The Five Stars of Online Journal Articles – an article evaluation framework

David Shotton

Department of Zoology, University of Oxford
South Parks Road, Oxford OX1 3PS, UK
david.shotton@zoo.ox.ac.uk

ABSTRACT

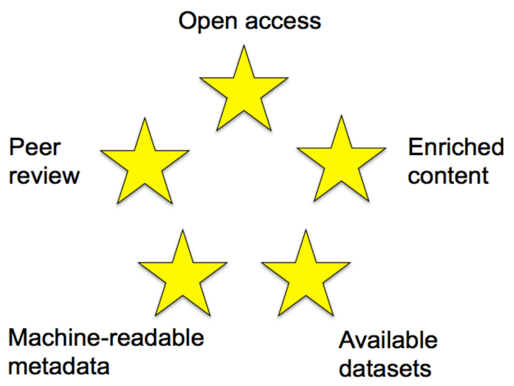
I propose five factors – peer review, open access, enriched content, available datasets and machine-readable metadata – as the Five Stars of Online Journal Articles, a constellation of five independent criteria within a multi-dimensional publishing universe against which online journal articles can be evaluated, to see how well they match up to current visions for research communications. Achievement along each of these publishing axes can vary, analogous to the different stars within the constellation shining with varying luminosities. I suggest a five-point scale for each by which a journal article can be evaluated, and a diagrammatic representation for such evaluations. While the criteria adopted for these scales are somewhat arbitrary, and while the rating of a particular article on each axis may involve elements of subjective judgment, these Five Stars of Online Journal Articles provide a conceptual framework by which to judge the degree to which any article achieves or falls short of the ideal, which should be useful to authors, editors and publishers. I exemplify such evaluations using my own recent publications of relevance to semantic publishing.

1. INTRODUCTION

Many people will be familiar with Tim Berners-Lee's five stars of linked open data [1]. These points provide incremental steps that categorise the publication of data on the web in levels of increasing usefulness, and encapsulate the present shared vision of the semantic web as a web of linked open data[1].

To complement these, I wish to propose **the Five Stars of Online Journal Articles,** in particular to characterize the potential for improvement to the primary medium of scholarly communication made possible by web technologies, including the semantic publishing approaches I have recommended and exemplified in recent presentations[2], blog posts[3] and papers [2-6].

The Five Stars of Online Journal Articles



- ★ **Peer review**
  Ensure your article is peer reviewed, to provide assurance of its scholarly value, quality and integrity.

- ★ **Open Access**
  Ensure others have cost-free open access to your published article, to ensure its greatest possible usefulness and readership.

- ★ **Enriched content**
  Use the full potential of web technologies and web standards to provide interactivity and semantic enrichment to the content of your online article.

- ★ **Available datasets**
  Ensure that all the data supporting the results you report are fully published under an open license, with sufficient metadata to enable their re-interpretation and reuse.

- ★ **Machine-readable metadata**
  Publish machine-readable metadata describing both your article *and* your cited references, so that these can be discovered automatically.

While Tim Berners-Lee’s five stars of linked data are hierarchical, all relating to the same thing and each building on the preceding one, the five stars of online journal articles shown in the diagram above are complementary, forming a constellation arranged along five independent axes within a multi-dimensional publishing universe.

[1] Linked open data: http://linkeddata.org/.
[2] http://imageweb.zoo.ox.ac.uk/pub/2011/presentations/.
[3] Open Citations blog: http://opencitations.wordpress.com/.

2. VALUATION OF THE FIVE STARS

Journal publication, as the primary dissemination channel and public record of new research results, is a vital ingredient of the scholarly workflow, and its key commodity, the original research article, is of primary importance, since it provides a peer-reviewed dated 'version of record' of the authors' hypotheses, data and conclusions at the time of publication, and as such becomes an immutable part of the scientific record.

Recent developments in web technology can be used for the semantic enhancement of scholarly journals articles, providing better linking to other resources; adding descriptive metadata that assist article discovery and specify the meaning of terms and concepts within the article; allowing users access to 'lively' content in the form of interactive figures, re-orderable reference lists, etc.; providing downloadable summaries and numerical datasets in which the data are both accessible and actionable; and supplying machine-readable metadata describing both the article itself and its cited references [3, 4].

However, at present, many publishers primarily use the web simply as a cheap and convenient distribution medium for PDF documents, ignoring its greater potential. As the electronic embodiment of the printed page, the static PDF document is both familiar and easy for humans to read. However, it lacks user interactivity and is difficult for machines to interpret, thus inhibiting the development of services that can automatically link information between articles.

The Five Stars of Online Journal Articles proposed above encapsulate a richer vision. Each star is highly desirable in its own right, but it is only by achieving them all in combination that we will really advance scholarly communication. Of course, the degree of achievement along each of these publishing axes can vary, equivalent to the different stars within the constellation shining with varying luminosities.

Let us now consider how we might score performance against each star. My comments are addressed primarily to authors, but is should be clear to everyone that realization of these publishing goals will require the active and enthusiastic collaboration of journal publishers and editors.

2.1 Peer review

> Ensure your article is peer reviewed, to provide assurance of its scholarly value, quality and integrity.

Anonymous pre-publication peer review is currently being challenged, for example by Cameron Neylon[4], yet stands at the heart of current scholarly publishing practice as the principle quality assurance mechanism applied to journal articles. While the peer review status of an article would at first glance appear to be either true or false, it can have different degrees of completeness and openness, here characterized using a simply five-point scale from 0 to 4:

0 **No peer review**
The article is published without pre-publication peer review, for example in *Nature Preceedings* or on a preprint server such as arXiv.

1 **'Light' peer review**
The article is subjected to thorough review for scientific correctness, but is not evaluated for innovation,. This type of peer review is undertaken, for example, by *PLoS One* and some data journals.

2 **Full peer review**
All aspects of the article are reviewed anonymously by at least two reviewers selected from a panel by the editor or an editorial committee. Most journals adopt this policy.

3 **Full peer review with author responses**
Authors may respond to the reviews before the editor decides whether to accept or reject the paper, e.g. as practiced by *PLoS Computational Biology*.

4 **Open peer review**
Reviewers' names and their reviews are published with the article, reducing the risk of the abuse of anonymity by reviewers, as in the *Semantic Web Journal* and *BMJ Open*.

2.2 Open Access

> Ensure others have cost-free open access to your published article, to ensure its greatest possible usefulness and readership.

The Open Access movement presents the largest challenge to conventional scholarly publishing apart from the web itself. As with peer review, varying degrees of access openness can be rated on a five-point scale:

0 **No public access**
A paper may be circulated privately among colleagues, but is not published.

1 **Subscription access**
The article is published in a subscription-access journal, inaccessible to those who lack personal or institutional subscriptions. The authors' copyright is transferred to the publisher, and preprint publication is not permitted.

2 **'Green' open access**
The subscription-access journal permits authors to self-publish preprints, or post-peer-review 'postprints', in their institutional repositories, preprint servers or elsewhere.

[4] http://cameronneylon.net/blog/peer-review-what-is-it-good-for/.

3 **'Author choice' open access**
For payment of a special fee, the author may elect to publish an article with open access status within an issue of what is otherwise a subscription-access journal.

4 **'Gold' open access**
The work is published as an article in an open access journal that provides full and free open access to all of its articles on the publisher's website.

For most scientific disciplines, 'green' open access is a poor third among the open access choices, because of the difficulty that potential readers have in finding the open versions of such articles. However, new text mining services such as CORE[5] are improving that situation.

2.3 Enriched content

> Use the full potential of web technologies and web standards to provide interactivity and semantic enrichment to the content of your online article.

Web technology can be used to provide various semantic enhancements of scholarly journals articles, links to external information sources of relevance to the textual context, and different types of user interactivity [3, 4].

0 **No enhancements**
The article is published online as a PDF document with no features beyond those that would be found in the print edition of the same article.

1 **Active web links**
The on-line article contains web links to information and web sites of direct relevance, for example authors' home pages, suppliers' catalogues, databases and cited articles.

2 **Semantic enrichment**
Key terms and concepts are identified and distinguished, with pop-ups providing definitions, formulae, database entries, etc. pulled by live web services; reference lists have citation typing.

3 **'Lively' content**
E.g. interactive figures, semantic lenses revealing numerical data beneath graphs, pop-ups providing excerpts from cited papers relevant to the textual citation contexts, re-orderable reference lists.

4 **Data fusions** ("mash-ups")
Integration of the article's data with pre-existing information (e.g. similar data from other articles), geographical location data plotted onto Google Maps, etc.

Of course, semantic enhancement is best integrated during authoring. While writing an article, authors can easily achieve quick wins in terms of functionality by ensuring links to external web resources are provided (e.g. to their own home pages, to reagent suppliers' catalogues, and to cited articles). An open-source plugin to Word 2007 has been published that permits semantic markup of named entities according to chosen ontologies [7], and it is hoped that other such authoring tools will soon become available.

Editors and publishers can also provide semantic enhancements, as exemplified by the Royal Society of Chemistry's Project Prospect journals[6], and by Elsevier's Article of the Future[7].

A post-publication alternative, that can provide semantic markup as annotations over articles presented as conventional static PDF documents, can be achieved by use of a 'smart' PDF reader such as Utopia Documents[8], while third-party web services such as Reflect[9] and OpenCalais[10] can provide automated post-publication markup of named entities in HTML documents 'on the fly'.

2.4 Available datasets

> Ensure that all the data supporting the results you report are fully published under an open license, with sufficient metadata to enable their re-interpretation and reuse.

Through the Brussels Declaration of STM Publishing[11], academic publishers have strongly endorsed the principle that research data relating to journal articles should be made freely available, to enable inspection of the data and validation of the claims made in the article, and to permit data reuse in other contexts. Particularly if the research has been undertaken with public funding, many believe that research data should be regarded as a common good [8, 9]. However, in this enthusiasm for openness, it is important to acknowledge the personal time and effort invested by the researchers who discover or create the data, and their moral right to have the first chance to explore, publish on and benefit academically from the data before publishing them.

The principles of how best to make data available on the web have already been described by Tim Berners-Lee in his five stars of linked data [1], and will not be repeated here. Rather, the following ratings reflect the nature of the data made available.

0 **No published data**
The only data available are those that can be obtained by the reader from within the text of the article itself. Figures and tables are not available for download, nor are any supporting datasets.

[5] CORE: http://core-project.kmi.open.ac.uk/.

[6] http://www.rsc.org/Publishing/Journals/ProjectProspect/.

[7] http://www.elsevier.com/wps/find/authored_newsitem.cws_home/companynews05_01979.

[8] Utopia Documents: http://getutopia.com/documents/.

[9] Reflect: http://reflect.ws/.

[10] OpenCalais: http://www.opencalais.com/.

[11] http://www.stm-assoc.org/brussels-declaration/.

1 **Figures and tables available**
The figures and tables within the article, which may have their own DOIs, are available for download, but only as images, e.g. as is the case from PLoS journal articles.

2 **Article data downloadable in actionable form**
The data contained within the figures, graphs and tables of the article are available in actionable form, for example as downloadable numerical spreadsheets.

3 **Underlying datasets available**
The full research datasets on which the published article is based are published, with sufficient metadata to enable their re-interpretation and reuse.

4 **Data available to peer-reviewers**
These datasets are made available to peer reviewers, to assist in evaluation of the article, prior to their publication at the same time as the article.

Where the data are published is of great importance. Authors should bear in mind the very unsatisfactory nature of journal supplementary information files as repositories for valuable research data, in terms of openness, discoverability, curation, and reliable persistence [10-12]. As safer havens for published data, they should look instead to institutional repositories or, better, subject-specific databases and repositories such as the Dryad Data Repository[12], that curates biological datasets linked to journal articles, makes them available pre-publication to peer reviewers, then publishes them either at the same time as the article, or after an optional embargo period, under a Creative Commons CCZero open data license, with DataCite DOIs[13] to permit proper citation.

2.5 Machine-readable metadata

> Publish machine-readable metadata describing both your article *and* your cited references, so that these can be discovered automatically.

To date, publishers have employed a variety of proprietary XML-based informational models and document type definitions (DTDs) to mark up component parts of electronic documents (author list, abstract, acknowledgements, etc.), but all too often even this basic metadata is not made available to readers, who are given only a PDF version of the article.

Modern web information management techniques employing W3C standards such as RDF[14] and OWL2[15] permit information to be encoded using standard vocabularies in ways that permit computers to query metadata and integrate web-based information from multiple resources in an automated manner. The SPAR (Semantic Publishing and Referencing) ontologies[16] are just some of the vocabularies being used for this purpose to describe scholarly publications [6].

Using these web standards and vocabularies, it is possible to provide semantic descriptions of the structural and rhetorical components of the article using DoCO, the Document Components Ontology[17], and to create and publish machine-readable RDF metadata that describe the journal article itself, i.e. that encode the standard bibliographic information defining the article (authors, publication year, title, journal name, volume number, page numbers, DOI, etc.) using FaBiO, the FRBR-aligned Bibliographic Ontology[18]. It is also possible similarly to encode bibliographic information for all references within the article's reference list, and to use CiTO, the Citation Typing Ontology[19], both to assert the *existence* of a citation between the citing and the cited papers (i.e. `<Paper A> cito:cites <Paper B>`) and also to characterise the type or nature of that citation both factually and rhetorically [5, 6].

Of course, machine-readable metadata need not stop there. There is a growing number of checklists and minimum information standards specifying the information that should be included in research publications within particular domains. One such example is MIIDI, a Minimal Information standard for reporting an Infectious Disease Investigation[20]. Metadata may be structured according to MIIDI to describe either a journal article or a research dataset. In the former case, the metadata can include statements about the main hypotheses of the research investigation, and the principle conclusions described in the article, in addition to providing factual statements concerning the nature of the disease, the number of patients, etc. Such metadata can form the basis for a structured digital summary describing the essence of an article in both human- and machine-readable form, which can be published as an Open Research Report[21].

Available metadata can be rated on the following scale:

0 **No available metadata**
The article is published as a PDF document only. The XML markup used by the publisher during the article production, editing and publication workflow is discarded.

1 **DTD markup available**
The XML markup of the publisher's DTD (document type definition) denoting 'Abstract',

[12] The Dryad Data Repository: http://datadryad.org.
[13] DataCite: http://datacite.org/.
[14] http://www.w3.org/TR/rdf-concepts/.
[15] http://www.w3.org/TR/owl2-overview/.
[16] The SPAR ontologies: http://purl.org/spar/.
[17] DoCO: http://purl.org/spar/doco/.
[18] FaBiO: http://purl.org/spar/fabio/.
[19] CiTO, the Citation Typing Ontology: http://purl.org/spar/cito/.
[20] MIIDI: http://www.miidi.org/.
[21] http://imageweb.zoo.ox.ac.uk/pub/2011/presentations/Shotton-ScienceOnlineLondon2011-OpenResearchReports.pdf.

'Acknowledgements', 'Authors', etc. is included in the XHTML version of the article.

2 **Bibliographic and citation metadata available**
Full bibliographic metadata for the article and full citation metadata for its reference list are published as open linked data.

3 **Rich embedded markup**
Additional structural, rhetorical and semantic markup is available within the online article.

4 **Structured article summary**
A machine-readable summary of the key facts, hypotheses, data and conclusions of the article is made freely available, based on a minimal information standard appropriate for the domain.

By using RDFa[22], it is possible to embed semantic markup within the HTML of web documents in such a way that these machine-readable metadata become part of the web of linked open data. Other possibilities of markup exist using microdata within HTML5 documents. Bibliographic and citation metadata can accompany the relevant journal article as supplementary online RDF files: such files accompany References [4] and [5]. However, as for the research datasets relating to the article, it is advantageous if the relevant metadata files are also submitted to appropriate open linked data repositories, such as those of the Open Bibliography Project [23] and the Open Citation Corpus[24].

3 Evaluating articles against the Five Stars of Online Journal Articles

While the criteria adopted for the evaluation scales presented in Section 2 are somewhat arbitrary, and while the rating of a particular article on each axis may involve elements of subjective judgments, these Five Stars of Online Journal Articles provide a conceptual framework by which to judge the degree to which any article achieves or falls short of the ideal, which should be useful to authors, editors and publishers, who should now ask themselves:

"How do my online journal articles rate against these five stars?"

As an exercise in 'drinking my own champagne', I have evaluated articles [2] to [5] in the following reference list, rating each article on the five-point scale for each star from 0 to 4, and presenting the results in the diagrams and tables that accompany each reference, each having a unique constellation of stars with varying luminosities.

Acknowledgements

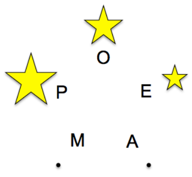
I am most grateful to Bob DuCharme [25] who, inspired by Berners-Lee's five stars of linked data, challenged me to come up with five stars of online journal articles, following a talk I gave recently on semantic publishing. I also thank Tanya Gray and Katherine Fletcher for wise comments after reading preliminary drafts of this paper, and wish particularly to acknowledge the specific input made by Silvio Peroni[26], who insisted that I specify evaluation scales for all five stars, and whose wise proposals concerning peer review and open access I have incorporated.

References

[1] Berners-Lee T (2009). Linked data. Available at http://www.w3.org/DesignIssues/LinkedData.html.

[2] Shotton D (2009) Semantic Publishing: The coming revolution in scientific journal publishing. *Learned Publishing* **22**: 85-94. doi:10.1087/2009202.

Publisher: Association of Learned and Professional Society Publishers.



Rating

Peer review (P)	3	Conventional peer review with author feedback.
Open Access (O)	2	Published in a subscription access journal that permits authors to publish postprints elsewhere.
Enriched content (E)	1	Plentiful web links within the text, and direct links to all referenced articles.
Available datasets (A)	0	Not applicable for this article.
Machine-readable metadata (M)	0	None. Article available in PDF only.

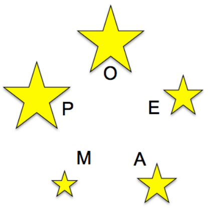
Overall rating 6

[3] Shotton D, Portwin K, Klyne G, Miles A (2009) Adventures in semantic publishing: exemplar semantic enhancement of a research article. *PLoS Computational Biology* **5**: e1000361. doi: 10.1371/journal.pcbi.1000361.

[22] http://www.w3.org/TR/xhtml-rdfa-primer/.
[23] Open Bibliography Project: http://openbiblio.net/2011/06/30/final-product-post-open-bibliography/.
[24] Open Citation Corpus: http://opencitations.net.

[25] Bob DuCharme: http://www.snee.com/bobdc.blog.
[26] Silvio Peroni: http://palindrom.es/phd/.

Publisher: Public Library of Science.



Rating

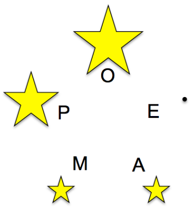
Peer review (P)	4	Conventional peer review with author feedback. (Peer review not open, but given the maximum score of 4 because of the exceptional lengths to which the editor and one reviewer went in helping the authors improve the paper.)
Open Access (O)	4	Published in an open access journal with a Creative Commons attribution license.
Enriched content (E)	2	Plentiful web links in the text, direct links to all referenced articles, figures bearing their own DOIs, and links to examples of semantic enhancement in our enhanced version of Reference [4].
Available datasets (A)	2	Two supplementary information files having their own DOIs, giving technical details of the semantic enhancements made to Reis *et al.* (2008) [4], published with this article.
Machine-readable metadata (M)	1	Structural markup available in the XHTML version of the article.

Overall rating 13

[4] Reis RB, Ribeiro GS, Felzemburgh RDM, Santana FS, Mohr S, et al. (2008) Impact of environment and social gradient on *Leptospira* infection in urban slums *PLoS Neglected Tropical Diseases* **2**: e228.

Original version (http://dx.doi.org/10.1371/journal.pntd.0000228).

Publisher: Public Library of Science.



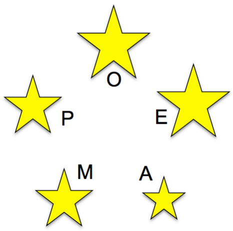
Rating

Peer review (P)	3	Conventional peer review with author feedback.
Open Access (O)	4	Published in an open access journal with a Creative Commons attribution license.
Enriched content (E)	0	Article lacks useful web links in the text and lacks direct links to referenced articles.
Available datasets (A)	1	Figures and table within the article have their own DOIs, but are only downloadable as images, so that the data in the graphs and table are not available as actionable spreadsheets.
Machine-readable metadata (M)	1	Structural markup available in the XHTML version of the article.

Overall rating 9

Semantically enhanced version (http://dx.doi.org/10.1371/journal.pntd.0000228.x001).

Publisher: David Shotton, IBRG, Oxford University.



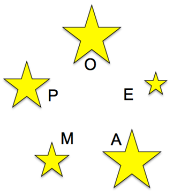
Rating

Peer review (P)	3	Conventional peer review of the original article, with author feedback.
Open Access (O)	4	Enhanced version republished with a Creative Commons attribution license.
Enriched content (E)	4	Of many types, as fully described in Reference [3].
Available datasets (A)	2	Data for table and some figures kindly provided by the authors, and made available as downloadable actionable spreadsheets with their own DOIs.
Machine-readable metadata (M)	3	Structural and semantic markup within the text available as XHTML; embedded RDFa provides basic bibliographic information; two downloadable RDF files with their own DOIs accompany the enhanced article, one giving full bibliographic metadata about the article and the second providing bibliographic details, citation typing information and citation frequencies for all the cited references. Article summary available as a separate document, but not in machine-readable format.

Overall rating 16

[5] Shotton D (2010) CiTO, the Citation Typing Ontology. *J. Biomedical Semantics* **1** (Suppl. 1): S6. doi:10.1186/2041-1480-1-S1-S6.

Publisher: Biomed Central.



Rating

Peer review (P)	3	Conventional peer review with author feedback.
Open Access (O)	4	Published in an open access journal with a Creative Commons attribution license.
Enriched content (E)	1	Plentiful web links within the text, direct links to all referenced articles, and figures and tables bearing their own DOIs, but lacking semantic markup of text and lacking any interactivity.
Available datasets (A)	4	CiTO ontology available online, with a downloadable human-readable supplementary information file with its own DOI providing further explanations, all available to peer reviewers.
Machine-readable metadata (M)	2	Structural markup within the text available as XML; downloadable RDF files with their own DOIs, one giving full bibliographic metadata about the article and the second providing bibliographic details, citation typing information and citation frequencies for the cited references; but no semantic markup, no embedded RDFa and no article summary.

Overall rating 14

[6] Silvio Peroni S and Shotton D (2011). FaBiO and CiTO: ontologies for describing bibliographic resources and citations. (Submitted for publication). Preprint available at http://imageweb.zoo.ox.ac.uk/pub/2011/publications/fabiocito_ontology_paper_PREPRINT.pdf.

[7] Fink JL, Fernicola P, Chandran R, Parastatidis S, Wade A, Naim O, Quinn GB, Bourne PE (2010). Word add-in for ontology recognition: semantic enrichment of scientific literature. *BMC Bioinformatics* **11**: 103. doi:10.1186/1471-2105-11-103.

[8] Boulton G, Rawlins M, Vallance P and Walport M (2011). Science as a public enterprise: the case for open data. *The Lancet*, **377**: 1633-1635. doi:10.1016/S0140-6736(11)60647-8.

[9] "Riding the wave: How Europe can gain from the rising tide of scientific data", Final report of the High Level Expert Group on Scientific Data; A submission to the European Commission, October 2010. Available from http://cordis.europa.eu/fp7/ict/e-infrastructure/docs/hlg-sdi-report.pdf.

[10] Evangelou E, Trikalinos TA, Ioannidis JP (2005). Unavailability of online supplementary scientific information from articles published in major journals. *FASEB J.* **19**: 1943-1944. doi:10.1096/fj.05-4784lsf.

[11] Anderson NR, Tarczy-Hornoch P, Bumgarner RE (2006). On the persistence of supplementary resources in biomedical publications. *BMC Bioinformatics* **7**: 260. doi:10.1186/1471-2105-7-260.

[12] Smit E (2011). Abelard and Héloise: Why data and publications belong together. *D-Lib Magazine* Volume **17**, Number 1/2. doi:10.1045/january2011-smit. (In particular, see Figure 12).