Open Society Institute

Creating Scientific Value with Open Access

A background paper for the Budapest meeting

January 16-18, 2003.

I
Introduction

The recent evolution of scientific publishing has led to a world of have and of have-not by creating a category of « core journals » that is supposed to stand at the heart of the whole of world science: while, indeed, the notion of core journals makes sense at the level of a specialty, aggregating all core journals into a whole that would be called « core science » is far more questionable.

This transformation of the science publishing system has led to several consequences that, on the whole, have been rather bad for science in general; chief among them is the fact that the so-called « core journals » have been reconceptualized by publishing houses as elements of an inelastic market[1] of scientific journals, thereby allowing them to create a very profitable business on the back of a process – namely communicating the results of science – which is largely financed by public money. Fundamental science is still largely supported by governments, including in the USA, and scientific journals are bought by libraries that are mainly located in public institutions. Making some money on this process might be justified if the added value were really significant; however, making a lot of money on such a process is far more questionable, to say the least.

After being erected as « core journals », in particular through the work of E. Garfield and his **Science Citation Index**, these publications have managed to turn into what they claimed to be, in particular thanks to their greater visibility. Indeed, « Core journals » are easier to find because they are listed in the **Science Citation Index**, better known nowadays under its commercial name of « Web of Science ». As a result of this improved visibility, « core journals » have demonstrated a growing ability to attract a fair fraction of the best scientific minds in the world, especially after the method of « impact factors » was developed as a way to create the pecking order of journals (and

1An inelastic market sees little change in demand despite steep price increases. This is a regular result of monopolistic or oligopolistic markets when goods that are absolutely necessary can be secured only from one or very few sources.

despite the fact that impact factors vary widely from discipline to discipline). Once impact factors began to be used by institutions to evaluate their own scientific personnel, « core journals » became a reality that could no longer be ignored. Being quantitative in nature, « impact factors » lent an air of objectivity to evaluations even when they were (and are) conducted carelessly. In effect, after having been described as « core journals », a particular set of publications managed to make the label stick.

Alas, the invention of the « core journals » category also led to a regrettable consequence: while science had long lived with a communication system that continuously extended from mediocrity to excellence, the appearance of the « core journal » category led to a sharp separation between the happy few and the rest. A boundary was set up, largely administered by the choices of ISI (Institute for Scientific Information), and later reinforced by the pricing policies of large commercial scientific publishers. At that point, no one could talk about a continuum of excellence any longer; instead, an elitist system had emerged.

The next stage in this evolution is relatively easy to reconstruct. Given the intense competition among laboratories, large commercial publishers have discovered that temporarily shouldering the cost of launching of new specialty journals could end up being profitable. In this manner, they can offer the possibility of an extra-promotion to some scientists – that of gatekeeper. By inviting them to be editors or members of the editorial boards of these new ventures, publishers bestow upon scientists the possibility of occupying a privileged position within science and, soon, within « core science ». In effect, publishers assist some scientists into becoming a kind of super elite.

Institutions such as universities and research laboratories often eye these developments favorably, if only for prestige reasons (and prestige may translate back into money, for example, from private donors in countries like the USA, to the point that it may justify the added expenses in the library purchases). It also provides a privileged position within a given field of research : the editor of an important esearch journal has a very good view of what is coming from where ahead of most of his colleagues.

Meanwhile, on the publishers' side, the idea is to help the new journal fly long enough to bring it to « core » status; once there, it becomes a profitable venture. Helping launch a new journal, in other words, is an investment decision; beyond the profit motive, it also bears strategic implications in the competition with other publishers; it also bears strategic implications in the kind of relationships a

publisher can maintain with a number of important laboratories. Thus, commercial publishers and a fraction of scientists and laboratories build a covert alliance, the consequence of which is a reinforcement of the elitist system. However, elitism in science wastes a good part of the human potential for scientific excellence. Because it is locked up by financial considerations, it is easily termed a ploutocratic system – a far cry from the « Republic of Science » that emerged in the Scientific Revolution of the 17th century.

In short, one can readily see that the present system of communication in science lends itself to various forms of interventions and manipulations that may skew the scales of excellence that ought to prevail if we wanted to establish the optimal system of science as a world-wide distributed system for the production of fundamental (and some applied) knowledge. Science rests on minds, information and equipment; however, if access to scientific information and equipment is organised to the benefit of a few, then important distortions emerge, the meaning of which may become clearer if we locate ourselves inside what is often described as a « knowledge economy ».

In this meeting, we will not deal with questions of equipment and we will deal with the question of access to the literature only indirectly, and only through the notion of « open access ». For example, other OSI programs try to improve access to some scientific literature through favorable licensing terms. Our main task here will be to examine science publishing with a view to making it better in every way. This means conceiving a system of science publishing that is better than the present one.

What does « better » mean in this context? Essentially, it means creating a system that is far more accessible to all practicing scientists, and not just a few from rich institutions in rich countries; it also means making good publications more effective by making them more visible and more easily retrievable; finally, it means making these good publications more valuable by increasing their ability to create authority, visibility, and even prestige, for their authors.

All the Academies are involved in science publishing, some modestly, others very intensely. All together, Academies represent a very significant fraction of science publishing in the world. The question is whether this publishing effort yields satisfactoy results. Are the authors publishing in Academies' journals easily found and are the articles easily retrieved? In what language(s) are they available/ Are the titles prestigious? Is prestige growing or decreasing?

The point is not to criticize, weaken or compete against the Academies' journals; quite the contrary, the point is to see how to help these journals evolve for the benefit of the Academies and their members. Having journals with well-recognized titles and a good reputation is a valuable asset. Our goal is to examine how best to improve this important and costly publishing effort.

We convened this meeting with only one initial thesis: **it is our belief at OSI that establishing an open-access system of scientific publishing is the necessary first step to achieve all these goals**. We also believe that this goal is now at hand thanks to computers and the Internet. However, we are also conscious that more than open access is needed to create a viable and credible system of science publishing. Open access publishing ought to be financially sustainable and it ought to rest on a system of evaluation which must at least start from peer-review, even though peer review, in itelf, is far from being a satisfactory solution to the various requirements that scientists expect from a good evaluation system. We also believe that in our desire to make things better, we must not hurt what is already good. This is where the input of the Academies will be particularly precious.

To conclude these few introductory remarks, it is important to note that we do not come here to tell anyone how to do things. The meeting is supposed to bring information to you, but, as will be quickly obvious, various solutions are available. Various levels of commitments are possible and no one needs to agree on much if anything at all, short of standards to make these digital documents visible in a universal way. Some solutions entail a public or institutional financial support, while others rely on new, somewhat experimental, business models. Commitments can vary from a minimal involvement in further discussions to being fully committed to a particular publishing scheme. In any case, the Budapest meeting of the Academies aims at bringing up-to-date information on open-access publishing and opening up the possibility of discussing and collaborating on a variety of projects that some or all may find important.

II
Fundamental elements of science publications

In the rest of this background paper, we shall examine the following points which will also be treated systematically in the various presentations that will follow.

1.Production

Producing electronic journals offers many possibilities. For one thing, many, if not most, scientific articles are written in some electronic format, and this means that publishing electronically means manipulating computer files rather than keying in the document anew. However, the choice of file formats is important. To make a long story quite short, two different solutions are widely present on the Internet: the pdf format derived from a proprietary standard developed by the company Adobe and some XML-based solution[2]. The first one is a simple, quick-and-dirty, solution that provides at practically no cost a pdf file that many liken to electronic paper[3]. Its disadvantages are many : full text searching is awkward at best; the files are relatively large and may be a problem for locations with limited bandwidth; also, it is a proprietary standard that, although well documented, depends on Adobe's well-being for its survival, etc.

XML-based solutions are much better from these perspectives : it is an open standard; it is supported by an international body that should be able to lend stability or at least forms of graceful evolution to it. XML files are not excessively large and cross referencing can be done very well in this environment. The downside is that it requires trained personnel that knows XML and it also requires a prior transformation of the original documents, whatever the source may be, and this can be something of a headache in some circumstances[4]. But, unlike pdf files, it is a good archival format and a good base from which to produce any kind of format desired by users.

2.Financing

The question of financing is crucial. Scholarly publishing has been the object of intricate

2Other formats are also commonly found, for example postscript documents that have the disadvantage of being often quite large but can be fed directly to a postscript printer; scientists in mathematics, physics and related domains often use a typographically-oriented system called TeX.

3A very cheap way to produce pdf file consists, on a Unix, and particularly on a Linux system, to intercept the printable file in poscript format and transform it with the open-source software ps2pdf which is readily available in the Net.

4The University of Lyon-2, in France, that was involved in setting up a pilot server in Minsk, offers an interesting path in this regard, which has been tested on doctoral theses. Author(s) using a common word processor, e.g. MSWord or WordPerfect, are asked to use a mall number of prescribed stylesheets that are provided to them. As stylesheets are good practice with word processors, this requirement is generally well accepted, although it may require a little bit of training in a number of cases (no more than a couple of hours at worst, in most cases). The word processor file is then saved in the rtf format, found on practically every wordprocessor under the sun, and is then converted into the format of the open-source office suite Open Office (available at http://www.openoffice.org). Open Office saves its files in a well-formed XML format from which it is quite easy to move to the particular DTD (Document Type Description) used by a particular institution. This becomes the archival format which can then be transformed again for web presentation (for example in the XHTML format) or some other format, including those suitable for professional printing. All the explanations and software tools are, or will shortly be, available to everybody. For details and news, contact Jean-Paul Ducasse (ducasse@univ-lyon2.fr)

financing schemes depending on the context, but most rely on subscriptions to articles bundled under a journal title. For some journals, including the « core journals », this is sufficient as profit is generated. For some journals linked with learned societies and scientific associations, the cost of subscription is part of the membership fee to the association and this amounts to forced sales to members. Finally, in a large number of cases, journals are supported partially by subscription, partially by subsidies from various sources, partially by donated work, office space, secretarial tasks, etc. In addition, some journals require a printing per-page fee from the authors when their article is accepted for publication.

The advent of digital, internet-based publishing has only multiplied the number of financial schemes. Commonly, a site-licensing scheme bundling several journals is offered and such bundles can run up to about 1,700 titles (Science Direct of Reed Elsevier, for example). At the other end of the spctrum, a pay-per-view scheme is sometimes available, allowing to buy access to a particular article even without susbscribing to the whole journal.

In the case of digital, internet-based, publications, open access means that the subscription revenues disappear; however, the printing costs also disappear and so do the costs associated with the need to manage a subscription mechanism (or the need to have someone manage it for you). Preparing the file into a format that makes it presentable on the Web can go from minimal (e.g with the pdf approach) to more expensive if an XML-based solution is adopted. Adding other functions (meta-data – more about this below), cross-referencing, etc. will also add costs. Refereeing costs ought to be about the same as before, although making good use of software available to manage manuscripts can bring the costs down while improving the editorial response time.

To meet these expenses, several solutions can be used and they are not mutually exclusive. For example, Biomed Central, a commercial outfit, proposes to do this kind of work with a per-article charge that covers the costs of refereeing: for $500.00 an accepted (after refereeing) paper will be tagged in a suitable XML DTD, put on line with the right kinds of meta-data (see below) and indexed in PubMed which is freely available. The full text is also freely available to anyone and the copyright stays where it should – namely in the hands of the author. A presentation of the Biomed Central model will be offered in the course of this meeting and, for this reason, there is no need to cover it further in this background paper, but it could be an interesting model to study for various institutions. Indeed, the Public Library of

Science is proposing a similar scheme with its two, newly formed, mega-journals in medicine and in biology.

3.Evaluation

It is a well-known paradox that, beside preserving the memory of science, journals tend to be more useful for the evaluation of scientists than for the communication of scientific research. As soon as a submitted article has been accepted, and in some disciplines, even before the acceptance stage, pre-prints circulate among members of a common, so-called, invisible college[5]. Journals are simply too slow to respond to the communication needs of scientists caught in intense competitive races. On the other hand, journals have come to play an important role in branding authors. In some fields, such as the main bio-medical disciplines, being published in **Nature**, **Science** or **Cell** amounts to a form of consecration; it certainly weighs a great deal in the exercise of peer review applied to grant submissions. In short, prestige, authority and the pecking order of scientists depends in large part on where they have managed to have their articles accepted; it is in this context that the transformation of the excellence continuum into a discontinuous, elitist, two-tier system of publications has been most important. It is hardly an exaggeration to say that it amounts to a pass/fail marking scheme because all that does not appear in core journal simply is negligible as a matter of first approximation[6].

Scientific journals rely on peer review to establish scientific credibility and ensure quality control. Peer review is central to scientific publishing, to the point that it has become the discriminating criterion between scientific (or scholarly) publishing and other forms of publication.

Peer review has also been submitted to a number of criticisms that can be briefly summarized as follows. The least interesting forms of criticism deal with the obvious mistakes that are

5The expression « invisible college » is linked with the pre-history of the Royal Society of London in the 17th century. It was used metaphorically by the sociologist Robert K. Merton to denote groups of scientists (rarely more than a couple of hundred people) that operate in closely related fields and know each other through their publications, conferences and other means. A Merton student, Diana Crane, has written a book-length study of this concept in the '70's.

6This does not mean that it will remeain negligible and the fact that fully 50% of the citations found in the Science Citation Index (SCI) originate in journals not listed in SCI demonstrates that there is life after « core journals », so to speak. However, the fact remains that articles in non-core journals have a harder time being recognized and that the simple fact of being incorporated in the core list immediately lifts the impact factor of any journal enjoying this promotion of sorts.

regularly mentioned as so many anecdotes. They amount to a form of criticism similar to what we sometimes hear about democracy: it is far from perfect, but it is the least imperfect system we know. And that does not teach us very much.

More interesting is the analysis that Paul Ginsparg[7] makes of the peer review process: if we focus on its function as a filter aiming at decreasing the noise in scientific publishing, then we must conclude it does not do its job very well. For experienced readers and beginners alike, the filtering process is too loose to be helpful: for the former, one's knowledge of the field is a much better guide; for the beginners, on the other hand, it is not precise enough to compensate for the lack of experience in the field. Moreover, in most cases, peer review is done under time pressure by people that receive essentially no reward for doing this job: doing peer review is a volunteer's job; it is done out of a sense of duty to the specialty. It may help to learn a few interesting details about coming work in one's own area of specialization, but the advantage is very slight at best. In effect, peer review says: the work is of academic level; it does not include any glaring error and the question raised appears of a sufficient interest to warrant being included in this particular journal. Consequently, Ginsparg argues, peer review amounts to an entry ticket into the scientific « game » at some level of prestige and little else. Better, he says, to allow « refereeable » articles to be allowed to circulate and let the quality be evaluated on the basis of the reception (or lack thereof) it enjoys.

Paul Ginsparg's critique of the peer review process is useful in that it clearly shows the distinction between a system aiming at allowing entry into the scientific game and the evaluation of the work itself through the way in which it is being received, used, criticized, etc.

In the open access perspective, it is clear that peer review should stay, at least for the time being, but that it should be clearly dissociated from other issues such as circulation and usage evaluation. Obviously, an institution has the ability to determine what is of « refereeable quality »; however, it does not have the capacity of imparting the universal scientific legitimacy to an article that peer review offers. This is because peers must be disconnected from the originating institution to ensure a modicum of objectivity and be free from local institutional contexts. As a result, institutions such as Academies could establish their own

7Paul Ginsparg launched an open access archive in high energy physics in 1991. He was then working at the Los Alamos National Library in the United States. Nowadays, he teaches at Cornell university.

internal policies as to what can be put in the public eye on the Internet. In effect, this amounts to putting the institution's good name behind the claim that the correspondin gpaper if of refereeable quality. Once put in circulation, international groups of scientists with the right credentials and credibility – editorial boards in effect -, could easily pick those articles they find most interesting and submit them to peer review; alternatively, authors or institutions could submit a number of refereeable articles to actual peer review from such groups of scientists. In the case of the Academies, such groups of scientists already exist: they are the members of the editorial boards of your existing journals. Presumably, these journals have moved to create editorial boards that are credible on a world wide basis and, because of this credibility, the name of the journal has acquired a certain branding capacity that is recognized at least in parts of the world. Academies can capitalize on these existing structures to exercize peer review. Alternatively, the present Academies could band together and invite yet other bodies to establish recognized and important evaluation boards that would play the rôle of virtual journals. The model of the « Faculty of 1,000 » evolved by Biomed Central is an interesting model in this regard, but it covers only the bio-medical sciences. Similar structures could be built in practically all fields of knowledge.

It is important to remember that branding evaluation is a competitive activity. Even within the Open Acces ambit, competing branding schemes are emerging: the recent announcements made around December 16th of last year by the Public Library of Science will lead to a competition between their mode of branding and that offered by Biomed Central, while both branding systems also run against the branding systems of the biomedical journals owned by various commercial publishers, particularly Reed-Elsevier. There is nothing wrong with this as it corresponds to the fundametal fact that scientific excellence is based on intelletual competition. The question here is not to remove scientific competition; it is to restore an even playing field to the extent possible so as to move away from scientific elitism back to the production of scientific excellence.

Usage evaluation is an entirely new field of activity which, so far, is essentially monopolized by citation analysis and particularly the impact factor methodology. With electronic storage and dissemination of articles, other forms of usage evaluation can be constructed, such as download figures, and citation analysis can be extended well beyond « core journals ». This will require a degree of coordination and cooperation, but the beauty of it is that, with open access, it can be achieved on a distributed basis and with minimal interference with each

existing institution. In this regard, the idea is to work along lines inspired by the Internet and the way it has managed to evolve a large set of complex standards without ever resorting to any centralized situation[8].

The issue of evaluation, in conclusion, is both central and complex. On it rests the present power of the large commercial publishers, based as it is on « core journals ». A frontal attack on this system makes little sense as too much of science and its workings has come to rely on the peer review process. It is better to work at improving it by completing it with other forms of evaluation, and thus ultimately decrease its importance through a process of dilution, so to speak. Such a gradual strategy will lead to a healthier state of world science – one where excellence is restored and elitism is weakened, hopefully to the vanishing point, - and it works particularly well with an open access strategy.

4.Preservation

Publishing in digital formats raises concerns about preservation and, particularly long-term preservation. This question is difficult and important, because of the fast evolution of digital technologies.

Preservation strategies of digital materials generally follow two paths: emulation and migration. In the first case, software emulation of obsolete equipment is produced so as to be able to create a kind of virtual machine on top of existing operating systems and thus gain access to the original files. Of course, such emulations are subject to the same technological instability as the target files and they too require constant updating. A second method is migration: migrating files decipherable by a particular kind of machine to a newer kind of machine can solve the problem, although one must remain aware that some losses may be incurred in the process.

Despite these difficulties, solutions are showing up and one of them, LOCKSS, will be exposed in detail in this conference, as it appears to be one of the more promising approaches in this regard. In effect, it emulates the strategy followed by living species: they aim at

8The only possible exception may be the Domain Name System, but this is largely due to the fact that Domain names have also acted as trademarks and logos and the necessity to regulate this (largely) commercial question has led to searching for the right kind of institution. Knowledgeable observers of the Internet know about all the difficulties and controversies surrounding ICANN. Thankfully, no equivalent of ICANN is needed in scientific publishing.

maintaining the integrity (and value) of a particular DNA combination as long as possible : reproduction, dissemination and selection are the basic tools of this strategy. Thanks to it, very complex messages traverse millions of years with relatively minor changes. We cannot say the same of our most important texts, and these are only a few thousands of years old at best!

5.Visibility and retrieval

The branding capacity of a scientific journal is dependent on its quality, but also on its visibility and the ease with which one can access it. Many factors feed into this question. For example, if you publish exclusively in your national language, you may lose a great many international readers; yet, the national language remains important if only for teaching purposes (and also local ease of access). Bilingual or even multilingual publishing is probably ideal in this regard: both the national language(s) and English appear important.

The beauty of electronic publishing is that it does not cost much more to provide several linguistic versions of the same text and it can often be done locally. In fact, one form of local promotion of local articles could take the form of some sort of access to a translation service allowing those articles to be read world wide. Through this simple device, any Academy can signal to the world that it attaches a special value to particular papers.

Creating eletronic journals must not amount to creating intellectual ghettoes; on the contrary, such journals must be widely available and visible in the world. To achieve this end, systems of meta-data are being designed in such a way that someone entering one particular site belonging to an international federation of scientific sites could then « harvest » every site in one single sweep. This work is largely being carried out at Cornell University under the name of Open Access Initiative (OAI)[9]. An outline of OAI will also be presented in this meeting.

Finally, citation linkage is a powerful way to ensure easy retrieval by colleagues. OpCit is a

9Some confusion has emerged between OAI and BOAI. The former is really concerned with open standards, applied to meta-data and allowing for a better interoperability between various collections of documents. These collections could be proprietary and locked up behind firewalls and password protections. However, for open access collections, OAI standards are eseential because they are the needed glue to prevent the whole effort to crumble into a dust of small intellectual ghettoes. BOAI or Budapest Open Access Initiative is a statement that was drafted as a result of a meeting that was convened by the Open Society Institute here in Budapest in December 2001. The statement was released on February 14th and t has become the rallying point of those supporting an open access approach to scientific literature.

project related to OAI and the Open Access Movement; it rests on software tools that allow each author to recordin the citations in his or her article. The idea is that each author, upon completing his or her article, would take the time to enter the list of citations into this system so as to create a vast, world-wide, distributed science citation index that would be freely available to all. This too would work toward diluting the excessive power presently monopolized by SCI.

All these instruments and many other yet to be invented aim at improving the visibility and retrievability of scientific articles. As mentioned earlier, these elements are crucial for the building of authority, prestige, and, ultimately, of a strong branding capacity. Reforming the science communication system amounts to wrestling a significant part of the branding capacity of scientific journals away from commercial publishers that pursue unreasonable pricing policies.

6.Transition from the existing publishing situation

This is probably the most complex question and it ultimately rests in great part on the specific characterisitics of local situations. This is where the Academies start from. However, to help think more clearly through this process, several examples of various kinds of projects and implementaitons will be presented. It is also a suitable coda for this background paper as it brings all of us to the ultimate and essential question: now what? The discussions and round tables that will take place in the next few days will aim at beginning a process that, hopefully, will only grow in strength in the coming years and will involve as many of your institutions as possible.

Welcome to the Budapest Meeting of the Academies[10]!

10This background paper was prepared by Jean-Claude Guédon with the help of Melissa Hagemann.