Guidelines for a supplement in relation to reproducible research

1) Introduction

Journal article supplements have become an increasingly indispensable resource for researchers, potentially useful not only for fully presenting the full extent of their research, but also as often overlooked alternate repositories of scientific information and data. These repositories are quickly becoming essential for the progress of science, necessitating efforts to develop substantial journal oversight in this heretofore otherwise unstructured area of publication.

Notably, online publication supplements can provide an important and dedicated space for related and relevant information that simply won't fit within the limitations of a particular printed or online publication,[1,2] i.e., *integral content*.[3] For example, in addition to allowing for more text, supplements also provide space for oversized items such as tables, equations, figures, and high resolution images or even unconventional items such as multimedia.

In other instances, the supplement further allows for the inclusion of *additional or associated content*; i.e., material that typically falls outside of standard presentation formats or their publication conventions and that serve to provide content to help provide context and further relevant explanation or background. These materials may include clarifying notes, sequence data, software and its accompanying notation, workflows, failed experiments, and additional multimedia content.[4]

In addition to their use as repositories of an ever-increasing amount of integral and additional data, supplementary material potentially also plays a growing role in the verification and reproduction process of published results. Given the oversized nature of many current scientific efforts, the ability to reproduce and verify research results could become more markedly more effective and efficient if supplementary materials included relevant supporting data, for example, information relating to workflow and computational efforts. In particular, workflow and computation related information frequently represent scientific decision making and assumptions that, if they were to be open to scrutiny, could improve the scientific process and allow follow-on researchers to better extend the results of the originally published research.

Moreover, these two integral components, reproduction and verification, are necessary to combat scientific fraud. This is particularly necessary given the growing realization that plagiarism, falsification and fabrication might be more pervasive throughout scientific culture than previously thought. The policing of science depend crucially on the availability and usability of the underlying raw data, and a sharing of the computational steps that lead from the raw data through to the finished published output. Thus, another reason to invest in properly produced supplements.

HAVE J INTROZ.

Dov Greenbaum 5/2/16 12:09 AM
Deleted: both

DOV 5/1/16 2:50 PM
Deleted: for …ubstantial journal oversight … [1]

Dov Greenbaum 5/2/16 12:10 AM
Deleted: less

DOV 5/1/16 2:50 PM
Deleted: lated

Dov Greenbaum 5/2/16 12:10 AM
Deleted: area.

Eli Greenbaum 4/19/16 11:09 AM
Deleted: These repositories are quickly becoming essential for the progress of sciencecesnecessitating for further journal oversight in this heretofore unregulated area..

Eli Greenbaum 4/19/16 11:09 AM
Deleted: O…line publication supplements … [2]

DOV 5/1/16 2:51 PM
Formatted: Font:Italic

Eli Greenbaum 4/19/16 11:09 AM
Deleted: [[footnote format??]]

Dov Greenbaum 5/2/16 12:11 AM
Deleted: r otherwise…oversized items such … [3]

DOV 5/1/16 2:51 PM
Deleted: some

DOV 5/1/16 2:51 PM
Formatted … [4]

Eli Greenbaum 4/19/16 11:27 AM
Deleted:

Dov Greenbaum 5/2/16 12:12 AM
Deleted: nt…and further relevant explanati … [5]

Eli Greenbaum 4/19/16 11:27 AM
Deleted: . These materials may include: [[too much list]]clarifying notes, video, audio, sequence data, dynamic data, software, or large dataset … [6]

Dov Greenbaum 5/2/16 12:13 AM
Deleted:

DOV 5/1/16 2:53 PM
Deleted: This is particularly necessary give … [7]

Dov Greenbaum 5/2/16 12:14 AM
Deleted: w…uld become more markedly m … [8]

DOV 5/1/16 2:53 PM
Deleted: .

DOV 5/1/16 2:58 PM
Deleted: T…ese two integral components, … [9]

Eli Greenbaum 4/19/16 11:32 AM
Deleted: of the scientific process

DOV 5/1/16 2:59 PM
Deleted: Typically, such information is of … [10]

DOV 5/1/16 2:59 PM
Deleted: . … [11]

Finally, the usage of supplements to provide access to the underlying raw data will become ever more relevant as supplements are used to fulfill journal requirements for the disclosure of the underlying clinical data.[8]

Given all these potential benefits, the deplorable degree to which supplements are overlooked by journals, (e.g., lacking editing, polishing and often even substantive peer review), contrasted with their exploitation by the scientific community as an additional source of important data and commentary, is particularly troubling in light of growing efforts by authors to appropriate this space as an important component of the grey literature, and particularly given the growing use of text mining methods and machine learning algorithms to analyze ever-increasing amounts of data.[9]

Essentially, with much of this data residing in its current state, unedited and unstructured, vast amounts of integral information may be unintentionally ignored.

2) Supplements are now an integral component of modern science

Their relegation to the metaphorical back of the journal notwithstanding, supplements ought to be an integral component of the modern scientific publication and data archiving efforts. Many, if not most articles in popular journals, and especially most genomics journals, include potentially useful if not necessary supplemental materials.

In general supplemental material can be seen as necessary both in terms of preserving and allowing subsequent access to structured and raw data, as well as cataloguing promising (and failed) ideas and directions for further research. Data archiving is also essential in the important goal of reproducing reported findings[10] and also, in developing follow-on research efforts and tangential, or even unrelated research efforts.

To the degree that journals can enforce standards in their supplements, supplements represent untapped potential as usable data archives, particularly for information that might forever be lost from the scientific record. For example, whereas negative results are typically not publishable, they can nevertheless become part of the scientific record through their inclusion in supplements.

3) Current Concerns with Supplements

The many positive aspects of supplements aside, for many journals the size and nature of these supplements are often overwhelming with some publishers now calling for curbs in their use[11] [12].

To some degree this is expected: currently supplements often contain a tremendous amount of data, facts, and analysis associated, sometimes tenuously, with their corresponding published papers. However, since there's no standard format for putting these things in, often many facts simply get lost. Lior Pachter has elegantly described

Eli Greenbaum 4/19/16 11:57 AM
Deleted: T

DOV 5/1/16 3:19 PM
Deleted: shocking

Eli Greenbaum 4/19/16 12:11 PM
Deleted: , both in their general manifestation

Eli Greenbaum 4/19/16 12:00 PM
Deleted:

Eli Greenbaum 4/19/16 12:11 PM
Deleted: but also in their

Eli Greenbaum 4/19/16 12:00 PM
Formatted: Not Highlight

Eli Greenbaum 4/19/16 11:58 AM
Deleted: ,

Eli Greenbaum 4/19/16 12:00 PM
Deleted: gray publishing area,[[[??]]

Eli Greenbaum 4/19/16 12:00 PM
Formatted: Not Highlight

Eli Greenbaum 4/19/16 12:11 PM
Deleted:

Eli Greenbaum 4/19/16 12:12 PM
Deleted: first tier

Eli Greenbaum 4/19/16 12:12 PM
Deleted: and beyond

Eli Greenbaum 4/19/16 5:12 PM
Deleted:

Eli Greenbaum 4/19/16 12:13 PM
Deleted: these areas of gray literature

Eli Greenbaum 4/19/16 5:12 PM
Deleted:

these missed opportunities in his stories from the supplement lecture series[13] where entire ideas, which are quite deep, are often purely contained within the supplement and difficult to find from the main text.

In some instances, references have been made to the Wild West in characterizing the current status quo for supplemental material; [14] for example, with some otherwise short papers including supplemental materials nearly 30 times their length. [15] We believe that these and other issues can be addressed with a more considered approach to supplemental materials, to be described herein.

In addition to the various scientific reasons described above, efforts to rein in and cabin supplements is necessary on a more practical level; given the often disorganized nature of current supplements, writers are often loathe to add important information down the rabbit hole. Perhaps because of this, authors regularly cram as much information as possible into the actual main text of the document. At the very least, this neglect of supplements and their scientific potential can result in making the main text very unreadable through overloading the limited space and writing very tersely. However, to some degree, these fears are founded: supplements often lack this extensive editing and mincing; they tend to be poorly edited and often make finding relevant data even more difficult.

Even with all these concerns, many journals support, if not outright promote the extensive usage of supplements.[16] Broad efforts, such as this one, continue to be made to set up a set of best practices to address a number of aspects related to supplemental material.[17]

4) Best practices for supplemental material

In general, best practices for supplements ought to be designed to deal with the above-mentioned concerns, as well as other pertinent issues particular to supplemental material. These best practices, varying to different degrees, depending on subject matter and audience, should include guidelines relating to the (i) size and format, (ii) scope, (iv) persistence and (v) accessibility of supplemental material. Additional best practices should relate to the (vi) curatorial responsibility of journals, focusing on remedying the general lack of peer review, lack of discoverability, and inability to cite substantial portions of the paper found only in the supplementary materials.

Ours is not the first effort to suggest better administration of supplements. However, a number of concerns specific to genomic oriented journals have been overlooked, particularly in the areas of interoperability, interpretability, reusability, organization, versioning, granularity in large dynamic data sets, and overall standards.

As such, with the growing relevance and importance of supplemental materials in genomic research, we propose a number of additional changes that can be employed in publishing supplements to help make the data and information published therein more useful for the researcher.

DOV 5/1/16 3:21 PM
Deleted: .

Eli Greenbaum 4/19/16 5:14 PM
Deleted: Because of

Eli Greenbaum 4/19/16 5:13 PM
Deleted:

Eli Greenbaum 4/19/16 5:19 PM
Deleted: ;

Eli Greenbaum 4/19/16 5:19 PM
Deleted: p

DOV 5/1/16 3:22 PM
Deleted: In

DOV 5/1/16 3:22 PM
Deleted: is abandonment o

DOV 5/1/16 3:22 PM
Deleted: f

Eli Greenbaum 4/19/16 5:20 PM
Deleted: Often this

Eli Greenbaum 4/19/16 5:19 PM
Deleted:

Eli Greenbaum 4/19/16 5:22 PM
Deleted: Conversely,

Eli Greenbaum 4/19/16 5:23 PM
Deleted: .

Eli Greenbaum 4/19/16 5:23 PM
Deleted: Particular journal inclinations aside,

Eli Greenbaum 4/19/16 5:23 PM
Deleted: b

Eli Greenbaum 4/19/16 5:23 PM
Deleted:

DOV 4/20/16 3:20 PM
Deleted: , among others, particular to ... [12]

DOV 4/20/16 3:20 PM
Deleted: .

DOV 4/20/16 3:20 PM
Deleted: issues

DOV 4/20/16 3:20 PM
Deleted: ,

DOV 4/20/16 3:21 PM
Deleted: ,

DOV 4/20/16 4:01 PM
Deleted: persistence,

DOV 4/20/16 4:01 PM
Deleted: and

DOV 4/20/16 5:53 PM
Deleted: We are not the first to

DOV 4/20/16 3:00 PM
Deleted: reigning in

DOV 4/20/16 5:53 PM
Deleted: W

5) Proposal

We provide, herein, proposed suggestions and suggested standardizations that will be useful in optimizing the usefulness of supplemental materials. As described above, scientific papers tend to become convoluted in their sometimes ineffective efforts toward conciseness. Supplements, if done right, can actually provide substantial clarity to the main published text, not only by providing often much needed annotation, but by also providing for an opportunity for a substantially expanded and understandable version of the published paper. With a recognized and useful supplement, authors need not jam as much raw data and related information as possible into the paper, and as such, the main text can be made all the more readable. This is particularly the case if each section and subsection in the main text can be directly tied to the corresponding expanded section or subsection in the supplement through an established, logical, and linked hierarchy.

Further, authors should not only focus on providing clarity in the main text. Even though the supplement will likely never be as refined a document as the main text, supplemental material ought to be better edited than what is often currently provided in most journals. In particular, without the constraints of space on a published page, online supplemental material can afford to be clearly written, allowing for an expanded and better defined representation of the research and results.

a) FAIR

Much of this proposal shares the goals of the recent FAIR data approach for scientific information that relates to both human and machine analysis of presented data.[18] Succinctly, under this paradigm, scientific data in supplementary material should be: Findable, Accessible, Interoperable and Reusable.

Data should be findable both for human researchers as well as computers— to some degree this requires unique and persistent identifiers for the data and its different parts. Data ought to also be accessible. Here accessibility relates mainly to good data stewardship, and in particular, this means that data should be actually accessible both in terms of long term electronic storage, but also legally accessible in terms of licensing and non-inhibiting access conditions that easily provide for authentication of authorized users.[19,20]

Accessibility also relates to making the underlying software code, often necessary for evaluating the paper's analysis, also accessible. However, while supplemental material should always strive to provide all the relevant information in one place, including a snapshot of the version of the software code used for the analysis, subsequent and further evolving versions of the code should be linked to, perhaps even indexed, but stored separately, perhaps on a specialty site such as github.[21]

Data stored in supplements should also be interoperable; human readers need to clearly understand the connection of the data to the main text. Further, readers should be able to appreciate the nature of the data from the presentation of the data, — i.e., how it can be

EARLIER

DOV 4/20/16 6:27 PM
Deleted: hinted to

DOV 5/1/16 3:29 PM
Deleted: clarity

DOV 4/20/16 6:28 PM
Deleted: put

DOV 5/1/16 4:29 PM
Deleted: information

DOV 5/1/16 3:30 PM
Deleted:

DOV 4/20/16 6:29 PM
Deleted: e

DOV 5/1/16 3:30 PM
Deleted: W

DOV 4/20/16 6:30 PM
Deleted: ,

DOV 5/1/16 3:30 PM
Deleted:

WHAT

vcs 1/28/16 5:47 PM
Comment [1]: I have some work on licensing here: http://stanford.edu/~vcs/papers/RRCiSE-STODDEN2009.pdf and http://stanford.edu/~vcs/papers/ijclp-STODDEN-2009.pdf

combined or compared with other data sets. Further, interoperability requires that the data also be easily able to digested by computational systems, e.g., in an standard that allows for straightforward data manipulation.

Finally, data needs to be reusable. For example, both humans and machines should be able to either apply the data to follow-up research or additional computational analysis.

b) Workflow

Supplementary material should also be designed to incorporate workflow related information. For example, outlining in depth the individual and collective workflows that resulted in the eventual dataset and the published conclusions. Workflows are especially relevant for *in-silico* analyses, as the exact particulars and parameters employed in a workflow can make all the difference between reproducible and non-reproducible data. In this regard, supplemental data should include both abstract versions of workflows, as well as flowcharts or similar representations of the actual executed workflows as they relate to the particular code and execution infrastructure of the lab conducting the research.[22]

Workflows should be directly linked to specific figures and files associated with the paper so that subsequent researchers can review and analyze what transformations, analysis, or other manipulations have already been done to a data set, and similarly, so subsequent researchers can understand the implemented processes that resulted in the figure. For example, a workflow should be able to trace the process from raw data to processed data, to a supplemental table of the processed table, to a figure in the primary paper and finally to the text describing that figure.

Workflows should also have their own standardized identifiers, such that those identifiers include references to the relevant datasets associated with the workflow, any relevant software applied to the workflow, dates that further help to describe the version of the data and the software, and any other relevant information that could be used to cross-reference different datasets and their associated workflows.

Finally, associated with being able to replicate the original workflow, there remains a need for veracity and verifiability of research data, in some instances, provenance of data, i.e., a complete description of the origins of the data, as well as the process by which that data arrived in its current database and current form (e.g., conversions, normalizations ect.) should be tracked as they are collected and repackaged in subsequent research.[23] Provenance is highly relevant to things like: (i) assessing data quality, which can often be estimated based on the source of that information; (ii) providing an audit trail that will allow for an appreciation of the resource usage in putting together the dataset as well as the locating the potential source of any errors in the data; (iii) providing the location of all the data relevant for replication of the results; and (iv) attribution of the resulting data and conclusions, an important issue in assessing ownership, copyright rights, license limitations, and liability, if any, ascribed to erroneous data.

c) Language in the Supplement

DOV 5/1/16 4:20 PM
Deleted:

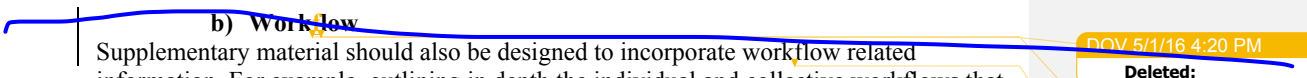
DOV 5/1/16 4:20 PM
Deleted: F

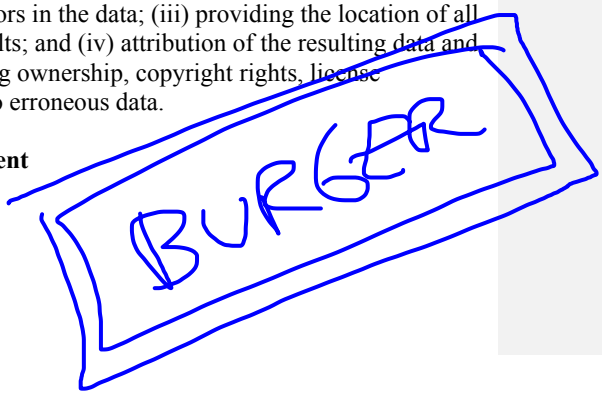
DOV 5/1/16 4:20 PM
Deleted:

DOV 5/1/16 5:07 PM
Deleted: in

DOV 5/1/16 4:20 PM
Deleted:

DOV 5/1/16 4:21 PM
Deleted: ’





A key aspect of scientific writing is language. The nature of scientific progress and the evolution of myriads of micro-disciplines, themselves within numerous sub-disciplines, have resulted in scientific writing that is difficult for the uninitiated to understand. To some degree, this jargon-filled language can be justified as it offers the necessary precision to properly present research, reproduce a result, or for effectively automatically parsing through text.

On the other hand, the broader scientific community would likely appreciate a simpler, more vernacular language that's easier for a more generalized audience to understand and potentially more communicative; allowing for cross-discipline fertilization and perhaps better reflecting the multidisciplinary nature of many current scientific efforts.

Overall, in terms of language, the supplement allows for multiple languages, allowing it to be both easily understood by human researchers, as well as being machine-readable. In some instances, this might be reflected in a standardized hierarchy and standard terminology, in other cases it may necessitate otherwise awkwardly composed machine readable text juxtaposed to human readable text.

The 'Goldilocks problem' of finding the level of jargon just necessary for accuracy, but without alienating the broader readership could potentially be overcome through the effective use of supplements. For example, the supplement can contain a section that provides a jargon-free schematic of the research. Additionally, the supplement can include easy to understand PowerPoint (or similar) presentations that an author might use in the scientific, or even lay, presentation. While the basic information provided in these types of slides is likely not ~~even~~ suitable for the main text, they remain extremely valuable in terms of communicating the ideas to broader audiences. This merger of presentation material with publication material has obvious benefits: The introductory slides, often part of a standard conference talk where the paper might be presented, contains important background information and even historical or scientific context often not included within the actual introductory sections of the published paper. The inclusion of this information is likely to be of substantial value to researchers from other fields. Further, providing additional components of the slide deck from a talk or a number of related presentations could effectively merge a dynamic presentation of the data with the heretofore more static published presentation of the data.

Further, ideally in the supplement, vocabularies, taxonomies, and metadata need to be standardized such that data can be easily read and manipulated across labs, fields and time. To this end, the supplement could also have a very precise glossary, translating language used in the paper into precise, database identifiers and standardized names so that machine text miners can learn for each supplement how to easily parse through that supplement and relate it to a database entry.

d) Presenting the Supplement using hierarchical information structures

DOV 5/1/16 5:38 PM
Deleted:

DOV 5/1/16 5:38 PM
Deleted: people

DOV 5/1/16 5:39 PM
Deleted: and would better

DOV 5/1/16 5:39 PM
Deleted:

DOV 5/1/16 5:39 PM
Deleted: .

DOV 5/1/16 4:53 PM
Deleted: ought to be

DOV 5/1/16 4:52 PM
Deleted:

DOV 5/1/16 5:52 PM
Deleted: a

DOV 5/1/16 5:53 PM
Deleted: the

DOV 5/1/16 5:53 PM
Deleted: is

DOV 5/1/16 5:53 PM
Deleted: but

DOV 5/1/16 5:54 PM
Deleted: are

DOV 5/1/16 5:55 PM
Deleted: P

DOV 5/1/16 5:58 PM
Deleted: all

DOV 5/1/16 5:58 PM
Deleted: the

DOV 5/1/16 5:58 PM
Deleted: a

To understand the overall structure of the supplement, one has to think of scientific writing both in terms of a hierarchy and, concurrently, in terms of parallel passes at increasingly greater levels of detail. Supplements can be both. They can provide a hierarchy in the sense that they divide the information into discrete chunks to allow readers to avoid reading through a tremendous amount of highly detailed albeit potentially irrelevant (to their present interests) text. Additionally, a hierarchy provides a roadmap: reading a scientific text can be seen as analogous to an information retrieval task, wherein a reader first peruses an introductory idea section and then jumps into a more detailed version of that section. To some degree, the current structure of a standard scientific manuscript implements this idea. A vague yet still informative title, a more detailed abstract, a somewhat expanding introduction, a detailed result section with even more detailed tables, and then moving back out, a conclusion that applies the details therein more broadly. The proposed supplement standard would expand on this age-old structure, building onto this preexisting hierarchy and providing even more detail.

This hierarchical structure would translate into the details of its construction: i.e., in a parallel fashion to the main text so that it essentially operates as a shadow text that directly tracks and corresponds to the main text, providing more detailed explanations for each part of the main text. A reader looking for more detail on a particular part of the main text could easily find and then consult the analogous part of the supplement, which would be similarly situated within the hierarchical structure. Using a literary metaphor, the published paper can be viewed as the primary classical text. The supplement reflects the annotation, gloss and other editorial content on that text adding both integral and associated, tangentially relevant content and context. Here however, the author and the editor are one and the same.

Effectively a shadow paper, this hierarchical mirroring can be readily extended to the figures and tables, which can have more detailed contents in the supplement. This idea, of course, of both a hierarchy with increasing level of details and a parallel shadowing flow to it can be extended beyond that of a single paper to a whole collection of papers, as often the case in a large multi-group project where a coauthored high level paper describes the overall structure of the project, and a succession of more detailed papers often across multiple journals describe single, specific, drilled down ideas. With big consortia science project publishing multiple interconnected papers, a global hierarchy for all the papers can be developed, with that global hierarchy then corresponding to various supplements associated with individual papers published in conjunction with a primary roll-out, or even later subsequent papers. This system would also provide a clearer picture of the interconnectivity of the individual papers. Further extending the literary metaphor. The supplement can act as a compiler and editor of a collection of works, providing relevant information to, for example, draw perhaps unseen connections across the body of work.

This proposed hierarchy 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.

Dov Greenbaum 5/2/16 1:09 AM
Deleted: s

Dov Greenbaum 5/2/16 1:09 AM
Deleted: natural

Dov Greenbaum 5/2/16 1:10 AM
Deleted:

Dov Greenbaum 5/2/16 1:10 AM
Deleted: while

DOV 5/1/16 4:55 PM
Deleted: ,

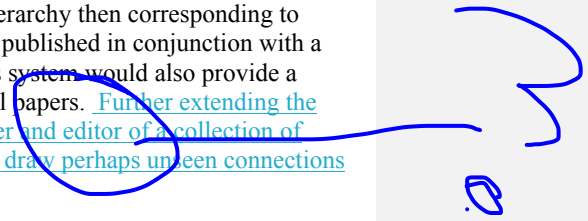
DOV 5/1/16 4:55 PM
Deleted: .

DOV 5/1/16 4:55 PM
Deleted: is

DOV 5/1/16 4:55 PM
Deleted: be

Dov Greenbaum 5/2/16 1:14 AM
Deleted: , which provide more detail

Dov Greenbaum 5/2/16 1:14 AM
Deleted: .



Dov Greenbaum 5/2/16 1:16 AM
Deleted:

Citation standards should 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 supplemental material and readers of the supplemental material will be directed back to the relevant portion of the main text. To some degree, this micro-referencing can be accomplished through an elegant hierarchical structure in the main text that would be shadowed in the supplemental text and/or vice versa. This should be further simplified through a standardized albeit dynamic numbering system, allowing for sections, subsections, and even further divisions if necessary.

Further, this citation standard can include additional information relating to super-sections, tying together published papers across multiple journals and even disciplines. Optimally, publication databases would provide identifiers to not only the main published paper, but would at minimum list the other identifiers associated with the paper.

e) Proposed hierarchy

In practice, in this proposed hierarchy, the primary text sits at the top of the virtual supplement page, synthesizing the entirety of the supplemental information in broad strokes. Local links, for example hyperlinks, therein point to more detailed descriptions of methods and data located further within the supplemental materials.

The detailed description expanding upon the top level primary text should be logically divided such that each division addresses one coherent aspect of the analyses. The order of these divisions would map onto the order of appearance within the top-level primary text. Additionally, the divisions would also map onto the actual published paper, allowing researchers to easily move between the supplement and the original paper. As a bonus a clearer the hierarchy that can be easily mapped onto the original published paper will make adding, editing or modifying hyperlinks by internally and externally that much easier.

In a secondary hierarchical structure, each of these individual divisions may relate to its own huge 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 top-level primary text. Moreover, to promote machine readability of the data sets, data should be provided in a standard tabular format, for example, Microsoft CSV format. Charts, graphs and other pictorial representations of the data should be decomposable, for example accompanied by machine readable files comprising the underlying the images.

Practically speaking, all data falling within the hierarchy should be localized to a single digital location. When necessary hyperlinks can be provided to outside sources, but all supplemental data should fall within the protectorate of the journal's supplement section. In some cases, the sheer size of intermediate or non-essential data sets may require that some data might reside in an off-site website, provided that the authors guarantee

viability to the links. Here, usage of standard widely accepted repositories, an institutionally supported and persistent website, a commercial cloud, or even a shared community repository might be best.

With an established hierarchy, different components of the paper and its supplement can be referenced intelligently, including, for example, distinct digital object identifiers (DOIs) for portions of the paper itself, as well as related identifiers, through the clever use of prefixes and suffixes for related portions within the supplement. The use of these DOIs need not be limited to just text, but can be expanded to include suffixes for related figures, tables, data sets and other related information. DOIs, or similar systems would also be useful given the dynamic nature of the supplement, allowing for the insertion or deletion of information without otherwise complicating the finding of other information. This use of DOIs is especially important in overwhelmingly large supplements that would be too time consuming to actually peruse through to find the desired section, text, figure or other source of information. Here, simply directing the reader to see the supplement, as is unfortunately, all too common, would effectively be a fool's errand without micro-referencing.

Unlike the published text, authors can take advantage of the nature of the supplementary section to provide for micro referencing of micro-authorship, noting which specific authors from the original publication, as well as perhaps, authors not included in the original publication contributed to each individual portion of the paper. Not only would this provide a more realistic accreditation of authors than standard author listings, but would provide interested readers with direct access, perhaps through published email addresses, for each 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 biological papers.

Supplementary material will also include an expand 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.

6) Conclusions

~~The age of Big Data 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 from the for the follow-on researcher to effectively use the original research.

Broadly speaking, as presented, supplements can be seen analogized to the Talmud on the Torah or an annotated version of Shakespeare. The brevity of the first requires substantial outlays in its successor both to provide for the presenting of all the data that can't fit within the precursor text, but also to provide follow-up relevant information for the downstream users of the primary text.

Although we provide a comprehensive wish list for a supplement to deal with the many issues inherent in current supplementary materials one outstanding concern relates to editing and peer reviewing of this unwieldly behemoths. To the extent possible, review of the supplements will be increasingly necessary as they become an integral part of the scientific process, however, given their large size, perhaps it is best to review random samples of the supplement to ensure quality without overwhelming the peer review system.

The popularity of consortia science and the deluge of data that it brings has created an ever-growing need for more structured supplemental data. This is necessary not only for providing FAIR access to important datasets, but particularly for the increasing use of machine learning tools to mine scientific literature. The proposals herein represent only some of the changes necessary to maintain the usefulness of supplemental data.

[1] Borowski, Christine. "Enough is enough." *The Journal of experimental medicine* 208.7 (2011): 1337-1337.
[2] Lior Pachter, Stories from the Supplement, Presentation at Genome Informatics, CSHL, November 1, 2013 available online at https://liorpachter.wordpress.com/2013/11/02/stories-from-the-supplement/
[3] Flanagin, Annette, et al. "Recommended practices for online supplemental journal article materials." *2014-05-07]. http://www. niso. org/apps/group_public/download. php/10055/RP-15-2013_Supplemental_Materials. pdf.*
[4] Kenyon, Jeremy, and Nancy R. Sprague. "Trends in the Use of Supplementary Materials in Environmental Science Journals." *Issues in Science and Technology Librarianship* (2014).
[7] Fanelli, Daniele. "How many scientists fabricate and falsify research? A systematic review and meta-analysis of survey data." *PloS one* 4.5 (2009): e5738.
[8] Taichman, Darren B., et al. "Sharing clinical trial data: a proposal from the International Committee of Medical Journal Editors." *The Lancet* (2016).
[9] Ratner, Mark. "IBM's Watson Group signs up genomics partners." *Nature biotechnology* 33.1 (2015): 10-11.
[10] Alberts, Bruce, et al. "Self-correction in science at work." *Science* 348.6242 (2015): 1420-1422.
[11] Maunsell, John. "Announcement regarding supplemental material." *The Journal of Neuroscience* 30.32 (2010): 10599-10600.
[12] Marcus, E. 2009. Taming supplemental material. *Cell* 139(1):11-11.
[13] https://liorpachter.wordpress.com/2013/11/02/stories-from-the-supplement/
[14] Carpenter, Todd. "Standards Column-Taming the World of Data: Pressures to Improve Data Management in Scholarly Communications." *Against the Grain* 22.6 (2014): 44.
[15] Pop, Mihai, and Steven L. Salzberg. "Use and mis-use of supplementary material in science publications." *BMC bioinformatics* 16.1 (2015): 237.
[16] Hanson, B., Sugden, A., and Alberts, B. 2011. Making data maximally available. *Science* 331:649.
[17] Schwarzman, Alexander B. "NISO/NFAIS Supplemental Journal Article Materials Working Group: A progress report." (2010).
[18] https://www.force11.org/node/6062
[19] Stodden, Victoria. "Enabling reproducible research: Licensing for scientific innovation." *Int'l J. Comm. L. & Pol'y* 13 (2009): 1.
[20] Donoho, David L., et al. "Reproducible research in computational harmonic analysis." *Computing in Science & Engineering* 11.1 (2009): 8-18.
[21] da Veiga Leprevost, Felipe, et al. "On best practices in the development of bioinformatics software." *Frontiers in genetics* 5 (2014).
[22] 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.
[23] Bechhofer, Sean, et al. "Why linked data is not enough for scientists." *Future Generation Computer Systems* 29.2 (2013): 599-611.

DOV 5/1/16 4:58 PM
Formatted: Font:(Default) +Theme Headings CS, 10 pt
DOV 5/1/16 4:58 PM
Formatted ... [13]
DOV 5/1/16 4:58 PM
Formatted ... [14]
DOV 5/1/16 4:58 PM
Formatted: Font:(Default) +Theme Headings CS, 10 pt
DOV 5/1/16 4:58 PM
Formatted ... [15]
DOV 5/1/16 4:58 PM
Formatted: Font:(Default) +Theme Headings CS, 10 pt
DOV 5/1/16 4:58 PM
Formatted ... [17]
DOV 5/1/16 4:58 PM
Formatted ... [16]
DOV 5/1/16 4:58 PM
Formatted ... [18]
DOV 5/1/16 4:58 PM
Formatted ... [19]
DOV 5/1/16 4:58 PM
Formatted ... [20]
DOV 5/1/16 4:58 PM
Formatted ... [21]
DOV 5/1/16 4:58 PM
Formatted ... [22]
DOV 5/1/16 4:58 PM
Formatted ... [23]
DOV 5/1/16 4:58 PM
Formatted ... [24]
DOV 5/1/16 4:58 PM
Formatted ... [25]
DOV 5/1/16 4:58 PM
Formatted ... [26]
DOV 5/1/16 4:58 PM
Formatted ... [27]
DOV 5/1/16 4:58 PM
Formatted ... [28]
DOV 5/1/16 4:58 PM
Formatted ... [29]
DOV 5/1/16 4:58 PM
Formatted ... [30]
DOV 5/1/16 4:58 PM
Formatted: Normal
DOV 5/1/16 4:58 PM
Deleted:
DOV 5/1/16 4:58 PM
Formatted ... [31]