Reports

Defining and Certifying Electronic Publication in Science

A Proposal to the International Association of STM Publishers

Originally Drafted October 1999; Revised March and June/July 2000

Prepared by an International Working Group

Mark S. Frankel, *Co-Chair, American Association for the Advancement of Science, USA*

Roger Elliott, *Co-Chair, International Council for Science, UK*

Martin Blume, *American Physical Society, USA*

Jean-Manuel Bourgois, *Magnard/Vuibert Publishers, France*

Bernt Hugenholtz, *University of Amsterdam, The Netherlands*

Mats G. Lindquist, *Lund University Library, Sweden*

Sally Morris, *Association of Learned & Professional Society Publishers, UK*

Erik Sandewall, *Linköping University, Sweden*

Synopsis

The peer-reviewed article will continue to play a crucial part in the certification, communication and recording of scientific research. However, in the electronic environment it represents one point on a potential continuum of communication. Other points on that continuum (such as preprints) are becoming increasingly common currency, and there is unlimited potential to add to or even change electronic content after it has been made available. All of these versions can be described as 'publications' in the sense that they have been made public. However, this does not necessarily assist the orderly development of scientific knowledge.

The working group was therefore asked to produce some definitions which might be helpful to scientists in this increasingly fluid information environment. We attempted to identify the most important fixed points in the continuum, and the criteria that would need to be satisfied in order to make them useful.

The crucial fixed point, in our view, remains the final published version of an article after peer review (or any future equivalent). We have called this the *Definitive Publication* and believe that it should be clearly identified as such. In the electronic environment, certain other characteristics are also required in addition to peer review:

- It must be publicly available.
- The relevant community must be made aware of its existence.
- A system for long-term access and retrieval must be in place (e.g. Handle).
- It must not be changed (technical protection and/or certification are desirable).
- It must not be removed (unless legally unavoidable).
- It must be unambiguously identified (e.g. by a SICI or DOI).
- It must have a bibliographic record (metadata) containing certain minimal information.
- Archiving and long-term preservation must be provided for.

This is the version to which citations, secondary services, and so forth should ideally point. However, we recognize that earlier versions of an author's work may be made available, and that in some disciplines these are already being cited by other authors. Such early versions might be all that is available to an author for citation at the time of submission of the author's work. However, versions that are not durably recorded in some form, or do not have a mechanism for continuing location and access, or are altered over time (without due provision for version control, as outlined below), should not be regarded as 'publications' in the sense that publication has been defined here, even if cited by an author. We recommend that a version which does satisfy the above criteria should be identified

as such. We have called this the *First Publication*. We recommend that a version satisfying the criteria of *First Publication* (and no other) may be referred to in citations or secondary services, but only until such time as it is superseded by the *Definitive Publication*. Versions which are made public by one means or another, but whose authenticity, retrievability, and permanence are not ensured as outlined above, should not in our view be cited, taken as the basis of claims of priority, or used for purposes of professional evaluation. That version of the *First Publication*, if any, which has been submitted for certification should be clearly identified as such in the bibliographic metadata.

We recognize that many journals (including some of our own) currently cite documents, such as preprints, which satisfy few if any of the above criteria. We would welcome debate on the desirability, and indeed feasibility, of introducing a greater degree of discipline.

We recognize that content can change after, as well as before, *Definitive Publication*. Absence of systematic version control will make life very difficult for scientists. We therefore recommend that errata should be recorded in the accompanying bibliographic record (metadata) and that substantive changes should give rise to a new publication, to which the bibliographic record should refer.

We acknowledge that many unanswered questions remain; no one yet knows exactly how new, dynamic forms of electronic communication can be permanently preserved. However, we would welcome discussion of these proposed criteria, and we would like to see scientists and publishers working together to establish the necessary framework. In particular, we would like to see joint work on the information (metadata) which should be associated with a publication, and on technological solutions for content protection and authentication.

The present task

In October 1998, the American Association for the Advancement of Science (AAAS), the International Council for Science (ICSU) Press, and the United Nations Educational, Scientific, and Cultural Organization (UNESCO) co-sponsored a workshop on developing standards and practices for electronic publishing in science. (A report is posted at http://www.aaas.org/spp/dspp/sfrl/projects/epub/report.htm.) On the basis of that workshop, two of its co-organizers were approached in February 1999 by the International Association of Scientific, Technical, and Medical Publishers to ask them to develop a position paper on how to define a scientific publication in the electronic era. (They had previously commissioned a report from a consultant on the topic.) They agreed, and formed a small working group from among the participants in the earlier workshop and an additional member representing the publishing industry to prepare the position paper. Members of the working group are collectively the co-authors of this paper.

Why it matters

The scientific journal plays a critical role in the advancement of science through its certification and communication of knowledge from author to reader. The electronic medium unquestionably creates added value in publication through the speed with which it can disseminate information, the size of the audience it can reach efficiently, its enhanced indexing and search capabilities, its hypertext linkages to a wide range of material, its ability to be updated and corrected as needed, its interactivity, which enables real-time exchanges between authors and readers, and its multimedia format, which can incorporate video and sound into text. These features are very attractive to scientists, and the number of refereed electronic journals in science, engineering, and medicine has increased dramatically since 1991.

The need to define what constitutes a 'publication' in science in the electronic era is of considerable importance. The enhanced possibilities of electronic publishing are challenging traditional norms and practices that equate scientific publishing with print articles appearing in peer-reviewed journals. Without a definition of publication that takes into account the many forms of scholarly writing found on the internet, the quality, integrity, and authentication of scientific

information communicated electronically will be difficult to determine.

Publication is the hard currency of science. It is the primary yardstick for establishing priority of discovery, making the status of a publication a critical factor in resolving priority disputes or intellectual property claims. Academic tenure and promotion decisions are based in large part on publication in peer-reviewed journals or scholarly books. To make these decisions fairly and with confidence, scientists and their institutions need assurances of what counts as a legitimate electronic publication.

The status of a published electronic document is critical in determining the trust that fellow scientists will have in it. This is increasingly important in the internet environment, where the explosion of information produces a pressing need for efficient and reliable means to distinguish between information that adds usefully to the knowledge base and that which does not. Scientists need to know the status of the information they encounter, whether they need to refer to it, critique it, or build on it to advance their own work. The document also needs to persist, since in science identifying a clear context for later responses is essential to maintain the quality and integrity of subsequent scientific discourse.

Our recommendations

A workable definition of 'publication' in the electronic era is needed to respond to these challenges. Such a definition should be useful to those evaluating the professional work of scientists, and to authors, publishers, librarians, archivists, and readers. The definition of publication that we are proposing has three primary objectives:

- To promote the advancement of science and the social good it serves.
- To contribute to the development of a system for managing scientific information in the electronic environment that will maintain and sustain an accurate and reliable record of science.
- To help resolve some of the existing uncertainty about the status, role, and function of electronic publication in science.

Our recommendations, therefore, are premised on what we believe would be most useful for science. They are not, however, intended to be definitive.[1] The issues are far too complex, the working group too small, and the time too short for producing such wisdom. Nevertheless, we hope that our proposal advances discussion of these matters, enough to push the issues forward to a new level of deliberation. We distinguish between informal notification of one's work (which we do not consider 'publication'), *First Publication* and *Definitive Publication*.[2]

To have any value to the scientific community as a whole, a document should, at a minimum, conform to the following characteristics:

- It must be durably recorded on some medium.
- It must have a persistent access mechanism so that it is reliably accessible and retrievable over time.
- It must be immutable (i.e. it should remain in the same form).
- It must be publicly available.

However, in themselves these characteristics are not sufficient to make the document one that can securely be referred to by other writers; the following additional essential features are required:

- Authenticity must be guaranteed (i.e. versions should be certified as authentic and protected from change after publication).
- Assignment and persistence of an identifier that identifies the work unambiguously.
- A bibliographic record (metadata) that describes the work and its various versions, and which must be public and freely accessible for any given address location.
- A commitment to continuing public access and retrievability.
- Notification of the community that the document is available.
- Commitment not to withdraw the document.
- That version of the document, if any, which has been submitted for a process of

certification should be identified as such in its bibliographic metadata.

In addition, to qualify as a *Definitive Publication*

- It should be vetted (e.g. refereed) to ensure quality, in order to maximize its usefulness for science and to establish a high level of trust among readers.
- There should be a more stringent requirement that the certified version of the document is not subsequently altered. Significant changes should be embodied in a new version with its own identifier and metadata record (the original and new versions should cross-refer). Errata should be registered in the metadata record.
- There must be a commitment to long-term archival preservation.

We realize that in making these proposals there are a number of challenges that lie ahead.

The challenges

Versions

In the traditional print-on-paper paradigm essentially the only version of a publication that merited that name, by virtue of being generally accessible, was a final definitive version which also had all the added value of editorial control, printing, distribution, and marketing. This was, inevitably, the version referred to by subsequent authors. But in the electronic environment this is no longer true, and there are a succession of versions that can be made publicly available without this full array of added value. Nevertheless, their wide availability seems to us to make them a 'publication' in the English language sense. It is therefore important to be able to distinguish among versions, and to identify which, if any, should be treated as definitive.

Quality

To establish its usefulness for science, a publication needs to have been vetted to ensure quality and to establish a high level of trust among readers. This process is equally essential for electronic documents – indeed, perhaps more so in view of the vast quantity of available information. Publication in a peer-reviewed print journal provides this assurance; a reliable equivalent of this 'quality stamping' is necessary in the electronic environment. Various more or less formal processes are being explored and we do not attempt to determine here which might or might not be valid.

Persistence

Methods for archiving and citing electronic publications are challenged by the electronic medium. An archival record of validated scientific work must be accessible for future use, since even the most innovative science is useless if scientists cannot identify, locate, or obtain the work. Yet the ephemeral nature of online publications and changing uniform resource locators (URL) makes citing and accessing information a moving target, unless additional discipline is added. Given the potential for multiple versions of the same document to be available electronically, decisions will also need to be made about practices for linking to and citing versions of a scientific paper.

Version control

We view the publication process as a continuum ranging from an initial 'public offering' of one's work, to claims of priority, to certification of knowledge, to subsequent updatings of work. This process occurs without regard to the medium used. For our purposes, however, we focus on how we believe this process should work in electronic online journal publication.[3]

The process may begin when an author offers his/her work publicly, perhaps by presenting it at a conference, posting it on a personal web page, forwarding it to an electronic listserv, or simply announcing it during a radio or television interview. We do not consider that these actions alone constitute 'publication' for the purpose of establishing the record of science. If scientists want their work formally recognized as contributing to knowledge, there are further steps they must take. In our view, the author has the exclusive right (and responsibility) to take these steps, or to arrange for them to be taken. Once an author decides to make a

particular work available in such a way that his or her community of peers can refer to it, critique it, or build on it, then in our view it must comply with the requirements that we are proposing. Once that is done, it becomes a 'publication'.[4]

First Publication

The author must identify the version that will be the basis for claims to priority.[5] We refer to this as the *First Publication*, and it must be marked by the following properties:

- Recording. The document must be durably recorded on some medium.
- Permanence. The document must be stored in such a way that it remains accessible and retrievable over time.
- Persistent identification. The document must be identified in such a way that can be located over time, even if its web location should change.
- Immutability. The document (including, where technically feasible, any links) should not be altered. (Minor amendments may be permissible to avoid unnecessary proliferation of different versions, but these must be clearly documented.)
- Version control. The document must be clearly identified as the version submitted to be considered for certification.
- Metadata record. The document should be associated with a record containing certain minimum bibliographic information (see below).
- Notification. The community of one's peers must be informed of the version attached to claims of priority.
- Commitment not to withdraw. To ensure an accurate record of science and to discourage a deluge of trivial material into the publication process, authors must agree prior to commencing the selection process that they will not delete the document and all record of its existence from the electronic literature unless there are compelling reasons for doing so. Authors may elect to retract (disavow) or may have to retract a document for scientific, legal, or other reasons, however. In cases of either deletion or retraction, authors should note the reason for doing so in the bibliographic record of that version.
- The version of the document, if any, which is submitted for certification should be clearly identified as such in its bibliographic metadata record.

Once the *First Publication* is determined, the process of selection and certification may begin.

Definitive Publication

Selection and certification is the validation process by which the scientific community identifies work that contributes to the production of useful knowledge. It requires a fair, organized, and recognized vetting process that leads to the definitive (certified) version of a publication. It includes a number of features:

- Peer review, which evaluates the scientific content of the *First Publication*.
- Feedback to authors from peers and editors intended to improve the quality of the publication.

In addition, formal publication (e.g. in a journal) will also include the following:

- Editorial judgements that help to determine the ultimate path taken by the document.
- Copy-editing and design, to improve the accuracy, readability and navigability of the publication.
- Collection, whereby related articles are selected and gathered together in a recognized (physical or virtual) journal or its equivalent for the convenience of readers.

These processes will add significant value to the *First Publication* over and above selection and certification. Some *First Publications* will not survive this selection process. But once certified, this version should be considered the *Definitive Publication* for purposes of establishing the record of science. The *Definitive Publication* must conform with the following additional requirements:

- There must be commitment to long-term archival preservation of the document. We make no assumptions here about how this might best be achieved and by whom; we recognize that substantial technical and funding problems remain to be resolved.

- The document should never be changed and should refer to all previous versions, whether or not retracted.
- Errata subsequently revealed should be appended to this version, with the dates that errata were recorded inserted into the publication's bibliographic record.

Further research that builds on and upgrades this version with new data and findings produces a new publication that must enter the system, secure its own bibliographic record, and earn its own place in the scientific literature. Authors and publishers should jointly develop criteria for determining when changes in content should mark a new publication.

The authenticity of all versions of the publication must be assured. This is critical, since electronic publications are easier to copy and alter than their print counterparts. At present, the technological solutions to achieve this tend to be costly and reader-unfriendly. However, technical and administrative measures may provide guarantees against changes to the content in circumstances where unrecorded change is absolutely unacceptable.

Appropriate technical and administrative measures should be implemented, once they are available, so that readers have confidence both that the version they read has not been tampered with, and that if it purports to be the *Definitive Publication*, it represents precisely the document certified by the selection (vetting) process.

Persistent access mechanism

Public availability and retrievability are essential; if scientists cannot identify, locate, and access the item, whatever version they are seeking, it is useless to the community. There must be a persistent means for locating and accessing the work (even if its web location changes) and, if applicable, for each of its versions. It is the responsibility of the publishing organization to guarantee this. Each publishing organization should have in place a back-up plan in case it is not able to continue to perform this function.

Making the work public means that searchers must be able to find it, whatever version they are seeking. It also means that if the address for the work as a whole is cited in another document, then the reader of that other document at a later point in time must be made aware of, and must be able to obtain access to, later versions of the first work, and not only to those versions that existed at the time the citation was made.

Whoever is responsible for making available the *Definitive Publication* makes a commitment to provide the persistent means to locate the current web address for the document. We recognize that URLs may change. In the future, it is hoped that systems will be developed that make this an intrinsic part of the process of identifying and locating a document. The long-term acceptance and viability of the addressing scheme must be credible, including the existence and proper functioning of a system that produces the bibliographic record when provided with the address.

Archiving and long-term preservation

To be optimally useful to science, publications must be retrievable, now and in the future. Archiving and preservation are necessary to help us identify prior ideas and prior disputes, and to offer a context in which to frame and conduct the debate. The author and other organizations involved must, therefore, make a commitment to archiving and long-term preservation.

An archive of electronic documents will not be static. Changes in technology may require format conversion of archived documents on a large scale. Other, as yet unforeseen, management and updating operations may also become necessary. The rules of archiving must, therefore, include provisions for the freedom to make digital archival copies. It must further be recognized that the continuous migration of technology to higher levels of efficiency and improved capabilities may mean that the format of archived publications will have to be altered in order to be preserved.

Bibliographic record

Unrecorded changes to a document to which scientists refer are not in the best interests of

science. Hence, modifications to the content of a publication should always be recorded in either of the following ways: (i) the creation of a new version, or (ii) the posting of an errata list that is attached to the bibliographic record for the work. In either case the new version or the errata should be dated.

We believe that the bibliographic record, which accompanies or is associated with each version of a given work through the publication process, must give not only generic information about the work, but also minimum information both about that version and about any other extant versions of the same work. The record should contain a subrecord for each published version of the work, indicating in particular the date of publication and location information for that version. Distinct versions should be identified in the following cases or in combinations thereof:

- On submission to a process of formal certification.
- After changes in the contents or presentation of the work.
- When the work is translated into another language.
- If a part of the work is selected as a separate publication.
- Optionally, if the same work is issued in both electronic and print form (these may or may not be considered as distinct versions).

Each subrecord for a version must contain links to the contents of that version in at least one, but possibly several, formats (e.g. PostScript, PDF, or XML). The contents obtained from those links must not be changed over time, but the locations where the contents are stored may change. For example, one organization may commission another to store the contents of some of its published works, thereby transferring contents from one location to another, but it must update the links in the version subrecords accordingly.

The bibliographic record may be, for example, a record in a conventional database sense, or an HTML page that can be viewed using a browser. It could also be an HTML page containing metadata or other hidden data that allow it to be processed effectively by software agents, thus combining the two previous alternatives. Regardless of how it is realized, the record contains references whereby the full contents of all existing versions of the work can be retrieved.

At a minimum, and when pertinent to any particular version, the bibliographic record should consist of the following:

- Author(s).
- Title of the work.
- Subrecords for all versions (at least one).
- Stipulation which of the versions is the *First Publication*.
- Stipulation which of the versions, if any, has been submitted for certification.
- Stipulation which of the versions, if any, is the *Definitive Publication*, by whom/what certified, and when.
- Stipulation of the version for primary citation. This will change as the manuscript moves through the publication process. Normally, it is the *First Publication* until the *Definitive Publication* appears, and then the latter.

Since electronic publications can be continuously updated, improved, and expanded, some system of version control must be in place so that readers are able to quote or cite them with certainty that they are referring to the 'right' versions. The following suggested requirements for subrecords are intended, in tandem with the full bibliographic record, to assist readers:

- Version identifier and date of publication of the version.
- Statement of why the new version has been created, according to the standard criteria for forming new versions mentioned above.
- Date of retraction by the author, if applicable.
- Statement of why the version has been retracted.
- Location(s) where the contents of the version can be obtained. These locations must be updated if the location of the content changes.
- Details and date of errata, if any.
- Reference to other version(s) from which the present one was derived, when applicable.

- Reference to other versions derived from the present one (e.g. translations, subsets, etc.) when applicable.

Who should do it?

For many of these features, it will be desirable to establish uniform standards to ensure as smooth a transition as possible to the proposed system. This task should be undertaken after broad agreement has been achieved on the basic characteristics of the system.

It is important to stress that our recommendations create functions and responsibilities that will require an infrastructure to carry them out effectively and efficiently. Stable and reliable organizations will be needed to undertake the tasks we are proposing; we do not consider it practicable for these tasks to be undertaken by individual authors on their own behalf. We offer no opinion as a group on what the most appropriate and effective infrastructure should be, but publishers, professional associations, research and archival institutions, libraries, and funders of scientific research will all have key roles to play in designing and maintaining this infrastructure.

It is implicit in our proposal, however, that a work should only be considered as 'published' for scientific purposes if the requirements specified above have been performed by an organization such as those outlined in the previous paragraph.

Summary of recommendations

All publications

- Recording. The document must be durably recorded on some medium.
- Publicly available (not necessarily free of charge).
- Immutability (i.e. should remain in the same form).
- Access mechanism so that the publication is reliably accessible and retrievable over time (i.e. through a persistent identifier).
- Version control (bibliographic record must be attached to each version; minimum details indicated above).

When available and affordable technology permits (the development of which should be encouraged) the following should be added:

- Authenticity (i.e. versions should be certified as authentic and protected from change).

First Publication

- Version control. The version of the item submitted for certification, if any, must be clearly identified.
- Notification (the community of one's peers must be informed of the version associated with claims of priority).
- Commitment not to withdraw (authors must agree prior to commencing the selection process that they will not delete the document from the electronic literature).

Definitive Publication

- Quality control/author feedback (it should be vetted to ensure quality).
- Version control (the bibliographic record should identify all previous and subsequent versions, whether or not retracted).
- Errata should be noted in the metadata record.
- Commitment to archiving and long-term preservation.

Notes

1. An analysis of how the law will affect or our proposals is beyond the scope of our original charge. We acknowledge, however, that the system we recommend will have to operate within international and national intellectual property regimes.
2. This is not to say that documents which do not meet the full criteria of *First Publication*, and the ideas which they contain, should not be treated with just as much respect as those which do. It may also be valuable to establish an agreed convention for referring to documents which do not qualify as *First Publications*.
3. Although we acknowledge the value of broadening this analysis to non-journal materials and to media other than online, because of constraints on time and resources we do not consider these here.
4. We recognize that some journals and publishers currently have policies that would preclude their considering for publication documents that have previously been made public by authors in one or more versions, for example by posting to preprint servers. We can only observe here that our definition of online publication is intended to facilitate the widespread dissemination of scientific work.
5. While authors may claim priority of discovery at this stage of the process, the validity of that claim remains to be determined by the vetting process that follows.

The proposals in this report will be of great interest to all learned and scientific publishers and to everyone else in the learned information chain. Reactions and comments, whether supportive or otherwise, will be welcomed by the Editor.