LSA

Linguistic Society of America

Guidelines for Supplementary Materials Appearing in LSA Publications

Introduction

As the LSA has developed a suite of online journals and proceedings, this has made it easier for authors to provide supplementary materials that support the research articles published therein. This document provides guidance on the format of supplementary files, and suggestions regarding what kinds of information to include in a supplement.

Why include supplementary files?

The purposes of publishing supplemental materials are multiple:

- to support the research arguments made in the article
- to enable replication and reproduction of findings
- to facilitate research that builds on the findings of the published work

Supplementary materials can also improve the readability of a research article by removing technical details that are not necessary to understand the article's argumentation or main points, potentially increasing the article's impact or breadth of appeal by making its findings more accessible to non-specialist readers.

What should be included as a supplement?

Generally speaking, supplementary material is information that is not necessary in order to *understand* the published article, but would be necessary to *reproduce* or *extend* the published research. For example, if an article describes an experimental study and gives one example of each stimulus condition, it might be appropriate to include all stimuli as a supplement.

Another type of information that is well-suited to a supplement is a preliminary analysis that supports the validity of the methods used in the published article: for example, graphs or numerical analyses that confirm the appropriateness of a particular statistical model, or tableaux or tree structures that illustrate alternative analyses that were rejected in favor of the analysis that is argued in the article to be the correct one.

Finally, supplements are useful for inclusion of material central to the main point(s) of a published paper, but that cannot be easily reproduced in a static, 2-dimensional visual form. Thus, although the multimedia capabilities of digital documents are becoming more sophisticated, it is recommended to include content such as audio, video, and interactive material as supplements rather than as embedded elements in the published article, since not all document viewing software fully supports such embedded multimedia, and even when such content is viewable it is not always possible to extract the embedded parts of the document for further analysis.

What should not be included as a supplement?

It is important to note that supplementary materials are governed by the same moral and legal guidelines and constraints that affect ordinary publications (e.g., oversight by Institutional Review Boards, plagiarism, anonymization of data, etc.). In no case should material be included in supplemental files that the author(s) do not have the legal right and institutional permission to publish.

However, even when those standards are upheld, it is not always necessary to include every single computer file associated with a research project as a supplement to the published article. For example, corpora that are already publicly available for research need not be included; it suffices to name the corpus in the published article and provide information on how to access it. Likewise, computer code that performs a statistical analysis that is crucial to the argumentation in the published paper may be appropriate for a supplement, whereas computer code that merely serves to typeset the article document may not be. Authors are encouraged to consult with the Editors if uncertain what to include.

General Guidelines

Metadata

All supplements should include appropriate metadata that defines the interpretation of the supplementary material. In most cases this can take the form of a README or manifest file, though for some types of textual data other forms of metadata are appropriate (e.g., Document Type Definitions, XML schemas, etc.; these are discussed further below). README or manifest files generally do not require extensive formatting and thus should preferably be provided in plain-text format unless there are strong reasons to the contrary.

File naming and directory structure

Supplementary files should be given filenames that are both human-readable and compatible with a wide variety of computer operating systems. This means using only ASCII letters and digits, and avoiding whitespace characters and punctuation marks (other than the underscore ‘_’). Filenames should also not rely on the distinctiveness of upper- and lowercase letters: ‘file.txt’ and ‘File.txt’ may be indistinguishable by some computer systems.

File names (but not directory names) should be given a single suffix, comprising a period character followed by ASCII letters or numbers, and following common conventions for file type extension naming. Examples are .txt (for text files), .tsv (for tab-delimited files), and .xml (for XML files).

When preparing supplements it is acceptable to compress large files (or large numbers of small files). Standard compression algorithms should be used, such as zip (yielding files with the suffix '.zip') or gzip (yielding files with the suffix '.gz'). Multiple files may be combined using the 'tar' program prior to gzip compression; the resulting filename may contain two suffixes, e.g. 'supplement.tar.gz'.

Provenance

In addition to documenting the structure and interpretation of supplementary data, README files should also contain information about the provenance of the data (i.e., where, when, and from whom the data were obtained). Note that ethical or legal considerations may dictate that some aspects of provenance be anonymized or obscured; when this occurs it should be indicated alongside other metadata regarding provenance.

Proprietary Formats

To maximize a reader's ability to use the supplementary files on a wide variety of computer platforms and operating systems, all supplementary files should be provided in non-proprietary file formats. For example, the audio format "Windows Media Audio" (WMA) is a proprietary format, as are the Photoshop image format (PSD) and the compressed "Roshal Archive" format (RAR). Non-proprietary alternatives to each of these formats are described below.

Identifying Language Variety

Metadata in the field of linguistics should include information identifying the languoid that the data represent. In most cases this will involve a language code defined in the ISO 639-3 standard; often more fine-grained descriptions of the language variety will be appropriate (i.e., descriptors of a particular dialect or sociolect). Glottolog identifier codes may prove useful for this purpose.

Textual Data

As used here, the term 'text-based supplementary files' refers to digital files which contain text in one or more languages, and the information included in those files is referred to as 'textual data.' Textual data takes many forms: examples include word or sentence lists, interlinear and other annotated text, dictionary entries, transcriptions of audio or video recordings or broadcasts, textual prompts or responses, word or affix paradigms, and tabulations. Textual data also includes text which is not strictly speaking in a language, such as numbers, artificial language stimuli, and reconstructed or other unattested forms. General guidelines for text-

based supplementary files are given below, followed by specific recommendations for several sub-types of textual data.
Excluded from this definition are statistical analyses, mathematical equations, formulae, and audio, video, and other multimedia files. Computer source code (e.g., a script or program used to perform analysis of the data) may be included in a supplement, but is exempt from the formatting guidelines expressed here due to formatting constraints imposed by the software interpreters of such programs or scripts. Instead, authors are encouraged to conform to whatever community standards exist for their chosen programming language (e.g., the PEP 8 standards for code written in the python language).

General Guidelines for Textual Data

Human- and Machine-readable Text

One of the motivating reasons for providing supplementary material is to facilitate the reproducibility and extension of published research. Consequently, text-based supplementary files should be prepared in a way that facilitates their use by both human researchers and automated systems. As one example, the visual display of a tab character is often indistinguishable from the visual display of a series of 4-5 space characters, and hence inconsistent use of spaces vs. tabs will make little difference to a human reader. Computer systems do not by default treat such characters as equivalent, so care must be taken to prepare text-based supplementary files such that whitespace characters are used in a consistent fashion, ideally in conformance to a publicly-defined standard file format.

Character Encoding

“Character encoding” refers to the system of numeric codes used to store and represent characters (letters, numbers, punctuation, spaces, etc.) in computer memory. In all cases, text-based supplementary files should be encoded in a Unicode-compliant format, such as UTF-8, UTF-16, or UTF-32, with UTF-8 being the format with widest support across applications and operating systems. When working with legacy data, it may be necessary to convert files from an older encoding to a Unicode-compliant encoding.

In the event that a text requires one or more characters which are not (yet) defined in the Unicode standard, those characters may be encoded in the Unicode Private Use Area (PUA). Inclusion of any characters in the PUA must be documented in a README or manifest file accompanying the supplementary data. It may also be worth informing the Unicode Consortium about the lack of support for the character(s) in question.

For unusual scripts (and for all cases of characters in the PUA), researchers should also document a typeface that can be used to properly display the text. Preference should be given to free (gratis) fonts whenever possible; examples of gratis fonts with good coverage of the IPA and non-Roman scripts are the various fonts published by SIL International, and the Noto family of fonts from Google.

If there is any doubt as to whether a text-based file format will be adequate to convey the necessary information, the textual data may also be submitted in Portable Document Format (PDF). However, because of the difficulty in automatically extracting text from PDF for computational processing, PDF should never be the only format in which textual data is made available.

Markup

Any markup included in text-based supplementary files must be documented by providing a description in the README or manifest file, and/or by including or linking to a document-type definition (DTD) or schema enriched by a human-readable description of the meaning of the tags and any attributes on the tags defined by the DTD or schema. If published DTDs or schemas are used, they may be named by reference. For example, schemas published by the Text Encoding Initiative (TEI) may be named by providing a URL; if a subset of such an XML schema is used, it is preferred to include a definition of the subset (e.g., by listing the tags used or by referring to a chapter in the TEI guidelines). If using a meta-standard — such as the ISO Lexical Markup Framework (LMF) for lexicographic data, which provides a Unified Modeling Language description but no reference XML schema — an XML schema that references the meta-standard should additionally be provided. For example, Relish LMF provides an LMF-compliant XML schema, which can be tailored to individual use cases.

Guidelines for Specific Types of Textual Data

Plain text

Many kinds of textual data can be adequately represented as plain text, with no markup or structure other than the arrangement of one datum on each line (i.e., data are separated by some combination of the end-of-line characters ‘carriage return’ and ‘line feed’). Plain text files should bear the file extension ‘.txt’. Simple lists of words, phrases, or sentences are well suited to the plain text file format.

Tabular data

Simple tabular data can be delimited and saved in file format TSV (Tab Separated Values) or CSV (Comma Separated Values). In such files, a second type of delimiter is defined in addition to the end-of-line characters used in plain text documents, which is used to separate each datum from other data on the same line of text. If the delimiter character also occurs within the data, an escape character (usually a backslash \) can be prepended to data-internal delimiters; another approach is to use quoting to group characters that belong to a single datum. The choice of delimiter and escape and/or quoting strategy should be documented in the associated README or manifest file.

XML markup may also be used for tabular data, and is generally a better choice if some table cells span rows and/or columns (as cells in grammatical paradigm tables often do). If XML is used, it is recommended to conform to a published schema; for example, the DocBook schema includes a schema for tables. If an XML schema is used that is not publically defined, the schema definition should also be included in the supplement.

Dictionary Entries

In some simple cases, lexical data can be represented adequately in tabular format. However, this is likely to be problematic if there are repeating values (e.g., multiple senses for a single head word, sub-entries, etc.). Dictionary entries are usually best represented with XML markup, preferably using a standard schema (and validated against that schema). The TEI encoding for dictionaries (chapter 9 of the current encoding) is sometimes used, but is intended to support detailed representations of print dictionaries. For most computational uses, a better choice is the Lexical Markup Framework (LMF), as implemented by an XML schema such as Relish LMF.

Transcripts and Annotations of Audio and Video Files

In some cases it is possible to bundle transcriptions and annotations in special container formats along with the audio or video files they describe. In such cases these formatting guidelines may not be applicable. Whenever possible, it is preferred that transcriptions and annotations of audio or video files should be provided as text-based files separate from the audio or video file that they describe, and should follow the general guidelines for machine-readability, character encoding, etc. In such cases, transcriptions and annotations of audio and video files are considered as textual data.

A time-aligned transcription should include the filename of the audio or video file of which it is a transcript, and links to the definition of the format, either as metadata within the transcript file itself, or in a companion README or manifest file. For example, if the Praat TextGrid format is used for annotation, it should include a link to the TextGrid specification.

Interlinear Glossed Text

Interlinear glossed text (IGT) typically consists of a language line, a gloss line, and a translation line, and is usually typeset with morphemes vertically aligned across lines. In some cases additional variants of the language line are included (items 1-3, as applicable):

1. Language line in standard orthography
2. Language line transliterated, for languages with non-Roman scripts

3. Language line transcribed in IPA
4. Morpheme segmented line
5. Morpheme-by-morpheme gloss
6. Free translation

However, when including IGT in a supplement, some considerations are relevant that do not apply when typesetting IGT examples in a published article. The most prominent difference is that IGT in a supplement need not necessarily be *visually* aligned in the supplementary file, but rather should include structure or markup that describes the intended alignment, in a plain text or XML document. A suggested XML format for IGT is Xigt.

Following are guidelines for the formatting, alignment, and glossing of IGT. It should be noted that this recommendation is about the formatting rules and not the list of suggested abbreviations for gram types (see the Appendix to the Leipzig Glossing Rules for a recommended lexicon of grammatical category abbreviations). Also note that, in service of the goal of data sharing to facilitate extensions of the published research, it is preferable that glosses include as much granularity as possible given what is known about the structure of the language being described.

Finally, authors and editors should be aware that maintaining consistency in glossing particular morphemes becomes difficult, particularly with larger texts, unless specialized glossing tools are used.

Formatting of Lines and Alignment

- The morpheme segmented line and the morpheme-by-morpheme gloss should conform to the Leipzig Glossing Rules.
- There should be the same number of morphemes and morpheme delimiters in the morpheme segmented line as in the morpheme-by-morpheme gloss line.
- There should be the same number of words (white space delimited tokens) on the morpheme-segmented line and the morpheme-by-morpheme gloss line.
- It is possible that the analysis of word boundaries or the conventions of using white space may differ between the standard orthography and the morpheme-segmented line. If not, it is expected that there be the same number of words in the orthography/transliteration lines as the others as well.

Glossing of Grammatical Markers

- Grammatical markers should be glossed with grams and stems glossed as uninflected English lemmas, such that German *gesehen* is segmented as *ge-seh-en* and glossed as ‘PTCP-see-PTCP’ rather than left unsegmented and glossed as ‘seen’.
- Grams should be used consistently across the IGT collection.
- Grams used in the IGT collection should be explained, either in the article the data is supplementary to, or in a README or manifest file associated with the supplement. Where appropriate, it is preferable to use grams that are

standard within the research community to which the article is most closely affiliated.

Language Identification and Provenance

- Each IGT instance should be associated with an unambiguous indicator of what language variety it represents, such as the ISO 639-3 language codes (if available). This can be in a field in the XML or as a tag at the end of the translation line.
- Examples that were not original to the present work should have their provenance clearly marked, including whether the glosses have been adapted.

Audio

Audio recordings of linguistic data or experimental stimuli should be provided in a format that is playable by a wide range of software (audio formats that meet this requirement include WAV, OGG, FLAC and MP3). Within that constraint, choice of format should be guided by the type of analysis that was published and the original fidelity of the recording. In other words, supplementary audio data should be provided in the same format in which it was recorded and/or analyzed. In no case should high-fidelity audio data be converted to a “lossy” format simply to save storage space, nor should audio data that was originally low-fidelity be “up-sampled” to a higher fidelity format (as this merely increases storage size without increasing quality or information content).

Video

Supplementary video files should be provided in a format that is playable across commonly used operating systems (Windows, OS X, Linux) and platforms (desktop and mobile). The free video player software VLC (http://www.videolan.org/) is recommended as a reasonable test of whether a video format can be easily viewed by a broad audience. As with audio files, researchers should submit files in the format used for recording and/or analysis and avoid converting between formats (sometimes called “transcoding”), because this almost always results in a loss of data quality. Researchers are encouraged to consult with the Editor in cases of uncertainty.

Images

Supplementary image files should be provided in a non-proprietary format that is viewable across commonly used operating systems and platforms. Examples are PNG and JPEG for raster images, and SVG for vector images. Within that constraint, choice of file format should be guided by the original type and fidelity of the image. In no case should high-fidelity images be scaled down, converted, or compressed. Likewise, low-fidelity images should not be scaled up.

Interactive Content

For types of multimedia or interactive content not discussed above, researchers are encouraged to contact the Editor for specific recommendations.