technical challenge becomes one of linking from locally-hosted content to remotely-hosted supplementary materials, a task that is relatively well understood and manageable. Ideally, the use of standard identifiers such as DOIs facilitates this approach, as DOI-based query services and searches on DOIs themselves relieve the need for literal URLs in source content and promote the general discoverability of supplemental materials wherever they are hosted.

4. ARTICLE MARKUP

The NLM Journal Publishing DTD version 2.3 offers the supplementary material element (<supplementary-material>) for modeling supplemental material. The element was designed to contain references to “additional data files that contain information directly supportive of the document, for example, an audio clip, movie, database, spreadsheet, applet, or other external file.” Acknowledging the multiple meanings of the word “supplementary material,” the Tag Library identifies as candidates for this element: extra tables, supportive materials “too voluminous” to appear within the narrative, and materials such as quizzes and forms that “enhance” the article content.

The content model of supplemental material allows one to describe any number of media objects, graphics, or tables, and the supplementary material element can be repeated in those locations where it’s allowed. As such, one may model multiple pieces of supplemental evidence as individual items or grouped into any number of sets.

Four points of association have been suggested for supplemental materials:

- Article-level association
- Structural unit or point within the article body
- Specific figure
- Separate section, often at the end of the article

The DTD accommodates all of these use cases, although support for associating supplemental material with a specific figure is limited.

The Tag Library describes two use cases for its supplementary material element. First, listing of supplementary materials in the article metadata so that those materials can be accessed from the article. Second, positioned inline like a figure (<fig>), in which case a position attribute may be applied so as to indicate whether the supplementary material is intended to be located at its point of reference or instead relocated to a point convenient for the particular display.

The first use case, listing of supplementary materials in the article front matter (<front>), models accurately the association of those materials with the article as a whole. Any number of supplementary materials elements can be included.

CONTINUED »

1 No major changes to the modeling of supplementary material were introduced in version 3.0 of the NLM DTD.
The **Tag Library** describes two use cases for its supplementary material element. First, listing of supplementary materials in the article metadata so that those materials can be accessed from the article. Second, positioned inline like a figure (<fig>), in which case a position attribute may be applied so as to indicate whether the supplementary material is intended to be located.

The Tag Library describes two use cases for its supplementary material element. First, listing of supplementary materials in the article metadata so that those materials can be accessed from the article. Second, positioned inline like a figure (<fig>), in which case a position attribute may be applied so as to indicate whether the supplementary material is intended to be located.

The second use case, associating supplementary material at a specific point within an article, may be accomplished either by positioning it (<supplementary-material>) within the narrative flow or by encoding a cross-reference to supplementary material tagged in the article front matter. The first model facilitates inline display but implies exclusion based on some constraint of the medium rather than the nature or role of the referenced material. Encoding the supplementary materials in the front matter and encoding cross references at mention in the narrative flow seems in keeping with the notion that supplementary items extend rather than support the basic argument of the article.

The content model of section (<sec>) includes supplementary materials as a child. Thus, a list of supplementary materials may be modeled as a (final) titled or untitled section in an article. As article navigation often keys on section titles, this content model may offer built-in support for navigating the supplementary materials without any additional development work.

Modeling supplementary material as an extension of a particular figure poses a challenge to users of the NLM DTD. The supplementary material element is not included in the content model of figure (<fig>). In most contexts supplementary material may be a sibling of figure, and adjacency might be used to imply ancestry, although that approach seems fraught with peril.

The content model of section (<sec>) includes supplementary materials as a child. Thus, a list of supplementary materials may be modeled as a (final) titled or untitled section in an article. As article navigation often keys on section titles, this content model may offer built-in support for navigating the supplementary materials without any additional development work.

Modeling supplementary material as an extension of a particular figure poses a challenge to users of the NLM DTD. The supplementary material element is not included in the content model of figure (<fig>). In most contexts supplementary material may be a sibling of figure, and adjacency might be used to imply ancestry, although that approach seems fraught with peril.

The NLM DTD also offers an inline supplementary material element (<inline-supplementary-material>) to model links to such materials. Allowed as a child of any paragraph, this element might be used within a figure caption to point to supplementary materials associated with a particular figure. The dearth of attributes allowed on the element limit the semantic richness with which the referred-to materials might be described, however. Presumably the text of the parent paragraph (part of the caption) could describe for the reader what she will find at the other end of the link, but reproducing that in a format actionable by the machine poses different challenges.

The content model of inline supplementary material does offer external link (<ext-link>), which in turn offers a type attribute (@ext-link-type) that might describe the reference material. The current Tag Library offers a hodgepodge of values from “doi” to label a DOI to “genpept” to identify a specific database. While “doi” is too general to identify the supplemental materials, values such as “genpept” might facilitate some inferences as to what’s at the other end of the link. The list of allowed values for this type attribute is not restricted, but the community might benefit from agreeing on some standard values.

A DOI may be associated with supplementary material by means of object identifier (<object-id>) child elements (with pub-id-type attribute values of “doi”). The list of allowable identifier types is not restricted, although the published list is under-descriptive. It’s advisable to register DOIs for supplementary materials, and that’s likely to remain a primary identifier. However, as (or if) the number of repositories of datasets and other supplementary materials grows, the need for additional values will likely grow. The catch-all content-type attribute may also be applied to object identifiers, allowing for additional labeling.

The supplementary material element offers a limited set of attributes for encoding metadata about the referred-to materials. MIME type and language may be defined using standard attributes. The catch-all attribute content-type is available to capture special semantic intent of the tagged content. Version 3.0 of the NLM Journal Publishing DTD adds the specific-use attribute, which “specifies distinctions in the applicability of a particular element.” The Tag Library advises using this attribute to specify audience for or medium in which the contents of the element are relevant.

Child elements such as attribution (<attrib>) and permissions (<permissions>) may be included to indicate authorship or intellectual property restrictions that differ.

It’s advisable to register DOIs for supplementary materials, and that’s likely to remain a primary identifier.
from that of the article as a whole. The content model of attribution does not allow for formal tagging of authors, however.

Supplemental materials may require additional metadata to support online display. For example, displaying an inline video may require a call to a player. That call may require a few basic display parameters, e.g., height, width, and duration. Depending on the format of the video and the preferred player, the available parameters and the syntax required when passing them may vary.

At HighWire Press, we’ve defined a separate manifest file that accompanies each video. This XML file accommodates a variety of metadata values designed to facilitate online display. The manifest knows the video by its DOI, as does the article XML. The article and all associated manifests are part of the processing context, so display variables are available during generation of the browser XHTML, for example.

This design acknowledges the relative stability of the NLM XML DTD and the relative instability of digital video encoding and players. The article XML refers to the video as an external object known via standard identifier. The manifest handles the specifics of that object. This model would allow a change in video format without a concomitant update to the article XML, for example. In short, this model externalizes the aspects of video content tagging that are most likely to change.

5. PRESERVATION

The questions of how to preserve supplementary materials and where they should be deposited are intimately linked. Online publishers have invested significantly in data models and storage architectures designed to preserve core article content. Accomplishing this requires, among other things, the adoption of some shared standards and basic formats. There is no such uniformity in the area of supplementary materials.

Consortiums and vendors have been slow to volunteer to take the lead in defining standards and offering centralized storage of data sets. Much has been invested in the adoption of XML and an industry standard XML DTD for journal publishing, but no similar effort has been undertaken for datasets. While authors can be relatively confident that the text of their articles will be processable (by a machine) for years to come, there is little reason to be as confident about datasets. The PARSE project (Permanent Access to the Records of Science in Europe) offers some hope in this area, although its scope is currently limited to European research.

Conclusion

Without any consensus on standards or best practices in the handling of supplemental materials, hosting services are left with setting their own requirements for publishers who submit content or with dealing with each publisher’s own policies and practices—a laborious and costly approach. Several projects currently underway are intended to develop standards and best practices for supplemental materials. The NLM Journal Archiving and Interchange Tag Suite version 3.0, the three journal article schemas, and the documentation are currently being shepherded through the NISO standardization process to become a consensus ANSI standard and attract a wider audience of publishers. A joint NISO/NFAIS effort to define best practices for supplemental materials is just underway (see article on page 33). The technical working group for that effort will no doubt consider the issues described in this article. As these standards and best practices are completed, both hosting services and publishers can look forward to adopting some common approaches to the handling of supplementary materials.

doi: 10.3789/isqv22n3.2010.03

ANDREA LAUE <andrea@highwire.stanford.edu> is Manager, Content Services at HighWire Press <highwire.stanford.edu>. She is a member of the NISO/NFAIS Supplemental Journal Materials Technical Working Group.