Learned Publishing (2002)**15**, 27–32

New practices for electronic publishing 2: New forms of the scientific paper

Joost G. Kircz
KRA-Publishing Research and Van der Waals–Zeeman Instituut, Universiteit van Amsterdam

ABSTRACT: We extend the analysis of our previous paper (Learned Publishing *2001: 265–72) to a broader discussion of the major features of electronic publications. We conclude that a distinctly different granularity of information will be required and that this will allow very different levels of authentication and quality control from those currently used for the traditional scientific paper.*

Introduction

In the previous paper,[1] we critically discussed the main features of a scientific publication. We based this discussion on the report of an International Working Group.[2] We argued that this report is the near-final description of a scientific publication within the traditional document paradigm. The authors carefully touched on all the important intrinsic issues of a scientific communication and listed the requirements that electronic publications have to fulfil. In our discussion of this report, we extended the argument beyond print on paper, and this resulted in a series of concerns. It illustrated that the transformation of scientific information from paper to an electronic carrier is not a simple projection but implies a complete reconsideration of the way in which scholarly communications are produced and read.

Below, the consequences of electronic preparation, handling, storage, retrieval, and reading are discussed, based on a model developed at the University of Amsterdam.

Towards an understanding of electronic publications

As indicated in the previous paper, we need to appreciate the differences between traditional paper documents and electronic documents, in order to arrive at a full understanding and new guidelines for electronic publishing. This means that we have to abstract from the current accepted daily practice of scientific communication in order, first, to define socially and scientifically acceptable rules of conduct, and subsequently to apply them within the context of a new environment.

The abstract notions of the International Working Group are, of course, correct overall; the problem is in the implementation. This implementation demands a better grasp of the nature of electronic documents. For that reason, we try to advance understanding on this issue in order to make the recommendations specified in the final section of this paper.

The most notable feature of electronic publishing is the integration of text, image, sound and simulations

The greatest step forward in scientific communication is that we are now able to use one carrier for all possible expressions of scientific knowledge. By translating knowledge into binary code, we create a mono-medium that allows us to integrate all kinds of representations. It thus becomes immediately apparent that text will play a less prominent role in the future, and other features, such as images and sounds, will come to the fore. Although language will remain the essential transfer mechanism for knowledge exchange, non-linguistic communication will regain some of the prominence it lost when written language enabled scientific communications to emerge independent of place and time. In the same way that high-quality prints enabled a breakthrough in herbaria and anatomical atlases,[3] the introduction of sounds and simulations will enable us to present relevant information to the reader in a much more realistic way. Fields like phonology, zoology, music and many others will benefit enormously as not only knowledge but also the underlying sensory data can be presented independent of time and place.

underlying sensory data can be presented

In the electronic future, still and moving pictures, sounds, simulations and soon also tactile information can be exchanged and experienced, and therefore analysed and interpreted, by different people in different cultural environments and epochs.[4] This means that a genuine electronic document will be a composition of text as well as different non-textual elements. All these components of the electronic document must adhere to quality and integrity standards. Thus, within the law of proper scientific discourse, all knowledge presentations are equal. To continue this political metaphor, we can say that we certainly need both a diversity policy, to replace the period of positive discrimination in favour of text only, and specific rules for each type of information unit.

This is not the place to dwell at length on the differences between intuitive understanding by means of non-textual stimuli and scientific understanding through linguistic reasoning, but we must realize that non-textual components will play a central role in the electronic document of the future.

As a first step towards creating an environment in which all this can be organized in a meaningful way, we have to consider all the various components as independent but interacting objects. This will lead to a modular approach to information. In a modular information system the various objects are well defined and can therefore be endowed with different sets of metadata, each set describing a different aspect of the information entity.

The next most notable feature of electronic publishing is multiple use

In a traditional environment, an author refers to an earlier author and cites part of the original work by inserting a reference to the original work, quoting part of the original work, or paraphrasing some of the text. This is a typical paper-based process, as it relieves the new author of the need to copy extensively from an already existing text. Only in the case of images, and then often only in review papers, do authors sometimes incorporate a full illustration from another article. In standard publishing practice, the author first requests permission from the original author, and subsequently the publisher requests permission from the original publisher.

However, in an electronic environment, introducing already existing information into a new work is trivial. This is exactly what makes the concept of well-defined modules crucial. In order to keep the integrity of the original work, introducing part of another work means introducing a coherent part, i.e. a complete module.

The difference between quotation and multiple use is that in multiple use the new author can rely on the completeness and integrity of the quoted work. Hence, if, in a

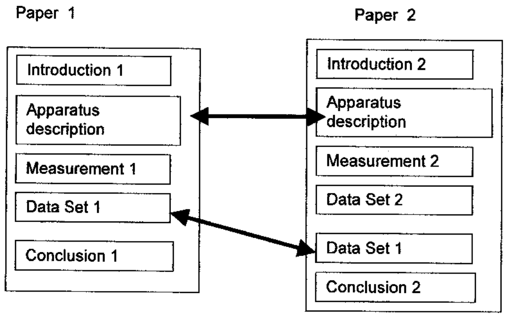

Figure 1 Multiple use

new work, a description of a machine, the working of a medicine, or a mathematical proof is needed, reference to another work adds a new dimension. Traditionally, in quoting another work, the author usually takes a few words and rephrases the quoted text. Now, we can seamlessly introduce the original text into the new work. The old work may be located elsewhere (e.g. in a library), but the network allows us to input this information right where it is needed (see Figure 1).

This means that a module must be compatible with usage in different environments. Thus a link does not *point* to relevant information elsewhere, rather it *transports* information located elsewhere into the present work. This implies that some modules can be represented in different forms, e.g. a data set in its basic form is a list of data; a derived form may be a plot or a histogram.

Modularity as a model for electronic documents

The idea of modularity as the next step in scientific communication[5,6] is further developed by Harmsze[7] who proposes the structuring of scientific articles in modular form. A module is defined as a '*uniquely characterized, self-contained representation of a conceptual information unit aimed at communicating that information*'. This means that a module is a textual, pictorial, or other representation of an amount of information that in itself is sufficiently comprehensive to convey meaning for a reader. Note that neither length nor size enter the definition of a module. Although Harmsze deals mainly with modules that comprise coherent texts, the model is perfectly capable of integrating non-textual modules as well. In the model, a distinction has been made between elementary modules and complex modules. Depending on the purpose, elementary modules can be merged to form complex modules just as atoms bind to form molecules. Two types of such 'bounded' complex modules can be distinguished.

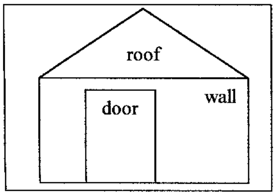

Figure 2 A compound module of a house

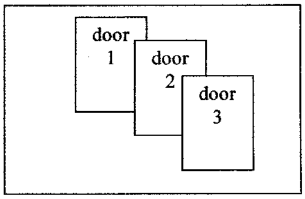

Figure 3 A cluster module of doors

we can seamlessly introduce the original text into the new work

(a) A compound module is a complex module that is an aggregate of (elementary or complex) constituent modules. This is the case if the complex module itself again represents 'uniquely characterized, self-contained' information of a new kind. An easy example is the complex module that describes a computing device and consists of a series of other modules comprehensively describing more-or-less independent components such as the cooling, the memory, the housing, the power transformer, etc. We can compare such a compound module (Figure 2) with a chemical molecule that is unique in itself, but can be analysed as a set of bound molecules and atoms.

(b) A cluster module is a complex module that focuses on a single concept which is a generalization of the specific concepts dealt with in the (elementary or complex) constituent modules.

In this case (see Figure 3), the complex modules host a multiplicity of the same kind of information. An easy example is the

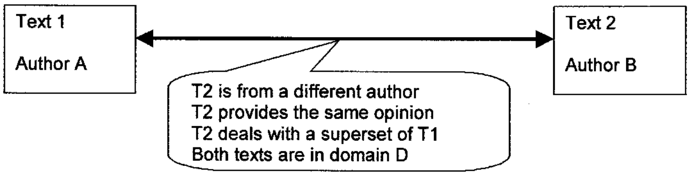
Text 1
Author A
Text 2
Author B
T2 is from a different author
T2 provides the same opinion
T2 deals with a superset of T1
Both texts are in domain D

Figure 4 Example of link types

complex module of a set of PET scans from a particular part of the brain recorded from various patients. Every scan is a module in itself, with its own specific metadata. The complex module disregards the specific, e.g. the patient's name, and concentrates on the common aspects.

We can compare this kind of complex module with the chemical example of a cluster, where we have many identical atoms weakly bound together.

Modularity allows for selected reading paths so that modules can be skipped or emphasized, depending on the reader's wish, expertise, or level of understanding.

Note that we store information units only once! The bottom line is SGML-coded objects that will change their appearance according to the document style definition tailored to the presentation medium.

the reader knows why a link is being suggested by the author

Unfortunately Harmsze's approach is not the end of the analysis. If we discuss multiple use, we have to incorporate other granularities of information as well, even down to a single number. This kind of more 'atomistic' approach is typical of the work on Mathematical Markup Language (MathML) and Chemical Markup Language (CML). Especially in chemistry, where many molecules are discussed and presented, multiple use is a prerequisite for truly electronic publishing. In this field the work of Murray-Rust and Rzepa is worth mentioning.[8,9]

In any event, full modules or a single datum must be identifiable as unique entities in a database. This means that all coherent objects must carry inseparable metadata with them.

Relations as information objects

Now that we have defined the electronic document as a collection of independent information units or modules, the next obvious step is to tackle the mutual relationships between these modules. As a database approach does not necessarily mean that we deal with one physical storage device but that the database objects can be distributed worldwide, it is logical to concentrate on the establishment of a system of relationships that not only connects the modules but immediately defines the type of connection as well.

It is crucial in the following to realize that links are considered to be anchored on both sides – source and target – and can be traversed back and forth. This means that, for example, the characterization 'part of' in one direction indicates 'contains' in the other direction. This is technically still a tedious problem, but within the XML (Extensible Markup Language) environment good progress is being made.[10]

In academic research, part of the game is to relate previously unrelated scientific findings within a new context. In a modular environment, this process can be enhanced. The way to do this is by naming hyperlinks in such a way that the reader knows why a link is being suggested by the author. At present, we have no clue as to why hyperlinks are added; we can only find out by clicking on them. In a structured environment, we will know what the reason for this link is and will be able to decide whether to follow it or not. This brings us to the tedious discussion on hyperlink taxonomies or typologies.

Unfortunately very little has been published in the literature. Most of the initiatives are attempts towards a more-or-less complete list of possible notions (tags). In some works, a distinction is suggested between structural/organizational relations and rhetorical or discourse relations. Our feeling is that, in a distributed database environment, we have to start with a clear differentiation between at least two, and maybe three, categories of relations (see Figure 4). (a) Organizational relations, describing the structural relationship of modules, e.g. hierarchical relations such as part of, etc.; (b) discourse relations describing the reasoning, such as argument for/against, an example, a clarification (the discussion on this issue is ongoing and part of current research – see Harmsze[7] and Kircz and Harmsze[11] and references therein); and (c) context relations describing the context in which a certain relation is valid. Obviously the structure of this last category might be domain dependent.

Conclusions

The goal is to establish clear and transparent understandings of what we mean by a scientific contribution – how we guarantee quality and integrity, and how we value the intellectual ownership of its originator. In these contributions, I have tried to evaluate critically the notion of a scientific document in an electronic environment. The result of my discussion is that we have to step back from the accepted practice of paper journals, but we must maintain societal and scientific standards *vis-à-vis* quality and integrity. People can cut and paste from each other's works much more easily than in the past. This dynamic co-operation needs to be accepted and appreciated as an advance in communications, but it has to fit within the regular framework of quality and integrity.

Instead of trying to curb history by taking a conservative approach, which some publishers try to enforce by their refusal to allow authors to post their own papers on personal websites, we need to be forward looking. The conclusion so far is that we face a transition after which the traditional journal article will cease to exist. This means that we have to reformulate our notions about scientific documentation. In my view, which I defend in this contribution, we have to go for information units with a distinctly different granularity to that allowed by the traditional paper.

1. If we define modules as conceptual units, we can apply strict rules about quality to each type of module. At present, a scientific article is peer reviewed without any discrimination between the various kinds of information in the article. In a world of well-defined modules, the refereeing standard for a module 'mathematical proof' will be distinctly different from that of the module 'data-acquisition'. Thus, quality control will go up as the rules will be more precise and we can even imagine that one referee will look at, for instance, the data-acquisition and reduction, whilst another scrutinizes the theoretical discourse.
2. If all modules are endowed with a set of metadata that clearly identifies the author and time of creation, integration of a module in another work is automatically taken care of with due credit being given. The DOI approach is promising in this respect.[12] Of course, people can always retype, steal, and add fraudulent data, but such misconduct is a social problem and not a scientific one.
3. Another interesting new outcome of this analysis is that relations which express themselves in hyperlinks become information objects in their own right. As relations in an electronic environment can be typified, they become objects with metadata. Thus, we have to add the bibliographic information of the originator and a time stamp. This way, the minimum scientific publication becomes the brilliant insight of a researcher who connects two separate information units by a specified link, without any further business.
4. For documents that are built from available and new modules, we will have two levels of authentication, one on the level of each module and the other on the level of the complete new work. As mentioned in part 1, an interesting scheme in chemistry has been proposed by Gkoukos *et al.*[13]
5. Modular publications will have a list or map of contents with links to all components as well as a new kind of abstract that reflects the content of all modules and serves as an orientation tool in the hypertext environment. Not only is the completeness of the information part of the integrity but so also is the overview

the traditional journal article will cease to exist and a description of the mutual relationships between the components.

Thus we may conclude by saying that electronic media enhance the integration of textual and non-textual knowledge representations, thereby enabling a proper conceptual segregation between various kinds of knowledge and allowing for more specific refereeing. The flip side of these new capabilities is that we have to develop a stable system of domain-dependent metadata for modules and relations that guide the logistics and storage of these modules and relations. We can think back wistfully to the stable situation of established peer-reviewed journals built up over the last century; however, the unknown is the object of science and we are entering a new and unknown phase in scientific communication. Therefore, we have to make sure that our social and scientific demands for quality and integrity are not confused with the latest fashion in technology. Technology is enabling us to expand scientific communication into a serious mix of textual and non-textual components. For most of the non-textual components we do not even have a good insight what the quality standards might be. Like all real advancements in science, the development of scientific communication will go through experimental phases. From the analysis of these experiments we will be able to develop new standards and rules. It is a matter of the highest importance that the scientific community takes this experimentation seriously and neither bows to conservative forces that try to restrict the developments to the known and established practices of the paper world, nor surrenders to the charms and advertising power of software package manufacturers that do not guarantee interoperability between different computer systems.

we will be able to develop new standards and rules

Note

This paper is an edited version of a presentation given at the 2nd ICSU–UNESCO International Conference Electronic Publishing in Science, Paris, 19–23 Feb. 2001 (http://associnst.ox.ac.uk/~icsuinfo/).

References

1. Kircz, J.G. New practices for electronic publishing 1: Will the scientific paper keep its form? *Learned Publishing* 2001:14(4) Oct., 265–72.
2. International Working Group. *Defining and Certifying Electronic Publication in Science. A Proposal to the International Association of STM Publishers*. Published as report. *Learned Publishing* 2000:13(4) Oct., 251–8.
3. Ivins, W. M. Jr. *Prints and Visual Communciation*. Cambridge, MA: MIT Press, 1996.
4. Kircz, J.G. Nouvelles présentations! Nouvelle science? In *L'écrit de la science, Writing Science. Forum Européen de la science et de la technologie* (DGXII), Nice 1998. Alliage no. 37–38, Hiver 1998–Printemps 1999. Pp. 14–24. An English version can be found on: www.science.uva.nl/projects/commphys/papers.
5. Kircz, J.G. and Roosendaal, H.E. Understanding and shaping scientific information transfer. In D. Shaw and H. Moore (eds), *Electronic Publishing in Science: Proceedings of the Joint ICSU Press/UNESCO Expert Conference*, Paris, Feb. 1996. Unesco Press, 1996, pp. 106–16.
6. Kircz, J.G. Modularity: the next form of scientific information presentation? *Journal of Documentation* 1998:54(2) Mar., 210–35. The final draft can be found on: www.science.uva.nl/projects/commphys/papers.
7. Harmsze, F. A modular structure for scientific articles in an electronic environment. Ph.D. dissertation, University of Amsterdam, 2000. The full text and appendices are available at: www.science.uva.nl/projects/commphys/papers.
8. Murray-Rust, P. and Rzepa, H.S. Chemical Markup, XML, and the worldwide web. I: Basic principles. *Journal of Chemical Information and Computer Science* 1999: 39, 928–42.
9. Murray-Rust, P and Rzepa, H.S. Chemical Markup, XML, and the worldwide web. II: Information objects and the CMLDOM. *Journal of Chemical Information and Computer Science* 2001: in press.
10. www.w3.org/TR/NOTE-xlink-req
11. Kircz, J.G. and Harmsze, F. Modular scenarios in the electronic age. In P. van der Vet and P. de Bra (eds), *Proceedings of the Conferentie Informatiewetenschap 2000*, Rotterdam, 5 Apr. 2000. Computer Science Reports 00/20 Eindhoven University of Technology, Department of Mathematics and Computer Science, pp. 31–43. An electronic copy is available at www.science.uva.nl/projects/commphys/papers
12. www.doi.org
13. Gkoukos, G.V., Murray-Rust, P., Rzepa, H.S. and Wright, M. Chemistry Markup, XML and the world-wide web. III: Towards a signed semantic chemical web of trust. *Journal of Chemical Information and Computer Science* 2001: in press.

Joost G. Kircz
KRA-Publishing Research
Prins Hendrikkade 141
1011 AS Amsterdam
The Netherlands
and
Van der Waals–Zeeman Instituut
Universiteit van Amsterdam
Email: kircz@kra.nl
Website:
www.science.uva.nl/projects/commphys/papers