

Structuring Supplemental Materials in Support of Reproducibility

Dov Greenbaum^{1,4}, Joel Rozowsky^{2,4}, Victoria Stodden⁶ & Mark Gerstein^{2,3,4,5}

1. Zvi Meitar Institute for Legal Implications of Emerging Technologies, Radzyner Law School, Interdisciplinary Center, Herzliya Israel
2. Program in Computational Biology and Bioinformatics, Yale University, New Haven, Connecticut 06520, USA.
3. Integrated Graduate Program in Physical and Engineering Biology, Yale University, New Haven, Connecticut 06520, USA.
4. Department of Molecular Biophysics and Biochemistry, Yale University, New Haven, Connecticut 06520, USA.
5. Department of Computer Science, Yale University, New Haven, Connecticut, 06520, USA.
6. Graduate School of Library and Information Science, at the University of Illinois at Urbana-Champaign.

I. Introduction

Journal article supplements are becoming an increasingly indispensable resource for researchers. They should be designed to provide essential meta-data and documentation, and act as stand-alone repositories for small data sets. Unfortunately, they often fail to live up to these responsibilities. Lior Pachter has elegantly described many of these missed opportunities in his “Stories from the Supplement” lecture where entire ideas are often contained entirely within the supplement and difficult to find from the main text. *(Please see the supplement for further details).*

Supplements contain a tremendous amount of information, including facts and analyses associated, sometimes only tenuously, with the corresponding published papers. Occasionally, entire projects are inaccessibly buried within.² With some supplements ballooning to multiple times their papers’ length, data within becomes nearly impossible to find. These issues are exacerbated by the often poor editing of the supplements. Further damage is caused when researchers, fearful of burying relevant data in inaccessible supplements, increasingly cram more data into their papers, eschewing the vernacular in favor of terse incoherent terminology. As a result, scientific papers have become more convoluted and unintelligible.

With all these problems, many are calling to curb the use of supplements.^{5 6} We believe this to be shortsighted. Instead, enforcing a considered and standardized approach can make supplements an effective and indispensable tool.

II. Proposal

Supplements have the potential to provide substantial clarity to the published text, not only by providing much needed annotation, but by also providing additional information and data. Even though the supplement will likely never be as precise or as defined as the main text, considerable improvements need to be made across the board. Without the constraints of space, online supplemental material can afford to be clearly written, better organized, and well-documented, allowing for an expanded and useful representation of the research and its results.

Universally accepted structures and standards will substantially expand the usefulness of supplemental materials. With an indexed, searchable, and useful supplement, authors need not jam as much into the main text of the paper, which will result in a more coherent and readable main text. Notably, both the published paper and its supplement can benefit from tying each section in the main text to its corresponding expanded supplement section that contains any corresponding raw data and related information through an established, logical, and linked hierarchy within a parallel structure.

a. FAIR Standards

Employing the FAIR approach for scientific information is essential for guiding the construction of supplements.⁷ Data should be: (i) **Findable** both for human researchers as well as computers, requiring unique and persistent identifiers (e.g., as provided by groups such as CASRAI⁸); (ii) **Accessible for the long term by using appropriate open licensing for** data, code, and workflow information;^{9,10} (iii) **Interoperable** via shared vocabularies, qualified references and shared vernacular; and, (iv) **Reusable** such that both humans and machines can easily use the data for follow-up research or additional computational analysis.

b. Provenance

Veracity of research data requires a complete description of the origins of the data, as well as the process by which that data arrived in its current form *(for example, any data manipulation such as*

- Deleted: As such, they
- Deleted: ,
- Deleted: vital
- Deleted: .
- Deleted: , as a result of the typical state of the supplement in research papers
- Deleted: , which can be quite deep.
- Comment [BC-SSU1]: I wonder if we want to elaborate on this a bit more here to tell the readers what they are looking at. It would be a good spot to put in a brief aside about the supplement that you have created, and how the reader should go about looking at it.
- Deleted: For further details here and herein, please
- Deleted: often
- Deleted: they tend to become too unwieldy to be beneficial
- Deleted: As a result of these and other typical characteristics, data becomes nearly impossible to find.
- Deleted: .
- Deleted:
- Deleted: eventually

- Deleted: In particular, without

- Deleted: To this end, universally
- Deleted: For example, with
- Deleted: –
- Deleted: ting
- Deleted: and expounded
- Deleted: ; i.e., tethering together text with its

- Comment [BC-SSU2]: Should probably spell this out here before introducing the acronym. I actually like the title in the corresponding supplemental section that you use, maybe we could use that here as well?

- Deleted: stewarded, and in particular,
- Deleted: n should be legally Accessible for the long term via appropriate open licensing and other methods of non-inhibiting access

- Deleted: and verifiability
- Deleted: the determination of the provenance of the data used in the research endeavor within the paper, i.e.,
- Deleted: e.g.,

normalizations).¹¹ Provenance allows for assessing data quality, providing an audit trail that could uncover sources of error, the location of all the data relevant for replication of the results, and attribution, necessary for assessing ownership, copyright rights, license limitations, any privacy restrictions, and liabilities, if any, ascribed to erroneous data.

c. Workflows

Understanding the data's provenance can be helped substantially by the inclusion of workflows within the supplement. Supplements should outline, preferably both superficially and in some depth, the individual and collective workflows that produced and employed resources and the final conclusions.¹² Notably, the workflows should be designed to work on at least two levels: as abstract general methods and as more specific schematic representation of a particular computer code. This is an important limitation, as workflows should not necessarily include the code itself as this paradigm sees supplements as important platform but not a repository of data.

Workflows are especially relevant for *in-silico* analyses, as reproducibility can turn on the ability to recreate the exact parameters employed. Abstract workflows, flowcharts and/or comments on the code, and execution infrastructure of the research are necessary.¹³ They should employ standardized identifiers that can be used to reference parts of workflow itself, the relevant datasets and software or any other information useful for cross-referencing workflows and their components. Alternatively, third party, open-source solutions, such as Galaxy,¹⁴ could be used with the supplement providing links to these solutions.¹⁵

d. Language in the Supplement

The supplement should be readable by both human and machine, optimally through the use of distinct formalized languages optimized for each audience.

Even in the predominant English science press, research is conveyed in multiple types of languages, including the simple vernacular language that provides a top-level simplistic understanding, a precise, technical terminology necessary to convey methods to experts and to aid in reproducibility, and increasingly, semi-structured English to aid in computer parsing and automatic text retrieval, indexing, summarization and search. This language is similar to what's been described for the structured abstract.¹⁶ and the structured digital table¹⁸

Length limitations often preclude the adequate provision of these novel aspects of papers, and they are rarely provided within the main text of a document. Since space constraints are less dire within the supplement, it is possible to express the same ideas in multiple iterations and forms. In particular, the same idea can be expressed in multiple "language channels" and additional aspects can be introduced. For example, supplements can provide for relatively simplistic schematic graphics and easy to understand intuitive text that might be unnecessary for the primary audience of the paper, but are necessary to make the information accessible to an increasing number of multidisciplinary outsiders or even the lay public. Likewise, the supplement could also contain paragraphs of excessively precise scientific detail necessary primarily for reproducibility and easier parsing.

To facilitate the use of machine parse-able sections, the supplement would contain a structured glossary connecting all the entities and their languages in the paper; this glossary would correlate with standard database identifiers.

Within the hierarchical structure proposed, one imagines that many of the headings of the supplement could also employ a highly standardized format, further enabling computer parsing and human usability.

- Deleted: (i)
- Deleted: ; (ii)
- Deleted: ; (iii)
- Deleted: ;
- Deleted: (iv)
- Deleted: an important issue in
- Deleted: The exact understanding
- Deleted: of

- Deleted: Both abstract
- Deleted: - ... [1]
- Deleted: both have and
- Deleted: or
- Deleted: (and their version, if pertinent)
- Deleted: and
- Deleted: In the alternative
- Comment [BC-SSU3]: Combine these two paragraphs?

- Deleted: - ... [2]
- Deleted: :
- Deleted: the

- Deleted: even more rarely, are they
- Deleted: However,
- Deleted: where the same ideas can be expressed
- Deleted: in a supplement,

YES

- Comment [BC-SSU4]: Would this be the same glossary that is in the introduction of the supplement or is it a separate glossary to be added to section II of the supplement?
- Deleted: below
- Comment [BC-SSU5]: Combine with previous paragraph

e. Hierarchical Information Structures

Reading a scientific text can be analogous to an information retrieval task, wherein a reader first peruses an introductory section and then jumps into a more detailed version of that section. The current structure of a standard scientific manuscript implements a simplified version of this idea: A short yet still informative title, a more detailed abstract, a somewhat expanding introduction, a detailed result section with detailed tables, and then a conclusion that applies the details, more broadly. The proposed supplement guidelines would expand on this age-old structure, building on this preexisting hierarchy and providing even more levels of information. In a parallel to the main text, the supplement should shadow the paper, providing more detailed explanations for each part of the main text, allowing a reader looking for more detail to easily find and then consult the analogous part of the supplement, which would be similarly situated within the hierarchical structure.

In this methodology, scientific writing would be presented both simply as a hierarchy and, concurrently, as parallel passes at increasingly greater levels of detail. Further, this hierarchy provides an essential roadmap that ought to be familiar across all fields (with well-known section heads such as "introduction," "results," and other standard research paper headings). It would include standardized headings for easy human and machine readability, with the structured headings directly corresponding to headings in the primary paper. Additionally, the supplementary material should be designed to include ample indexable metadata relating various elements within the paper's hierarchy.

Employing an apt literary metaphor: the published paper represents primary classical textual sources, and the supplement mirrors the annotation, gloss and other editorial or academic content on that original text, adding integral, associated, and tangentially relevant context. However, the versatility of the supplement allows it to be more than simply like the annotator's close elucidation on a Shakespearean sonnet, supplements can be useful at the other end of the spectrum as an expansive and sometimes meandering, albeit hierarchically organized Talmud to the published paper's succinctly presented and sometimes cryptically presented Torah. Notably breaking with these metaphors however, supplement authors are also the original authors who can draw perhaps otherwise unseen connections and information.

In some instances supplement hierarchical paradigms can extend beyond that of a single paper to a whole collection of related papers. This becomes all the more relevant as a result of Big Consortia Science where research projects result in high level papers and a succession of more detailed related papers, often across multiple journals. Here all papers can conform to a single global hierarchy with a top-level main paper and more detailed companions.¹⁹ This, in turn, corresponds to various interconnected supplements associated with each individual paper, for example, similar to the structure of the Encode rollout.²⁰ Importantly, this would help illuminate the interconnectivity of each of the individual papers within a series.

f. Proposed Hierarchy

Within the proposed hierarchy, the paper, the supplement and all associated data are each seen as interrelated elements within the larger expansive architecture of a stack or research platform. Thus, in our proposed hierarchy the primary text sits figuratively atop the supplement, synthesizing the entirety of the supplemental information in broad strokes. Other elements sit beneath the supplement within the stack, including software, databases and other elements associated with the research. Local links point to more detailed descriptions of methods and data located further within the supplemental materials.

NOT SURE

Comment [BC-SSU6]: I wonder if you want to move this section earlier. I think that it nicely sets up the supplement hierarchy, and this would be good to have at the start of your proposal section to give the readers a broader outline of what you are thinking before you get into the technical aspects and workflow ideas.

Deleted: seen as

Deleted: moving back out,

Deleted: therein

Moved (insertion) [1]

Deleted: Supplements can provide both: they divide the information into discrete hierarchical chunks allowing readers to avoid reading through a tremendous amount of highly detailed text, and they provide access to expansive relevant and related data.

Deleted: e.g.,

Deleted: How so?

Moved up [1]: In a parallel to the main text, the supplement should shadow the paper, providing more detailed explanations for each part of the main text, allowing a reader looking for more detail to easily find and then consult the analogous part of the supplement which would be similarly situated within the hierarchical structure.

SHORTER

Comment [BC-SSU7]: I like the idea behind these metaphors, but I'm not sure they are needed here. It seems to be a bit repetitive.

Comment [BC-SSU8]: This section also seems to repeat a lot of what has been just said in the previous section. Would you be able to combine these two sections, moving both earlier (to section a, instead of e and f)?

CDISC

INTRO

The detailed description within the supplement that expands upon the top level primary text should be logically sub-divided with each corresponding original paper division addressing a coherent aspect of the analyses. The order of these divisions would map onto the order of appearance within the top-level primary text, allowing researchers to easily move between even a physical version of the supplement and the original paper.

In a secondary hierarchical structure, each of these individual divisions may relate to its own potentially vast amount of supplementary calculations and data sets. These calculations and datasets would be further linked such that they relate back to each division within the supplement and then to the top-level primary text. To promote machine readability of the data sets, data associated with the paper should be provided in a standard tabular format (e.g., CSV), and charts, graphs and other pictorial representations of the data should be decomposable, i.e., accompanied by machine readable files comprising the underlying data. One also can envision shadow tables and figures, which would parallel those in the main text but provide a more expanded layout, with additional detail. *See supplement and figures*

Practically speaking, all data falling within the hierarchy should be localized to a single digital location. When absolutely necessary, for example with regard to sensitive data, hyperlinks can be provided to outside sources. In some cases, the sheer size of intermediate or non-essential data sets may require that some data reside in an off-site website. Here authors should guarantee link viability as has been attempted in other disciplines.²¹

g. Citation Standards

All references in the supplement should be indexed in the standard indexing databases. In some cases citation systems will need to be broadened to allow for pinpointed referencing between the primary text and the supplemental text such that readers of the primary text will be directed from the main text to the relevant section in the supplement, and vice versa using micro DOIs or other referencing systems. To some degree, this can be accomplished through the hierarchical structure, and further simplified through a standardized numbering system, allowing for DOIs of sections, subsections, and even further divisions if necessary. This citation standard can include additional information relating to super-sections, tying together published papers across multiple journals.

h. micro-referencing and micro attribution

With an established hierarchy, different components of the paper and its supplement can be referenced intelligently: clever use of prefixes and suffixes can provide DOI (or similar systems) links to important portions within the supplement.

Unlike the published text, authors can further take advantage of the nature of the supplementary section to provide for μ -referencing of micro-authorship, utilizing ORCID IDs or other persistent unique identifiers to note which specific author contributed to each individual portion of the paper. Not only would this provide a more realistic accreditation of authors than standard author listings, but this would provide interested readers with direct access, perhaps through published email addresses, to the appropriate author for the particular area, text, figure of interest.

Figures would not only include captions and links to relevant parts of the text, but might also include additional information related to the relevant contact individuals for each figure, and access to the source code and data that generated the figure. Again, this would be particularly important with the growing trend to have tens if not hundreds of authors on genomics papers.

Comment [BC-SSU9]: Could this be linked with your section g above?

Supplementary material should also include an expanded bibliography. This bibliography can be designed to provide contextual information both with regard to the paper itself as well as the supplementary material. Additionally, the bibliography can be annotated to provide substantive information as to how each source relates to the presented information. It may be useful to have separate bibliographies for each section of the supplement.

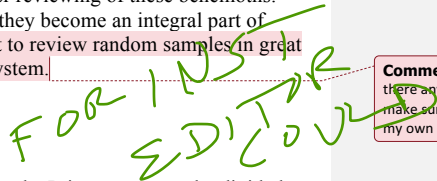


Comment [BC-SSU10]: I wonder if it is worth noting that these references can't yet be input into Pubmed or counted as official citations. I think this is a great idea, but right now there are some potential limitations here.

III. Conclusions

The age of Big Data and Supersized Papers is here. Supplements have become a necessary part of conducting regular scientific business, both from the original researcher's standpoint of presenting their research in its entirety, and also to allow others to effectively use the original research.

The proposals herein represent only some of the changes necessary to maintain the usefulness of supplemental data. One outstanding concern relates to editing and peer reviewing of these behemoths. Detailed review of the supplements will be increasingly necessary as they become an integral part of science. One useful tactic may be detailed sampling: perhaps it is best to review random samples in great detail ensure overall quality without overwhelming the peer review system.



Comment [BC-SSU11]: What do you mean by this, exactly? Is there anything that you think journals could be doing right now to make sure that the supplements are reviewed? (This is mainly for my own information, you don't have to include this).

Figure 1

Figure 1 is a schematic representation of a proposed supplement. Here the Primary text can be divided into one or more subsections, for example, the subsections of common research papers including abstract, introduction, results, discussion, and conclusion. The primary text can be subdivided into other subsections as well related to content or research methodology. The primary text can be further subdivided into additional elements or components, for example tables and figures associated with one or more subsections. The Primary text has, to some degree, a parallel supplement, in that the sections of the supplement should parallel the sections of the primary text, and in most cases, expand on those sections. The supplement expands in these instances, through the addition of related information, including information related to workflows, data, software and additional background. The Supplement should optimally provide both human readable as well as machine readable and parsable text as well as technical information that can be provided without the extraneous vernacular.

¹ <https://liorpachter.wordpress.com/2013/11/02/stories-from-the-supplement/>

² <https://liorpachter.wordpress.com/2013/11/02/stories-from-the-supplement/>

³ Pop, Mihai, and Steven L. Salzberg. "Use and mis-use of supplementary material in science publications." *BMC bioinformatics* 16.1 (2015): 237.

⁴ Newton-Cheh, Christopher, et al. "Genome-wide association study identifies eight loci associated with blood pressure." *Nature genetics* 41.6 (2009): 666-676.

⁵ Maunsell, John. "Announcement regarding supplemental material." *The Journal of Neuroscience* 30.32 (2010): 10599-10600.

⁶ Marcus, E. 2009. Taming supplemental material. *Cell* 139(1):11-11.

⁷ <https://www.force11.org/node/6062>

-
- ⁸ Haak, Laure, David Baker, and Thorsten Höllrigl. "CASRAI and ORCID: Putting the pieces together to collaboratively support the research community." *Procedia Computer Science* 33 (2014): 284-288.
- ⁹ Stodden, Victoria. "Enabling reproducible research: Licensing for scientific innovation." *Int'l J. Comm. L. & Pol'y* 13 (2009): 1.
- ¹⁰ Donoho, David L., et al. "Reproducible research in computational harmonic analysis." *Computing in Science & Engineering* 11.1 (2009): 8-18.
- ¹¹ Bechhofer, Sean, et al. "Why linked data is not enough for scientists." *Future Generation Computer Systems* 29.2 (2013): 599-611.
- ¹² Donoho, D., and Stodden, V., "Reproducible Research in the Mathematical Sciences," in *Princeton Companion to Mathematics*, Edited by Nicholas J. Higham Mark R. Dennis, Paul Glendinning, Paul A. Martin, Fadil Santosa & Jared Tanner, Princeton University Press 2016.
- ¹³ Garijo, Daniel, and Yolanda Gil. "A new approach for publishing workflows: abstractions, standards, and linked data." *Proceedings of the 6th workshop on Workflows in support of large-scale science*. ACM, 2011.
- ¹⁴ <https://galaxyproject.org/>
- ¹⁵ Deelman, Ewa, et al. "Workflows and e-Science: An overview of workflow system features and capabilities." *Future Generation Computer Systems* 25.5 (2009): 528-540.
- ¹⁶ Seringhaus, Michael, and Mark Gerstein. "Manually structured digital abstracts: A scaffold for automatic text mining." *FEBS letters* 582.8 (2008): 1170-1170.
- ¹⁷ Gerstein, Mark, Michael Seringhaus, and Stanley Fields. "Structured digital abstract makes text mining easy." *Nature* 447.7141 (2007): 142.
- ¹⁸ Cheung, Kei-Hoi, et al. "Structured digital tables on the Semantic Web: toward a structured digital literature." *Molecular systems biology* 6.1 (2010): 403.
- ¹⁹ 1000 Genomes Project Consortium. "A global reference for human genetic variation." *Nature* 526.7571 (2015): 68-74.
- ²⁰ ENCODE Project Consortium. "An integrated encyclopedia of DNA elements in the human genome." *Nature* 489.7414 (2012): 57-74.
- ²¹ Zittrain, Jonathan, Kendra Albert, and Lawrence Lessig. "Perma: Scoping and addressing the problem of link and reference rot in legal citations." *Legal Information Management* 14.02 (2014): 88-99.

Workflows

Unsurprisingly, even