SECURING AUTHENTICITY OF SCHOLARLY PATERNITY AND INTEGRITY

Anthony Watkinson

September 2003

Resource: The Council for Museums, Archives and Libraries

BOOK INDUSTRY COMMUNICATION

Library and Information Commission Research Report 148
British National Bibliography Research Fund Report 106

BR/008

ISBN 1 873671 31 8
ISSN 1466-2949 and ISSN 0264-2972

Library and Information Commission Research Reports are published by Resource: The Council for Museums, Archives and Libraries and may be purchased as photocopies or microfiche from the British Thesis Service, British Library Document Supply Centre, Boston Spa, Wetherby, West Yorkshire, LS23 7BQ

This report is co-published by Book Industry Communication and is available in PDF format free of charge at the following URL:
http://www.bic.org.uk/securing authenticity.pdf

Book Industry Communication
39-41 North Road
London N7 9DP
UK

www.bic.org.uk

Contents

Executive Summary

This study is concerned with scholarly communication, in particular in the sciences, where an article in a learned journal is the main vehicle for formal communication of knowledge. It seems that, in the digital environment, there are serious problems in being sure that a message that is sent from one author to another and so on is not altered, be it either deliberately or otherwise. The reader also faces difficulties in determining whether the version that they have is the authentic version that has been certified through the peer review process. In this study, authenticity is viewed as a complex of associated concepts s that are not the main concerns of the author or reader.

It is established that the academic community and the publishing community — who act as agents to the academic community in the establishment and protection of authenticity — are not sure what procedures and actions are needed in the digital environment. In this context, moral rights are often invoked. However, at least in the UK, the legal system does not provide a clear point to refer to, when, for example, web-users cut and paste as common practice, and publisher policies incline towards slicing and dicing.

The understanding of the definitive article is examined at some length, both in traditional publishing and in various alternative scenarios. There is no consensus about the relationship between formal and informal communication and even, in some circles, whether this distinction has the force that it once had in the academic community. It is interesting that in true e-only publishing, when the functionality that is available on the Web is made use of to provide multimedia options for the author, there is some serious interest in the establishment of policies regarding versions.

The identification of versions, and in particular the establishment of a definitive version, is a central concern in those sections of the community investigating standards and the metadata associated with them. It is clear from this study, that the driving forces leading to the development of schema and also to the protection of entities are essentially economic, as one would expect, and that at present the establishment and protection of authenticity is not a central issue. This could be because, in the area of scholarly communication, these issues are as yet mostly theoretical. There is little in e-only form of real importance.

Nevertheless, the determination of authenticity has found theoretical expression in high-level discussions of archiving and preservation of digital objects. However, this study shows that the implementation of these ideas in practical procedures and archival policies have little , there is as yet less interest in these questions than one might expect. It is the view of the author that the fact that the two central intermediaries in scholarly communication, publishers and librarians, are currently projecting and promoting very different scenarios about its future is currently discouraging a proper understanding of what is in fact central to both the progress of knowledge and the true interests of scholarly authors and readers. In practice archiving and preservation discussions may enable joint understanding of the issues concerned, which could impact on policies over a wider front.

1. Introduction and explanation

This section explains the purpose and context of this study, and outlines the content of the preceding sections. First, the general themes are set out. Next, some methodology is explained, and the final section summarises the findings of this study.

In the preparation of this study, the author has realised that he could understandably be accused of yoking together a number of disparate concepts under a convenient heading. What is certain is that the relationships discussed in the following study are not clear, but there is little doubt that relationships do exist, and that the concept of authenticity is a useful, central principle. It seems that we are confronted essentially with a problem that is a product of the digital revolution itself. One aspect of this problem is that the use of the Internet encourages lateral or even, one might say, 180 degree thinking — thinking that at any rate is not linear. During the years immediately following the widespread adoption of the World Wide Web, there were several evangelists for what they saw as a mode of thinking, which would replace traditional linear modes: we would all think in a non-linear way because the Web encouraged such thinking and the transmission and growth of knowledge would benefit. Where are they now? A challenge that we all face is to handle the progress of scholarship, the advancement of knowledge, which is essentially linear in a new context that is not. That is in part what information overload is all about.

1.1 THE PURPOSE OF THIS STUDY

The original proposal for this study (as accepted by the funding body) was explained in the following terms:

> "The advancement of knowledge in any academic discipline involves one scholar communicating a vision/message to another scholar. It is important to the reader that the message is received in the form sent by the author, complete and without distortions, additions and subtractions, and that the attribution to the author is correct."

It should (obviously) be added that the maintenance of the paternity and integrity of the message is equally important to the author. The assumption is that the author wants the reader to receive his or her authentic message.

1.1.1 A problem of the digital environment

It was and is argued that in the digital environment there is greater opportunity for this message to be distorted and more difficulty in detecting the distortion. It is even encouraged by the medium. It is part of what we all do when we work with one window open for composition and one open to the Web or some other document that we are holding (**Dorner** p134–135). No particular skill is required to merge or conflate, and almost no decision is made.

The sort of distortion that was envisaged (when the project was first thought of) was slicing and dicing by the publisher as well as cutting and pasting by the

reader. In either case the resulting document could well be passed on without attribution or in a derogatory manner by whomever it is sent on to.

Much of this study is concerned with how publishers and librarians think, and particularly what librarians think should be done to 'free' scholarship. One can be brought up short by the succinct statement of Kircz and Roosendaal (**Kircz 2**), whose other contributions we will come back to later:

> "The main issue to be addressed in the context of electronic publishing is: '*How can it support and enhance the science process'?* (authors' italics)

It is our aim to bear this in mind throughout this study, but we might not entirely succeed.

1.1.2 A problem for scholarly communication

The assumption here is that the author is the sufferer, but the reader can also be the sufferer. In the scholarly context, scholarship is the sufferer.

Shoffner examines the problem of authentication and sees it primarily as an issue for scholarship. He sees the author as an enemy of authenticity. He writes:

> "When published in traditional print form, a book is produced in multiple copies, and so we can reasonably sure that the copies will not be changed after publication ... this is not the case for electronic information. Although great care has been taken to avoid unintended changes of data within computer systems, it has also been an objective of the computer community to make it easy to change data intentionally ... for example, by judicious editing after the fact, a page of predictions could always be true. An author may go further than changing a page; it may be removed entirely.

Actually, the publisher usually controls the certified output of the academic author. Would it be in the interest of the publisher to make such alterations? One answer is given in section 9.4.2. The position could be characterized as one that is hostile to the changes following a digital transition.

1.1.3 Use and misuse in scholarly communication

It is obvious from the previous subsection that there is a serious opportunity for exaggeration about both the importance of the problem and the exacerbation of the problem in the digital environment. There is an element of a continuum here in that an action inimical to scholarship or against the reasonable interests of the author is not always or even usually easily differentiated from a perfectly justified, natural and even helpful action. For example, it could be argued that, in the case of a reader/user taking part of the work of another scholar from a web-site and using it, unattributed, in their own work, the action could be a tribute rather than a theft. Whether this is plagiarism depends not just on the amount of content that is taken up in this way but also on the nature of the content. Throughout this study we are concerned not so much with authenticity in the sense that a piece of work is associated with its author, but instead with authenticity in the sense that the content is not altered in a way inimical to its real meaning.

This is a truly complex area. The purpose of scholarly communication is to communicate. An author wants his or her ideas to be taken up and absorbed into the mainstream of discourse. The author wants to be recognized (usually) and the ideas to be represented faithfully (usually), but there are modest authors and ideas that only flourish when they are adapted through scholarly dialogue. The criteria, which could be established when considering the validity of such transmission, are not our concern here.

1.1.4 Identification and preservation of scholarly content

So far, the thread of the overall discussion has been misuse of content, leading to loss of authenticity. Another concern is recognition and identification. Questions of identification arise with the immediate need of the users/readers to establish that they are receiving authentic content.

However, questions of how to archive and preserve such 'messages' as they are — either only in digital form or in digital form only as far as their complete representation is concerned — has come to the forefront of the authenticity discussion. Such questions are probably of greater interest — certainly to librarians and (to a lesser extent) publishers. They are certainly more discussed and at a higher level of intellectual concern. How is the authenticity of these messages to be preserved? If it is not possible to preserve all the features that constitute authenticity, what features are most important to preserve?

Associated with this wish to preserve the authentic message comes a series of questions about what sort of 'messages' are regarded by a scholarly community as worth preserving. In the digital environment, there are several versions of many documents. Is there a definitive version and, if so, what is it, how is it arrived at and who is responsible for it? Do the other versions available have a status and, if so, what is it? Any intermediary, whether publisher or librarian, working with and for scholars is aware that this is an issue of concern to many, although there has been remarkably little research done on situations where this concern is taken into account. For example, it is a truism that electronic-only journals are handicapped by the fact that authors are doubtful about whether their contributions will be preserved for posterity. When this sentence was first written in 2000, the evidence was mostly anecdotal, but surveys such as those compiled by **Key Perspectives** now confirm this.

1.1.5 Limitations in the scope of the study

The rest of the content of this study is associated with these two threads and their expansion into real situations. The aim of the study is to examine the implications of issues of authenticity for the scholar and of the advancement of scholarship not as a philosophical and/or legal concern. It has not proved easy to relate this discussion with those philosophical concerns that relate to archiving and preservation and with legal concerns based around the copyright legislation. It could be argued that some sections of the study do not represent an extension of the fundamental concerns about authenticity already mentioned. Section 5 could be seen as an example of something of an excursion into an area that is of adjacent rather than derivative interest. However, where questions of authenticity

do come up in a practical setting, the study will try to follow. The boundaries of the study will be explored further in section 1.3.

1.1.6 Standards in authenticity

The sponsor of this study is Book Industry Communication/Editeur (http://www.bic.org.uk). Their mission is standards, memorably designated by the Wired guru Kelly as equivalent to laws in the digital environment. Without standards authenticity cannot be preserved or protected, and for that matter the authentic document cannot be identified or its security maintained. As has already been mentioned, the aim of the study is to be pragmatic, which means in this context to look for standards that have a chance of acknowledgement and implementation. This approach is needed even if the standards are currently difficult to visualize and require much work to be done on them (for a good example see section 3.2). Standards are both highly technical and require consensus. It is not for a study like this one to propose standards but rather to attempt to provide the context in which proposals for standards can be initiated. We will, however, in some sections point to areas where the establishment of authenticity does seem to cry out for definitions, and subsequently standards, and we will return to this question in section 10.

1.2 TERMINOLOGY

The distinctions made below are made for the purposes of this study only — to help clarify intention and meaning. The author of this study is not claiming any significant contribution to the sort of semantic debate that discussions of this type of topic bring up, important though they are.

1.2.1 No standardization of terminology

Because of the arcane nature of some of the subject matter, it is easy to appreciate that some of the terminology used will not be understood immediately. There is also no consensus about which word it is customary to use in which context. Different scholars use different terms to describe the same object, or is it the same object, or rather concept? There are subtle differences in approach to the concepts that are outlined here, and the terminology used by key authorities has often been preserved to prevent a glossing over of such subtleties. I have already been guilty of appropriating the word 'message' as used above, but hope that the meaning is made clear through the context in which it is used.

1.2.2 Terminology embodies assumptions

Most of the sections in this study approach the usage of 'authenticity' in different ways, but an attempt at definition is made in section 3. I recognize that, because of my own background and interests, it is possible that the assumption made here about the way in which scholarly communication works and (to a large extent) seems to be continuing to work might not be obvious. Almost all the writers who have provided the foundation for work in this area come from a background in information science or information technology, either currently as an academic or who apply a training that has been conceptually derived from these disciplines in the context of libraries. Section 3 summarizes what this means in practice. On the contrary, I have derived my interest from working with

scholars as a publisher: see section 2.1 for an explanation of how this interest developed. Section 5 treats, at length, an area that has yet to impinge on the consciousness of those from the disciplinary background described above.

1.2.3 Preferences in and decisions relating to terminology

Reference is also made to 'documents' in many cases where another term, not so associated with texts, is probably preferable. The word 'document', in its derivation, has nothing to do with print as such [personal communication by Sally Morris]. However, it seems to me that the association with print is often assumed, which is of course valid where there is an associated print version. In such cases the usage preference is for the word 'version', to make this distinction clear.

Section 7 is headed 'Informational entities' because it is entirely concerned with content in digital form, and this is the term used by **Kircz (2)** on whose work much of the discussion is based. The term is, however, cumbersome to use throughout. The key reference for the meaning of this term is a paper by **Rothenberg (1)**. In section 3 we have tended to use 'digital objects' because Lynch favours this particular expression **(Lynch 3)**.

1.3: THE BOUNDARIES OF THE STUDY

In this subsection, distinctions about the scope of the study are defined by what is not covered, but we start by looking briefly at 'knowledge' as it is understood in this context and, in practical terms, by the usage of the term 'chapter-by-chapter'.

1.3.1 The transmission of knowledge

This study is concerned with the transmission of knowledge (defined as information to which value has been added) rather than information *per se*, but it has to be recognized that others use the term 'data' when referring to information and 'information' when referring to knowledge. Essentially, it is concerned with knowledge — the substance of the 'referred' to, which is current, currently discussed, currently used and currently considered saving for posterity. The knowledge under discussion, or rather the message containing the knowledge, is not different in nature and function because the format is different. The study asserts that the role of knowledge, irrespective of the format, has the same function in scholarly communication This point has to be underlined because in the early days of the adoption of the World Wide Web there were widely canvassed views emphasizing that what could be called lateral thinking would replace a causal linear approach. It could reasonably be argued that the user searching the Web uses the Internet in a different way from the reader working through a book or article, but this study assumes that the way in which scholarly communication is conducted has not changed and will not change. This is a potentially big topic, considered at length for example by **Meadows**, but serious discussion does not belong within the confines of this study.

1.3.2 Knowledge: digital not digitized

It might seem obvious that we are discussing content in the form that has been decided on by its creator. However, as we will move on to discussions concerning digital archiving, it is necessary to distinguish between born-digital material and digitized material, even though this is not always done. This study is concerned with born-digital material. Librarians tend to think in terms of digitized material. A whole new range of questions arises when one seeks to determine the integrity of a digital copy (**Bearman 3**).

Nevertheless, it is impossible to demarcate strictly. As discussed in section 6, the treatment of images, to give one example, presents similar problems whether one is attempting to preserve the message that is expressed through an image in a born-digital entity or questions of quality of reproduction in a digitized entity. Another reason for demarcation lines being difficult to draw becomes apparent if our concentration is on the reader rather than the author/reader relationship that has been set out above. When the reader accesses digital resources, does he or she distinguish between what is original and what is a copy, and does such a reader assume different criteria for judging the authenticity of a copied object from that used in consideration of an original?

In addition, the following point needs to be made. Much of the interest in the difficulties of establishing authenticity lie in the special problems relating to content which are not possible to express in print, this is the burden of section 7. However, in practice much e-only content that is currently available does not take advantage of the extra functionality that is available to the author. The publication might be e-only in form, but it could just as well have been p-only (print only).

1.3.3 Knowledge not data

The study is also not concerned with data. Databanks and databases are important to a range of scholarly communities and the relationship to primary research communication is subtle and complex. It is interesting that journal publishers are only now taking this relationship seriously, though it has been implicitly understood for a long time by authors and readers. You can see this new understanding of the place of the journal article in the overall research environment both in links from journals and links across journals within the new 'joined-up' environment that we are working towards.

The distinction between knowledge and data is not of course a completely obvious one. If you take a publishers viewpoint, one can note that many databases are now refereed or at least there is gate-keeping, which prevents inappropriate or defective submissions. To take one example from the biosciences, GenBank submissions undergo various automatic checks including syntax checks, checks for common contaminants, but also computational validation of some types of features. The individual submissions, characterized by my informant as non-high-throughput (there may be, for example, up to 30–40 sequences in a submission but usually only 1–3) undergo examination by a trained Ph.D.-level biologist. They run additional analyses, including comparative analysis, as well as checking nomenclature, etc... They communicate with submitting scientist to deal with any necessary corrections. They also provide a

finished record-to-submitter for their review [personal communication from Dr David Lipman].

There are clearly important issues of authenticity here but they are different issues and for someone else to deal with. The literature relating to these different authenticity issues seems to be either non-existent or different from the literature that we are dealing with. It is interesting that there have been discussions within the International DOI Foundation about how to handle this type of content, but serious engagement with the problems involved has been put off for the moment.

1.3.4 Knowledge and 'electronic records'

Librarians are much concerned with the authenticity of 'electronic records' and the preservation of this authenticity. Once again we are in the presence of another, different type of endeavour. The distinction between knowledge, as we are defining it, and such records might at first glance seem obvious but for someone researching the literature on authenticity it is not always clear what the focus of a project is. A good example is the ambitious InterPARES project **(Gilliland)**. This project, which is frequently cited, deals with electronic records. However, as is the case with much of the research on digitized content being distinct from born-digital (see above) there is a lot of relevance in the outcomes, as long as they are used with caution.

1.3.5 Authenticity not authentication

It is also worth making the point that this study is not concerned with authentication, when used in the sense of recognizing those who have the right to access a particular piece of content. This point needs to be made particularly because, in early versions of the specifications for the project, a chapter on authentication mechanisms was listed. There is a lot of important work on standards in this area that are mostly prompted by the desire to control access for commercial reasons. Particularly useful is the work by **Bide (1).** Bide is also a senior protagonist in the <indecs> project, which will be discussed and cited in section 8. This project makes it clear that questions of authenticity are linked to questions of authentication because the same sort of questions have to be asked, about the entity as well as the person, and they are equally, in the last resort, as impossible to handle definitively. In general, however, reference to this literature is only made when some interest is shown in the nature of the content to which access is controlled. This is rare.

1.3.6 Scholarly communication and education

Scholars communicate research but they also educate. There are a lot of visible authenticity questions within the educational context. For example, electronic course-packs are identified more clearly than scholarly communication, and this is an issue central to section 4 and implicit throughout. The facilitation of plagiarism in the digital environment is becoming a serious topic of concern to the educator at all levels and has attracted a commercial interest in exploiting the opportunities revealed **(Herman)**. Questions relating to the educational process are obviously of considerable interest to many, but they are not our concern in this study, though from time to time references to the specific concerns within education will be made.

1.3.7 Plagiarism and scientific misconduct

Plagiarism is used without definition in the previous section. Here, the student is passing off an essay, available from a commercial site, as his or her homework. The term is used from time to time throughout the text of the study. The prime meaning of the word is 'wrongful appropriation' but in the instance cited the concept of 'passing off' is more relevant. The problem with the word is its vagueness. It is used to signify any sort of wrong doing that relate to the misuse of someone else's words or ideas, not necessarily even by stealing, as we have just seen. It is interesting that a recent major textbook on intellectual property law does not even index the word (**Bently**). That is not surprising as the term is not a legal one (see 4.1.4). For the purposes of this study, the word takes on the definition that is implied by its context, and it represents a wider and more diffuse concept than is our concern.

Scientific misconduct, once called 'fraud', is in principle part of the remit of this study as it is an enemy to authenticity. It is a subset of the distortions that we discuss particularly in the legal context of section 4, and is subject to some of the legal remedies mentioned in passing. This study will not deal with this area in detail or specifically. There is an excellent short article by Richard Smith (**Smith 3**), which is easily available. He suggests ways in which electronic publishing could increase the integrity of the scientific record, though he also recognizes that there are, at the same time, new ways to corrupt it. Some of his recommendations are quoted below (section 5.4.2) in the subsection on peer review.

There is a lot more in the literature about both these topics than about the subject of this study, and some of the respondents who replied to the author would have agreed with the following e-only publisher:

> "Attacks on paternity and integrity may be a problem, but, frankly, scientific data fraud is a much bigger and dangerous one."

1.3.8 A sector of scholarly communication

We will examine scholarly communication further, but in a section that is concerned with the boundaries of the study. It is important to emphasize that almost all the text that follows is concerned with scientific scholarship, and particularly primary papers in scientific technical and medical journals. The digital revolution is much further advanced in this particular area of communication than it is in others (**Watkinson 1 and 2)** but one cannot always generalize. There are problems of authenticity that are specific to scholarly communication in the humanities and social sciences, but there is less emphasis on, for example, matters such as priority (in any case only touched on), than there would be in biomedicine. Also, for most of the constituent disciplines, special problems for digital informational entities are not so evident, though there are exceptions such as archaeology — mentioned in section 6. There are experiments in monograph publishing, which are bringing up authenticity questions, but which the present author has discussed elsewhere (**Watkinson 1**).

1.4 THE CONTENT OF THE STUDY

The three subsections that follow this introduction deal with context, perception (2), conceptual (3) and legal (4). The following section (5) deals with questions relating to a separate but related debate concerned with the definition of a publication. These sections raised questions about standards to which answers cannot be given. Other standards issues are the subjects of section 6 and 7. The next two sections (8, 9) deal with applications of issues raised earlier. Can the 'message' in digital form be protected against misuse (8) and can it be preserved for posterity (9). The conclusion (10) draws together the threads.

1.4.1 Digital transition

The purpose of this study is to examine an aspect of scholarly communication, though the concepts involved are important to all creative endeavours. Scholarly communication is in a period of transition, with the result that taking stock of the role of authenticity might be viewed as timely. The transition is from the transmission of knowledge in a print format via physical vehicles, such as journal issues sent by post, to the expression of knowledge in a digital form and the making available of the digital entity, currently online. "Digital is Different" — a banner under which the publisher community fights copyright wars — clearly cannot be denied in the current context, but exactly how and how much is the burden of the discussion, and is background to much of the argument throughout.

1.4.2 Concentration on science journals

This is not to indicate that print is dead. We have to qualify by suggesting that print might mean printing out rather than delivery of print. But in any case the argument is that the new circumstances bring up new problems to be answered and highlight older, and often neglected, concerns. As we have already explained (1.3.8), there will be special emphasis on the role of learned journals because electronic availability of journal content is now the dominant mode of access to that content in many disciplines. Much of the analysis and many of the examples will focus on journals in science (including medicine under that heading) because the communication of science is where the phenomena described below show themselves most clearly. There will be many occasions in this text when scientific communication and its practices are treated as if they represented the operation of scholarly communication in a more general sense. Qualifications should be understood.

Meadows (in page x of his preface) warns us to note the differential adoption of technology. He points to the fact that "changes affecting the world of research as whole do not necessarily have identical impacts on research communication in the sciences and on that in the humanities". This is taken into account: indeed there are a whole range of different practices within science — see section 5. In some of the sections of this study the nascent e-book initiatives in the scholarly arena will be examined. One specific reason is that, whereas the journal article is the basic level of granularity in that particular mode of primary research communication, the book chapter is not (or, more correctly, has not been) where books fulfil that function.

1.4.3 The author community

A lot of the discussion in this study is concerned with actions being taken or not taken by intermediaries between author and reader. These intermediaries are primarily but not only publishers and librarians. The reason for this is that the thinking about the issues involved at present, and as a generalization, impacts little on the scholars adapting to the digital environment. In section 2, what we can say about the expressed concerns of the author community (which is not much) is presented. Intermediaries, particularly librarians and apostles of what is called alternative publishing, often claim to speak for authors but there is remarkably little written by scholars themselves, and any evidence tends to be anecdotal. When it is not anecdotal it is drawn from surveys of pre-selected groups, and, alas, no attempt is made to arrange proper sampling or any of the other prerequisites of research in the social sciences. A good example is the 'evidence' that is gleaned from focus groups behind assertions in one recent report (**RSLG**). It requires an act of will to remember that studies like this one are concerned with scholars and scholarship, and the advancement of knowledge.

1.4.4 The authentic entity

The trigger for the modest upsurge in consideration of the concept of authenticity has been the consideration of the preservation of digital items that has become a serious area of concern to librarians and others involved in archiving. The research concerned with what can be preserved in the digital environment has thrown back the question of what the authentic entity is. In section 3, the conceptual underpinnings are set out insofar as they are relevant to the rest of this study. It has not proved easy to discern what is relevant and what is not. It is also artificial to separate entirely a discussion of where the authenticity of a scholarly informational entity resides from a discussion of the theoretical basis of the archiving enterprise. Nevertheless, the more practical applications of the concepts to questions of archiving and preservation are discussed in more detail in section 10.

1.4.5 Moral rights

The title of this study draws on the wording of the specific moral rights legislation embodied in the Law of England and Wales. The legal understanding of moral rights in this and other jurisdictions needs exploration. In the electronic environment, as we will see, it is easy for users to cut and paste from some formats. In addition, publishers are either making downstream arrangements to sell 'fragments' to a greater degree than was the case in print, or are actively planning the slicing and dicing of content from databases. In section 4 the obvious tensions between the exploitation of the content and the maintenance of its authenticity and the interests of user/reader and author will be considered. However it is made clear that moral rights do not have the force that they might have been expected to have.

1.4.6 The definition of a publication

In the two sections following, as described in outline here, we are moving into the territory of a wider debate, which the author and some others have been involved in, but where the literature is mostly informal. It is also a debate which

seems so far to have had little impact on the publishing community in general, and even less on the other intermediary communities and especially authors and readers. Nevertheless, it is the view of this study that the debate is not only important but that it has only just got under way.

Section 5 of the study is concerned with this ongoing debate on the definition of a publication. The question being raised is whether or not the changes that are a consequence of the adoption of the World Wide Web as the main vehicle for scholarly communication has changed the nature of formal communication as previously understood in print. Is there one 'definitive' publication, which alone is considered authentic? What are the hallmarks of a definitive publication? What is the role of peer review in the process? Certification is central to the claims of traditional publishers and their assertion of a continued role in adding value in the digital environment is a central thread. Nevertheless, peer review can be conducted in other ways and by other players. Insofar as peer reviewed = authoritative = authentic, this area will be outlined in the next section.

Also in this section, the implications bought up by the fact that the Internet renders it easy for what used to be considered informal communication to be made public in a way not possible before are examined further. What used to be called pre-prints (indicating their relationship to the definitive publication) are now e-prints and the Open Archives movement aims to provide what could be characterized as an alternative virtual database of academic knowledge. The attempt to produce a new formal entity — the 'first publication' — will be examined in the light of an analysis of the practices of different disciplines and sub-disciplines. What is the status of 'unpublished' but stable manifestations? Are they part of the record of science? In what way are they 'authentic' and are they to be protected, archived and preserved and, if so, by whom? It is clear that there is no single answer that is appropriate for all scholars, but there are trends that can be discerned.

1.4.7 Informational entities

Because most scholarly content available on the Web is essentially identical to content available in print, there is a tendency — not resisted here — to write of 'documents'. The word is used as a shorthand term even when it is recognized that the use of the term is a misnomer when applied to versions that contain additional matter or dynamic components, or that are not available in print in any or in full form. It is not a matter of the origin of the word but a matter of the associations it carries. Section 6 looks at informational entities square in the face. There is consideration of the concept of the definitive or 'normative' version, which becomes a more complicated matter when the question is examined closely. There are also questions raised about how a digital version can be prepared in such a way that archiving and preservation can be achieved, which is examined in section 9.

1.4.8 Certainty of identification

Much of the rest of this study will be concerned with the application of standards to this part of the scholarly process. Identification is necessary to admit retrieval in the digital environment of the desired entity. The publishing community has invested heavily in the digital object identifier (DOI) and the implementation of

the technology developed in the CrossRef project. The Open Archives movement is developing protocols to enable interoperability. There is a lot of literature available on these related topics. The reader wants to be able to discover what he or she is looking for and be sure that what they get is what they want. In section 7 the aims of these systems are examined insofar as these aims are concerned with the purpose of this study. The central question is how far there is scope for the nature of the sought entity to be described in the metadata, especially insofar as this nature relates to concepts of certification. This is a highly technical area, which will be considered in this section as far as discovery and recognition are concerned. Deposit or submission metadata, as it is linked to considerations of archiving and preservation, will mainly be considered in the penultimate section (9).

1.4.9 Mechanisms for protection

In section 1.4 above the question of how far it is the obligation of publishers or other intermediaries to protect the moral rights of the author has been posed. In section 8 the physical mechanisms are examined. It will be established that most of the mechanisms available are intended to facilitate digital rights management rather than protect authenticity, but corruption and distortion must clearly lower the value of the rights that are being sold. E-commerce is not necessarily at odds with the interests of scholarship. In both this section and the previous one, there is a recognition that someone has to pay for the developments that are described. The conclusion is that protection of authenticity by mechanical means will probably not be effective.

1.4.10 Authenticity, archiving and preservation

As has already been mentioned in section 3 the new interest in questions of authenticity results, in the main, from the need to determine what can be and what should be preserved for perpetuity. The criteria used for deciding what are essential and what are peripheral components of the informational entity vary a great deal from programme to programme and are often naïve, for example in the characterization of 'look and feel'. This section will not constitute an exhaustive account of the archiving and preservation of non-print material, but will concentrate on the determination of authenticity and the standards associated with that determination, particularly insofar as they relate to deposit or submission metadata.

1.4.11 Concluding comments

The concluding section sweeps up some of the questions raised in previous sections. It is deliberately selective and discursive. Essentially we return to scholarly communication and how it will work in the digital environment. The Internet makes possible direct communication between author and reader, and some commentators have seen this direct communication as freeing the central participants from shackles that have been imposed from outside the process. Concentrating on the protection and recognition of authenticity, this final section of the study considers whether the traditional intermediary functions still have a place, and whether the same players exercise them.

2. Authenticity in scholarly discourse: the perceptual context

This section is concerned with the way in which the Internet, as a vehicle for scholarly communication, is perceived and the perceived dangers to authenticity in the Wild West of the World Wide Web. Perceptions are as important as facts in terms of their impact on decisions. This study in the end is concerned with decisions if they are to be useful. Both needs that are elicited and actions that are proposed have to take into account the reality of where we (those involved in scholarly communication) are in our thinking about the digital environment.

The first subsection is concerned with my own experience as a publisher and the context in which my own interest in authenticity has arisen. It seems to me that it is relevant to this study. The way in which my own conceptual framework has become established and has developed may well have resonances for the undocumented developments of others.

The remaining subsections look at different attitudes to and aspects of scholarly communication in the digital environment. In the second subsection, the suspicions of all parts of the information chain concerning the Internet are charted. In the third subsection what we can glean about the attitudes of the academic community itself towards concepts of authenticity are laid out.

Essentially this section frames a number of statements. Scholar communication is now digital. Scholars distrust the digital environment. Scholars, and intermediaries working with them, have not really come to terms with the questions of authenticity, which are components of this distrust. Scholars are not at ease with the environment in which they find themselves but they have generally not analysed why this might be.

2.1 A PERSONAL VIEW

In this subsection I am forsaking the authorial passive. My justification follows.

Most historians have for long recognized that their approach to their chosen subject matter is determined by their own world view. History is about selection and selection is made on the basis of what seems to the author to be important. I am therefore exposing my own experience so that the reader can see where I am coming from.

2.1.1 Putting journals online

It also seems worth bringing into this study a piece of the history of the development of online versions of print journals, which, in the nature of things, has already begun to disappear. Almost all the documentation (essentially grey literature) will soon be difficult to assemble. Why publishers began to put journals online when no business models were available and libraries were not ready to receive them is a story to be told but this is not the place to do it.

Only those aspects of this one small slice of history that impact on the subject of this study will be covered. Questions of integrity (completeness), quality of reproduction, protection against corruption, identification of versions and archiving of e-versions came up early for one company at least in a context relevant to the present study.

2.1.2 The CAJUN project

My own epiphany, my realization of the importance of authenticity began in 1993, when, as publishing director of Chapman & Hall (a company no longer in existence) I authorised the funding (jointly with John Wiley & Sons Ltd) of a project known as CAJUN. The summary prefacing the article, which describes this project (**Smith 3**), begins:

> "The publication of material in 'electronic form' should ideally preserve, in a unified document representation, all the "richness" of the printed document."

This principle is (to my mind) central to much of the discussion in the first half of this current study.

2.1.3 The richness of the message

The concept of 'richness' was important to me at the time. In the mid 1990s there was an obsession among Internet gurus and visionaries of all sorts with multimedia and interactivity. Not taking advantage of the opportunities represented by the World Wide Web was seen as almost morally wrong. At the same time there was also a tangential assumption that text was all-important and that images of all sorts were mere illustrations, whether they were vital half tones, graphs, line drawings, or even chemical structures or four lines of mathematics, and were essentially optional.

It was obvious to any experienced publisher that in many disciplines (and not just in science) the image is as important as the text in communicating the message – and is essential. This is what 'richness' meant to me. The relevance to the debates about 'essentials' in authenticity (see below in section 3) and the position of 'look and feel' in archiving (see section 10) is obvious. To some extent the same battle lines are being drawn up.

At the time I plumped for Portable Document Format (PDF) because it preserved that richness, which meant for cost reasons rejecting SGML and SGML derivatives. For many at the time Portable Document Format (PDF) was regarded with distrust (as proprietary) and with loathing by many because it did not take advantage of the opportunities.

2.1.4 Opportunities presented by the Internet

Like many others at the time I did envisage a serious and rapid take-up by academic authors of the opportunities for expressing their message using such tools as video, audio and simulations, though I saw the appropriate route as 'clip-on' to a PDF file. Part of the aim of the SuperJournal project (**Pullinger**) was to take advantage of these new opportunities and to find out how users reacted. In

fact, the publishers involved in this project could not then persuade their author communities to invest in multimedia of any sort. Some of us (see below) might have recognized that different versions would come to pass, but, as it turned out, such questions could at that time be put on the back burner. Now (see section 7 below) the situation has changed.

2.1.5 History and security

Much indeed has changed since the mid 1990s. Increased access became recognized as the single most important contribution of communication over the Internet. This is still the case.

PDF is recognized by most as having a place alongside the latest SGML derivative. PDF is for printing. Most journal publishers offer both, recognizing different user needs. This may change. Over at least five years, and possibly longer, there have been rumours of an impending release of what can be called structured PDF under various names. This release (when it happens) points to a future where only one format may eventually combine the advantages of what are now two basic approaches.

The other advantage of PDF was, and is, the simple built-in security, which for the ordinary reader makes cutting and pasting a difficult task. This will be discussed in a later section. It is still very relevant. Back then in the mid 1990s it was much easier for the publisher to sell the process of going online to successive editorial boards of journals who are more concerned about security of content than positive about any opportunities.

2.1.6 More lessons from the past

Nevertheless, publishers (and learned societies) were really impressed with the Journal of Laser Guided Surgery published by Wiley, which as a journal was a failure – or so I have been told. It was not a failure as a prodigy and a public relations exercise. Authors might not have wanted to publish in it but it certainly influenced competitive publishers. As a spectacle it was said to be very costly to set up and maintain, as procedures were developed on the fly, but it was much envied by competing firms.

I have unearthed an internal memo from 1995, which records my reaction at the time to this phenomenon. I wrote:

> "I know that those who looked at this journal on the Net are not impressed ... in terms of delivery it is poor and the half tones are not of a high standard ... compared with what we are doing with Acrobat files.
>
> However the journal is significant because it is designed to have an electronic equivalent of the print version, which is not actually just an equivalent. Wiley are putting up multimedia and also extra colour plates in the electronic version which means that they are beginning to exploit what the Internet offers (and) which is different from what is possible in print.

> They are also presenting a real problem for bibliographers ... which is the definitive version? If there are two versions should there not be two different ISSNs? Who is to archive the non-print version?"

It will be seen from the above references and quotations that for one publisher at least, some of the issues discussed in this study were in the air. It is also the impression of that publisher that scholars were as yet not engaged in what was at that time seen by most as an irrelevant debate. The thinking of the library community meanwhile was dominated by their picture of a serials crisis caused by publisher pricing and, for them, the central hope and indeed belief was that e-publications would mean lower costs and lower prices. Successive ICSU/UNESCO conferences and workshops conveniently document the disappearance of this illusion and the new interest in the sort of concerns on which this study concentrates (**Shaw**).

2.2 SUSPICION OF THE DIGITAL ENVIRONMENT BY SCHOLARS

As has already been mentioned this subsection is concerned, not specifically with authenticity, but with one major part of the context. This is the generalized suspicion and distrust of the use of the Internet for the transmission of knowledge.

2.2.1 Scholarly communication is digital

The special nature of scholarly communication, the rules connected with committing knowledge in the form of primary research into the 'record of science' (to give an example from one group of disciplines) is broadly treated elsewhere (for example, **Meadows** passim and **Mabe 2**). There will be a discussion of the rules relating to this transfer of knowledge and the relationship between formal and informal publication in sections 5 and 6. The particular problems relating to the period of digital transition, where we are now, are touched on below.

It will be obvious, in this section in particular but also throughout this study, that the assumption is made that the digital environment is the environment in which scholars now exist. Back in 1997 **Butterworth** wrote of his own discipline:

> "In my own particle physics research group at Imperial College the young Ph.D. students never, but never, look at a printed journal. They get all their information from the display screen ... this is the way the future will soon be in all subject areas."

Professor Butterworth eschewed the provision of a date for the completion of this process. Most scholars are careful in their predictions. We will return to the latest predictions (by non-scholars) later.

The fact that there are scholars who do not use e-mail for informal communication or who expect to access much of the research that inputs into their own work online, does not contradict the fact that all the evidence points to the change being one way. This is digital transition — the movement from one phase to another. The suspicion of this type of author is that printing out will

always be preferred for many types of scholarly communication for reading if not for searching, that the delivery of printed books will continue to produce a better result than downloading from the Web, and that browsing remains easier in print. As librarians have come to recognize that, however much they digitize, they are likely to run 'hybrid' and not digital libraries for the foreseeable future. However, at the same time, most of their energies have to be devoted to enabling electronic access to as much content as possible and for as many of their patrons as they can.

In a way, it could be argued that the concentration on digital transition has obscured the probability of a continued role of print. There is considerable evidence that the behaviour, which Butterworth observed, is not the behaviour in all disciplines and for all purposes. Experience with e-coursepacks, for example, does not demonstrate a necessary preference for this form of accessing knowledge over standard printed textbooks. The jury is out. The pressure for this particular mode of delivery might be economic rather than an actual concern for what the student wants. That is not to say that an economic reason is a bad one if resources (as they are) are finite.

In a later section (5.5.2), this study points to a major scientific journal which was among the first to take advantage of the opportunities of the Web. The editorial group running this journal has now apparently decided that, although the electronic version is normative, the print version has also to be usable without the specifically non-text content – and so probably a definitive version too. The word 'apparently' is used because the information has yet to be confirmed, but the decision would not be an unreasonable one. Is print fighting back and what does this mean as a portent?

2.2.2 Digital transition and its implications for scholars

There is a lot of literature on the changes wrought in the nature of scholarly communication by the movement of scholars towards communicating online. Curiously, much of this literature is not concerned with the nature of the scholarly message as such. One has to really search for any consideration of authenticity and related question, even implied consideration, in works that are specifically about digital transition. There is a lot of optimism about how the digital environment can help scholars, and the downside is downplayed or ignored. The fact that the drivers for some changes are not the communication of scholarship but economic motives directly is the theme of several of the later sections (for example section 8).

One important recent book looks at the digital revolution in this way (**Quandt** and Ekman pp 2-3):

> "The argument in favor of the wholesale adoption of the new information technology (IT) in universities, publishing houses, libraries and scholarly communication rests on the hope — indeed the dogma — that IT will substantially raise *productivity*" [my italics]

They go on to ask the question, in the same introduction:

> "What is the output contribution (that is, the contribution to producing 'truth' in particle physics) of Ginsparg's preprint server in Los Alamos"

A more important work on scientific journals, perhaps the most important book on the topic for a decade or more, is essentially concerned with the cost of the process and not the content that is processed (**Tenopir** and King).

2.2.3 Suspicions of the digital environment

Nevertheless, it is generally accepted that the relationship of the scholar with information or knowledge received online is not straightforward. There is plenty of evidence that both the author and the reader are suspicious of the Internet as a vehicle for formal communication, however much they use the medium. For once, the scholar in his or her guise as both author and reader does not present the schizophrenic approach so familiar to publishers and librarians alike. They are suspicious in both roles. Scholars in the humanities, who have not yet much experience of electronic communication, express their concern most strongly as authors (see **Watkinson (1)** section 7.1 pp 59-63), but many scientists also have similar doubts.

Projections on the growth of and growth in use of electronic-only journals have always turned out to be optimistic. Even up-beat 'alternative' publishers like BioMed Central (http://www.biomedcentral.com) feel the need to offer an annual print archive to librarians. I have been told, however, that take-up has been minimal (BioMed Central, personal communication). There are various arguments about why scholars in all fields are reluctant to submit to these journals, ranging from references to the problems of any new journal, such as lack of a ranking in the Science Citation Index, to a presumed fear of their important contributions disappearing because archiving is not in place. As recently as 1999, an article about the SuperJournal project by **Mabe (1),** a publisher strongly involved in the project, argued that lack of access and therefore readers was the prime reason for suspicion. How different from the situation two years later when it is possible to argue that it is online or invisible, that articles which are not made available online lose out in citations by authors or readers. However, being available online, achieving greater access, is not the same as being available exclusively online.

It is interesting, however, that fears about quality and of plagiarism were also referred to in the SuperJournal article as reasons for suspicion. Another comment from Willis G. Regier — an old hand in humanities publishing — expands on these issues. This quotation is not only graphic in style, it also covers much of the background further explored in section 5 (**Regier** page 164):

> "The fluidity of the Web, gushing with nautical metaphors, often seems a murky sea ... the Web seems unstable, engulfing, and founded on the premise of perpetual replacement. Scholars care about speed, but they care more that their work endures ... Scholars who use the Net frequently encounter defunct URLs, obsolete references, nonsense, wretched writing and mistakes of every kind."

The extent to which this is a reasonable approach is immaterial. Historians recognize that the perceptions about a situation are as important as a reality as is the reality of the situation that is demonstrated by research: a classic case is the

perception of the intentions of and possibilities for George III as viewed by the revolting Americans.

Clifford Lynch (**Lynch 1,** page 145) strikes a note that recognizes both the problem and offers hope:

> "What is striking is the level of distrust with which some people view the electronic information environment, and the stringent demands that are being placed on a system of publication in this environment, which goes far beyond what a current print-based system can deliver. As we begin to migrate to a networked environment and become more comfortable with this new world, it seems likely that some of these expectations will become tempered with realism."

2.2.4 The positive view: scholars want digital publication

It is important to balance the suspicions and doubts of many in the scholarly communities with the positive picture. A perceptive commentator, looking at the state of play from outside the industry, writes about PubMed Central and E-BioSci back in September 2001 in a privileged briefing:

> "Neither publishers nor their market can now deny that the needs of users in the 'hot' science areas like genomics are becoming crystal clear. While published, peer-reviewed text remains vital to researchers, articles themselves are insufficient in length and format to present all the vital evidence needed by researchers. E-Biosci (http://www.e-biosci.org) points out for example that non-invasive spectroscopic or microscopic techniques create images which can only be viewed digitally and interactively."

Alongside the concerns expressed by many about the inherent unreliability of information found on the Internet, the obverse is that many scholars in some areas and some in others actually want to put their communication online. They also want to use it as a vehicle for non-text content and not just as a delivery mechanism. We need to keep both processes in mind as part of the context of the study.

2.3 ATTITUDES TO AUTHENTICITY QUESTIONS

I have attempted to extract feedback from those concerned with scholarly communication regarding the topic of the study. Few questionnaires were filled in by any sector, and even librarians, who are usually so helpful, were stumped by some of the questions. There was no possibility of extracting any results that are statistically significant, but there were some quotations, which reveal some serious thought, that have been used. There are three subsections below, dividing up the relevant material to reflect the views of librarians, publishers and authors or their representatives but, as will be made clear, the divisions are artificial. All the responses are concerned with the perception of the respondent of where the author (and to a much lesser extent) the user interests lie.

2.3.1 Responses by librarians

One generalization that can be set out on the table is that, at present, librarians are singularly uninterested in these questions and report little interest among their academic patrons. There were a limited number of responses to a set of questions that included:

> Do you see any role for librarians defending (scholarly content on the Internet) from distortion?

None of the librarians answered this question directly and most of the respondents ignored it altogether. A related question that was also asked was:

> Are you aware of any concern among your patrons about the perceived problem?

This did elicit one concrete reply:

> "Yes, there is concern and bewilderment how to proceed."

The question was prompted by the strong traditions in library training that indicate that an important role that a librarian has is to select the most important resources for their patrons and to help their patrons use them optimally. Librarians are certainly interested in, for example, making access available to the 'best' text of a Shakespeare play, even when access is costly, rather than proposing that inferior texts that are freely available are to be preferred on that count. My own experience on the JISC e-book working group in the UK (http://www.jisc.ac.uk) has confirmed this attitude.

However, many librarians do not usually seem to feel the need to look inside the covers as it were. Of course, on the whole, neither do publishers. Many of the library inspired projects mentioned later in the text are curiously unconcerned with issues relating to the authenticity of publications, except where the library community is itself the author (as it were) through the digitization of sources. For evidence of this see the programmes of library meetings. Archiving and preservation is now seen as a central question, but unless Clifford Lynch is speaking (as he often is) authenticity does not play a big part in the content of the discussion. Section 9 goes into such questions further.

2.3.2 Responses by publishers

As far as publishers are concerned there is much more evidence in terms of the way they handle moral rights. This is explained in section 4 of this study, particularly in section 4.3. In this subsection, the context of author/reader perceptions is provided insofar as is possible.

What was clear from the responses there were given was that plagiarism was the biggest concern. This is the concern that publishers find to be evident in the print environment also: most publishers have had some experience of plagiarism, particularly book publishers, and have devised strategies for satisfying authors without having to go to the law.

The publishing director of a large learned society publisher (himself an academic) wrote in response to the questions:

> "Certainly plagiarism causes us occasional problems when we detect it. We have well-tried procedures for dealing with it, once found. Finding it is the problem, though computer-based text-correlators can help in this process. Referees are not particularly good at seeing plagiarism — the older and more experienced ones seem to do better at this, as might be expected. As for the community telling us it is a problem — frankly, no, we don't hear anything from authors. I do observe that 'cut and paste' plagiarism — sometimes 'wholescale' in nature, seems to be on the increase — but I'd find it difficult to quantify with reasonable error bars. One recent book chapter we received is almost completely a 'cut and paste' of large blocks of other people's work — we do consciously look for that all the time with books."

It is noteworthy that he mentions books, though the publisher concerned is primarily a journal publisher. The structure of the scientific journal article is so concise and formal that it is difficult for one author to insert material from another without the joins showing, though it is said to be much easier and a much more common problem in the social sciences.

2.3.3 Responses from authors and those who represent them

An underlying contention of this study is that the majority of authors/creators and readers/users would accept the centrality of authenticity once it has been pointed out to them. This is not a denial of the differing ways that different disciplines communicate (**Meadows**, particularly page 48ff), but in a Postmodernist age it does represent an assertion. It also represents an assertion that does not resonate widely in the current thinking of most scholars.

All researchers find it difficult to get responses of any fluency from authors and readers and, with this is in mind, the following set of questions was sent to learned societies and author representative organizations.

> "As you will see from the attached I am interested in the preservation of the 'authenticity' of the 'message' of the author in the digital environment, where cutting and pasting makes distortion and plagiarism so easy. I cannot find any evidence of much interest among authors in this problem. They do not seem to be demanding that their publishers protect their interests in this matter.
>
> Are you aware of any concern among academic authors and their representative bodies?
>
> If so could you tell me about it?"

The group from learned societies that was approached to answer on behalf of their members was a different group from the representatives of their publishing wings, quoted above. In order to help them, the following further help was added to the questions already listed.

> "In order to give you a better grip on what I mean in practical terms I am passing in below a comment from a prominent authority on scholarly communication on the web about his own practice;
>
> *I have many articles on the web, and some have been copied by well-meaning supporters to other web sites (when a link would have sufficed). Some have been altered. When I detect either situation, I ask the copyist to take the copy off the web because it harms my work. As an author, my concern with authentication is that readers don't attribute to me assertions or omissions that I didn't authorize. When the copyist's copy is exact, I still worry because I want the right to revise the document in the future and want readers to access only the most recent authorized version. I don't mind when old copies are accurately archived (say, in the Internet Archive's Wayback Machine), because readers can be presumed to understand they are old copies and might have been superseded by newer ones. I copyright my web documents in order to give me a leg to stand on to make these requests.*" [personal communication]

As already mentioned the responses were patchy and did not lend themselves to statistical treatment. Nevertheless, some are worth considering. The concern of one author, a professor of philosophy who was quoted in italics above, was uniquely thought through, even though he disclaimed a serious consideration of the subject.

Representative author organizations mainly deal with book authors and not usually with authors of scholarly books, though they claim that academic authors have a growing interest in their help. When an employee of such an organization wrote "authenticity and integrity of text are one of my hobby horses" it was disappointing to discover that the interest was not in the area covered in this study. Instead, it lay in the possibility of income coming from downstream audit trails rather than in the integrity of the document as such. However, to dismiss this concern would inappropriate. Another author from a different author organization wrote apropos a recent meeting of a (then) newly formed Academic Writers' Group:

> "[Although] their primary concern was in having to give all rights to the publisher, rather than what subsequently happened further down the chain, it seems to me that the two issues are in fact closely linked." [personal communication].

The same person wrote subsequently in an e-mail, the text of which has been altered slightly to make the argument clearer:

> Now you have jogged my memory, maybe not our academic members specifically, but certainly our educational writers (writing school textbooks, EFL, that sort of thing) regularly bemoan the fact that creating course-packs, for example, like researching via the Internet has the following result. It means that students nowadays tend to get to the one sentence (or whatever) that they think answers their question. The sentence has no background context (which of course can radically alter the meaning of an extract). The student has no understanding of how the conclusion has

been reached, and no chance of picking up general learning, which may or may not be related to the topic in question, on the way.

This off-the-cuff analysis combined a reference to the problems of selection of appropriate material on the Internet with the idea of an inappropriate level of granularity distorting authenticity. It is interesting that the same sort of approach is taken by librarians who are concerned with educational uses or rather misuses of the services for students offered at the time by various 'dot.com' companies, either now defunct or working under a different business model.

A senior manager in a major book publisher that recently committed to putting all its books online and finding some difficulty in retrospectively clearing rights, has found some authors exercised about their legal position but:

> "The instances have been more to do with authors and indeed author organizations and agents when we have contacted them about the digital rights of existing contracts rather than [about] abuses coming to light."

These are almost certainly authors who are prompted by publicity about publisher behaviour, for example relating to the Tasini Case in the USA, that has generated the Public Library of Science and related movements. It could be argued that such authors are rarely prompted by worries about attacks on integrity but more about either barriers to dissemination of their ideas or, alternatively, a worry that there is money to be gained from downstream rights and that they should share in the additional income.

Finally, in a quotation echoing that which concluded the previous subsection, Jane Dorner, in her excellent book advising authors in general on how to tackle publishing online, writes in a section suggestively entitled *Information chunks* (**Dorner** p.99):

> "Pundits fantasise a world where all published information is stored in the computer and can be retrieved again in small usable units by other writers, who can manipulate them differently to give different conclusions. This suggests an alarming world of plagiarism combined with information overload. If publication now includes chunks of information forever stored in computers, will that leave us in a world where technological progress dominates and alienates us from human control."

This is a particularly satisfying quotation because the imagery brings us into direct contact with the more familiar concerns about fair use and fair dealing, which is the substance of publisher/library relations in the digital environment. This is familiar territory. Authenticity is not.

3. Concepts of authenticity

The trigger for the modest upsurge of consideration of concepts of authenticity has been the thinking about the preservation of digital items, as this has become an object of concern to librarians and others involved in archiving. In this section, problems relating to archiving are relegated to the background (because they are discussed elsewhere) and the philosophical issues are bought forward.

This section will be drawing heavily on a central source for considerations of authenticity in scholarly communication, a conference and a publication organised by the Council on Library and Information Resources (CLIR) in a field that is not highly populated (**Smith 1**). The online version has been used. It is paginated by chapter and, where page numbers are given, they are numbered from the start of each chapter as downloaded. Where the work of Lynch is quoted, the source will be **Lynch 3**, unless otherwise stated. I have been helped in his understanding of the issues at stake by conversations with both Abby Smith and with Clifford Lynch, but they of course take no responsibility for the extent of his comprehension or the use of the ideas.

Lynch's understanding of the issues are particular important for our purposes. While he is clear that "Authenticity and integrity ... are deep and controversial philosophical ideas", his approach is essentially pragmatic and grounded in a wish to face up to problems of the digital environment. He sees these problems as not just intrinsically interesting ("because they are there") but because they need to be solved for the promises of the digital revolution to be fully realised.

Not all the topics, which lead naturally from the CLIR conference, are covered in this section. Among other topics discussed later include cryptographic technology, which will be taken up instead in section 8. Although these are obviously central considerations, this study will delay discussion of identification and metadata until section 7. Within this section, where topics have been isolated, pointers have usually been provided to applications and developments later in this study.

3.1 SOME ASSERTIONS ABOUT AUTHENTICITY

In this subsection, we are concerned with some central assertions. In the subsequent subsections, some particular concepts, perceived as being of special importance to the subject of this study, are examined. These are questions about the determination of authenticity depending on the purpose of the object concerned, on the way in which the object is constructed and presented, and on its provenance in the broadest sense of the word.

3.1.1 Starting with print

Part of the aim of the CLIR conference was to bring together scholars from a number of different constituencies, but many of them came from a print environment — distinct from having a background in information technology. There is little doubt that concepts of authenticity tested in the print environment,

or indeed further back in the pre Gutenberg era of diplomatic and allied skills, have a long intellectual pedigree and pose some of the central questions.

Cullen writes:

> "When objects are presented digitally, deciding what is required to authenticate them may be informed from past practices with non-digital objects."

It is natural to think in this way. Those of us involved in questions relating to the preparation for forthcoming covering the legal deposit of non-print materials (as they are known in this particular context) have found ourselves time and time again starting from existing practices in the print environment. These have been developed for over a century.

But digital is different. In the legal context, practices for archiving of printed publications are not a blueprint for the future archiving of digital objects. The dangers of visualising digital informational entities as somehow much the same as printed documents are ably explained by Kircz, referenced in section 7.

3.1.2 The importance of provability

Bearman begins one of his articles by the brave statement, with which I have much sympathy:

> "At its extremes, authenticity carries with it all the philosophical problems of truth, but ... we will try to confine the assertion that something is 'authentic' to a number of 'provable' claims." **(1)**

The emphasis on 'provable' is a useful yardstick for what follows both in this section and in the rest of the study. Without a concept of proof, there cannot be standards in this area.

Cullen looks at the same question in another way. He posits 'tests'. He writes:

> "Because digital objects bear less evidence of authorship, provenance, originality and other commonly accepted attributes than do analog objects, the former are subject to additional suspicion. Tests must be devised and administered to authenticate them." (**Cullen**, page 2).

3.1.3 Authenticity equals integrity

Cullen writes:

> "An authentic object is one whose integrity is intact – one that is and can be proved or accepted to be what its owners say it is. It matters little whether the object is handwritten, printed or in digital form." (**Cullen** page 1).

This statement should be an epigraph to all discussions of the archiving of digital objects. Any loss of integrity because the complete message cannot be preserved for posterity must be a source of sorrow and not perceived as a simple matter of

discarding inessential aspects. However, as we shall see, the preservation of integrity is not (alas) a simple matter and is by no means the only concern the scholarly community has.

3.2 AUTHENTICITY DEPENDS ON PURPOSE

The introductory essay (**Smith 1**) identifies the following concept as a key understanding derived from the conference on which the collection is based. What is deemed intrinsic to an object is determined by the purpose for which it is created. We can bring this concept to bear in the arena of scholarly communication in obvious ways. The following paragraph represents my own experience and as I see it is a practical example of the general approach set out in the CLIR volume.

In certain disciplines for example, the message might have optional features, like illustrations in a primary work of history. They illustrate the point at issue. They are not an intrinsic part of the message, and are not part of its authenticity. On the other hand, in many scientific works the illustration, for example a critical half-tone, is intimately bound up with the text and the loss of such an illustration or indeed an inferior reproduction severely impairs the message. There are however examples of journal articles, particularly in medicine, where the editors of the journal provide an opportunity for additional illustrations to be mounted on the server of the publisher. It could be argued that these illustrations are not central to the message and might even be additional to the message. There will be further consideration of this issue in the treatment of versions in section 5 and in the section on archiving and preservation (9).

3.3 THE NATURE OF DIGITAL INFORMATIONAL ENTITIES

This topic has already been touched on, but the position of this subsection is that of Lynch — a different position from that taken by Cullen (already quoted in section 3.1.1). But is it really different? Cullen recommends an initial approach. Lynch gives warnings. For Lynch, digital objects have different characteristics from the documents with which most of us are familiar. We know how to identify a print document. **Lynch** provides a checklist, which reflects a process that is obvious and familiar and then sets the digital object in this context. In section 7 we will examine digital objects practically as objects to be archived and preserved.

3.3.1 Sequences of bits

One point is signalled here as particularly relevant. Although digital objects are sequences of bits, which can be checked to make sure they are correct, such sequences are not directly apprehended. They are “rendered, executed, performed and presented to people by hardware and software systems that interpret them” (**Lynch**). The implications for the preservation of authenticity in an archival environment are obvious and will be taken up below. The easiest example given by **Lynch** is that text, “marked up in HTML and displayed through a Web browser, takes on a sensory dimension”. This is what we see and what we learn from.

3.3.2 A hierarchy of digital objects

Lynch also introduces a taxonomy. He posits a hierarchy of digital objects, which runs in order of complexity from data through documents and sensory presentations to (interactive) experiential works. As we shall see in section 9.2.1, this taxonomy is particularly useful when addressing questions of authenticity in the archiving situation, because the problems relating to each level of the hierarchy differ. Scholars work with data and create and transmit data but we have decided not to deal with its special problems (see 1.3 above). Most primary research is in document form, whether or not it is e-only, but much of the real interest lies in sensory presentations (see section 10) which is as yet, Lynch admits, poorly understood. That being said, however interesting the higher levels of complexity might be, practical considerations for those concerned with archiving decisions mean that relatively simple objects are those that can be approached first and in any case there are more of them. Most archiving projects have to put on one side any sort of dynamic object.

3.3.3 The essence of digital objects

Lynch writes:

> "Often we seek to discuss the *essence* of a work rather than the exact set of sequences of bits that may represent it in a specific context; we are concerned with integrity and authenticity as they apply to this essence, rather than to the literal bits."

He admits the problems of troublesome imprecision even at the lower levels of complexity in his hierarchy but has introduced *canonicalization* as an organizing principle to make sense of the problems and bring them into the realm of verification by computational methods. In the paper, in which he introduced this concept (**Lynch 2**) Lynch looks at the implications of a change to format for the preservation of the essential characteristic (essence?) of that object concerned. It is particularly relevant to the purposes of this study that, in Lynch's view:

> "The use of a canonicalization approach offers insights that may be useful in efforts to standardize, or at least to develop requirements as a preliminary to standardization, for provenance, authenticity and integrity metadata and the practices that archives might use to manage such metadata over time."

The concern here is archiving and preservation but, because canonicalization enables corruption of the essential elements to be detected, other areas where authenticity is important can have call on the concept. The work of the World Wide Web consortium on the canonicalization of structured XML objects, likewise of wider relevance, is clearly worth following up in another study (**XML**).

3.4 TRUST

We have chosen this heading rather using the word 'provenance' as the presiding concept for this subsection because it seems to me that in the digital environment trust is crucial for practical purposes.

3.4.1 Trust necessary to establish provenance

Whatever technical means are used to establish that an object is what it is said to be, external tests such as those that can be used on a printed book, are not really possible. You have to trust the source of the assertion of provenance. It is interesting how important the concept of 'trusted repositories' has become for different reasons in the context of archiving and preservation. For the library sector, publishers cannot be trusted — a perfectly reasonable position in that archiving is a new role for them. At the same time publishers should be proving to authors of e-content that they have made arrangements for archiving that can be trusted.

Lynch writes that "virtually all determination of authenticity or integrity in the digital environment depends on trust". His discussion of this issue is central to his thought as expressed in the CLIR conference. The statement by Cullen, already quoted in 3.1.1 gives a lower level expansion of this line of thought.

3.4.2 Trust and metadata

Where trust comes in becomes clear when we look ahead in our study to the crucial role of metadata. Metadata is only as reliable (a key demand) as the person or body that created it. In a crucial passage, Lynch, writing of provenance, suggests that we do not have a clear understanding and certainly not a consensus concerning of "where provenance data should be maintained in the digital environment, or by what agencies". When we come to contemplate the problems associated with different versions of an entity (sections 5 and 6) this is a highly relevant observation and it will also be taken up again in section 7 when we look more closely at metadata issues.

3.4.3 Trust as a relative probability

Finally, Lynch points out that it is "important to recognise that trust is not necessarily an absolute, but often a subjective probability that we assign case by case". In all discussions of authenticity there is a tendency to think in terms of absolutes, so this warning is apposite. Whose statement of authenticity do we trust and how much and in what circumstances? Lynch writes:

> "If we are to trust a claim of authorship, whom do we expect to sign it? The author? The publisher? A registry, such as the copyright office, which would more likely sign a claim stating that the author has registered the object and claimed authorship."

In the context of this study questions of watermarking and/or digital signatures or digital audit trails (**Levy)** have to start with questions of trust.

4. Authenticity, moral rights, and other legal questions

This section is primarily concerned with the relationship of moral rights and ideas of authenticity. There is some discussion of copyright legislation. Another theme, which has legal implications and which is discussed in this study, is the definition of a publication. This will be dealt with in section 5.

Legal questions should be hard-edged but in Common Law jurisdictions statutes are interpreted by case law, which can make legal concepts practically uncertain. The relevant law is part of the context rather than a driver. Concern about moral rights is not central to the way authenticity is handled. It is also clear that moral rights legislation, when investigated, does not seem to be drafted with the concerns of scholars and scholarship in mind. That is of course the case.

The way in which the title of this study is associated with the moral rights clauses of the 1988 **Copyright Act** (see below) is therefore somewhat misleading. As we will see, moral rights in law do not automatically lead to a protection against the misuses described in 1.1. The relationship is not a straightforward one.

In the first subsection the moral rights of authors under the law of England and Wales will be outlined. There will also be some consideration of other jurisdictions and a little history. In 4.2 there will be some comments about recent copyright directive and about practices in other jurisdictions. Finally, in the last two subsections those dangers to scholars and scholarship in the Internet environment, which form the subject of this study, will be set against the law as it stands. Perceptions of what the law prescribes and what the law can/should deliver will be considered.

As elsewhere in this study, the attitudes of and activities by the publishing industry is centre stage because the norm is that the creator/author transfers his or her copyright or at least publishing rights to the publisher.

4.1 MORAL RIGHTS IN THE LAW OF ENGLAND AND WAYS

Throughout the first subsection extensive use has been made of the standard book on publishing law (**Jones 1**) and to a lesser extent the recent textbook on intellectual property law written by Bently and Sherman (**Bently**). Throughout this section when Jones refers to the lawyer Hugh Jones (**Jones 1**)

4.1.1 The World Intellectual Property Treaty

It is customary in writings on copyright law to go through the history of those clauses in the currently operative legislation that are relevant. This is not an appropriate exercise for this study. In the overall context of digital transition, transition towards digital transmission of scholarly information as the norm, it is worth referring to the WIPO Copyright Treaty of 1996 (**Owen** page 6). There is a long history behind this, which we shall mostly ignore. WIPO is both the progenitor of the attempts to regulate copyright in the digital environment leading on to the Digital Millennium Copyright Act (**DCMA**) in the USA and the EU

Directive on Copyright (see below), but also the creator of the new right of communication to the public.

4.1.2 The Berne Convention

In the exclusive context of moral rights however, the relatively short sub-clause from the text of the Berne Convention 6*bis* (**Berne**) is worth mentioning here:

> (1) Independently of the author's economic rights, and even after the transfer of the set rights, the author shall have the right to claim authorship of the work and to object to any distortion, mutilation or other modification, or other derogatory action in relation to, the said work, which would be prejudicial to his honor and reputation.

As Hugh Jones points out, British remedies to the sort of actions that are legislated against in Berne have traditionally and characteristically been founded on breach of contract or remedies such as defamation action, passing off etc. (**Jones** p. 44). However, again characteristically (compare the situation relating to the famous three step test) when moral rights were introduced into English law, they came with new wording in the form of sections 77 to 89 of the Copyright, Design and Patents Act of 1988. Of these four statutory moral rights, three are relevant. Two are enshrined in the title of this study, how appropriateness of which will be seen. These are the right of paternity, "the right to be identified as author", and the right of integrity, "the right to object to derogatory treatment of work". The wording in quotations is taken from the text of the Act. The other relevant right is concerned with false attribution. In section 84 (6) "the right is ... infringed by a person who in the course of a business — (a) deals with a work which has been altered after the author parted with possession of it as being the unaltered work of the author". Unfortunately or fortunately this does not apply to literary works but instead to artistic ones.

4.1.3 Paternity and integrity in the law of England and Wales

The rights of paternity and integrity are hedged around with qualifications.

Paternity has to be asserted. Publication has to be 'commercial but the right of paternity does not apply to every sort of commercial work. Those exceptions, which apply to scholarly communication, include contributions to a newspaper, magazine or similar periodical or contributions to collective works such as an encyclopaedia.

Hugh Jones writes:

> "Where individual articles or entries are significant contributions to the literature in their own right, it will probably still be advisable to identify the author in accordance with the Act." (p.47)

For the purposes of this study the right of integrity is equally interesting. The crux is what 'derogatory treatment' means. Treatment is defined in such a way as to be fairly comprehensive. It means (section 80(2)(a)) "any addition to, deletion from, or alteration to or adaptation of the work". There are qualifications, which do not concern us here. This is fairly straightforward but derogatory treatment is

another matter. The definition is strong and uses such words as 'distortion', 'mutilation' and 'prejudicial to honour or reputation'.

Jones provides some examples from case laws previous to 1988 but points out that "we still await a fully decided case on derogatory treatment". (**Jones** p. 51). Meanwhile in his view:

> "Editorial 'improvements' in general are now highly unsafe, particularly where the risk of breaching the author's right of integrity is accompanied by the risk of defamation or breach of contract."

It is the duty of lawyers to play safe. Their clients expect them to advise on the worst possibility. In the quotation above any alteration is equated potentially with mutilation or distortion. It could also be and indeed is argued that derogatory treatment is possible, when, for example, an article or chapter is taken from its published context and re-used in another context, such as an anthology alongside another document offering views that are strongly disapproved of by the article or chapter author.

Moral rights do not exist in a legal vacuum and there are other types of law to fall back on.

Both moral rights can be waived by way of agreement in writing (section 87). Bently and Sherman (quoting Dworkin) comment:

> "It has been said that most 'objective observers would acknowledge that such wide waiver provisions, both in theory and in practice, erode significantly, indeed drive a coach and horses through the moral rights provisions". (**Bently** p.251).

This is hardly a vote of confidence.

4.1.4 False attribution

As previously mentioned, the other moral right with which this study is concerned is the right to prevent false attribution (section 84 of the Act). This right (as now framed) is not important for our purposes as it deals with the whole work not extracts from it. A person, who issues copies of a work to the public, on which there is a false attribution (**Bently** p.242) infringes the right. When a scholarly author complains to a publisher that another scholar has been guilty of plagiarism (apparently not a legal term) the author is not usually asserting that a whole article or book has been passed of as the work of another. The complaint is against paragraphs, pages or illustrations being lifted, copyrighted content permission for the use of which has not been asked for or given by the rightsholder. We are in the environment of general copyright law — the control of the right to copy, not the subset of moral rights.

4.2 MORAL RIGHTS IN THE EUROPEAN DIRECTIVE AND IN OTHER JURISDICTIONS

Some of the content of this subsection (and the following subsection) is put together as a result of advice received from publishing lawyers based in Europe

and also in North America. The interpretations are, however, the work of the author. There is no attempt at a comprehensive approach. The reason for looking at moral rights in other juridical traditions is that scholarship is global and, in the digital context, transmission of scholarship recognizes national boundaries. How legal systems can come to terms with the Internet is a huge question that obviously impacts on the theme of this study. However, for the purpose of this study, only part of this can be looked at.

4.2.1 The European Directive on copyright

The European Directive "on the harmonization of certain aspects of copyright and related rights in the information society" (**EU-CD**) was an attempt to adapt copyright concepts to the new needs of a networked society. For the purpose of this subsection, this document is worth examining. It is, after all, concerned with learning and culture, and the protection of content in these areas (note 14). E-commerce and the recognition of what successful e-commerce demands, balanced by the needs of users, including those engaged in research, are some of the primary objectives of the directive.

There are no direct references to moral rights or an overt attempt to harmonize the differences in practice between the traditions embodied in English copyright law, the concept of copyright as a property and the droit d'auteur traditions of most continental European jurisdictions.

Article 2 of the directive is concerned with the right to reproduce, and instructs member states to establish proviso for "the exclusive right to authorise or prohibit direct or indirect, temporary or permanent reproduction by any means and in any form, in whole or in part ... for authors, of their works". It is interesting and possibly also important to note that in the 1988 Act, the relevant section (16(3)(a)) refers to any 'substantial part. It will be interesting to see whether any adjustment is made in the implementation. This broad and general approach is followed through in article 3, concerned with the right of communication to the public, and the distribution right (article 4). In the section concerning exceptions and limitations, article 5.3 insists on attribution in most circumstances.

Finally, in article 7 there is a proviso against the circumvention of encryption, which is not unlike that already established in US jurisdictions as part of the Digital Millennium Copyright Act. Questions of encryption are considered in section 9 of this study.

As far as the law of England & Wales is concerned, implementation will almost certainly be in terms of an amendment to the 1988 Act rather than a new Act. It is not clear whether any changes will have serious impacts on the nature of the moral rights legislation that was mentioned earlier. However, the broad treatment of integrity, which seems implicit in the quotations made, and the emphasis on protection of rights throughout provide a framework in which scholarly interests are more than tolerated. The nature of the exemptions that are provided for scholarly researchers as users are likewise encouraging from the viewpoint of this study.

4.2.2 Moral rights in other jurisdictions

In continental European jurisdictions, moral rights include the right to divulge or disclose (the right for the author to decide when the work of the author should be brought to the public), and the right to retract. It is often said that authors cannot waive their moral rights, but as far as the right of integrity is concerned, under most jurisdictions, the right can be 'temporarily' waived under certain conditions to allow adaptations of works. These adaptations are defined in contracts.

Copyright law in the USA is superficially similar to the law of England and Wales. Many of the concepts are similar and of course case law has much the same interpretative function. However, the origins of the concept of copyright is much less based on property rights and is more concerned with the protection of the rights of the creator. When the USA joined the Berne Convention it did so whilst asserting that it did protect moral rights under general principles of unfair competition — someone could sue someone else if the latter person was misusing the name and reputation of the former.

In practical terms, where matters such as plagiarism are concerned, consideration of a possible infringement is often a mixture of what the moral rights would be if they were recognized and straightforward issues of copyright.

4.3 THE PROTECTION OF AUTHENTICITY AND PUBLISHING CONTRACTS

In this subsection and the one following I quote a number of legal counsel and others working in the area of intellectual property, who have given their insights into condition of anonymity.

4.3.1 The limits of moral rights legislation

Bently and Sherman, in considering theories of copyright and the justification used by lobbyists for their interpretation of the law, comment that "problems arise when people begin to believe the rhetoric and assume that copyright law is determined and shaped by ... philosophical ideals". Copyright law is essentially pragmatic and this pragmatism includes the interpretation of moral rights. Pragmatism is particularly obvious in English law because of the role of case law in interpretation, but it is clear that under continental jurisdictions, publishers could not exist if there were not compromises. Where there are even tougher author rights, as there are in countries previously part of the Soviet bloc, it is almost impossible to run a business.

The question implied by the previous subsections and now considered in this one is: do the moral rights of paternity and integrity (as enshrined in law, particularly the law of England) enable the author to stop the misuse of his or her content? The short answer seems to be a negative one, but let us look into the matter further. As scholarly content is characteristically entrusted to publishers, much of this subsection is going to be taken up with their attitudes and their actions.

4.3.2 Publisher policies

Our concern is with both cutting and pasting by users, and with slicing and dicing by the publishers or their licensees. As these concepts do not appear in the contracts and licenses discussed below, we are equally concerned with what is implied as with what is actually set out. It is appropriate at this point that we are seguing from the realm of copyright law into the realm of contract law. It is a truism that publishers, in the licenses they have offered, have moved to licenses because the copyright law did not define exemptions in the digital environment. Librarians, often the other signatory to the licenses involved, have tacitly accepted this approach. The history and progress of this shift in the way these matters are regulated is well documented on the Licensing Digital Information site based at Yale University (**Licensing**). The new copyright laws in Europe following from the European Copyright Directive already mentioned, as well as the Digital Millennium Copyright Act in the USA are not likely to change this significant cultural shift.

It needs to be pointed out that publisher motives in an area like this one are complex. Most scholarly publishers are aware that it is important for them to work on behalf of their authors and act for the authors, as they usually pledge to do in their contracts and in other circumstances. Part of their rationale for the assignment of copyright or the transfer of publishing rights is that the interest of the author will be protected, as we see below.

4.3.3 Clauses in publishing contracts

In the past, the commitment by the publisher was indicated in legal terms in the contract offered and also probably mentioned in passing in whatever instructions to authors the publisher handed out. In the current climate, scholarly publishers have become aware that they compete for authors, of journal articles as well as of books, and in addition the web-sites sometimes prominently display their policies of protection as well as exploitation of the rights they have taken. A particular good example is Emerald (formerly MCB University Press) whose long standing Literati Club (**Literati**) contains in its Author Charter the following sentiments under the heading 'Copyright principles':

> MCB seeks to retain copyright of the articles it publishes, without the author giving up their rights to use their own material. Authors are not required to seek MCB's permission to re-use their own work. As an author with MCB you can use your paper in part or in full, including figures and tables if you want to do so in a book, in another article written for us or another publisher, on your website, or any other use, without asking us first. We believe that this copyright policy benefits our authors by ensuring that we can:
>
> - Develop our electronic publications and their delivery to meet customer needs and create maximum dissemination of authors' work.
> - Protect authors' moral rights and their work from plagiarism, unlawful copying and any other infringement of copyright.
> - Recoup copyright fees from Reproduction Rights Organisations[1] to reinvest in new initiatives and author/user services...
> - Provide an efficient service for permissions

The document goes on to describe "your moral rights as an author" as follows:

- To be acknowledged as the author of your work and receive due respect and credit for it
- To be able to object to derogatory treatment of your work
- Not to have your work plagiarised by others

The sentiments are not unique even if expressed in more detail in this case.

Many contracts with scholarly authors specifically list moral rights but others do not. Here is a clause from a contract drawn up for book authors from a different publisher:

> 14 **The integrity of the Work**
> 14.1 The Publisher reserves the right for itself and its sub-licensees and assigns after consultation with the Author:
> 14.1.1 to alter the text including illustrations (either prepared by the Author or for which the relevant third party has given consent) of the Work after consultation with the Author in an appropriate way to exploit any of the rights granted by this Agreement
> 14.1.2 to delete anything which in the judgement of the Publisher or on the advice of the Publisher's legal advisers is considered objectionable or capable of being actionable at law.
> 14.2 The Author irrevocably and unconditionally undertakes not to maintain or support any claim for infringement of the Author's moral right of integrity in any part of the Territory by reason of alteration to or deletion from the Work made by the Publisher, its licensees and assignees.

The framers of this document specifically did not mention the assertion of the right of paternity, because what advantage is there in that assertion for the publisher? The last clause asks for a waiver of the right of integrity for reasons which will be explained later. Author organizations and agents would be unlikely to accept these clauses. Notice also that the word 'derogatory' is omitted from the final subclause because it has been found that even authors, not aware of such matters, are liable to protest over agreeing not to claim against derogatory treatment.

The following extract is from the same publisher but from the assignment of copyright for contributors to a journal:

> I/we recognise the need of the Publisher to be able to make available the Article without restriction, and irrevocably and unconditionally waive all moral rights to which I/we may be entitled under any laws relating to moral rights which may be in force in any part of the world. Notwithstanding the above the Publisher shall identify the Author(s) as the author(s) of the Article and shall not alter the text of the published article without the agreement of the Author(s).

What does the protection offered by an author's charter or its equivalents, contractual or otherwise, mean in practice?

A good indication of what publishers understand by protecting author interests in this area are the terms and conditions for customers and users that are often displayed on web-sites. An example of an intriguing turn-of-phrase is the following:

> "You may not integrate material from the Electronic Journal (s) with other material or otherwise create derivative works in any medium. This is not to prohibit quotations for purposes of comment, criticism or similar scholarly purposes."

The first sentence quoted seemed to have cutting and pasting in mind, but the legal counsel responsible for the framing was not able to confirm this. Cutting and pasting will be examined below (4.4.1)

4.3.4 Downstream licenses

It is also instructive to look at specific 'downstream' licenses that are negotiated by publishers with licensees such as library consortia or aggregators. 'Upstream' licenses are agreements between authors and publishers. In terms of downstream enforcement, almost every electronic license for published information has restrictions against alteration and modifications. The following example is an extract from a contract of yet another large publisher:

> Neither Subscriber nor its Authorized Users may modify, adapt, transform, translate or create any derivative work based on the Licensed Products, or otherwise use same in a manner that would infringe the copyright or other proprietary rights therein.

A legal counsel comments that this mostly concerns modification of content, not necessarily the protection of content authenticity, even if he also believes that it amounts to “nearly the same thing”. This is a contract written in terms of a legal jurisdiction in the USA, so it is not surprising that moral rights are not mentioned. However, a straw poll suggests that in European jurisdictions, including contracts referring to the law of England, downstream contracts or licenses of this type do not usually refer to moral rights. Neither do they typically ask for waivers of the right of integrity even when they are, as quite frequently, drawn up by the licensee. The question of the waivers will be returned to later.

The same counsel also highlights another consideration. Typically, in copyright law where commercial interests and the interests of the author as scholar run together, there will probably be an interest in enforcement. In his view:

> "There is a considerable incentive for publishers to act with their authors on such matters for commercial reasons as well. I think that part of the 'bargain' that many STM authors consider they are making with STM publishers is that the authors will transfer rights, but the publishers will enforce rights and police for infringing activity... enforcing such rights should help drive potential users to the official/authentic source, which should have commercial implication for the publisher...
>
> Most information providers consider these restrictions (in the license) important not only for the commercial reasons noted above but also out of concern for potential liability. Especially in STM publishing, some of the information provided has significance for health treatment, could be used in chemical experiments, etc. The idea that such information could be corrupted and modified in a way that could be dangerous is troubling" [personal communication].

With these statements in mind, a number of senior publishers and publishing lawyers were asked if they thought that publishers had an obligation to prevent third parties from infringing the author's moral rights. Those who replied were negative. A European lawyer writes:

> "Only if it could be said that a publisher actively encourages and in so doing participates in the infringement of moral rights is he in violation of his obligations" [personal communication].

Another lawyer has provided a personal opinion, applying to European jurisdictions in the droit d'auteur tradition, which raises a number of issues both in this section and in section 9. An extended quotation is relevant here:

> "Publishers have an obligation to respect moral rights when publishing a work. Of course, if they decide to protect the work they published they would also protect the integrity right, which is essential in the digital world. There is no statutory obligation to encrypt or otherwise protect the work against destruction of integrity, although I would not be surprised that a number of contracts would have such provisions.
>
> Concerning downstream rights, in the contract between author and publisher you should define to what extent the adaptation of the work is allowed.
>
> So to resume there is nothing in continental laws that force you to put in place technical measures protecting the integrity of the work. There is in the Copyright Directive the obligation that, when Digital Right Management (which may include encryption etc) is attached to a work, it is forbidden to remove it (but that applies more to users)" [personal communication].

The concerns of the above final paragraph will be the raised in section 9. The middle paragraph could refer to an action by the publisher and those it licences, which is frequently defined as slicing and dicing and which we consider in the next subsection. It could also refer to cutting and pasting by users, a consideration of which will form the first part of what follows.

4.4 THE PROTECTION OF AUTHENTICITY: AUTHOR CONCERNS AND PUBLISHER PRACTICES

Although it has been suggested earlier in this study that scholars as authors and scholars as user/readers are at one in their concern with authenticity this is not the whole story. Admittedly the assertion has been qualified further but now further qualification is necessary.

4.4.1 Cutting and pasting

It can be argued that those scholars who are also users, who work in the digital environment, are not different in the way they play on screen from the rest of the population. Cutting and pasting from the work of others is the easy way to put together ones own message and it is easy too to lose track of what has been

borrowed. There are as yet no clear conventions. As usual Dorner writes perceptively from the point of view of the general writer (**Dorner** p.139):

> "Technology — and nothing more so than the Internet — provides effortless ways of pouring other people's ideas and expressions into your own computer. Once lodged on hard disk, cuttings can be neatly subject-sorted, keyword-searched, text-retrieved and amalgamated.
>
> At this point control of the text passes into the user's hands and it is the huge potential for manipulation, recycling and re-ordering that makes digital texts more vulnerable than their print counterparts. Electronic words, detached from their contexts and comfortably sitting on your own screen, begin to feel like your own, don't they ... it is difficult to regard electronic words as property because they appear insubstantial. It is difficult to realise that unauthorised cutting and pasting is theft."

Dorner goes on to point out that every text is the result of borrowing and explains about technical solutions. However, there is a point embedded here that is central to this study and that is difficult to come to general conclusions about.

Firstly, it is not clear that all scholars actually want to have their name attached to all their ideas, all of the time. They want the ideas to be transmitted, not the source of the ideas. Several publishers have noted this reaction from the community of scholarly authors. Perhaps more important is what happens downstream. The first user can evaluate the sources used, as all scholars do and in ways that are touched on in section 5. However, the user disseminates the ideas that he or she has absorbed into his or her own work and passes them on to others, the second level of users and so on. The first user becomes the author. This has always happened and always will happen. Dorner's point is that it is so much easier to cut and paste in the digital environment than in print — and this fact itself results in new situations and possibly a need for new conventions. In the consultations preceding this study, I proposed to a number of author representatives and publishers the idea that there might be a tension between the wishes of the author to make sure that what he or she writes gets to the reader uncorrupted and appropriately ascribed and the cutting and pasting, which is so easy in the Internet environment. Whether there is such a tension among scholars is not proven.

4.4.2 Publisher policies regarding cutting and pasting

Those publishers that were consulted about questions of cutting and pasting had obviously considered the issue. The approach adopted by one main publisher of scholarly books was that generally mentioned. The publisher writes:

> "We are as strict as possible about the use of material and how it must be represented. We do not allow wide use or reuse without our consent. On our ebooks etc we do not allow pasting tools to be used (and) on eBrary any chunk of material comes with a full reference attached."

The modus operandi of eBrary (http://www.ebrary.com) has changed somewhat since this personal communication was provided, but the offer of the company still revolves around downloading content on a page by page basis.

Another publisher writes:

> "If we or one of our authors came across a case of blatant misuse, we would do what we could to get it withdrawn, relying on the law."

Is this always what the scholarly author wants and is it what the scholarly user wants? Is scholarship served best by the greatest diffusion of content or by the protection of the authenticity of that content?

Publishers have always responded to the taking of another of large chunks of copyrighted content in print. But in the scholarly environment publishers have rarely, if ever, gone to the law and have instead relied on strong pressure to withdraw, or even naming and shaming. If such misuse is so much easier in the digital environment, what will their position be in the future?

4.4.3 Slicing and dicing

This subsection continues with a consideration of slicing and dicing, actions taken or sanctioned by the publisher. A writer on copyright issues makes a strong point:

> "Slicing and dicing potentially infringes two moral rights — the right to be identified as an author and the right to object to derogatory treatment. If the author remains named on each slice/dice piece, there is no problem with the first moral right. Slicing and dicing is (however) potentially an infringement of the moral right against derogatory treatment and it is a foolish publisher indeed that engages in slicing and dicing without getting the permission of the author first."

The concept of slicing and dicing is one much admired by corporate strategists and it has come into scholarly publishing from areas, like much of business-to-business (B2B) publishing, where the publisher deals with data. The strategist envisages, not a series of individual publications, but a database from which material owned by the publisher can be run off in various different manifestations, both of content and of format. It can be re-purposed. Each piece of data can be used in a number of different ways, and the costs of acquisition and processing in this way can be leveraged to the profit of the publisher.

Unfortunately this model does not work in scholarly publishing or for that matter in educational publishing. Scholars produce knowledge not data. Databases in B2B publishing were/are highly structured to a detailed level of granularity, making it possible for specific nuggets of data to be run off as needed. Databases in scholarly publishing, or (internal) archives as they are now often known, have been common in journal publishing for some time but publishers have only just begun to construct them for non-journal material. It is indeed remarkable that the many scholarly publishers, intending to make monograph content online, have just not thought through the problems of holding such material on their own site, which is usually their eventual aim (**Watkinson 2**).

There is a distinction between journals and books, with, it should be added, encyclopaedias treatable and treated as journals. In journal 'packages', as they are sometimes called, the level of granularity is the individual article and, as we

see in section 8.3.2, this is the object that is identified and can be sold. Clearly the entry in an encyclopaedia is somewhat similar. Over the last decade there have been many discussions about lower levels of granularity being exploited, for example medical illustrations, but so far the level of commercial interest has not justified more than marginal activity at this level.

For many books however, the level of granularity is the book itself not the chapter, which is not yet written to stand-alone. Traditionally, book publishers encouraged their authors to think of the book as a single unit rather than as a collection of chapters, although anecdotal evidence is that there is a change in editorial advice. There is as yet no serious experience in commissioning electronic-only monographs, even though the occurrence of such entities is rising. Not only is the chapter not the natural (in the sense of scholarly) level of granularity, but at the same time book chapters, together with review articles in journals, lend themselves to subdivision. It is possible to envisage the separate selling of segments of a chapter, as will be discussed in section 8.

4.4.4 Publisher policies relating to slicing and dicing

Scholarly publishers are edgy about what they might wish to slice and dice. The advice given by lawyers to journal publishers is not uniform. In spite of the apparent lack of applicability of moral rights clauses in the 1988 Act to contributions in journals, waivers to the right of integrity were insisted on by some lawyers, which explains the wording in the assignment of copyright given as an example in section 4.3. Waivers are customarily asked for by book publishers and are resisted by author's agents and those bodies that represent the author communities. Obviously neither of these groups represents many scholarly authors.

On the other hand, as the database approach to scholarly publishing mentioned previously has been replaced by what one might consider an approach more in tune with the nature of the scholarly endeavour, many publishers in the UK are keen on working with their authors over matters relating to integrity. As we have seen in other European jurisdictions, such an approach would be compulsory in any case.

In my experience, this represents something of a change in policy for many publishers.

Because the issue is so central to this study, quotations from two large publishers concerned with scholarly communication follow:

> "1. We actually take very seriously respect for integrity and paternity. We consult with and get adaptation/re-engineering of their work we propose ourselves, and for any licenses we grant (print or electronic).
>
> 2. We are not slicing and dicing at anything below chapter, article and encyclopaedia article level. Technically of course you can slice and dice at any atomic level but in practice, I do not see any drive to do this.
>
> You can stop cutting and pasting by publishing in PDF, but you cannot stop it technically with HTML, except by legal and cultural means i.e.

people will not do it because they have signed an agreement not to do it or they do not do it because this is wrong."

This final paragraph looks back to the previous discussion. Is it a cultural question? It also looks forward to section 9 where the questions of protection by technical means are taken up.

5. Authenticity, certification and the definition of a publication

This study has so far considered authenticity in the philosophical sense, which mainly relates to and that leads on to archiving and preservation criteria, and relating to authenticity in the legal sense, which mainly relates to integrity and paternity. In this and the subsequent section, the concern of this study is with the definition of a publication. As we have seen, concerns about authenticity cannot usefully be separated from the purpose of the entity to be authenticated (section 3.2). In section 10 we see that it is publications that are central to discussions of what we need to archive and preserve in the digital environment.

As is explained in the second subsection, I have been closely involved in preparing papers on the issues involved in these sections. It was and is my contention that in scholarly communication the definitive publication has a special and continuing role and that this definitive publication is certified or validated. I started out writing section 4 with the assumption that moral rights concepts had a direct relationship to authenticity. As we have seen, this is not the case. During the course of the writing of this section, I have come to realise that my assumptions as expressed in section 5.3 need to be qualified. In particular, the special role of certified publications is seen to be less straightforward and the question of whether or not the digital revolution has impacted on formal publishing has been re-thought. Unfortunately, as we will see, particularly in the latter subsections, there is no new model that works, and more questions are raised than answered.

In the first subsection, the study further examines the relationship of concepts of authenticity to the question of publication. The second subsection is dedicated to the debate about the definition of a publication mentioned earlier. In the latter group of subsections, some of the central points in this debate are unravelled and discussed further, specifically the definition of a publication, the centrality of certification by peer review and the question of versions. Finally, the impact of alternative approaches to publishing is discussed. Alternative publishing approaches are discussed under a number of headings in an attempt to see how the varying models themselves, together with their execution, deal with certification and other aspects of authenticity. As we have mentioned in the preface, much of this section was written before the movements towards institutional repositories became 'flavour-of-the-month'. If it was written now, the balance would be different.

5.1 HOW DO QUESTIONS OF AUTHENTICITY RELATE TO DEFINITIONS OF PUBLICATION

The concern of this study is with scholars and with the progress of scholarship. Regarding the 'Proposal' entitled *Defining and Certifying Electronic Publication in Science* (**Frankel**), with which we are primarily concerned with for much of this section, one academic sector writes "Publication is the hard currency of science". Another author already quoted (**Mabe2**) writes of the "minutes of science". These statements can be applied to all scholarly endeavours. Publications are

what are passed down to posterity. These are certified publications. What this means is examined later subsection.

5.1.1 Archives are for publications

The assertions above are not just assertions by a publisher. National collections in print and projected for digital content have as their primary aim the collection of publications produced in the country concerned. It is recognized that ephemera in print are worth holding on for social historians and others, and some national libraries, for example in Sweden, are attempting to capture pieces of the Web in the non-print environment, though this is a secondary concern. In this context, the national archives know that something is a publication because it has been published: it has, for example, an International Standard Book Number, which comes with it and on it. This is just an example. Offline digital entities tend to be delivered to the national depository just because they have an ISBN. Actually, this is a statement made for the purposes of the argument being developed in this section, but which cannot be substantiated in such a sweeping way in the general context of archiving and preservation, as we will see in section 9. It does, however, reflect the relationship in print between publishers and the national libraries.

5.1.2 Publications need to be authenticated

This study will be considering archiving in the digital environment in a future section (9) but enough has already been written in a previous section (3) to make it clear that determination of authenticity is not something that will be entered into lightly. It is the unspoken assumption that these items to be archived and preserved are publications. More than this, they are publications worth preserving. What this mean in terms of definitions will be examined later. Our concern here is just to establish links between this section and the thrust of the whole study. It could reasonably be argued that a certified publication is in some real sense a more authentic expression of a scholar than an informal publication. That is what he or she wishes to be judged by. The criteria for certification and how it is decided, is a different matter and, judging by our findings later in this section, is probably not something that can be accomplished.

5.1.3 An alternative view

It could be argued, and indeed has been argued by one of the authors of the Proposal in a private communication, that my position is based on:

> "A fundamental and dangerous confusion between the following things:
>
> Authenticity – i.e. it is what it says it is, and by who is says it's by
> Definitiveness – i.e. it is a version which overrides all others
> Essential-ity – (there must be a word for this) – i.e. it is the irreducible minimum, which should be preserved".

The argument of this study would be that these are different aspects of the same entity for those charged with archiving and preserving such entities, but there is undoubtedly something to prove.

The question of how one decides what to archive if it is not possible to archive everything, definitive or otherwise, us another related argument that will come up in future sections, particularly section 9. The concept of definitiveness elides into the concept of authoritative. They are often used as if they are synonymous.

5.2 THE DEBATE CONCERNING THE DEFINITION OF A PUBLICATION

The word 'debate' is used because this section in particular, and to a lesser extent section 6, is based on a debate surrounding a report from an International Working Group already mentioned (**Frankel**). This section is based on my own contributions to this debate. The debate as such has only happened within various conferences, and my contributions, commissioned by the International Association of Scientific Technical and Medical Publishers (STM), have had restricted circulation and have never been published. In the current circumstances this is ironic. However, I have been given permission to make the substance of the reports available. The actual published document from a group of individuals encouraged by the International Council of Science (ICSU) is subtitled *A Proposal to the International Association of STM Publishers*. It is referred to subsequently as the Proposal. The unpublished reports referred to above were written as responses to that proposal, and in that sense there is a debate.

5.2.1 The antecedents of the Proposal and subsequent developments

It is also ironic, particularly in view of the discussion of 'versions' that will follow, that this report is now available in three different locations, and no doubt there are others. It is an appendix to **Sandewall**, on the web-site of the American Association for the Advancement of Science and in *Learned Publishing* 13:4. There are slight differences, at least in the headings, and for convenience the last mentioned version is taken as definitive.

The antecedents of the document are now explained. ICSU Press, now the ICSU Committee on Dissemination of Scientific Information, has for many years been concerned with the impact of the electronic environment on the communication of scientific knowledge. Two important meetings were the Expert Conference held in co-operation with UNESCO in 1996 and the subsequent international workshop on economics, real costs and benefits of electronic publishing (**Shaw**). The organisers of both of these meetings took considerable care to make sure that not only were working scientists represented but also that delegates from all parts of the information chain were invited. The participants included some working publishers.

A further international workshop to discuss normative issues was held in October 1998 (**AAAS**). The recommendations, which are key for our purposes, are given below.

a) “The existence of multiple versions brings the possibility of confusion in citation and referencing, and the Workshop RECOMMENDED that each publicly available version of a document carry a full specification of its status laid out in a visible and readily understandable manner”

b) “Because of the possible existence of multiple version of a document, there is need for a convention on the citation of electronic material... The Workshop RECOMMENDED that the scientific community become involved in the development of standardised citation practice that are friendly to science, include appropriate metadata, capable of automatic assignation and easy to use.”

c) “Formal peer review was regarded as essential in arriving at the final version of a scientific publication... The Workshop ... RECOMMENDED that scientific societies and/or journals establish and distribute guidelines in order to maintain the quality and integrity of the review process.”

Obviously, recommendations a) and b) are of great importance, especially from such a source, to subsequent sections of this study. Recommendation c) will be considered in a later subsection.

The Proposal makes clear that the subsequent smaller group who produced this later document were working within the conceptual framework of the workshop.

Since the Proposal was produced there have been further gatherings convened by ICSU and UNESCO that considered its thesis. At present, the debate seems to be in abeyance — which is a pity.

As far as we can tell, the publishing model espoused, which we perceive as emerging mainly from practices in high energy and particle physics, has not been embraced in most other subject areas — contrary to the expectations of some within that part of the physics community. It is also not our impression that learned bodies representing specific disciplines and subdisciplines have come to a positive view of the proposal. We note that the editorial board responsible for the excellent programme on electronic publishing in science did agree to follow the recommendations contained in the proposal. We also notice that the record of the conference (http://associnst.ox.ac.uk/~icsuinfo/proc01fin.htm) does not contain much discussion of the ideas we are commenting on, with the exception of the important paper by Joost G. Kircz.

We will make use of the ideas of Professor Kircz in section 6. His critique of the Proposal is summarised in section 6.1.1. Some of the other points raised will be discussed, particularly those that embody a pessimistic view of the model proposed. The view of the STM Council in deciding not to respond was probably influenced by a conviction that there was, as it were, no case to answer. Personally, I do not necessarily agree with this view and I will examine it further in this section.

A more recent item on the web-site of ICSU are recommendations of the Paris conference of 2001 (**Recommendations**), which have appeared since the above was written. Some of these are relevant and will be returned to below, for example, recommendations 2 and 5 on peer review in 5.4.2., and both 4 and 6 when the position of e-prints (**preprints**) are considered in 5.5.1 and 5.6.

5.2.1 The content of the debate

The basic position of the Proposal is that impact of the Internet on scientific communication has challenged the way in which scientific publication has been understood in the past. This understanding leads to the recommendation that two types of formal publication should be recognised and identified by all those involved in the communication of science. The two definitions have been given the tentative names of 'First Publication' and 'Definitive Publication'. The latter definition represents a renaming of what is currently regarded as a 'publication' by scientists. A generally acceptable description could be that this is a document that is peer-reviewed and stable, made available in a recognised journal and archived as part of the 'record of science'. The former definition refers to a document that is also stable and 'notified' but that is not peer-reviewed or published in the sense that is currently and has previously been accepted in the world of scholarship. This succinct summary is expanded on in section 5.3.1.

This document is written about scientific communication rather than scholarly communication in general. In subsequent subsections there will be some consideration of the different ways in which different groups of scientists rate different types of communication, but the expressed need and the expressed solution set out above is a relevant and worthwhile attempt to solve common problems.

The basic positions, which were set out by the author in the unpublished responses, are that the digital environment has made it possible for much more efficient and effective ways for informal scholar communication, but it has not altered the boundaries between informal and formal communication that are so important to scholars. The actual definition of 'First Publication' is discounted as only acceptable to a small group of sub-disciplines. In any case, a new definition such as this one can only cause confusion in a situation that is already confused.

It is worth adding to this short summary that one point became clear during some informal interaction. The link between First Publication and Definitive Publication is not one of an early draft and a final draft. The majority of e-prints subsequently become articles in a journal, but under the scheme envisaged above, the e-prints that were not offered to a journal and, more important, the ones that were offered to and rejected by a journal would remain on the site in perpetuity. First Publications are considered stable.

Although the problems remain as stated, I am now even more appreciative of the exercise that is attempted by the Proposal in subsections 5.3 and 5.5.

For the purpose of this particular study, it is our intention to tease out the ideas previously set out, in the context of the studies overall aim.

5.3 WHAT IS A PUBLICATION?

This is an intrinsically difficult question to answer. Scholars on the whole think that the answer is obvious and, on the whole, this confidence is shared by 'traditional publishers, but almost any other commentator finds the nature of a definition problematical to establish.

5.3.1 Publication defined in the Proposal

Any content made available on the open Web is consequently made public or published in a way that was not possible in the print environment. One idea that came out of the working group responsible for the Proposal was the distinction between publication with a small 'p' and Publication with a large 'P'. Publication is reserved for defined Publications. It was decided that the following characteristics were required for a Definitive Publication. Remember that this definition is concerned with the realities of the digital environment. I have slightly compressed the list in the Proposal itself, but where the definitions is particularly relevant, I have left it complete and in quotation marks.

- Peer reviewed
- Publicly available
- The relevant community “made aware”
- “A system for long-term access and retrieval must be in place” Not changed and preferably with technical protection
- Not moved unless legally unavoidable
- Unambiguously identified, e.g. by a Digital Object Identifier (DOI)
- “It must have a bibliographical record (metadata) containing certain minimal information”
- “Archiving and long term preservation must be provided for.”

This is undoubtedly a sensible list of characteristics, but as a benchmark comprising both a complete and necessary description it does not, in my view, fit in with how the communities actually work. It is an ideal (perhaps to be aimed for) rather than a reflection of practice. What is clear, if nothing else is, in this section of the study is that almost all aspects of scholarly communication are more difficult to define in the digital environment than one would have hoped. The implications of these problems of definition will become apparent in later sections.

Consideration of the proposed First Publication is part of the content of section 5.5.1

5.3.2 Definitions of a Publication in the legal context

The working group, who produced the Proposal, wrote the following footnote:

> "An analysis of how the law will affect 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. We also recognise that how publication is defined can affect contractual provisions."

This is a timely recognition that there is something of a minefield here. In particular, the definition of First Publication', which we will return to in section 6, is especially dangerous to attempt. Legal counsel, consulted by the author, has strongly advised on any sort of 'definition' that might conflict with what the law has come up with. The relevant term in the actual legislation is 'work', which tells us little. It is in patent law that the legal system impinges most alarmingly on this whole question. First publication, as defined by the working group, is not

necessarily first publication as defined by patent lawyers. There is little case law on publication as recognized by the scientific community. It would be best if the two classes of definition remain separate. It is our view that we are in the area of recognized conventions, arrived at through consensus rather than legal formulations. It is interesting that in the United Kingdom, what is viewed as a publication under the terms of the 1911 Act and what is therefore liable to deposit, is in practice determined by an ISBN. This is assigned by and claimed for the publisher, and none of the characteristics listed above are demanded as part of the claim. In the digital arena there seems to be no pressure to change this pragmatic approach.

5.3.3 Informal communication and publication

Much of science has always been done informally. Scientists have always corresponded. Such correspondence and notebooks, when they have survived, are an important part of the history of science. Travel to attend meetings has always taken a significant part of the time of most scientists and, in spite of e-mail, conference calling and teleconferencing, this urge to have a face-to-face exchange of views does not seem to diminish. The idea that the scientific process depended in the past on print, and that the reader had to wait for the journal to wend its way round the world by surface mail before they knew what the author was up to, is an erroneous construct. Current communication has never been confused with the establishment of a peer-reviewed record.

The argument put forward in the responses is that the Internet has made informal communication much easier, that the network of interacting scientists can be much wider and more international than it used to be, and that real-time interaction by e-mail speeds up discovery. There is also access to data. Scientists can draw directly on electronic resources for huge amounts of data, and more importantly can manipulate it, in a way that would have been unthinkable before. The availability of informal communication to those outside the rather narrow network — a feature of most subdisciplines — can be exaggerated. How many scholars actually surf the Internet for information about fields outside their own immediate area of knowledge, rather than going to the literature? The crucial point is that the Internet has made such communication 'public' in a way that it was not before.

5.3.4 Definitions emerging from the scholarly community itself

When I described this project to a biologist, now working as a librarian, he was scandalized by the temerity of those seeking to define such matters from outside the community. At the time, such sentiments seemed a bit extreme but the material set out below seems to demonstrate part of the difficulty of finding any definitions that suit all.

The following publisher viewpoint summarises what this study asserts and is set out in more detail later:

> "Each field is different and in the fields where non-peer reviewed stuff is cited, haven't they always done that? The digital environment is making some distinctions more important and has highlighted the need for a peer reviewed, authoritative article — even the Public Library of Science doesn't

argue with this, their issue is with the terms of access to the peer-reviewed articles."

The last point made in this quotation will be taken up again in section 5.6.

As part of the preparation of this study, a number of specialist publishers from science-based subject areas, usually from learned societies or other not-for-profit organizations, were asked two questions. These organizations were all in science, broadly interpreted to include, for example, engineering. The respondents were chosen carefully on the basis of personal knowledge. They are all publishers with significant experience in the discipline concerned. The suitability for archiving was used as a touchstone because this brings a practical way of eliciting an understanding of the attitudes of a particular discipline. The questions were:

i. Do the communities you publish for view the refereed article as the only authentic utterance of an author in the sense that only such an utterance can be expected to be archived as part of the record of science?

ii. Do they conceive of any change taking place and, if so, would they expect e-prints, for example, to be archived for posterity because they are worth saving – even if not refereed or 'certified'?

Unfortunately there were not enough responses to produce a statistically worthwhile table, but some general and specific attitudes can be discerned and will be set out below.

Broadly speaking, the answers were distinctively positive to the first question and generally negative but less so in answer to the second. The answers were qualified but distinct. It was quite possible for a respondent to view peer-reviewed publications as special but at the same time suggest that e-prints, as well as journal articles, should be archived. The more detailed answers (looking behind the literal question) demonstrated that actual practice in different disciplines does differ a lot. It seems that there is a gradation of respect rather than a distinction between one type of publication (peer reviewed and certified) and others. This came as something of a surprise to me.

For many biologists, a publication that is not in a journal recognized as respectable should be cited as 'personal communication'. It is a pity that this approach is often taken to be a model for science. Even other biologists, for example, in applied areas such as forestry, are happy to cite in-house reports, which are clearly part of the grey literature.

What is also clear is that in most disciplines scholars view publications that do not offer the same level of refereeing as a journal, perhaps only approval of an abstract, or sometimes no refereeing at all, as usable and worthwhile citing. It is interesting that conferences (not mentioned in the question) were often mentioned by respondents — though not by biologists.

Engineers are often mentioned as using conferences a lot for getting out their results. A civil engineering publisher confirmed:

> "Our sector does make great use of conferences. They are certainly worth citing, but they are perceived as second rate – our editors are very twitchy about including papers in our journals that have previously been published in conferences, even when they have been reworked" (personal communication).

The reference to editorial policies is most interesting and compares with the attitude of many journal editors to e-prints, which we touch on in section 5.6. There is an ambivalent attitude in this case. If the e-print is not a publication, why not accept it for consideration.

A mathematical publisher gave an interesting response to the first question:

> "Conference proceedings and other works that are not refereed to the same high standards (as journals), but weighted according to the care taken in their preparation and review".

A senior publisher in chemistry had a different angle, which is particularly interesting because of the reference to archiving and the contrast to the longer comment by the physicist below:

> "In chemistry the refereed article (or book chapter?) is by far the most valued vehicle for scientific communication. Papers delivered at professional meetings are regarded as authentic, but they are not systematically preserved, although most abstracts are captured in secondary databases. A lot of this so-called grey literature just disappears, while some of it turns into journal articles and book chapters."

Finally, one answer from a physics publisher deserves to be quoted at length, because it amply demonstrates the divergence between stated principles and actual practice. The final paragraph is particularly interesting:

> "The global answer to this question is "yes, no, and sometimes." In certain communities, where electronic preprints form an essentially complete record of the research literature — even though most preprints eventually enter the formal literature — the refereed article is not the most important portion of the research record. "Communities" in this category include, of course, high energy theoretical physics (particle physics), some segments of the astronomy community, and mathematical physics. Together, these communities represent far less than 10% of the broad physics community and probably no more than 2% of the greater physics and engineering communities.
>
> Another segment of our community, which could be identified as "basic physics," considers the refereed article as the definitive publication. While this community, especially the theoretical (as compared to experimental) portion of this community, might make use of electronic preprints, they would not cite preprints in their peer reviewed articles. (There is one exception to this; in the case where a preprint has not as yet been published in the peer reviewed literature, it may be cited in the same sense as a "private communication" or even "in press.") This community is much larger than the "preprint-powered" community specified in the

> preceding paragraph. I would estimate its size at 50% - 60% of the physics community, and perhaps 25% of the greater physics and engineering community.
>
> There is a third community, however. This community can be defined (loosely!) as "applied physics and most of the engineering community." This community includes the remaining 40% or so of physics and the vast majority (80% or more) of the engineering community. For this community, conference papers — which may be peer reviewed, but often are not — are as important as peer reviewed literature. Indeed, for the engineering segment of this community, the conference paper represents the definitive publication for most engineers. Both applied physics and engineering heavily cite published conference proceedings in their peer-reviewed articles. So, for this community, conference proceedings are considered to be equally important as the traditional peer-reviewed literature.
>
> Certainly the first of the communities described above will insist that e-prints be archived for posterity. If the Ginsparg archive (arxiv.org) were not to be archived, there would be a high percentage of "dead links" in this community's literature. I believe that the second community I described (i.e., formal physics) would also like to have e-prints be archived, but the community would not be extensively harmed if the e-prints were not archived. The applied physics and engineering community would not care about preserving e-prints."

Obviously some of the points made here will be relevant to the next subsection. There is no attempt at comprehensiveness here, but it might be possible to produce a chart showing what each discipline or subdiscipline recognizes as a 'publication'. However, such a chart would have a short lifetime. A final point made by the physicist quoted above was that the picture is changing all the time.

5.4 THE CERTIFICATION PROCESS AND PEER REVIEW

Bearing in mind how difficult it is for any one scholar to agree with another scholar over what the defining characteristics of a definitive publication are, it might be thought that an analysis of what certifies or validates an informational entity would be even more problematical. It is. We are thinking of scholarly journals in this context. In this subsection, we first examine the process of certification in scholarly communication — another way of looking at the achievement of a definitive publication. We then look at peer review, and finally reflect on how the whole process is actually perceived by the communities themselves.

5.4.1 The place of certification in scholarly communication

The (laudable) aim of the authors of the Proposal was to be “helpful to scientists in this increasingly fluid information environment”; the main objection to their new definition of a First Publication was that it made the situation more confusing. It seems to me that the well-known phenomenon that where there is a period of change secure solutions have a premium is at work here. The way in

which impact factors have become so important is another example. We consider this in the concluding section.

This reflection is relevant to the question of certification because it is certification that is the gateway to the record of science (see 5.4). It is the way in which the process devised by the scholarly community to decide what should currently be paid special attention and handed down to posterity. The system is particularly well articulated in science, in the STM information system (**Watkinson 3**). The system has been much criticized as inefficient and inequitable (see http://cogprints.ecs.soton.ac.uk/~harnad/ for one way into the extensive literature). The big problem for those concerned with alternative approaches is that currently it is scholars, and not just scientists, who have handed over the organization of the certification process to publishers. The focus on what publishers say and do in some parts of this section, and indeed elsewhere, might be argued against but can, in my view, be justified because in practice they make or encourage most of the decisions.

There is no doubt that the system of certification has been looked at as never before. Essentially, this is because of the digital revolution. **Roosendaal** and Geurts, in their seminal work, analyse:

> "The transformation of the familiar, linear scientific information chain into an interactive scientific communication network in response to concomitant changes in scientific education and research" (**Roosendaal**).

A particularly important framework is laid out as follows:

> "The scientific communication market is described in terms of four main forces and their interplay. These forces are the actor pair (author/reader), accessibility, content and applicability. Scientific communication is described in terms of four functions: registration, awareness, certification and archive."

Although I have taken a different route by which to examine the concepts of concern, throughout this study the framework is implicit.

On the whole, the concept of certification is not challenged. It is the way it is run (too expensive) and who runs it (part of the Faustian bargain) that is seen as inappropriate in the post-Gutenberg age. These are terms invented by Stevan Harnard, whose presence lurks behind any discussion concerned with the future of scholarly communication in the digital environment. The details of his views are probably to be found at greatest length at http://www.arl.org/scomm/subversive/toc.html, and more formally and in a more succinct form at the reference under **Harnad**. It is my impression (and this impression has been confirmed by private communications by those sympathetic to my views) that Harnad is not interested in all issues with which this study is concerned. At any rate, his views are not analysed in the study.

For the purposes of this study, it is sufficient to recognize the fact that this system of communication exists. The way that this system operates in practice leads to a whole range of questions relevant to future sections — those on

metadata and on archiving in particular. For this reason the longer part of this subsection is concerned with peer review.

5.4.2 Peer review in a period of transition

There is an amazing agreement in almost all that is written now about the central importance of peer review. If there is a certain ambivalence in this subsection, and even an element of contradiction, it reflects the literature. There can be few subjects where the analysis by the information scientist differs so much from the strongly held views of the practitioners. We have already seen that in practice peer review is not so necessary. Indeed and in addition, it could be pointed out here that some of the most important journals are not peer reviewed in any traditional sense. Proceedings of the National Academy of Science (PNAS) springs to mind.

As far as peer review is concerned there are plenty of ways of doing it, but however the process is conducted, the fact of peer review seems to give confidence to scholarly communities. For a particularly interesting discussion from within the e-journal community see www.niwi.knaw.nl/ccsc/summary.htm. Even within this community, open peer review is not necessarily accepted in the sense that the names of the reviewers are disclosed to the authors.

The actual usefulness of peer review as a guide to whether it leads to good decisions or not — decisions of acceptance or rejection — is much doubted. It is, however, generally agreed that a well-organized peer review system improves papers that are actually accepted if the reviewers' proposals for revision are taken into account. There is much literature on this topic, mainly from a medical viewpoint, which is most easily accessed at bmj.com and then by using 'peer review' as a search term. It goes without saying that certification in the form of making sure that units and dosages are correct is an obvious example of why some sort of intermediary process is important to the user/reader.

Is there a real need for an official (coming from a body representing scientific unions) code for peer review as envisaged by the recommendation lettered (c) above in 5.2.1? The answer here, looking forward to future sections, would probably be that it is a good idea. The answer, looking at the different practices discipline by discipline, which extend to peer review also, is that it would probably have little impact. It should, however, be added that the recommendation might suggest action by individual scientific unions. There is some ambiguity. If this is the aim, perhaps there is some chance of action to create standards. In 2001 (Recommendations) there was a further recommendation concerning peer review:

> 2. Peer review is essential to ensure the quality of scientific information. A standardized approach across all disciplines for peer review would be inappropriate. There should be further study of alternative approaches to peer review (including more open variants) in order to assess the impact of such processes and associated behaviour. The results of this experimentation should be widely communicated.

The view, which would seem to arise naturally from this study, is that such a recommendation is less likely to have any practical use

A second recommendation is more difficult to understand:

> 5. When technically feasible, publicly available and particularly peer-reviewed versions of articles should be authenticated to guarantee that they are the correct version

Presumably this relates to metadata, which will be considered in a future section.

Among the results of the digital revolution there is a speeding up of the reviewing processes by using e-mail and other, more sophisticated, systems, together with an actual improvement in the way the refereeing is done. It could be argued that as publishers are encouraged to look more closely at their core competencies, peer review is treated more seriously by both publishers themselves and the editors or editorial boards that they support. The range of, and changes in practices of peer review are well documented in a survey by two bodies in the field (**ALPSP**).

Various ways of operating open peer review (disclosing the names of referees) have been adopted by some journals and has been promoted by those who support such approaches. Richard Smith of the British Medical Journal has argued (**Smith 4**) that open peer review could increase the integrity of the scientific record in a rather convincing piece. He is writing of the digital environment and his recommendations are worth quoting in full because so many of them are relevant to one or other of the topics we are investigating:

1. Scientific papers could be published together with the full raw data plus the software used to analyse the data.
   Much fuller methods can be published than is usually possible in paper journals.
2. Authors could be obliged to complete standard forms — for instance, the CONSORT criteria for the publication of randomised controlled trials — to increase the chance that all essential information will be included.
3. Study protocols could be peer reviewed and published. Journals that have accepted the protocols would then be obliged either to publish the final study results or give a reason why not.
4. There need be no problem getting "boring negative" results published because space would not be at a premium.
5. The peer review process could be conducted openly on the web, increasing, for instance, the chances that somebody will recognise something being published twice.
6. All discussions that took place in the peer review could be posted on the web, allowing those interested to reassure themselves on the integrity of the process.
7. Corrections could be posted almost immediately. Indeed, the study could be modified--with some record that a change had taken place.
8. Much more space would be available to describe who contributed what to studies, avoiding the problem of gift authorship. Standard proforma might be used.
9. As the full text of studies becomes available on large databases it will become more difficult to publish the same data twice and redundant or duplicate publication will become easier to detect.

10. The linking of papers to full references will make it easier to detect the many cases where the "supporting references" do not actually support what is published.

The experience of publishers is that most journal editors do not accept these arguments.

The lack of support for open peer review is interesting. If we are concerned with the definitive version as the authentic version, how we reach the definitive version is important. There was a period in the middle of the last decade, when activists from the physics community promoted a consensus as to the worth of a communication that naturally evolved from exposure on the Internet. Clearly this could only work where the user and the reader community was more or less identical, but in practice, even in physics, all of the main journals use traditional approaches. There is a flavour of this discussion to be found in Paul **Ginsparg**'s (**1**) contribution to a 1996 conference but this grumbles about traditional peer review rather than denounces it. The following sentence in the abstract gives something of the flavour:

> "Is there an obvious alternative to the false dichotomy of 'classical peer review' versus no quality control at all?"

By 2001 **Ginsparg** (**2**) seems to have accepted classical peer review for physics for the same reasons (identity of author and reader communities) as he had once favoured open peer review. Now he continues to argue that in other disciplines there is a real need to assess systematically the whole system to determine its utility, and he quotes from a study in medicine (**Godlee**) that:

> "The process (has) so many flaws that it is only the lack of an obvious alternative that keeps the process going."

No satisfactory alternatives have in fact emerged. It is also true that journal editors to a man or woman believe in the system, whatever surveys may say. In my view, it is not even clear that the *modus operandi* is becoming more relaxed — for an opposite viewpoint see **Meadows** (pages 203–4). The question of who controls the peer-review and the certification puts traditional publishing models under threat.

One final point needs to be made, and that is to draw attention to the way in which some of the most visible alternative publishers emphasise their observance of peer review conventions. For more information on this point see subsection 5.6.

5.4.3 The uncertainties of certification

The point being made in this section is that scholars do not decide whether or not an article is more or less authoritative based only on what they know of the processes of certification only, but because of other more complex reasons.

In the previous subsections, it has become clear that the distinction between peer-reviewed articles and other forms of content is not a clear one, and that the importance of formal certification can be over-emphasised. It is clear that, in

relation to conference proceedings, individual scholars weight different proceedings differently, depending on their provenance and context or what they know of them. The same is true of refereed journals themselves. In medicine for example, there is the New England Journal of Medicine and its peers, and at the other end of the spectrum the numerous journals published by medical communications companies. All these journals claim to be peer-reviewed and to a place in the record of science. Some librarians like to distinguish between the important and worthwhile journals and the rest, usually published by commercial companies, which should be discontinued. A senior American librarian, well known for his emphasis on the importance of selection, told the author that very specialist journals should go to the wall. He saw no obligation to buy them for research groups in his university. These journals are characteristically published by commercial publishers and are in areas where there is no learned society or (more likely) where the learned society has yet to take ownership of an emerging field. All journals, however, whether they are poor journals with low standards or just highly specialized journals, publish some good papers judged by any criterion and of course many highly specialized papers may be needed by posterity. It is how science works. The reason why poor journals publish good papers is because of the way the scholarly community works. A scholar who starts a journal brings in his friends to populate the editorial board and leans on them and his dependants to contribute to his enterprise. The journal may not flourish and may die, but the papers remain and are needed in the future.

Each scholar has his or her own way of deciding what journals are likely to include appropriate information for his or her own research. Each community has its own hierarchy of importance, which does not necessarily follow the impact factors from The Institute of Scientific Information (ISI). The process is a complex one and has been ably treated in an infuriating monograph full of serious insights (**Guedon**, see in particular chapters five and six). In most disciplines, scholars know most of the journals where work of importance to them is likely to be published, but they do have problems with journals they do not know. In the digital environment, their problems are greater because searches bring them so many more possibilities to assess. The way they handle this situation has little to do with formal mechanisms of certification.

Guedon also treats the problems of those assessing research, which is not in their own specialist field. The Research Assessment Exercise in the United Kingdom has raised these questions as in need of serious consideration (because it involves future funding) and the activities of the panels for different subject groupings has been much debated. Are they rating articles that are submitted for consideration at a higher level (automatically) if they come from journals with a high impact factor, and how do journals with an impact factor (but a low one) or journals that are peer-reviewed but have no impact factor, or other types of publication not certified become rated? In many cases they claim that they will actually read all the publications submitted for consideration and assessment. This is just an example of the problems facing all panels for appointments, promotion and tenure. Guedon's solutions are discussed in 5.6.

Within the broad church of 'certification' some journals are prestigious to the researcher (author or reader) and some are less so, or not at all. The branding has always been seen to reside in the journal itself, its title, its history, its current editor and editorial board. Lynch has suggested that it is important that the

context of the journal, not just the individual paper, should be carried over into the digital environment so that posterity can see who the editor and editorial board were, and who was the source of certification.

There is another process at work. In the digital environment more papers are seen and downloaded outside their context than was the case in print, although the novelty of this divorce can be over-emphasised. The same complaint was directed at the reprint culture. We will look again at this process, the divorce between the individual article and the journal issue, and its significance for our study, in section 5.6.2.

5.5 VERSIONS

This subsection is centrally concerned with two separate matters both of which relate to versions and version control, and which both have relevance to the underlying question (already expressed) — is one communication by a scholar more authentic than another, and why? The first part is concerned with how to handle the continuum of versions of a message in and after the traditional publication is made available and fixed. The second part looks at how publishers handle the different versions of the same article that they are responsible for.

5.5.1 The continuum of communication

The Proposal begins:

> "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".

As early as 1995 one learned society, the Association of Computing Machinery (ACM) made a serious and sustained attempt to look the new world in the face. The Plan is still worth reading in its original form (**Denning**). The authors wrote:

> "Publishing has reached an historic divide. Ubiquitous networks, storage servers, printers, and document and graphics software are transforming the world from one in which only a few publishing houses print and disseminate works, to one in which any individual can print or offer for dissemination any work at low cost and in short order. This poses major challenges for publishers of scientific works and for the standard practices of scientific peer review".

Looking back, eight years later, the judgement of this author is that some of what was then prophesied has not come to pass, and that the traditional journal has proved a tough survivor for reasons implicit and explicit in much of this section. Two concepts in particular, exciting at the time, have not moved from theory into practice (my lettering):

a) "Journals will become streams flowing into the society's database and will retain their identities as database categories."

b) “Publishers will distribute notices of availability rather than journals or documents; readers will locate and obtain copies on demand using new software tools.”

Nevertheless, the Plan was important particularly because it emphasized that informal and formal content has become accessible in the same way and through the same channels, and that there are a multiplicity of versions available (at least currently) through any search engine. These versions represent the continuum indicated in the heading.

There has been some debate on how to handle corrections to the Definitive Publication in the digital environment. It is obviously easy to alter the online version, while the usual issuing of corrigenda in subsequent issues is appropriate for the print version. This is what some publishers do. Mostly however, the problems caused by any interference with claimed stability of content has made this approach unacceptable to most organizations or companies. Bernard Schutz, the editor of an e-only journal, set out one approach (**Schutz**). This is quoted at length because it demonstrates the problems of when there is no convention:

> “We want the information on the website to be as accurate as our authors can make it. Therefore, when an author corrects even a small error, such as a spelling mistake or a mistaken reference, we correct the current version but we note the change and make the original text accessible in the history list. We feel that the ease with which information can be changed on the web has its dangers, and should not be allowed to become a mechanism for sanitizing controversial or mistaken statements. If something has appeared on the website, then someone may have referred to it in a publication elsewhere, and we feel we must make it possible for the text that was referred to, to be reconstructed. In line with this, we ask readers who want to refer to Living Reviews articles to give the date on which they last read the article. This is to allow easier reconstruction once the review has changed. However, this style of reference is unusual to readers, and many do not use it”.

It is the impression of this author that most publishers do not alter the text, but instead provide a link to the article. We shall return to this question when we look at ‘trust’ in 9.4.

The existence of multiple versions has been suggested as one objection to the definition of First Publication in the Proposal. Why choose this version when there are other versions publicly available? In the meeting, from which the Proposal emerged (**AAAS**), one of the contributors drew attention to at least seven versions in what she called the evolutionary track of the e-article (**Fleming**). The fact that more than two versions of any article can be reached is attested by anyone who has sought some piece of writing through the use of a search engine.

This is not a valid objection — what First Publication represents is what elsewhere we could call preprints or (preferably) e-prints, or at least e-prints in some e-print archives. This qualification has to be made because the rationale and standards vary from archive to archive, although a more consistent approach is under development. E-prints do have a real culture, as we have already seen from

comments of scholars in different disciplines. We will look at some related questions in the context of a discussion of institutional repositories in section 9.

In subsection 5.6 we will discuss the position of e-prints further but for the present it is worth recording some more of the Recommendations mentioned above (5.2.1):

> 4. When citing preprints, authors should be encouraged to identify the version referred to and should provide a reference to any subsequent published version.
> 5. When technically feasible, publicly available and particularly peer-reviewed versions of articles should be authenticated to guarantee that they are the correct version.
> 6. Rights holders and publishers should facilitate linking for all references. It is desirable that systems for reference linking be bi-directional, interoperable, and open to all authors and publishers.

Recommendation 5 has already been quoted in the section on peer review. Recommendation 4 demonstrates by implication the serious version problems in what is essentially at present an unregulated area of activity. Is that changing? This is one of the questions we shall examine in 5.6. The problem of versions and corrections in the regulated environment of published articles is considered below. The regulation referred to is of course custom or (more properly) customs.

5.5.2 p-versions and e-versions

This subsection is concerned not with various versions of the same message, of which one is the Definitive Publications, but with versions of the Definitive Publication. It will be seen that it is possible to have two definitive versions of the Definitive Publication, but the matter is more complicated than that. Much of this subsection is concerned with responses to some questions, but three general points, which do not directly arise from the answers, can usefully be covered here. The assumption implicit here is that hybrid journals (electronic versions and print versions) are the norm in scholarly communication at the moment – as they are – but there will be further consideration of e-only journals in 5.6.3.

In the first place, the certified entity, whether in a journal or in a book, is not what the author has written but what the publisher has made available. The actual publication has not only been through a refereeing process but through copyediting. The author may well not have seen the end result. Copyediting can be very intrusive, yet many contracts insist on the publisher having the final word on the exact form of what is released. The differences are usually minimal, but not always, and there are plenty of instances of last minute serious disagreements between author and publisher over the introduction of distortions of meaning.

In the second place, publishers are under pressure to bring the electronic version of an accepted article to the readership as quickly as possible. After all one of the big advantages of the digital revolution is the possibility of speeding up communication. It is one of the advantages that BE Press (see below) and similar e-journal publishers make much of. Publishers of hybrid (print and electronic) journals began by releasing the electronic file as soon as the typeset version went

to press – thus gaining a few weeks or months. They are now experimenting with what are in some cases called Express versions, where the accepted article is released online before copyediting. Copyediting in good journals can be useful in improving the clarity of the final published version. There is bound to be some difference, including subtle differences in meaning, between the two versions. Which is cited? Which is definitive? Undoubtedly, the first is cited as well as the second. Should both versions be preserved for posterity?

Finally, there is an interesting version problem revealed in a talk by the current editor-in-chief of The Astrophysical Journal, the pioneer of the normative electronic version. He reveals that:

> "Although the full electronic articles, including supplemental materials, are now the formal archival versions, we insist that the paper versions of papers be scientifically self-contained (**Kennicutt**)."

In my opinion, this statement seems to imply two versions. If the message of the electronic version depends on e-only material, the print version will have to be rewritten to account for the fact that this material cannot be accessed in print. Differences in interpretation are bound to happen in some cases, however careful the contributor is.

Most journal publishers make their publications available in at least one online form, often part of a larger platform. Reference publishers are also online and some of the questions raised over journal publishing are relevant to them. Book publishers are in general lagging behind, but the way they handle their electronic interface to the world is even more interesting in the context of authenticity for the following reason. It has become clear that the journal article (once seen by some as a dying means of communication) is here for the medium-term future and it is as a unit of granularity that it is being made available. Books are made up of chapters and subsections of chapters, and it is clear that the possibilities of breaking up what was usually conceived of as a unit by the author of the work in question are interesting publishers. Such an enterprise goes back a long way, to the Primis project of McGraw-Hill for example. There are real challenges for authenticity here but we can sidestep these as they belong to the realm of education rather than the transmission of scholarship. For some pointers to what this means in the relevant world of scholarly monographs see **Watkinson (1).**

As far as journals are concerned, e-versions of journals, from the point of their conception, were intended to diverge from p-journals at least in the sense that additional matter could and (it was expected) would be attached to the PDF files. What were the author's own expectations in the middle of the last decade are mentioned in 2.1 and there is also the testimony of the SuperJournal Project (**Pullinger**). However, until recently and with some serious exceptions, the e-versions were much the same as the p-versions, except that an e-version based on an SGML derivative would have different presentational values. The most serious exception was the Journal of Astrophysics under the guidance of Peter **Boyce**, whose relevant archive can be found at www.aas.org/~pboyce. In this collection, the most significant item is the Plan of 1992, which already envisaged additional data (**Boyce**). In the last year, the situation has noticeably changed in that many authors are now ready to and actively offer material in e-form only.

This does not mean that these authors yearn to publish in e-only journals, rather they are now ready to take advantage of the opportunities of the medium. When the author asked publishers a question about versions some years ago, the general response was one of bemusement. Those publishers, who provided an e-version of their print journals, were asked to explain which version was the normative one. The few, who had thought about the question, all plumped for print (**Watkinson 5**)**.** However, as long ago as 2001, with a smaller sample, almost all assumed that the e-version was definitive (the word used this time) or 'canonical'. What this means in practice will be explored below, though actually as far as the publishers were concerned the implications were far from having been thought through – as we will see.

The questions were as follows and mostly but not entirely these questions followed from the answers to the questions that are recorded in subsection 5.3.4:

iii. If you publish journals in more than one format, which format do you regard as being the definitive version?
iv. If you publish two versions which differ, would you see both versions as worth archiving?

This information is not usually available on the web-sites of the publishers themselves. For example HighWire Press, which occupies a special position in digital transition reveals nothing of its web-site (http://highwire.stanford.edu/). Information about the twice-yearly publisher meetings, at which policies are established, is for members only. However, I was reliably informed two years ago that:

> "What publishers receiving services from HighWire said or agreed to was that the Internet editions of their journals were the editions of record. This occurred at the October 2000 HighWire Publishers Conference held in DC at AAAS headquarters".

It is significant that, even now, although the decision stands it is not at all clear how seriously it is treated.

Some actual responses to the questions are worth recording. A senior chemistry publisher writes:

> "I am not sure that there is a consensus yet in this question. I regard the electronic version as the canonical one and I believe that view will prevail eventually."

This is a fairly typical answer. The rationale is that the e-version has material not available in the p-version, and is therefore the normative version.

There is however more to be said about this question. There is not a single e-version. Two other chemists are worth quoting. Their answers to the first question (iii) are as follows:

- The electronic version (the HTML version rather than PDF, which is Identical to the print version). Increasingly the HTML version contains or is linked to information that is not available in the print or PDF versions. Where

this is supplementary information this is all sent to referees along with the article. We are beginning to distinguish two types of supplementary information though: essential (and therefore needing to be archived) and non-essential. This will cause us some problems as we need essential supplementary information to be captured in XML if it is to be archived and at present we do not have this (it is largely Word and CIF files). Reference linking will presumably push authors more towards accepting the HTML version as the definitive, as it will be much more functional.

- Our thinking follows the OAIS model. The PDF, HTML and print (made from high resolution Postscript) are just distribution formats. The definitive version, subject to preservation, is the SGML version used to produce the distribution formats. There is no commitment to preserve the HTML or PDF versions. New distribution formats, as they become popular, would be generated from the SGML. The SGML is not made available to subscribers. It uses our own in-house DTD. The SGML contains more information than is available in any one of the distribution formats and it probably contains more information than the sum total of the information in the current distribution formats. An essential item for preservation is the DTD itself without which the SGML can not be interpreted.

The answer of the second respondent also includes his answer to the second question (iv) because it was a natural complement. The significance for later consideration of archiving is obvious. For the moment, the point is made in this context. It is not just a choice between p-version and e-version.

The first respondent also touches on another complication — supplementary material. Quite a number of publishers offer to host additional material, which could be data, additional colour, video or audio clips. Are they part of the definitive article? There are no rules here and no consensus; and policies are stated, if in fact actually worked out. If they are not part of the definitive article identified by a DOI, how are they to be retrieved? It might be added at this point that, although the rules of the International Standard Serial Number (ISSN) Centre makes it clear that each version in each medium should have a separate ISSN (http://www.issn.org:8080/English/pub/faqs/issn), by no means do all publishers follow this advice.

This is not the place to cover in any detail the answers provided to question two (iv), but it is worth recording here that many of the respondents felt that both p-versions and e-versions should be archived. None of them, however, apart from the answer quoted above, thought through the implications.

5.6 ALTERNATIVE APPROACHES TO PUBLISHING

There is a large amount of literature on alternative publishing. Almost all of it comes not from a mainstream scholarly environment nor from publishers, but from library and library-associated sources. This is because alternative publishing is perceived as one way of overcoming the serial crisis and declaring independence from commercial publishers in particular, but really from all 'traditional' publishers. There is a lot of cross-referencing and overlapping membership in library and library-associated organizations concerned with the web of initiatives. There is a strong agenda, the transformation of scholarly communication, but it cannot be described as a conspiracy because it is not

secret. There is no publishing equivalent of this movement, mainly because of the fears of anti-trust in the publishing community, and because each company or organization is in a competitive position.

The tendency to self-regarding rhetoric, e.g. the 'subversive proposal', should not disguise the seriousness of some of the arguments. A good way in is through the web-site of Stefan Harnad at http://cogprints.ecs.soton.ac.uk/~harnad/. The following sources for alternative publishing are useful:

- The Free Online Scholarship Newsletter and various archives and resources associated with this enterprise - http://www.earlham.edu/~peters/fos/index.htm
- The Scholarly Publishing and Academic Resources Coalition (SPARC) at http://www.arl.org/sparc/home/index.asp?page=0 and in particular the list of provided in stage 2 Exploring Alternative Options of Declaring Independence at http://www.arl.org/sparc/DI/stage2.html.
- PSP Bulletin Summer and Fall 2001 available through http://www.pspcentral.org./. This is the viewpoint of "traditional" publishing.

The rest of this section first works out what 'alternative' publishing is offering and then examines what alternative offerings might mean for the concerns of this study.

Does 'freeing the scholarly literature' mean loss of standards and new threats to authenticity?

It should be noted here that alternative publishing does not mean self-publishing by authors. In the e-environment this is not more common than it is in the p-environment as far as scholarly communication is concerned.

5.6.1 Alternative to what?

Alternative publishing is alternative to traditional publishing. The web of initiatives mentioned above tends to present itself as an alliance between librarians and scholars, but in practice this does not mean an alliance between librarians and organizations representing scholars. This has to be qualified slightly. The publishing arms of learned bodies tend to be at one with and work alongside traditional commercial publishers, although the policy arms (where such exist) might not. I have written on this topic in Learned Publishing (**Watkinson 6**) for October 2001 and this 'letter' plus related items in the preceding and succeeding issues will indicate how far my contention is a reasonable one.

Traditional publishing is characterised as based on print. As Kate Wittenberg, herself a major innovator in the e-book field and later director of the Electronic Publishing Initiative at Columbia University Press, writes:

> "One of the big advantages that these new organizations have over traditional publishers is an absence of pre-conceived notions about what the market wants. Rather than attempting to recreate traditional print publications in digital form, many transformational publishers are instead focussed on disseminating information and services that respond to users'

needs in whatever forms seem appropriate to the content" (**Wittenberg** p.3).

Whether the new players bring anything new to the table in the area of scholarly communication will be examined in the next subsection

5.6.2 How alternative is alternative?

There are a range of new models that vary in their probable sustainability, and in the way in which they take responsibility for certification and for providing the context for certification convincing to the scholarly audience.

There is, for example, a huge literature on Open Access models. The fact of Open Access does not concern us here. The jury is still out on whether it is sustainable as a business model, but it is now as a model having serious impact.

One new venture, which does have a business model, is BioMed Central. Our concern here is not with the model (although it might have an impact mentioned below) but what the company says about certification. In spite of it being a commercial enterprise, SPARC and some similar organizations embrace it.

As we have seen, BioMed Central is strongly supportive of traditional peer review. Both Berkeley Electronic (http://www.bepress.com) and BioMed Central (http://www.biomedcentral.com) are insistent that not only are all the articles they publish 'properly' peer-reviewed. The only difference in what they do compared with what 'traditional publishers do is that they, the new boys, are more efficient and of course quicker — for a particular convincing list of arguments see http://www.bepress.com/faq.html. Any e-only journal can of course get accepted articles out into the world quicker as can those publishers who offer some sort of 'express' service online — although there are version problems with this sort of approach as we will see later.

The emphasis on peer review means that the owners think it counts for something. If there was one single indication of the importance to scientists in biomedicine of the importance of certification, this would be it. This is in spite of of the senior managers of BMC being the co-editor of perhaps the most devastating critique of the peer review system – quoted above in 5.4.2. Guedon writes of this project (**Guedon** page 35):

> "It, like SPARC, is apparently moving in the direction of creating new journals; however, these 'journals' really act as specialty or disciplinary boxes, while the branding through peer review is really attached to the whole BioMed operation".

Berkeley Electronic Press makes an even bigger point of a centralised reviewing procedure and the fact that the author may be offered publication as one of a family of journals depending, it would seem, on the quality.

> The Berkeley Electronic Press's quality-rating system (patent pending) allow authors to submit simultaneously to several journals at once, giving authors a better opportunity for publication without having to resubmit to another journal. (http://www.bepress.com/advantages.html)

In a sense, BioMed Central is resurrecting the principles set out by **Denning** and Rous back in 1995 (see 5.5.1). The playing down of the brand of the journal goes against much conventional wisdom.

The really ambitious move by BioMed Central is reflected in its business model. The costs are partly supported by a sophisticated range of page charges. This model has not worked in the past, but this is not the place to go into the arguments for and against the strategy. Open access could, however, result in a significant influence on the impact factor for the BMC journals, when the time comes for the first occurrence in the Science Citation Index. It would be interesting to learn if any of the older journals working with this model (although without success in getting a critical mass of authors) have achieved impact factors that build on the downloads they have experienced. As far as this author knows, none of them have yet to achieve an impact factor. This was true in 2001 and is still true but BMC is confident about its own journals. It could be that the business model does not work, but the journals do. By that I mean that they get plenty of good authors to write good articles in them. If that happens, it will have transformed the current lack of success of e-only serials (see below).

Declaring Independence, a SPARC publication (see above), recommends the publications of university presses and learned society as a whole. It would obviously be invidious to pick out some and reject others. However, it is difficult to see these organizations as a body, bringing any 'alternative' practices to bear in the way they serve scholarly communication. At least they do not do so at the moment. There are 'however' new library-led so-called electronic university presses. Representative of such presses is a European venture – Signal Hill (http://www.signal-hill.org). What follows was written in 2001. The project is now mainly subsumed into the bigger FIGARO project – http:// www.figaro-europe.net/news.html. It is our understanding that the model is much the same.

The mission is well worth reading:

> Signal Hill is a European partnership for academic publishing. The aim of the partnership is to create a community of practice for organizations engaged in electronic academic publishing to enable them to combine forces and share their experiences.
>
> Initiatives have been introduced in several countries in Europe to support academic publishing without involving traditional commercial publishers. Information technology is being used to create an infrastructure to facilitate and promote academic publishing by scientists and scientific communities, with an emphasis on communication. New business models are still being developed and elaborated.

The fact that the business model follows the technology is characteristic. The technology is there but the money to pay for it has not been found. There are other points that could be made about the type of approach, but which are not appropriate in this study (**Watkinson 7**). What is appropriate is the question; do these new organizations take seriously their mission to certify? The following points indicate where a possible failure to take on these responsibilities impacts on this study. Are the 'alternative publishers' of this type offering the type of

certification that scholars recognize as serious, as indicating that the publication certified is likely to be authoritative?

The sort of thinking demonstrated in the Signal Hall enterprise and some similar enterprises does seem to fail on the three main counts, when judged in this way.

In the first place, the approach tends to be parochial. During much of the last decade the mantra that the certification function should be independent from the dissemination function has been repeated. The role of certification has been assigned to universities in many of the scenarios. However, on the whole there has been no enthusiasm for this approach from scholars. A leading characteristic of most of scholarly communication in science, though less so in other sectors, is that it is international and the assumption is that judgement of worth should be international. The internationalism of scholarship — scholarship judged by international standards — is certainly regarded as a good thing (if not essential) in some disciplines (see above). What appears to be on offer, for example in Utrecht, may be a form of the sort of research publication series. Such an approach may of course be acceptable in some disciplines, but not in most. An example of a similar approach is described as Guild Publishing (**Kling**).

Secondly, there is a related argument worth considering. Generally, these presses also tend to take from and publish for faculty anything that is presented to them. No commercial press, particularly a book publisher, could work in this way. The interests of the marketplace have to be discerned (by market research) and heeded. Selection at this level is an important part of the publisher role (**Watkinson 7**).

Finally, such projects tend to stop when the funding stops and no 'commercial' business model has been found. What happens to papers published in the journals that finish with the project? The same sort of objections could be (and was) made about some of the experimental e-journals devised and financed by the JISC e-Lib project for a while.

It could be argued that straw men are being set up here, but the author, on the basis of personal communications and conversations, would argue that there are dangers in the sort of developments described — due essentially to ignorance. The position of **Roberts** (a sympathiser) is positive as well as realistic:

> "Money is, for most universities, always difficult to come by and there will be invariably a shortfall between what academics would like to do and what is practical or financially feasible. Setting up and maintaining *rigorous, internationally refereed electronic journals* (my italics) may, however, be a domain of academic activity worthy of increasing institutional recognition in the future. Declaring independence means finding the money, learning how to run the country with it, and investing now for future generations".

The initiatives described tend to present themselves as facilitators rather than publishers, but some take more responsibility for central publishing functions than others do. We will return to this contention in the next section.

5.6.3 E-only journals

We have looked at problems of authenticity insofar as they relate to the electronic version or versions of hybrid journals. What is the situation where we are dealing with a journal only published in an electronic form? In the previous subsection the enterprises we discussed were working with e-only journals, but our concentration there was on the alternative models. We shall return to e-only journals in section 6 when we look at the particular technical questions.

As we have seen, most alternative publishing is either e-only from the start or struggling to throw off the shackles of traditional print. On the whole there is an enthusiasm for e-only publishing that traditional publishers do not currently share. **Odlyzko**, whose views are always worth considering, writes:

> "Will the free electronic journals dominate? Most publishers claim that they will not survive ... and will be replaced by electronic subscription journals. Even some editors of the free journals agree with that assessment. My opinion is that it is too early to tell whether subscriptions will be required. It is likely that we shall have a mix of free and subscription journals, and that for an extended period neither will dominate".

This author does not share this optimistic view of the large and increasing number of e-journals, for reasons set out in this section. Some titles are exceptions. One has to admit that a glance at the references at the end of this study will reveal a significant number of them come from a small number of free e-only journals. Perhaps information science is different.

When I did a survey of UK 'traditional' publishers (**Watkinson 5**) my conclusion was that there was a renewed interest in starting such journals, but, the work for this study indicated that among the really big publishers there was no longer a wish to experiment with them. Presumably, lack of authors as well as lack of subscribers had killed any interest that there had been. This is in spite of the fact that all commentators agree that any cost savings or price reductions, depending on how you look at it, cannot be achieved with the current hybrid situations, where costs are in fact heavier than for print only.

In this subsection, we bring together some interesting comments about the publishing of e-only journals drawn from a small sample, but leave specific technical questions to later sections. For the moment, one generalization could be made under this head. Where there are experiments relating to the inclusion of different types of non-print content in a scholarly journal, it is likely to be in essential experimental and often avowedly alternative journals that are conducting such experiments/ making available these potentialities. There are some exceptions of course: a good example, already mentioned, is The Journal of Astrophysics.

The remaining part of this subsection is concerned with contextual points rather than with the way the journal is run.

For **Guedon** (page 34) print necessitates selection. Other versions are not saved for posterity. Where there is a print version it is the version of record, whatever

other versions may be circulating in an informal version. There is one print version. When there is no print version, the e-version of record, the definitive e-version is just one of many. **Silverman** (page 61) expanded by Gregory approaches the same topic from a different angle:

> "Electronic journals may cause more problems even with the notion of authorship. If there is no 'final' paper, but an evolving one, if there are uncertainties regarding whether the original authors should accept the challenges or recommendations of their peers, or even whether they have a choice, then the meaning of authorship becomes more complicated. An author may not survive by doing, or knowing how to do, research only in a certain way when it is likely that one will be confronted in a very public way with suggested alternatives for performing the inquiry, or elements of it. The publicness of 'scholarship in the making' will require that the process of crafting skills and understanding be more finely understood and more easily articulated than at present".

Guedon takes this argument further as we shall see in section 5.6.5. For the moment, and for the purposes of this study, this fact does place a special responsibility of identification, if nothing else, and making it possible to archive and preserve in addition. The traditional view of the scholarly record as "a series of discrete, permanently fixed contributions of readily attributable scholarship" (**Lynch 1**) is no longer tenable in the digital environment or at the least has to be rethought.

There is some doubt that the responsibility is being recognised. So many of the thousands of e-only journals bear an experimental appearance and function. The bad news about the barriers being lower is that much more material gets into the public arena that can be passed off as part of the scholarly record.

Peek takes a strong position, which many of her colleagues should listen to:

> "... [electronic publication] requires the same commitment to the object just as if it appeared in print. The technologies only change the vehicle. We will still look for the same values in cyberspace that [we] do now. We will look for authority and authentication. We will want to know when a work is complete, not a work-in-progress. We still need to know when the author feels ready to let something not merely to go public, but to belong to the public within the bounds of copyright".

Roberts, already quoted above, adopts a similar viewpoint. For him:

> "Refereeing may, at times, be a nasty, interest-serving exercise, but the benefits of peer review still outweigh a situation where 'anything goes'".

5.6.4 E-prints

In the previous subsection, we have seen that the foundation of e-only journals is an important activity of alternative publishing, of providing a means of scholarly communication not controlled by traditional publishers. We have also suggested that questions of authenticity, to some extent answered by the ways in which

traditional publishers go about their business, needs to be thought through again by new entrants.

In this subsection, this study will examine the relation of the e-print movement to the purpose of this study. This is not only because it is clear from what we have established earlier in this section that e-prints do have a place in scholarly communication, which is different from and more important that the similar position of offprints and reprints in the print environment. It is also because:

> "E-prints are seen as the catalyst for the freeing of the scholarly and scientific literature from the cost barriers imposed by journal publishers" (**Day**).

There is a large literature and a lot of activity in the area of e-prints even though they are important for scholars in a relatively small number of scholarly disciplines. There is no certainty that the e-print movement will sweep over the whole of scholarship. **Herbert Van de Sompel (1)** gives a list of the main e-print archives, now out of date. It gave the impression of being padded out then. A the time of writing there has been little progress in subject-based e-print archives as such rather than institutional repositories in general. There have been initiatives in chemistry (see Guedon chapter 11) and medicine (http://clinmed.netprints.org) that have not taken off. It could be a generational thing or it could reflect, as has been suggested, significant differences in the way scholarship is pursued.

Obviously there is real danger of building a structure and assumptions based on a structure, and applying them to the way all scholarship works when it does not and may not. I suspect that the e-print movement will be taken up by other disciplines, but how quickly? It is difficult to see why some academic communities should perceive it as appropriate for their purpose, while others do not. Other, more distinguished writers have, in the past, been over-optimistic about the speed of take-up and the implications. Sir John Maddox, for example, wrote in 1993 after a visit to the Frankfurt Book Fair:

> "There is every likelihood that it is a matter of a few years only before the practice long-established in particle physics has spread to other fields. (The frequency with which biologists now sport Internet addresses is already conspicuous). But the question will then arise whether the formal publication of research articles in what are at present called journals will continue to make sense ... In short, when access to the networks is universal, it seems unlikely that the concept of the journal as an authenticating agent will survive even in an electronic form" (page 689).

As we know, what was predicted here has not happened (yet?), but is the spread of e-prints out to other disciplines important?

Ginsparg (1) has always anticipated a wider movement and there is a lot of optimism among those involved about recent developments:

> Stevan Harnad just made it easy to join him in his quest to free the scholarly literature. Download Eprints software from http://www.eprints.org and build your own repositories quickly and

> cheaply (SPARC E-New 08-09/2001 at www.arl.org/sparc/core/index.asp?page=g20).

As we will see, anyone can join in with any sort of content as long as the metadata allows interoperability. One would assume that traditional publishers, such as ChemWeb (Elsevier), the British Medical Journal (BMJ) and the American Institute of Physics (AIP) will take part, and indeed IOPP (Institute of Physics Publishing) certainly has. The ChemWeb chemistry preprint server became compliant with the OAI protocol early on (http://preprint.chemweb.com).

According to **Guedon** (chapter 11) Ginsparg was the first to set up an e-print (then preprint) server in 1991, now archiv.org and transferred from Los Alamos to Cornell. Ginsparg has certainly been in the forefront of the movement. Much of this chapter is profoundly unhistorical, for example, in the suggestion that Biomed Central, BEPress (both already discussed) BioOne and HighWire Press were 'commercial responses', which were 'not slow in coming', but this should not obscure a serious insight:

> The advent of Ginsparg's pre-print server has demonstrated that the act of publishing could easily and safely be dissociated from evaluation and from long-term archiving (page 33).

This single sentence looks back to the status of e-prints and the First Publication of the Proposal and forward to our future consideration of Archiving and what should be archived. It leads on, in this subsection, to further questions about how e-prints are understood, some of which have already been raised earlier in this section in 5.3.4 (the long quotation from the physicist) and 5.5.1 (recommendation 4 about citation of "preprints").

Odlyzko (page 6) points out the simplicity of the Ginsparg approach. There is no filtering of submissions nor any editing, "the features that distinguish a journal from a preprint archive". Actually this author understands that there is some facility for removing some classes of content such as pornography, though the site itself tells us nothing of the way it is conducted. **Ginsparg (3)** glories in the 'rawness' of his database. Some sectors of the physics community, as we have seen, perceive the content on this server as 'publications' that need to be preserved for the good of science at least to avoid 'dead links'. These 'publications' include almost anything anyone wants to put on the site including articles subsequently rejected. A senior publisher consulted in 2000 finds this too much to bear:

> "A paper rejected for publication by an editor of a serious journal should not be called a publication even when it is put on a private or university server with a message that this text is officially offered for publication. This would create a stream of literature worse than the so-called grey literature".

Are such attempts to codify or control as proposed here or in the recommendation appropriate, necessary or impossible and, in any case, what do they tell us about the authenticity of these writings if the advancement of knowledge is to be a touchstone?

For the purposes of this study the author did look at the policies of other e-print initiatives mentioned as 'notable' by **van de Sompel (1)** already mentioned above. Are there different degrees of 'rawness'.

CogPrints does explain its filtering policy:

> (Question) Are there any limitations on what we can deposit in CogPrints? What is there to stop students clogging the archive up with all their essays? Is there any refereeing or quality assurance criterion before we post material to the archive?
>
> (Answer) Incoming submissions do not go straight into the archive; initially they are placed in a "buffer". The papers in this buffer are then reviewed by Stevan Harnad, and only those with content suitable for the archive are installed. This prevents the archive becoming clogged up with unsuitable material. (http://cogprints.soton.ac.uk/faq.html#quality)

The description of its aims seems to assume that communications posted will be submitted for refereeing or have been refereed.

The medical site (see above) has the terrifying warning (red and yellow lettering) posted before you can get in:

> Articles posted on this site have not been accepted for publication for a peer reviewed journal. They are presented here mainly for the benefit of fellow researchers. Casual readers should not act on their findings and journalists should be wary of reporting them (http://clinmed.netprints.org/home.dtl)

The reader has to click "I accept" before entrance into the Netprints™ site is allowed. The implicit assumptions here about the efficacy of peer review are seemingly at odds with the hostility to the process evidenced by many senior staff of the journal.

The other sites listed seem mainly to be concerned with different types of content such as dissertations and (in the case of computer science) research reports.

It is difficult to characterize an e-print as any defined sort of communication.

Back in 2000, the comment from a publisher quoted above was part of a highly confidential exercise conducted by the author in the context of meetings at that time going on between representatives of ICSU and of STM. The survey was of publishers, but it is very likely that the views of the senior publishers reflected the views of the journal editors they worked with. That is how publishing works. There was at the time a strong divergence between publishers serving different disciplines in their attitudes not just to e-prints as such but to any exposure on a web-site before traditional publication, whether in an 'archive' or through a personal or institutional URL. Many strongly opposed the consideration for publication in one of their journals of some content already exposed on a web-site and thus published. However, two years later there are few journals that do not accept for peer review material previously made available in this way.

There is a paradox here. At the same time as publishers and their journal editors are coming to accept that that e-prints are 'communications' and in no way publications, many in the same scholarly communities are moving in the opposite direction and treating e-prints as publications, well worth citing and important to preserve.

There is one approach to regulation and regularization that has really important consequences for future developments, both immediately as will be demonstrated in 5.6.5 but also (as we shall see) in relation to both archiving and metadata as dealt with in subsequent sections. This is the Open Archives Initiative. For this initiative (now usually OAI rather than the original OAi) see **Van de Sompel (1)** already quoted and www.openarchives.org.

The purpose of the whole initiative is only gradually becoming clear — probably to the originators as well as those observing from the outside. **Van de Sompel (1)** sets out the aims:

> To contribute in a concrete manner to the transformation of scholarly communication. The proposed vehicle for this transformation is the definition of technical and supporting organizational aspects of an open scholarly publication framework on which both free and commercial layers can be established. This framework (is) the Santa Fe Convention. This convention is a combination of organizational principles and technical specifications to facilitate a minimal but potentially highly functional level of interoperability among scholarly e-print archives.

The content coverage seems to assume a whole range of different types of communication, in line with the way in which the e-print movement appears to have developed (see above) but probably even wider — see **Van de Sompel (2)** on the annoyingly named Bison-Fut< Model. As the OAI site explains (www.openarchives.org/documents/FAQ.html):

> "The roots of the OAI lie in the E-Print community ... It soon became evident, however, that the concepts in the OAI interoperability framework ... had applications beyond the E-Print community. Therefore the OAI has adopted a mission statement with broader application: opening up access to a range of digital materials".

We will return to the question of metadata and interoperability in a future section.

Does the breadth of coverage matter? It might do if for purposes of the record of science a scholar (in an unfamiliar field) has to make determinations of the status of e-prints in different archives, interoperable in terms of metadata but not consistent in terms of aims, scope and filtering mechanisms. However, for the moment it is the implications that count and these are considered briefly in the next subsection.

5.6.5 Building on e-prints

Since the importance of the World Wide Web for scholarly as well as other forms of communication became apparent, commentators have anticipated a revolution in the nature of formal scholarly communication — not that they have always

understood the difference between formal and informal contributions. No-one has been sure what form it would take and, so far, it has not happened. As we have seen the e-print movement has only become important in a small group of specialisms and 'traditional' publishers have continued to be the gatekeepers of scholarship.

The various attempts by librarians with a mission to subvert, influence or correct have not had a major impact on the STM publication system. The current author discussed some of these challenges three years ago (**Watkinson 3**) and not a lot of happened since then — until now.

Herbert Van de Sompel does envisage a new structure being built on the building blocks of OAI. In my view, his concept is both visionary and manageable and both potentially inclusive as well as disruptive. At the time of writing, the author has not found a good reference for these views except a set of slides (**Van de Sompel 2**).

It is worth setting the argument in this respect succinctly: it is based on a report written by the author for another purpose following a meeting in November 2001.

- He argues, as others have, that the clear functions in the print environment, which operate the scholarly communication system in a linear way, are now disintermediated and blurred.
- The functions are registration (claiming a new finding), certification (certifying the claim), awareness (ensure information throughput) and archiving (preserving the heritage). Van de Sompel following Roosendaal (see reference) and Geurts adds the following – rewarding i.e. what the scholar gets out of his or her publication in terms of career advancement etc.
- It is legitimate to look for a new system because there is much wrong with the old one e.g. the slowness of the present system and its cost (the serial crisis). He sees the current peer review system as "suppressing new ideas".
- The inspiration for a new system is the established e-print initiatives [which van de Sompel was much involved in] but they only cover the registration and awareness functions and not the certification, archiving and rewarding functions.
- He argued that the Internet is a disruptive technology [from the work of Christensen, which is much used by people with these types of views]. Established companies have a real problem dealing with disruptive technologies, because new companies can come along and create cheaper and more efficient solutions.
- He sees the e-print approach as such a disruptive technology and they will form the starting point for the new value chain. Lots of different players can fulfil the registration function: it can be distributed. The important element is interoperability so that information can travel easily across the system.
- The physics model does not have to be the only one. Different communities can work in different ways.
- In some communities certification (as now) can be handled subsequently but in others some kinds of secondary databases can certify by selection from the e-prints.
- What is important is that the information that the certification has been done and the way in which it has been done can flow through the system and can flow back through the awareness function.

- The rewarding function currently done by citation can in the future results from other metrics such as metrics based on usage.
- There is a lot of talk about economic models and policy but it is the underlying technology issues that are most important.

This is the basic argument. Others are looking to extrapolate from it. For **Guedon** (page 34):

> "The evaluation process stands ready to be reinvented in a clear rational way by the relevant research communities themselves".

He calls to librarians to:

> "Develop strategies favoring the outcomes best corresponding to the deepest values of their profession, in particular the desire to keep the knowledge commons open" (page 40).

Specifically (page 42) he sees the Holy Grail as follows:

> "With a well designed principle of distributed intelligence, with the help of scientists self their work, with the help also of selections that do not rest on the prior reputation of the brand, but on the actual quality of each selected work, librarians hold the key to developing a total, global mapping of science".

Guedon is proposing a new publishing structure in which librarians work with scholars to do the evaluation, but there is no reason at all why publishers should not adopt the same strategy. They do currently work with scholars on evaluation, and the project is so organized that there is no reason why any interest or sector should not add the value (**Van de Sompel 1** envisages commercial exploitation).

Roosendaal (back in 1997) had already proposed a much more complex model, which relates to the wider concerns of his referenced article. He does envisage firstly, that there will be an "integration of formal and informal communication" and that:

> "A result of these developments will be that the now distinct roles of publishers and libraries will be merged to become notes in the overall management of scientific communication" (page 12).

Hitchcock and his colleagues, working outside publishing, have already tried an experiment with a new "service that dynamically interconnect(s) material in the archives" – a sort of review journal titled *Perspectives in Electronic Publishing.* The journal is to be found at http://aims.ecs.soton.ac.uk/pep.nsf where it is described as:

> a **journal-centred portal**, with enhancements for exploring selected **full-text papers** on a focussed topic - in this case, on electronic publishing. You will find papers linked from the *original sources*.

This is data mining. The enhancements are the framework, including the fact of selection, but will also include reviews (though it does not seem to do so yet).

Nothing will be lost at present if this experiment goes to the wall, but once the reviews start appearing we have the usual problems of evaluation and the obligations of archiving. Hitchcock has done some preliminary thinking about such questions (see page 3 of the original reference). This journal no longer exists; it was an experiment.

There are a lot of questions here in the models, which seem to be striving for development. Obviously, there is the question of how evaluation is handled in the new publications mining the open archives. There are clearly issues of authenticity. Where does certification come in? Selection is one thing, but selection only after revisions have been suggested and accepted is another. The way is opened, perhaps usefully, to lots of selected and revised versions, which may or may not cause problems of discerning the author's message. It does not need to cause problems if the identification of the versions is handled properly and if the author agrees to the different versions but it certainly complicates life.

7. Digital informational entities

The title of this (shorter) section indicates that we are trying to avoid an association with print. Print may or may not go away, but we are certainly now dealing with e-only definitive scholarly communication. As we have mentioned earlier, the difference between the situation now and the situation a few years ago is that many authors want to include e-content in their submissions.

Work on digital informational entities is scattered. Descriptions are usually highly technical, which is reasonable. The conventions which govern print are understood by many, but not the conventions that need to be established in the digital environment, except in a piecemeal way and in the way connected with what can be done rather than what can be preserved for posterity. We will look at these questions, in the broadest sense, in section 9. This section looks back to section 5.5, where some of the issues are hinted at.

6.1 NEW ANIMALS IN THE ZOO

The heading is explained by a quotation from Joost Kircz (below). His contribution to thinking about this area has already been mentioned in 5.2. The quotation below is from an earlier version of a paper (**Kircz 3**) that was later published as **Kircz (4** and **5).** The latter version does not contain this conceit:

> "In my view we have to make a difference between documents that look, smell and sound like a paper document but are stored and transmitted by electronic means, and documents that are originally created for an electronic environment, and hence are the new animals in the zoo of scientific communications.
>
> The discussion on the value of electronic documents is often hampered by the fact that one starts from what one is accustomed to in the paper world and attempts to impose that on an electronic environment".

When this study comes on to discuss archiving and preservation issues in the context of questions of authenticity and integrity we will recognize this sentiment as relevant in that context too.

6.1.1 A critique of the proposal

This is the proposal explained in 5.2, the attempt to define both Definitive Publications and the proposed new category of First Publication. Kircz critiques the Proposal from the viewpoint set out in the quotation. He takes the main characteristics, proposed as necessary, one by one. For the purposes of this study, only some of the characteristics will be considered. In the rest of this subsection quotations are taken from **Kircz (4).**

i. **Fixation or permanence**

For Kircz working from within the digital environment, we have to interpret fixity:

> "As a demand for a well-defined descriptive standard about the content of the document – a standard that enables the storage and maintenance of the integrity of the information independent of the carrier of the information" (page 267).

This looks forward to section 9. An unaltered entity is not much use if the 'message' cannot be accessed

ii. **Persistence**

This is the question of retrieval. The scholar must be able to find the entity again in the same place. Kircz comments on the DOI and the OAI standards, which we shall return to again in section 7, but, writing of digital entities, he proposes:

> "The persistence aspect can be covered by introducing a complete list or map of contents as an integral part of every document. We have to maintain not only the bitstreams of every component of the document, but also the mutual relations between the various components. We also need a mechanism to check that all components are present" (pages 268-269).

This demand looks back to the philosophical distinctions of section 3 and forward again to later sections, but we will disclose more of his pointers towards a solution later in this section.

iii. **Version control**

For Kircz, version control in the print environment is relatively easy. The problem is that for many e-only entities links are integral to the message and the links may be to dynamic sources. We are entering into the realm of metadata, which we shall discuss later. For a different way of looking at the same question from within the same community see section 6.2.4.

iv. **Authenticity**

Kircz uses the term in a very restricted sense. He comes to a novel position when considering the question of protection from change. He distinguishes between re-use and multiple use. In the case of multiple use he envisages the future use of the article extending, as it were, outside the article. This concept is further examined in the next subsection.

He instances the swapping of datasets and methods and cites the work of Rzepa — well-known as a pioneer of e-only journals in structural chemistry — where many of the problems of e-only entities have been much discussed. **Rzepa** argues that "data must be regarded as a critically important part of the publication process, with documents and data being part of a seamless spectrum". The message is incomplete without the data. The actual instance is surely blue-sky material, but the concept is highly relevant.

Kircz summarizes:

> "The conclusion of the above discussion is that the scientific article will change its form considerably but that, in its new more composite form as an ensemble of various textual and non-textual components, it will retain the cultural and scientific demands with regard to editorial quality and integrity" (page 271).

Authenticity in the wider sense is still relevant even when the semantic web beckons. In the next subsection, we will look at the insights derived from real e-only journals that have achieved some sort of position in their disciplines, though on the fringes. In the final subsection we will look at the solutions proposed by Kircz.

6:2 THE PROBLEMS WITH REAL JOURNALS

Much, though by no means all, of the content of this subsection comes from the proceedings of a conference (CCSC) held in Amsterdam in June 2001 (**Kircz 1).** Not only is manageable information about the individual presentations available on the web-site, but the outcomes have been summarised subsequently. A footnote to the programme explained the purpose and scope:

> "The idea of the conference is to invite speakers from many different scientific domains representing initiatives, which clearly divert from the classical model".

6.2.1 Why do some electronic-only journals struggle?

The journals described at CCSC, despite being experimental and all (we think) dependent on grants, subsidies or time freely given, had clearly established some sort of position in their field. As we have remarked before, too many e-only journals are 'experimental' in that they are not intended for scholarly communication but rather to investigate models. This practice seems to me to be indefensible as long as all the contributors were consenting adults and did not mind that their messages were potentially lost forever. The journals mentioned in the next few paragraphs were not intended to fail.

We have already (in section 5) touched on the fact that the number of e-only journals available is not reflected in their importance in scholarly communication, but the quotation that heads this subsection is from an article by Vincent Kiernan that lists some serious journals that never got off the ground. He instances in particular the Chicago Journal of Theoretical Computer Science. The present author has worked on three e-only journals in this field, two of which have, as far as he knows vanished without trace. He, his colleagues and the editors of the journals worked hard to get authors, but they would not come. **Kiernan** quotes from the web-site of the Journal of Digital Information (jodi.ecs.soton.ac.uk/ but inaccessible to this author at the time of writing):

> "JoDI will not always be free. Development costs money, and it is our aim to provide you with the best journal in its field".

This journal does seem to be free — presumably subsidised by the British Computer Society. However, obviously journals that contain 'digital informational entities' (articles that have intrinsically necessary content not available in print)

will be more expensive to run and more difficult to sustain. Those that we mention later fall into the category, as far as we can judge, of true e-only journals where the e-content is not additional but intrinsic.

6.2.2 Print look-alike or true e-only journals?

It is not clear from the article cited whether the journals that are not working were print look-alike, had the sort of features we are discussing, or whether this might or did make any difference either way. Some quick sampling suggests that the great majority is in the former camp and present no special problems relating to authenticity. These are examples of commonly stated contention that barriers preventing journals start-ups being lowered. Questions of certification are considered in 5.6.3 but we have also discussed special versioning problems within the 'real' e-journal context in section 6.2.4.

Many of these journals are free, both to authors and to subscribers. The free journals are listed on the excellent web-site http://informationr.net/fr/freejnls.html#j. Admittedly this also covers newsletters and other types of serials, but it is made clear in the short blurb that accompanies the link what the status of the journal is — at least insofar that it claims peer review.

What seems clear is that the number of e-only journals is growing. Chemical Abstracts (CAS) monitor over 100 (http://www.cas.org/EO/ejourn2.html). There are over 50 serious mathematics journals in this category according to the American Mathematical Society (http://www.ams.org/mathweb/mi-journals2.html).

A significant number of the journals monitored by CAS are from BioMed Central. We have discussed their offering before. They seem to be essentially text journal allowing for additional/supplementary material. Their instructions suggest this:

> "Supplementary/additional files:
> These may consist of larger tables or other files, such as movies, PDF files, etc, that are not intended to appear within the body of the article".
> (http://www.biomedcentral.com/info/edgr-preacceptcheck.asp)

If that is the case, this new tranche of journals are not of interest in this context. The problems they present in questions of authenticity in the digital environment are relatively minor if this is the case. It is not always easy to discover the status of non-print content. The question of the status, whether the 'additional content is part of the 'message' of the article or additional, is a question that has been ducked by the publishers interrogated for the purpose of this study.

The following subsections mimic the scientific publication process following the summary of the CCSC conference already mentioned.

6.2.3 Authoring

The point is often made that publishers in the digital environment push some costs back on to the authors, who are asked to supply files using templates and may even be asked to key in alterations proposed by copy editing. When dealing

with digital content, the need for authors to provide what the CCSC summary (http://www.niwi.knaw.nl/ccsc/summary.htm) calls 'computer friendly content' becomes even more important.

There would seem to be particular problems where interactive elements are involved. In the middle of the last decade, there was much speculation about interactivity in learned journals; Routledge, now an imprint of Taylor & Francis, investigated such a publication from a theoretical rather than a practical standpoint for some time. By this I mean that the initiative came from theorists in information technology rather than any pressure from an author community. It was assumed by some that they would come to resemble that constant interaction characteristic of the Web. Questions of authenticity and the need to preserve a scholarly record were ignored.

A lot of thought seems to have gone into the Journal of Interactive Media in Education (JIME), which, by definition, deals with interactive content. There is a description of its policies both on the journal web-site (http://www-jime.open.ac.uk/) and in an article, the latest version of which is in Learned Publishing (**Shum**). The page references below are to the article. The abstract (page 273) explains that the journal is interactive in two ways, its peer-review process and some of its content:

> "This innovative review model and the resulting enriched digital documents illustrate some of the possibilities of promoting knowledge construction and preserving intellectual products in digital scholarly publications".

This is a fair claim, although it is likely that mostly other journals concerned with scholarly communication concerning education strategies where this particular approach is used and is useful. However, simulation — certainly tried in applied mathematical journals — is likely to come into the same category insofar as questions of authenticity are concerned.

What needs to be stressed here is that the journal does seem to be looking for interactive material that is an intrinsic part of the message rather than just additional illustrative content. The guide to authors suggests this:

> "If the description of new interactive media forms a substantive part of the submission, the article must be integrated with illustrative extracts of the media, which convey to readers its interactivity".

The range of possibilities presented to authors includes the following:

- a demonstration version or even the full system, which readers can download and run on their own machines;
- a website used in your study
- an interactive extract via the WWW (e.g. using Shockwave™, Java™ or VRML);
- screen recordings of the system in use, with audio commentary, e.g. as a QuickTime™ movie using a utility such as Snapz Pro (Mac OS), or Lotus ScreenCam™ (Windows).

Obviously, when we come to look at archiving and preservation, we will see that any dynamic content must cause difficulties, but what is proposed here seems to be mostly (as it were) stable dynamic content. The template is 'published' and remains stable, but others can interact with the content. The Definition of 'interactive' seems to confirm this interpretation:

> **Interactive** - refers both to interaction *through* the media with other people (e.g. teacher-student, student-student, researcher-teacher), and to interaction *with* the materials embedded in the media (e.g. control of a simulation or educational game).

Clearly such material presents different problems of authenticity from material that is constantly changed by new reader interaction.

Regarding the interactive review, there does seem to be a 'final' document (see explanation of Peer Review), even though interaction continues but it is very much a point in a process rather than what all the process aims to achieve. The nature of the journal is such that it would seem more appropriate for all the process to be archived as authentic in an ideal world.

It should be added that this particular journal is a real journal, contributing to scholarship with real contributions, though not many, even if it is part of an experiment and not in any sense self-sustaining.

The CCSC summary raised two other questions under the heading of 'authoring'. The first of these two (second of three) was:

> How do you coordinate, process and integrate the various multimedia elements that authors wish to add?

It is clear that in some fields, for example astronomy, metadata standards are agreed, but in others, for example chemistry, they are not. It was agreed that "authors need to learn to incorporate metadata in their texts during authoring" — presumably where there is a consensus about its form. It was also noted that, even in these special conditions, where one has to assume that only highly motivated authors do contribute, it was difficult to get them to follow rules.

Of interest here are some of the comments on non-text elements made by the editor of Internet Archaeology (http://intarch.ac.uk) in her introductory material

> "Although a scientific discipline, archaeology might be said to differ from other disciplines because each 'experiment' is unrepeatable" (www.niwi.knaw.nl/ccsc/talks/winterstalk.htm).

It could be argued that other scientific disciplines, for example the earth sciences, share this problem, but it is the implications that concern us here:

> "The journal is of course concerned about the integrity of non-text elements that are published (datasets, VRML sections) but there is only so much that can be checked, due to the unique and unrepeatable nature of archaeology". (www.niwi.knaw.nl/ccsc/talks/winters.html)

Flaws in methodology cannot be sorted out because the evidence is now gone. However:

> "Journal experience has shown that the technical process of making datasets available online can sometimes highlight additional inconsistencies at the editing/mark-up stage".

The interaction that are involved can have serious implications for a better understanding of the data — as of course it does in print. The non-print content, which is part of the e-version of a print journal or an e-only journal is not content just to be shoved into a space. Editorial processes lead to a more authoritative or authentic version.

Some of these questions will be discussed further in subsequent sections.

The final question raised under this heading by the CCSC summarizers was a complex one:

> "Is there a future for an entirely new type of article (composed of multimedia, hyperlinked information modules) or is a narrative still the best way to convey scientific information?"

This is a difficult question and it was not really answered. What did come out was the interesting comment that if readers want to print out what happens to the 'nifty multimedia agenda'. What happens indeed? Is seems that the pioneering Journal of Astrophysics gets over the problem by having two self-contained, and presumably authentic, versions (see 5.5.2).

6.2.4 Editing

This section of the summary mainly concentrated on two areas, refereeing and impact factors but the biggest debate was concerning peer review. The argument was mainly about the merits or otherwise of anonymous peer review (see 5.4.2 above)

It was recognized that for e-submission reviewers have to be trained as well as authors. This has been the experience of all those traditional publishers with hybrid policies, now faced with real e-content rather than hypothetical content. From the point of view of this study, the key thing is that it is expected that the e-content will be refereed.

The question of how you judge the importance of a paper is one, which belongs under another heading, except if you are deciding that there are journals that you cannot afford to archive because they are not worth it. In this case, a replacement of or an addition to impact factors based on author decisions by impact factors based on user decisions (downloads) is going to change part of the picture.

6.2.5 Publishing — versioning

In the summary, there are sections under publishing on versioning and archiving. Some of the comments on archiving are interesting if not important for our

purpose and will be dealt with in section 9 below. Under versioning, there is a question relating to updated articles and different versions side by side. The comments recorded are so interesting that they are recorded in full with a few alterations to make them clearer. They hark back to discussions of the definitive publication:

> "Living Reviews in Relativity (LRR) asks authors to update their reviews to maintain compliance with advances in science. The question was raised how to motivate authors to update their articles, given pressure in academia to publish new articles all the time. The editor of this journal argued that each new version could/should count as a new publication. To allow a reference to be made to an earlier version, several versions of an article are available simultaneously.
>
> With interactive publications, they can also be seeing information presented in a different way. In other words: two readers can be reading different versions of the same article, and there is not always a "definite" version.
>
> To a lesser extent, this problem also exists when a preprint version and a journal version of an article coexist. If the preprint version if free, there is a chance that the unrefereed version of an article prevails, thus leaving room for errors that the referee might have picked up".

Here are three paragraphs making different points. The web-site seems to show that LRR is a review of a sort not unfamiliar in biomedicine, for example the Current Opinions stable. We have not previously discussed special problems relating to review journals because we have concentrated on the transmission of primary research. The LRR web-site seems to indicate that each update is indeed a separate publication. Each update seems to be discrete, but purists might argue that the following taken from the web-site does show some blurring which could bring problems in citation and in archiving an authentic version:

> "*Living Reviews* is committed to provide most accurate and up-to-date information to its users. Errata or small, important additions will be published within the original article when requested by an author, without waiting for the next major article update. See section IV of Using an Article for details on how these changes are documented. For such small changes though, no new publication number is assigned to the article. Hence we strongly recommend to include (sic) the cited on date in your citation to specify the state of the review you are referring to".
> **(http://www.livingreviews.org/Info/AboutLR/AboutLR.html)**

The word 'important' is significant. It has always been a policy with textbooks, for example, to reprint with corrections between editions. If this is a more substantial alteration ('important'), where the idea of using the date of citation does help identify the original text (or are all versions really kept even if the alterations are small), there must be problems of authenticity here. There does seem to be a confusion that goes to the heart of the sort of problems dealt with in the earlier part of this study and which relates to integrity. The editor writes elsewhere:

> "When an author corrects even a small error, such as a spelling mistake or a mistaken reference, we correct the current version but we note the change and make the original text accessible in the history list. We feel that the ease with which information can be changed on the web has its dangers, and should not be allowed to become a mechanism for sanitizing controversial or mistaken statements".
> (www.niwi.knoaw.nl/ccsc/talks/schutztalk.htm page 2)

The last sentence of the second paragraph of the Summary does raise a very important point, which we shall return to in section 9.

Finally, the last sentence illustrates the sort of problem discussed in section 5. If preprints were really preprints, the version would be removed or altered once the submission was accepted by a journal, but they are not. Sometimes they are altered but not always. As we have seen from the Proposal there are reasons to aim for fixity and reasons not to do.

6:3 A MODULAR STRUCTURE FOR ELECTRONIC SCIENTIFIC ARTICLES

We have already examined in this section Kircz's critique of the Proposal, which is a critique (an examination) rather than a criticism. At the heart of this critique is his suggestion that attempts to define a publication are starting from the print environment and adjusting to the digital. He wants to start from within the environment, from where publications are in the environment and where they might be. This study is concerned not with the present transitional phase, just one phase of continued transition we should point out quickly, but with the way questions of authenticity have to be handled when we have moved on to exploiting the opportunities presented by the Internet in a more thorough-going way. It is worth repeating here that there are signs that the author community is beginning to actively desire the inclusion of e-content in the messages they wish to send to the community. The age of experiment and surmise may be past. Some of the considerations from the cutting edge have been presented in earlier parts of the section

6.3.1 A definition of the problem and the possibilities

The rest of this section is based heavily on the seminal work of Kircz and his colleagues. The locus is the paper in Learned Publishing this year (**Kircz 5**). It is a work of great richness and difficult to summarise. For the remainder of this subsection, references to Kircz will be to this paper. Most quoted will be:

> "Genuine electronic documents will be a composition of text with different non-textual elements ... By translating knowledge into binary code, we create a mono-medium that allows us to integrate all kinds of representation" (page 28).

The assertion here can be separated into two parts. There is the definition of what is 'genuine'. I want to hark back to my concerns from 1993, which prompted my espousal of PDF (see section 2). The difference now, as we have seen above, is that I am now willing to accept the word 'genuine' as used here, although I would prefer a rather less emotive term. From the point of view of this study, what is interesting in the digital environment, now that the tools are there,

are entities, which make use of what is possible. Kircz also uses the argument, also much used in the middle of the last decade that the image will gain much of the power lost to language. This loss of power can be exaggerated. One of those truisms that happen to be true about scholarly communication in many areas of science is that the image has always been predominant. Readers go to the image before the text. The point is that this is the image in print or in the print look-alike form of PDF.

The above paragraph is a gloss on-and-an expansion of Kircz's assertion about the "most notable feature" of electronic publishing. The second assertion is more radical and controversial:

> "The next most notable feature of electronic publishing is multiple use".

Kircz argues that in electronic publishing in its 'genuine' form, I will not point to relevant information elsewhere but instead will transport information located elsewhere into the work in question. It is a re-definition of the concepts of links, hardly touched on so far in this study but to come into play in section 9. Not everyone would agree that this is a valid way of looking at the issues in terms of current practice. However if we look back at the previous subsections on publishing practice at the cutting edge we shall note a number of resonances, for example in the description of JIME policies.

6.3.2 The answer is a modular approach

It does not matter if part of the analysis in the assertions, the second in particular, is incorrect or at least premature, because in the context of authenticity and archiving the answer is most interesting and potentially very productive.

Kircz cites earlier work by his own colleagues and cites particularly the article by Harmsze in 1999. **Harmsze** writes:

> "We develop a structure for modular articles, based on the idea that an electronic article can be made up of well-defined modules and links that, following the SGML-philosophy, can be identified with tags. We define a module as a uniquely characterised, self-contained representation of a conceptual information unit that is aimed at communicating that information".

For- and against- arguments can be made about this vision.

In the first place, the thinking seems grounded in the realities of scholarly communication. Lindquist makes a point about the situation as it is, in another context, which Kircz seems to assume rather than state:

> "Multimedia documents that are 'born digital naturally appear as parts that are linked together, since the parts are usually produced by different tools, or are collected from different sources or data-capture devices" (**Lindquist** page 5).

The concept is also developed to take into account questions of authenticity. It is envisaged that the individual modules and the assemblage of modules are separately refereed.

In the second place, as Kircz himself says (page 32):

> "Electronic media enhance the integration of textual and textual knowledge representation, 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".

Harmsze admits of the project that:

> "Rather than concentrating on the capability of present-day software, we choose an analytical approach".

This is blue skies stuff, not inappropriate in itself but perhaps difficult to build on when looking to archiving solutions. We will examine this doubt further in section 9.

I am also worried about whether this way of looking at the way science is communicated actually fits in with how science is done or should be done. The scientific article is a structured document, which works as a means of communication. Selection of evidence is careful and dedicated to the message. To introduce a module provides greater richness and makes it possible for the reader to make up his or her own mind, but it also dilutes the power of the argument.

Surprisingly, there is an interesting analogy, though the structure is different, with the Darnton pyramid, a concept being investigated by the American Historical Association with funding from the Mellon Foundation. Robert Darnton has projected an exploitation of the digital environment, which provides for the traditional specialist monograph but provides the documentary evidence in digitized form as the bottom part of the pyramid (**Watkinson 1**). There will be similar problems of archiving and at the same time advantages. Consideration of this project leads one to a concern as to whether there is a distinction between 'raw' data and 'worked-up' material, which happens to be in a non-text format. In terms of the determination of authenticity, are the criteria for judgement not different?

Kircz (page 32) pleads for experiments so that new standards and rules can be developed. It is not clear how such experiments can be conducted. Nevertheless, he is certainly making a reasonable point in pleading that the scientific community takes these ideas and the projected experimentation seriously.

E-journal publishers of both e-only journals and of e-versions have not scrupled to demand the use of their templates and their standards. We have seen some hints of this in the subsections above and in the general statements about the need to educate authors. There is a long tradition of educating authors in the print environment. It is not just a matter of publishers demanding certain

conventions for their own convenience. A request for consistence (of citation for example) comes from the communities themselves. Journal editors are often much more prescriptive than a publisher would wish to be.

In this section we are writing about the concept of modules and how it is introduced to the scholarly disciplines, but education could go further. If a consensus about authentic versions could be achieved, albeit discipline by discipline, there will need to be enforcement for it to work in practice so that the reader can be sure of what he or she is getting. The publisher, as gatekeeper, will need to do this enforcement to set acceptable standards.

The next section discusses standards from a technical viewpoint. It is not at all clear how far the subject of this study lends itself to the level of definition, which is necessary to implement standards in a valid way.

7 The identification of the authentic entity

The heading of this section was chosen deliberately both to narrow down the focus of the consideration of metadata and also as an assertion of its importance for identification. Metadata exists in part to help people find out what they want. Identification does not actually need metadata. The DOI, for example, can just point to an entity.

In what follows, we set out where we currently are in the argument of the study insofar as it relates to identification. There will be some repetition, or, more correctly, some restatement of conclusions set out at greater length in section 5. We will then discuss metadata, and the various schemes providing metadata frameworks or which depend on metadata, with a view isolating characteristics relevant to the concerns of scholarly communication rather than as objects of intrinsic interest. Finally, the last subsection will attempt to bring the needs and the possibilities together.

This section has caused the author more problems than any of the others in spite of the help he has had from a number of people, who are acknowledged in the preliminary matter of the study. He has in the end decided not to attempt to dive into the complexities of the subject, which are considerable, but to approach them from within the context of the rest of the study, using the vocabulary he has used throughout and avoiding a shift in usage. There will undoubtedly be problems of clarity as a consequence.

7.1 PROBLEMS OF DEFINITION

Those who work in this area aim for definitions that are concise and fool proof. When it comes to identifying and retrieving an entity, there should be no doubts about what is being looked for and no possibilities of a mistaken result of a search.

7.1.1 The blurring of boundaries

However, as we have seen, scholarly communication in the digital environment suffers from the blurring (as **Meadows** calls it) of boundaries between the formal and the informal aspects of communication — but this is nothing new. The big change is that informal communication is now public in a way it was not public before.

In some disciplines, more than one version of the message is considered worth archiving, which means a serious level of interest in more than one expression of an idea. The word 'version' is used rather than 'manifestation' because in our understanding version refers to a difference in content, whereas manifestation usually refers to a difference in format. Again the distinction is not totally straightforward. A difference in format may produce subtle differences in the message, which is of course not only a problem for the current version but also,

as we shall see, what bedevils all schemes of preservation in the digital environment.

7.1.2 The definition of the authentic version

In addition, if we try to equate the authentic version with the version which results from a process of peer review, the version certified by a community, we have found that the equation only works for some disciplines and not others. In some disciplines, a lot of material is cited, which is not certified that way. Indeed such material is not certified in any sense other than that it is published by being made available, and one could add in a circular argument, because it is cited. We have also seen that in disciplines where e-prints are important, they are not always just preprints — preliminary to acceptance in a learned journal. They are not certified and some never will be certified (see the discussion of the concept of a First Publication in section 5).

7.1.3 The expressed wishes of scholarly communities

It is worth going back to what the community (or at least a sub-set of the science community) seems to have proposed as an appropriate approach, expressed in the recommendations following the 1998 workshop on Normative Issues, and recorded in this study in section 5.2.1. The 'experts' at this workshop wanted a "full specification of its status" to be attached to "each publicly available version of a document". They also regarded "formal peer review as essential in arriving at the final version of a scientific publication". Finally they noted the need for a "standardized citation practice" in the digital environment. In the context of this section, citation was read as identification. Citation is what authors do, when they relate to the scientific record. For further consideration of this point see **Paskin (2)** and subsection 7.4.

These recommendations represent what it is reasonable to hope to achieve in maintaining authenticity in these particular circumstances. It is at least part of what this particular group (supposedly representative) wants for their members. The recommendations are, to our mind, curiously ill-defined. This is not to blame those who drew them up. Digital transition is difficult and confusing, and means different things to different people and, for that matter, different scholarly communities.

7:2 RELEVANT CHARACTERISTICS OF METADATA SCHEMES

The reason for this subsection is not to provide a history of the development of metadata as it relates to scholarly communication but to pick out elements of this history, which are relevant to the purpose of this study. Nevertheless, we do have to begin with some general history and then isolate some specific themes.

7.2.1 How and why metadata schemes have developed

Metadata has not had a long history. What did we call metadata before the term was invented? There was plenty of data about data about but it is only in the digital environment that the concept has come into its own. This author remembers that in 1996 the translators in French at the UNESCO building in Paris sent out to the assembled 'experts' during the UNESCO/ICSU Conference

mentioned elsewhere in this study and asked them to provide the French translation. There was no consensus, in spite of the fact that by then the Dublin Core process was well under way (**Bearman 2**). Many spell checkers still reject the word.

In the scholarly environment the most fundamental reasons for the creation of metadata, the purpose of metadata, has been well set out in the following (abbreviated) explanation (**Morris**):

> **Resource Discovery** - identifying and locating a piece of information, the location (and possibly identity) of which was not previously known. An example from the print environment would be a search in a secondary database. Library cataloguing is one specific use of a subset of Resource Discovery metadata (such as Dublin Core).
>
> **Rendering** - realising a specific information object on the user's computer. To do this, the receiving computer needs technical information about the characteristics of the object.
>
> **Rights** - in order to carry out <u>any</u> operation on an information object, the user needs the right to do so.

The concern of this study is with the first two purposes, or at least within the context of these purposes. In practice, much of what we shall cover in this section will be driven by the third purpose.

The section entitled *A definition of metadata* in the DOI Handbook (Paskin 1) indicates the fact that a simple definition of metadata is not sufficient:

> "An item of metadata is a relationship that someone claims to exist between two entities.
> *[From The <indecs> Framework]*
>
> The word metadata means many things to many people. So we begin from this definition of metadata that provides us with a concise paraphrase of much of the <indecs> framework. This definition stresses the significance of relationships, which lie at the heart of the <indecs> analysis. It underlines the importance of unique identification of all entities (since otherwise expressing relationships between them is of little practical utility). Finally, it raises the question of authority: the identification of the person (individual or corporate) making the claim that a relationship exists is as significant as the identification of any other entity.
>
> At the same time, it underlines the essentially boundless nature of metadata -- the relationships between any entity and other entities are potentially infinite. Attempting to define a metadata schema for "all the metadata about something" would literally be an infinite task". (Appendix 3.4)

The emphasis on both uniqueness and relationships is important for our purposes.

Dublin Core (http://purl.org/dc) grew out a need for an improved information/ resource discovery in the networked environment (**Bearman 2** page 2) perceived by what one might call the 'greater' library community including museums, archives and government agencies. Dublin is Dublin Ohio, the headquarters of the library co-operative OCLC.

<Indecs> (www.indecs.com) comes from a commercial background and has an avowedly commercial aim. The acronym stands for "interoperability of data in e-commerce systems". The <indecs> metadata framework is laid out in detail in a handbook and data dictionary (**Rust 1**).

'Commercial' in this context is used as many librarians use the word, not to distinguish between commercial and not-for-profit 'vendors' of content, but between people selling such content and free material or material generated within and by libraries – not quite the same of course. The purpose is crucial.

The schema is explicit (**Rust 2**):

> "The aim of the project was to address the need, in the digital environment, to put different creation identifiers and their supporting metadata into a framework where they could operate side by side, especially to support the management of intellectual property rights".

It follows that concerns of authenticity are only going to important to the <indecs> community if they have relevance. Similarly, Digital Rights Management lies behind questions of the protection of integrity as we will see in the next section.

Although early work in the <indecs> context did involve criticisms of Dublin Core methodology and results (**Rust 2**), what has emerged is (to a large extent) a consensus embraced by two different communities. **Bearman (2),** page 19, explains that:

> "Libraries want to share content; publishers want to sell it...What (both) share
> is the need to identify content and its owner, to agree on the terms and conditions of its use and reuse, and to be able to share this information in reliable ways that make it easier to find."

Our concern, as we shall see, is primarily with descriptive metadata enabling resource discovery and not with the rights metadata, which is central to <indecs> and those initiatives making use of its framework, such as development of the Digital Object Identifier (DOI). In that sense, the purpose of the Open Archives Initiative with its concentration on discovery would seem to have more in common with the enabling of scholarly communication on the Internet as pursued in this study. However, as we intend to show below, such would be a simplistic view

7.2.2 The place of commerce and the role of the publisher

A few years ago Thomas wrote:

> "Most metadata framework developers misjudge the degree to which schemes will be implemented. In the scenarios currently envisioned by Internet planners, the burden of resource description falls upon those who create online content. Such an assumption is flawed, because information providers in many sectors will encounter strong disincentives to generating metadata for the Internet. Most of these disincentives relate to money. Whether electronic information is provided by governments, academic communities, or profit-driven enterprises, the key to promoting a metadata explosion is financial incentive".

An amusing squib by **Doctorow** makes a different but related point:

> "A world of exhaustive, reliable metadata would be a utopia. It's also a pipe-dream, founded on self-delusion, nerd hubris and hysterically inflated market opportunities".

This is a little unfair on the nerds, who always get blamed. As Mark Bide has said on various occasions, the M-word does drive listeners out of the room, but the limitations of the whole exercise are recognised by all those working on schemas.

Doctorow's first straw man or insurmountable obstacle is headed "People lie". The summing up under this head is instructive:

> "Meta-utopia is a world of reliable metadata. When poisoning the well confers benefits to the poisoners, the meta-waters get awfully toxic in short order".

It is not difficult to see the relevance of this statement to the word of peer-review and certification. We have seen how those in any discipline arrange claims for impartial peer-review in quite an elaborate hierarchy. Lots of journals are peer reviewed but the papers that emerge certified from the process are not ascribed the same status. This is not quite the same as an evaluation of the merits of the paper. It is a matter of relative trust. We have seen in section 3 that, in matters of authenticity, we have to accept relativity, however painful it is. This hierarchy is particular important in medical research and is a side effect of the uneasy relationship between the pharmaceutical industry and the medical profession. It seems to us that, in the larger context of scientific communication and the advancement of science, the real problem lies in how those outside the discipline are supposed to know. A journal might be certified and certifying, and the metadata may reflect that fact but the content is not taken seriously by those in the know. But that is also a problem of print, and it is difficult to see how digital transition is going to help solve it.

Doctorow makes another assertion that is also relevant. Doctorow has headings "People are lazy" and "People are stupid". In the previous section, we postulated that education was a necessity. This study considers that the author community can be educated. It has been educated in print. However, **Doctorow** is less hopeful:

> "To believe that J. Random Users will suddenly and en masse learn to spell and punctuate — let along accurately categorize their information

> according to whatever hierarchy they're supposed to be using — is self-delusion of the first water".

In the scholarly environment is this too negative?

One final quotation will round off this group. As I finish writing this section, a newsletter is emerging that seems to encompass metadata research. It is the Digital Document Quarterly and is available (free of course) at http://home.pacbell.net/hgladney/ddqstart.htm. The (self-appointed?) editor's definition of metadata is:

> "Document information that is considered not part of the document itself, but is often essential to the correct management of the document. Metadata is mostly added by people other than each document's authors".

This is a crude but rather relevant definition. The word 'management' points back to the last subsection, but it is the second sentence that concerns us here.

These quotations point to the following scenario. Metadata is expensive to produce. Without some reason for its production no-one is going to pay for it. Publishers or similar intermediaries, and not the author/creator, are likely to produce metadata in the digital environment as they do in the print environment. There are issues of and problems with trust. Publishers are producing metadata in order to make money out of exploiting the rights in the content that they own or are licensed to exploit. Any commercial organization is going to make sure that the nature of the metadata reflects the purpose for which they are creating it.

We will return to this scenario in the last subsection.

7.3 INDENTIFICATION APPLICATIONS

In this subsection we will look at applications. Both Dublin Core and <indecs> are concerned with the conformation of the metadata rather than the objects that could be described by it. We are taking the simplistic view that the DOI (www.doi.org) and the Open Archives Initiative (www.openarchives.org) are applications of <indecs> and Dublin Core, respectively. They are both concerned with discovery and identification. Both these initiatives are well documented on their sites and, on the whole, the documentation is accessible to the interested layman. In what follows, we look at both these projects, especially the former as it is further developed. We will also look at the ONIX and CrossRef as developments of (to some extent) and respectively applications of the DOI initiative.

7.3.1 The Digital Object Identifier

The DOI is concerned with the identification of an entity in the digital environment for the purpose of facilitating e-commerce. References below are to sections of the excellent Handbook (**Paskin 1**). There is no need to rehearse the way in which the scheme operates. What is worth looking at is the way in which the DOI works, as this is relevant to our purpose here.

The DOI was created by the publishing community because the Uniform Resource Locator (URL) was deficient "not least because it was not used to identify content but rather location":

> "The location is transient, whereas what was necessary was a means of identifying content itself, persistently and *without ambiguity* (2.2)". [The italics have been added].

There is a recognition here that ambiguity about content, that the content is what it is claimed to be, is both possible and to be avoided.

The DOI is indeed not intended to be and never has been a simple identifier:

> "If the DOI were simply a system providing persistent single point location on the Internet, then metadata would not be essential to its function. However, the DOI is conceived as much more than that. In order for the DOI to be able to fulfil its wider potential in providing the basis for a full range of services relating to intellectual property in the network environment, metadata becomes an essential component of the DOI System as a whole". (5.1)

The metadata envisaged is that metadata required for the purposes for which the DOI was created. However, that does not mean that the concerns of this study are not worth considering. If the purchaser asks questions of the content, which relate to questions of authenticity, metadata relating to questions of authenticity must be worth considering.

<Indecs> covers all envisaged DOI metadata and provides a framework for creating any new structured metadata. (5.5). This is important because the kernel metadata, the essential metadata for the system to work:

> "Supports only the simplest of applications — the discovery, from a DOI, of enough information to be able to recognise what it is that the particular DOI identifies".

If we want to add other metadata there are rules. The data has to be well-formed (appendix 5) and all metadata other than simple labels (what things are called) has to be:

> "Drawn from a controlled vocabulary of values, which are supported by a data dictionary in which those values are concisely defined".

As the <indecs> data dictionary is not set in stone but is undergoing constant and productive development, such values can be added. The demand for concision might be more of a problem, and there is the question of ambiguity. The following quotation from 3.3 of the Handbook is worth giving at greater length:

> "Identification requires that we understand precisely (unambiguously) what it is we are naming. And if we are to communicate and trade with others, we need to have a way of sharing that understanding -- is what I call 'a chapter' the same as what you mean by 'a section'? This doesn't

> require that everyone use the same vocabulary (an impossible demand) but to accommodate the multiplying entities being traded we need an ontology ... which could describe each precisely: 'Alice in Wonderland' as an abstraction, and the printed edition and so on as "manifestations" of that work in various forms. This could then provide a common analytic tool — a Rosetta Stone for the various ways in which things are described. That is now possible through further development of the indecs activity."

This section of the DOI Handbook goes on to describe other developments under way. One of these (ONIX) will be treated briefly in the next subsection.

7.3.2 ONIX

ONIX stands for ONline Information eXchange. It refers to a standard format that publishers can use to distribute electronic information about their books to wholesale, e-tail and retail booksellers, other publishers and anyone else involved in the sale of books. This information comes from the FAQ section of the ONIX site at www.editeur.org/onix.html. We are moving into the area of actual selling of publications rather than the selling of rights. Why?

The reasons are that the conformation of ONIX is more recent than that of DOI, it is now specifically concerned with e-books as well as p-books, and it is being extended to serials. It also has some interesting features.

ONIX documentation at www.editeur.org does demonstrate a way of accommodating the recognition of authentic entities. The reference is to Guidelines Release 2.0 at 8 August 2001. In section 4 of the XML message specification document ONIX defines XML 'attributes' to encode a type of source, to name an actual source and to datestamp its creation or last amendment. It is probably that currently no-one is making use of this capability. There is also the question of trust that comes up again and again. You have to trust the person (individual or corporate) who is making these claims.

We will return to these concepts.

7.3.3 CrossRef

CrossRef is described on its site (www.crossref.org) as "**a collaborative reference linking service** — through which a researcher can click on a reference citation in a journal and immediately access the cited article". It says more than that because it is essentially selling a service. It is the application of the DOI. It adheres to DOI rules. The owners are the 100+ publisher members, but there are affiliates in the library world. We will discuss their outreach later.

The site reveals little of planned developments because many of them are secret to the owners, but it is undoubtedly in a state of flux as those responsible seek what their market wants. There is one interesting entry that points to one relevant development:

> "Will individuals be able to use the CrossRef system to do resource discovery searching or look up DOIs?

> Initially the CrossRef system is being used by member publishers to enable reference linking using Digital Object Identifiers; however, there are plans to add functionality so that libraries, individual scientists and others can locate and access information on known references" (http://www.crossref.org/faqs.htm#What_is_CrossRef).

What information will these enquirers find?

It is worth noting that CrossRef is assigning a DOI for a Work, not to a manifestation of the Work. It is said that this course was adopted because of the need to ignore the different manifestations and avoid multiple citations. However, the publisher member of CrossRef will be selling, or more correctly in most current cases, making access available to an entity they have already sold as part of a subscription, a manifestation of a work. Using the terminology of this study it will be a version of this Work, the one which the publisher has certified. There is nothing wrong with this but, given interoperability with OAI, there will very likely to be two or more versions in play. One of these incidentally is likely to be free. We consider the discussions between CrossRef and OAI in subsection 7.3.5.

7.3.4 Open Archives Initiative

We have suggested that in the context of this section OAI can be seen as a development from Dublin Core, but that is only one way of looking at it. It is, compared with DOI, a relatively simple scheme embodying a simple concept, which is easily accessible through the site and especially the FAQ at http://www.openarchives.org/documents/FAQ.html. It is simple because the aim is narrow and straightforward. It seeks to make it possible for a scholar to find out what is of interest to him or to her from across a range of member e-print 'archives' or (preferably) repositories. The discovery is done by the use of the Open Archives Protocol (http://www.openarchives.org/OAI/openarchivesprotocol.htm). What is actually 'harvested' is the record: it is not clear to me how many records currently lead straight through to the actual content but presumably that is the aim. The FAQ explains the situation as follows:

> "The Open Archives Metadata Harvesting Protocol defines a mechanism for harvesting records containing metadata from repositories. The protocol does not mandate the means of association between that metadata and related content. Since many clients may want to access the content associated with harvested metadata, data providers may deem it appropriate to define a link in the metadata to the content. The mandatory Dublin Core format provides the *identifier* element that can be used for this purpose".

Of course, unlike CrossRef, it is still an experiment and (as it is admitted in the FAQ) there are a lot of questions still to answer.

We mentioned earlier that it is a simple scheme. The metadata set in use is simple and is described as follows:

> "The fifteen elements of Dublin Core has over the past several years developed as a de facto standard for simple cross-discipline metadata and is thus the appropriate choice for a common metadata set".

However this is not the whole story:

> "The metadata harvesting protocol supports the notion of parallel metadata sets, allowing communities to expose metadata in formats that are specific to their applications and domains".

This and the previous quotation are from the FAQ. It is too early to be sure what this means in practice. Is there an allowance for elements in the datasets that might be possible, which could have reference to questions of authenticity? In the current Protocol there is one interesting item which relates to datestamping;

> "A *datestamp* is the date of creation, deletion, or latest date of modification of an item, the effect of which is a change in the metadata of a record disseminated from that item".

The wording does seem to indicate that the latest version is what you get, and does not seem to assume any sort of definitive version, but perhaps that is reading too much into a definition.

7.3.5 Linking to the appropriate copy

When the DOI was first announced at the Frankfurt Book Fair, the problem now known as the 'appropriate copy problem' was obvious to intermediaries present and subsequently to librarians. The group that set up the DOI Foundation envisaged all identifications resolving to their own site, but in practice almost all publishers now have downstream contracts' and the intermediaries licensed through these contracts have multiple relationships with libraries. This is not the place for explaining how developments in CrossRef are seeking to handle direction to the appropriate copy.

What is particularly interesting is the way that that one approach to resolution (**Beit-Arie**) proposes an architecture that is a combination of the OpenURL framework (part of the way in which OAI works) and the DOI resolution system. OpenURL is an interoperability protocol that enables the context-sensitive resolution of service links for information objects (www.sfixit.com/OpenURL.html) . The picture from within the library community (www.jisc.ac.uk/dner) suggests that:

> "The DOI, CrossRef, SFX and Open URL are complementary frameworks and components that can be integrated. Collaboration between the SFX community and the DOI community is under way, to integrate the Open URL framework and the DOI framework" (Journals in the Information Environment).

We are presumably here in the realm of the appropriate copy, but once such 'integration' has been found to be possible, and there has been motivation to make it work, discovery can be taken further. The talk is certainly of navigation in

general, and not just from the OpenURL community. For example on the site of SFX, who provide the linking technology, a CrossRef spokesperson (Amy Brand) is quoted:

> "CrossRef intends to provide the scholarly community with a comprehensive network of linked publications, eventually covering all content types and content areas and serve as a vehicle for true collaboration among publishers, librarians, and vendors. Our alliance with Ex Libris [the company owning the SFX technology] is a very important advance, in part because CrossRef acts here as a vehicle for OpenURL-compliance on behalf of its member publisher" (http://www.sfxit.com/news/press_initial.html).

If you look at this statement in the context of the motivation behind CrossRef — making money out of content — and that of OAI — freeing scholarly communication — it is a remarkable one. The sort of co-operation and integration is possible with respect to finding authentic scholarly messages is examined in the last subsection. As Mark Bide has pointed out (www.indecs.org/london/ParallelUniverses.pdf) there are lots of ways of viewing stuff (he suggests three but admits to more). He writes:

> "In a persistent, distributed digital environment, we need these ... views to interoperate seamlessly".

7.4 FINDING AN AUTHENTIC ENTITY

During this section and in what precedes this final subsection, we have reviewed metadata schemes and applications of schemes that have been developed for purposes of identification or rather identification for various purposes. We began with a short overview of some of the expressed wishes of the science communities. In this subsection we will draw together some of the threads. How can scholars find what they want to find in the digital environment as it is now being fashioned?

What we can ask of metadata is part of the question. **Martin** wrote perceptively back in 1998:

> "What do we mean by *high quality* product information? Accuracy and timeliness are fundamental, as is consistency in following industry standards for the content and format of data elements. Richness of content will also be increasingly expected, not least because with today's technology we can afford it, and we have the tools to store, process and display rich metadata. Above all, product information must fit its purpose of supporting resource discovery".

This was of course written from within the environment of rights trading rather than that of scholarly communication, but the emphases are as appropriate to the purposes of this study.

7.4.1 The Work and its manifestations

In his document concerned with what he calls e-citations **Paskin (2)** has considered some of the issues that the concern of this study in the context of one aspect of the scholarly communication process. As he points out "scientific communication is founded on dependable links between articles (i.e. references to earlier work)". This is citation. For this purpose "linkage only between digital entities is insufficient".

He is interested in the question of differing versions. In his view (rightly):

> "For citation purposes, a scientist wants all reference to his 'work' to be counted irrespective of manifestation format".

In his (<indecs>) terminology, it is a matter of a Work, the intangible creation, and the different manifestations, the physical (paper) and the digital. He is looking to find a way to recognize that these are manifestations of the same Work and that they might be available from a range of locations in their digital form. For Paskin, working within the DOI Foundation looks to the work of the CrossRef Consortium, which "assigns an identifier to the 'intangible Work' entity and uses resolution to locate a manifestation instance". We previously discussed one aspect of this process when we touched on the 'appropriate copy' problem.

In passing, it is worth recording that I am is not as optimistic as **Paskin (3)** was then (2000) when he wrote this article about the 'sameness' of the different manifestations, even when they are supposed to be the same. For example the expression of an intangible Work in a digital form is not usually one expression. Most publishers produce a PDF version and an SGML version. We use SGML here to represent SGML derivatives also, for example obviously XML. The PDF version is by definition a digital version of the print original, but it is likely that the presentation of the message in the XML version will differ, and in some cases there will be more than one XML version. Some hosts want to convert to their own DTD.

Paskin (**3**) argues that we are not concerned with "differences in manifestation appearance (rendering or format)", but as we will see in section 9, this might be important in the context of archiving and preservation. His view is that, in these circumstances:

> "We have no problem saying that a PDF file and an HTML file of a paper are 'the same paper' even though a bit-by-bit comparison would show marked differences".

A new version, as he suggests, does require a "substantive change", but what is a substantive change? He writes perceptively:

> "This issue of 'granularity' (at what level of detail do we distinguish?) is a matter of social judgement, not technology. <Indecs> has coined a useful phrase of *functional granularity* as one of its guiding principles: "*an entity needs to be identified only when there is a reason to distinguish it".*"

There are always going to be small differences. They can be due to changes in format or they could be irrespective of changes in format. The author of an article in practice usually concurs with the judgement of the publisher in the print arena. For example, he or she accepts (usually) copyediting changes. In the electronic arena there is rightly a little more nervousness about publisher practices. I have suggested some reasons in an earlier section (2.1) but now that the electronic file is the way in which the message of the author is held definitively, any such nervousness has a new strength and validity

7.4.2 Handling different versions

In the subsection above, we have been concerned with manifestations of a Work or different expressions of the same (essential) version. However, as we have also seen earlier in this study, there are different versions of what is essentially the same Work. In some cases we can see a series. There is an original version and there are changed versions that embody corrections. We have to recognize the actualities of the situation in which it is the publisher that controls such a series, and it is not difficult to envisage metadata requirements that indicate the latest version so that identifiers can point to it. It is not, however, just a matter of date stamping. The reader will want to know that they are looking at a 'corrected' version. The corrections might be matters of substance or at any rate will involve more than putting in a comma (to use one of the examples Paskin gives). In any case the reader will want to know that he or she is looking at a version that might different from a version cited by another.

In the digital environment however, a version of a certified article might differ substantially, as we have seen in section 5, depending on its format or rather whether it is in a print format or some digital format. As we have seen, publisher practice has yet to be formalized. The ISSN authorities are clear about identifications of separate versions insofar as they relate to the title of the journal:

> "Are different ISSN numbers assigned for the different versions of a publication (paper, online, floppy disk, CD-ROM, microform...)?
> Yes, each separate edition on a different medium should have its own ISSN, even if the title is identical. Only reproductions issued as substitutes to the original retain the same ISSN".
> (http://www.issn.org:8080/English/pub/faqs/issn).

Many publishers ignore this and many also have yet to decide on the status of the additional matter relating to the article, which they may hold on their servers. In a sense, however, this is a problem not of the digital environment as such, but of the current and possibly permanent 'hybrid' situation where scholarly communication is both print and digital. It is probable that we should be closely examining how hybrid scholarly communication is handled. One cannot assume, as many seem to do, that print as a separate stream will go away and be restricted to print outs. Paskin recognizes this (page 2) and speaks of the danger of creating "two worlds that don't interconnect". For the purpose of this study the digital object identifier points to the digital entity (not the print entity) and as yet no-one has suggested that any metadata should indicate a different from the print version.

7.4.3 Handling certification

It is clear that scholars want to know that an entity they are citing or seeking is certified. There is a clear understanding of what certification means (section 5) even if there are non-certified entities that a scholar might want to cite or find (see the subsection below). If there is an authentic entity — one version among other versions — it is going to be the certified or definitive version: that is the way that scholarship works.

But how can a version be uniquely and unambiguously identified? It seems to me that there can only be one way of handling identification of a certified version, and that is setting up a simple field. The entity is either 'certified' or 'not certified'. Exactly how this is handled as part of an extended metadata kernel is not part of my expertise. It would seem to be related to origination.

What we do get into is the question of who decides whether an entity is certified, and the answer is clear. It is the authority that assigns the identifier and composes the metadata associated with it. In other words it is, in practice, the publisher. As we have seen the organization or company that publishes is a publisher whether it is a library or even an individual running his or her own alternative journal. The role and function of a publisher carries with it the responsibilities of the publishers, and that includes assertions of certification. As usual <indecs> provides guidance in its articulation of the principle of designated authority - see **Rust (1)** page 10:

> "The author of an item of metadata should be securely identified. Well formed metadata must provide mechanisms for declaring the authorship and for authenticating claims of veracity in any item of metadata".

In the context of rights trading, the identification of the supplier is key. The exact relationship between the supplier of the entity and the author of the metadata is something that could be explored further but not in this study.

We are returning to the territory of section 3 where provenance and trust play an important part in establishing authenticity. All the evidence is that the body that asserts certification has a part to play from the point of view of the scholar, although for most purposes it is the branding of the journal in which the certified article is included that is more important. The scholar is constantly making qualitative decisions about what is worth looking at or for, and what credence to place on the message embodied in a particular entity. It is difficult to see how such qualitative decisions can easily become part of metadata, but already questions of this sort have been considered (see the next subsection). The authentic entity does not of course have to be a certified or definitive entity. We will look at this in a further subsection.

7.4.4 Handling ranking

As we have seen, each discipline ranks sources of information in different ways. It is difficult to see how any such ranking can be handled by any sort of metadata. This is essentially a matter of a qualitative decision from within a community.

Within the context of one learned discipline, electrical and electronic engineering, there has been an attempt by a learned society in the field to use PICS specifications to indicate peer endorsement of articles. The society is the IEEE and the IEEE Computer Society, heavy-weight professional associations. PICS, or Platform for Internet Content Selection, was developed by the World Wide Web Consortium to provide a common foundation for definition of rating systems and rating services that can be applied to objects on the Internet.

A description of the project appears in **Armstrong** (1997), and there is said to be further information at www.computer.org/standards/Internet/peer.htm (although the site seems to be currently unavailable). The project did not fly and seems to have been forgotten in some relevant quarters. However, the fact that it was attempted or at least discussed is important in itself. Armstrong quotes lists of quality selection criteria from other projects that design subject gateways that look sensible but that are very qualitative. In the UK, at a time when subject gateways are to be heavily funded by the government (www.jisc.ac.uk/dner), the approach has a lot of merit. However, it can be argued that the initiative cannot be led from within the JISC bureaucracy, responsive though it may be, but from institutions representing the learned community.

7.4.5 Where do e-prints fit in?

E-prints have been discussed in an earlier section of this study. We have seen that in the digital environment they are no longer synonymous with preprints in the sense that they do not necessary represent earlier versions of what will become refereed articles. We have also seen that in some disciplines they are treated as a useful part of scholarly communication and not just for the contemporary scholar seeking knowledge at the cutting edge. The place of preprints in the informal part of the system is not new. However, it is now clear that in some disciplines e-prints are worth saving for posterity. The recognition of the new status of preprints/e-prints in the digital environment is part of the motivation behind the attempt at definition in the Proposal — as discussed at length in section 5.

E-prints can be preliminary statements exposed on the Web for various reasons such as a claim to priority or a request for feedback. If they are published, they might be amended to reflect the changes brought about by peer-review or they may be linked to the peer-reviewed article. They may never be submitted for publication or they may remain in the 'archive' even if rejected as a contribution by a reputable journal. In practice, their current stability is not secure and their status is in the archives that are known to me is not made clear in any controlled way.

The whole thrust of the OAI is to provide a simple means of harvesting the metadata associated with the entities in the archives. Under the heading *Is the Open Archives Initiative only concerned with metadata?,* the following information is provided:

> "The current OAI technical infrastructure, which is specified in the Open Archives Metadata Harvesting Protocol, defines a mechanism for data providers to expose their metadata through an HTTP-based protocol. There is nothing in the OAI mission that restricts the work of the OAI to

> metadata alone. However, we are guided by the goal to define a low-barrier and widely applicable framework for cross-repository interoperability and believe that exposing metadata is plausible route to such a goal. We may, in the future, explore and define other mechanisms for interoperability". (http://www.openarchives.org/documents/FAQ.html)

The metadata associated with this mission is simple cross-discipline metadata based on Dublin Core. There is nothing to stop the 15 elements of Dublin Core being extended to provide some information about the status of the entity within the context provided by those notions of authenticity that we are exploring in this study. However, there is no evidence that this is as yet part of the mission.

It is a curious but not a surprising fact that the driving factor behind sfx is not just a wish to enable free access to scholarship as is frequently presented. This is part of the story, of course. However, economic motives, but library economics this time rather than publisher economics, are prominent is discussions of the importance of the enterprise:

> "No more dead links whereby the user clicks on a link to navigate to a new information space but finds that they do not have rights of access to the resource to which they have linked and are therefore blocked from access.
>
> SFX allows the librarian to define the library's electronic collection, including both licensed and freely available resources; and to determine the manner in which the component resources can be linked to best suit the library's users". [quotation from the SFX site referenced below]

If some level of interoperability between CrossRef and OAI is achieved, as seems to be signalled by information coming out of SFX site (www.sfxit.com), a link might lead to a publisher or an archive and in neither case the status of the content described by the metadata is made clear. At least the DOI is likely to lead to the database of a publisher, and what the publisher invests in is almost certainly going to be in what can be perceived by the community as the definitive version. The value added will usually revolve around a process of certification and the context will be one that traditionally asserts fixity and authenticity, although with only partial justification. As we have seen, most if not all archives have not bought into this tradition. It will be interesting to see if as/if the e-print movement does extend outside its heartlands in physics and mathematics. Scholarly demand to know what they are discovering will lead to some rethinking within the movement.

7.4.6 Finding out about what has been identified

As we have already seen earlier in this section, there has already been some thinking within the DOI environment about the possibility that scholars might wish to find out something about the entity that a citation points them to. The concept is of course not alien in the general Internet environment. There is for example, reverse lookup from telephone numbers to addresses, which is widely offered. It is not alien either to the DOI Foundation (see **Paskin (1)** 5.6). It is interesting, but not surprising, that the presentation of a potential facility is linked to commercial opportunity rather than to scholarly needs:

> "Reverse look-up (from metadata to a DOI) is *not* a function of the DOI system itself. Reverse look-up may be offered by other services as a value-added feature. Individual applications or registration agency services will offer this service by agreement with their registrants and suppliers on commercial terms; this will not be determined by IDF. In many areas of intellectual property, extended metadata and reverse look-up via sophisticated searching techniques is an important business activity. As a matter of policy, the IDF will not consolidate DOI state data or kernel metadata for resale or re-use. This data, where held by IDF, is solely for the purposes of permitting look-up from a DOI to the declared metadata by any user".

Someone has to pay for both the system and the enhanced metadata, which in practice we are envisaging in this section, although the exact nature of it has been articulated. Publishers will pay only if authors and readers want further information than it is planned that they will currently receive.

However the developments that have been touched on pan out, in practice there is little doubt that there is no scope for any sort of centralized control, either from publisher or library interests. **Cockerill**, writing in defence of PubMed Central but using arguments that carry conviction in the situation described here, insists:

> "Multiple communicating archives makes transparent what should already be obvious — that existence of one or more central indices and archives in no way implies a central point of control for what can be published".

Standards will depend on what scholars want and how they communicate.

8. The protection of the authentic entity

When this study was originally scoped, we had assumed that there would be a lot to say under this heading. One of the five sections of the original proposal were intended to answer the question "How is the published article transmitted securely?" We will see that the lack of interest in general questions of authenticity among those involved in scholarly communication either as author/readers or as intermediaries of various sorts is paralleled by a fundamental lack of interest in protection — at least the protection of integrity and paternity.

It is interesting that the authoritative discussion of the relevant <indecs> principle the principle of unique identification, seems to assume protection as one of the four most important properties that enable an identifier to make possible wider interoperability.

These properties are:

(1) uniqueness within a given domain
(2) stability (identifiers should never be changed or transferred)
(3) *security, whether through protection by watermarking of encryption, and/or by internal consistency through the use of check digit algorithms* (our italics)
(4) the public availability of some basic descriptive metadata for the entity identified, without which the identifier has only limited use (Rust)

We have already discussed the other three properties in previous sections, particularly the last.

Protection costs money. The context in which money is going to be spent by publishers, who again take centre stage, is digital rights management (DRM). In this section we first elicit the relationship between DRM and the issues which concern us. We will then look at formats and how they relate to protection. Finally, we will look at the realities of watermarking and encryption, following up the promises implicit in the already quoted principle.

8.1 DIGITAL RIGHTS MANAGEMENT

The relationship between digital rights management (DRM) systems and the concerns of this study is a subtle one and the language used by DRM evangelists can mislead. For example, the words 'integrity' and even 'authenticity' have been much used in their literature, be it less frequently than they once were, according to recent survey of some sites surveyed. In this subsection, we will try to unravel what is really being talked about.

DRM is more and more a matter of interest and concern to the larger scholarly publishers, and therefore comes into the area of scholarly communication and its concerns, but the systems have been built up with B2B (business to business) as

the driving force. As we will see, there does not seem to have been any adaptation of the scholarly environment.

It is appropriate to make one other point here, which relates to the whole of subsection 8.1 but not to the rest of the section. DRM is mostly concerned with book content and not with the journal content — but this is changing. For example, see the quotation on the DocuRights brochure (www.docurights.com). Further examination of what lies behind the interest of journal publishers will probably reveal that it is associated with the 'reprint' business and monitoring the activities of the pharma companies than with more central dissemination of content.

We have already explained that DRM is the context in which the protection of a digital entity is mostly considered at the present time. Protection is only part of the operation of DRM systems. A useful survey of the situation in 2000 explains:

> "While digital rights management (DRM) is currently a buzzword in electronic publishing, there is much confusion surrounding the term. DRM is often used when describing security technologies, although this is simply one vital aspect of an end-to-end digital rights trading solution. Strictly speaking, DRM refers to the provision of back-end services relating to the trading of digital intellectual property such as the tracking of content usage and clearing house services which monitor the various types of rights purchased and distribute payments to the parties involved" (**EPS 2**).

It could also be added that some of the security technologies are concerned not just with tracking but also with the prevention of re-use in another product, without permission.

8.1.1 The commercial imperative

We will concentrate on this 'vital aspect', but first we need to look further into the reasons why DRM is being taken up, or not as the case may be. It is part of the structure we have previously touched on in section 7. What is being traded is intellectual property.

Another report, originating from the same consultancy already quoted above explains the relationship between publishers and users as follows:

> "It is not entirely clear what publishers need to do with regard to securing their content, nevertheless is irresponsible not to have a plan in place. Users should ideally see security measures as part of a partnership arrangement, not as an issue of conflict. The consumer experience needs to be satisfactory in terms of cost, content and convenience" (**EPS 3**).

The first sentence of this quotation points forward to the measures to be discussed in subsections 8.4 and 8.5. There is a recognition that there is a joint concern for the seller and the buyer in the security of the content, but the nature of the relationship is clearly between these parties rather than between the author and the reader.

8.1.2 The DRM process

One of the reports (EPS2) already quoted provides a useful description of the process involved in protecting and exploiting the intellectual property:

- The publisher produces a module of content.
- The publisher packages the module with associated usage rules, cross-promotional material and encryption. The user receives the secure content package from a range of possible sources, i.e. directly from the publisher or from a distributor
- Coding within the package interrogates the user's hard drive to see if there is a permit indicating a prior relationship with the publisher, for example an ongoing subscription or evidence that a certain amount of content has been purchased which entitles the user to a discount.
- If no permit can be found, the user is prompted to request one.
- The request goes to the clearing house which handles payment for and delivery of the permit and collects usage information for each particular piece of content.
- The user accesses the content, viewing, printing or saving depending on the usage rules.
- The user may pass the content on in its secure package to further potential users, referred to as superdistribution.
- Payment and usage information are transferred to the appropriate departments of the publisher.

> "Players in this market fall into two groups, those who facilitate the packaging of content and those who handle the clearing process. Members of each group form partnerships with each other to offer a complete solution to the process.
>
> Players in the packaging and encryption bracket of digital rights management offer customers the licence of software, which enables them to package modules of their content in secure containers. Business rules governing the usage of the content and the price may be included within the container. The creation of these containers allows publishers to sell small modules of their content and gain revenues through superdistribution. These players also handle the secure passing of information between customers and the clearing house and between the clearing house and the relevant departments within the publisher". [There have been several truncations within the quotation.]

As we can see from the above, the creation of the module is perceived as something very much in the hands of the publisher. This is not a big deal, as most scholarly communication involves a final version determined by the publisher, but the wording does seem to refer to information rather than knowledge with the creator essentially synonymous with the publisher. More relevant is the other end of the process. The secure container is secure in some cases only until payment has been made. In some cases, payment is made for only part of the content; this is examined this below. Security is only important where there is money to be made; where the rights are linked to the content. Integrity is important only insofar as the customers want to make sure that they get what they bought.

8.1.3 DRM offerings

When this study was scoped, part of my aim as the author was to examine those commercial offerings of digital rights management on the market in the light of the concerns of scholarly communication. This survey has been done but in the end it has proved to be irrelevant. In some of the literature the words 'authenticity' and 'integrity' are used. Further investigation however, invariably leads, as one would expect, to a recognition that the concern is to deliver to the buyer that which they have been sold.

It has been made clear to me that on a number of occasions there was no demand from either publishers or users (other publishers or other intermediaries in this context) for any further level of security of content. Some of those who were interviewed were DRM houses but others were intermediaries concerned with building up databases of content.

Protection of the integrity of the content is possible, even after the final payment has been made, but it does cost more. The same goes for protection of paternity. Making such protection possible where there is no economic driver, cannot be the concern of DRM vendors, whose offerings have to reflect the demands of their customers. Tracking of usage is the important aspect of the systems. In subsections 8.4 and 8.5 we look a little further into what is possible with encryption and watermarking.

Unfortunately, none of the companies concerned felt able to commit those general assertions that have been mentioned above to paper. As a result I did not consider it appropriate to quote from these conversations. The names of relevant players in DRM can be found in reports quoted above and in the useful material available on the site of the specialist company Rightscom (http://www.rightscom.com).

8.1.4 Superdistribution

As we have seen above, this term seems to be used in two different senses. It refers to the rights adhering to the content being tracked through successive users and usages. That definition does not concern us here. It also seems to suggest slicing and dicing of content as a concern. The concern in the DRM context is that the sliced content continued to represent a source of revenue commensurate with the amount of the original content used.

At first glance, this seems to be a translation into the digital environment of the concern of any rights manager with payments for permissions. However, this is publisher generated slicing and dicing. The publisher produces a module. We are looking primarily at slicing and dicing information (again) rather than in the realm of scholarly communication, but the approach does not fit in well with some of the concerns that we have been discussing in earlier sections.

One can however, be unnecessarily pessimistic or optimistic depending on one's viewpoint. An experienced publisher (November 2001) wrote of the policies of his own large company in the scholarly arena:

> "We are not slicing and dicing at anything below chapter, article or encyclopedia entry level. Technically, of course, you can slice and dice at any atomic level, but in practice, I do not see any drive to do this". [personal communication]

8.2 PORTABLE DOCUMENT FORMAT

In this and in the next section, we will look at the ways in which publishers themselves directly make content available online, and how the processes relate to the concerns of this study. For a number of reasons (already discussed in a different context in section 2) portable document format (PDF) has emerged as one of the main ways in which scholarly content is presented on the web. Rather surprisingly, some level of protection of integrity has been built into the technology of creating PDFs ('distilling') since its inception.

The publisher quoted already in the previous subsection also wrote at the same time:

> "I think the main guarantees of authenticity are the format — PDF is unalterable, location — the publisher's or reputable aggregator's site, and identifiers such as the DOI, which will resolve to an authentic location".

We have already discussed the latter two aspects of his solution in section 7, though from a different angle. In the subsections below, we will look first at the nature of the security of integrity provided by PDF, then how it can be evaded, and finally what publishers are actually doing. The first two of these subsections rely heavily on insights and information provided by experts who remain anonymous

8.2.1 Protection of integrity using PDF

PDF files, if viewed using the Reader, cannot be edited. What you cannot do is wholesale editing of the original file and save it as the same file. As we will see, there are plenty of ways of interfering with the content when it is transferred to another file, but for the layperson at least, the file produced using standard settings prevents a challenge if tampering is wanted.

There is surprisingly little information about the security provided by this property on the home site for PDF (www.adobe.com), but in general this site is a particularly difficult one to search for non-technical implications of all the technologies offered. Indeed "browsing is not allowed in this directory". The general statement can be made that, judging from their site, Adobe Systems are more concerned with virus attacks than alterations to standard files, but there are enhanced ways of encrypting files provided both by plug-ins and in new releases.

The author is advised that if you look under Document Info within File Menu for a PDF file you will see that security can be set to disallow printing, changing and selecting text and graphics. This enhanced level of security has to be set (by the typesetter at the time of distilling) when the file is created.

The author is also advised that encryption of the content of a PDF file can be encrypted using a particular technology supported by Acrobat reader or printer

technology. If an encrypted file is distributed it will of course need to be decrypted before being downloaded or printed out.

It is highly unlikely that scholarly communications are going to be encrypted in the way described in the immediately preceding paragraph. There is the question of costs. The offerings of such an encryption technology bear all the marks of being aimed at the B2B publisher.

More general points on publisher policies are made in 8.2.3

8.2.2 Evading protection provided by PDF files

Standard settings result in a file, which can be altered in the following sense. One of those consulted writes:

> "It is possible to select the content and copy and paste into another file. You may lose some of the special symbols *in itself a cause for the loss of some of the authenticity of the message* [my addition and my italics], but you get the bulk of the text to work with. If you hold the PDF file on your local machine you can overwrite it with the new version and then send the amended version, which has the same filename on. The time and date stamp on the file will be different to the original, but unless these details are compared no-one will know it is a later version". [personal communication]

In other words, the message of a scholar transmitted in PDF format can be altered, but it is not something to be done lightly and it is possible (as stated here) to recognize that the file has been changed.

Another person who was consulted makes the following comment about files protected as described in the third paragraph of 8.2.1:

> "Acrobat Distiller offers the options to make a document read-only and also to prevent copy-and-paste. As I understand it, no application that can read the PDF format will override these options". [personal communication]

8.2.3 Publishing policies

The above two subsections explain what can be done. Publishers and a publishing consultant have provided the information. However, the facts of what can be done do not in themselves give any indication of what is being done by publishers and why. This subsection will not provide the answers. It is not easy to elicit technical information from publishers, because, in addition to the reluctance (already mentioned) of publishers to answer questionnaires, there is the additional problem that the executive deputed to answer such queries is not the technical expert. In a previous work, I found it difficult to get sensible replies to technical questions in (**Watkinson 5**).

Arguments that I have used in section 2.1 to encourage acceptance of online journals are described. It is now my impressions that the security is so accepted by the academy that publishers do not need to make the point. From the very

beginning of online journals, it was discovered that anything more than the basic level of security described above interfered with the work of the intermediaries and aggregators, for example, when full text indexing was involved. The same process happened earlier with CD ROM security, so it did not come as a surprise. It is our suspicion that typesetters are being instructed only to apply basic security, which, as we have seen, does help maintain integrity in some circumstances, but further research would be needed to confirm this fact. Scholars as users are unlikely to be happy with any more stringent settings that might prevent them from cutting and pasting.

8.3 EXTENSIBLE MARKUP LANGUAGE

Extensible markup language (XML) combines the virtues and avoids most of the problems of both SGML (the original) and HTML (SGML on the web).

In the earlier subsection on versions and elsewhere, we have discussed the relationship between PDF and SGML or its derivatives. For most publishers looking to the future, PDF is for printing and XML (now emerging as the e-format of choice) is emerging as the main, definitive vehicle.

In the previous subsection we have described protection of PDF files. In the following subsections, we examine the role of XML and the dangers from XML.

8.3.1 XML as the definitive format of choice

Peter B. Boyce has contributed so much to the development of strategies for taking advantage of the functionalities of the Web to enhance scholarly communication. In his work with the Astrophysical Journal, he realized many of his hopes. He wrote last year about the place of PDF files in the publications of the AAS as follows:

> "According to the feedback we have, the major value of our journals lies in the abundant links to the references we provide to abstracts, to full text, to machine readable data tables, to astronomical databases, and to supplementary material relevant to the articles. These links, along with the electronic-only information, are lost when only the PDF files are available. Only the electronic version of our articles contains the complete set of information. The paper version is not the version of record. The distribution of the emasculated version via PDF files is not a threat to our journal circulation. On the contrary, such distribution serves as advertising for our journals. We also let authors post the PDF files on their papers both on their own Web site as well as on the Los Alamos e-print servers. This has not proved an economic problem for us". (pboyce@aas.org on liblicense-l@lists.yale.edu on 20 April 2001)

This interesting statement raises a lot of questions, not least of which is whether, from the point of view of scholarly communication, it is a good idea to encourage the dissemination of an incomplete message. There is also the possibility that the long-awaited structured PDF or PGML will one day appear and enable all the linking and all the other functionality to be provided in the context of this particular technology.

8.3.2 XML in the wider publishing environment

A report from a few years ago sets the scene well:

> "By the start of the year 2002, most textual content served on the web irrespective of the format in which it is served will have been converted to XML and will be stored in XML content management databases in highly granular form. These will be fully integrated with multimedia content and with complex metadata relevant to the data structures. By the end of 2004 the delivery of pre-rendered documents in formats such as PDF will have largely disappeared for all but the most highly styled products. PDF may however survive as the major format for print on demand purposes". (**EPS 1**)

This is a prediction in the context of B2B publishing. It could be argued that much scholarly communication, especially in science as loosely understood, comes into the 'highly styled' category. Nevertheless, it carries weight. It certainly fits in with the other aspects of the drive towards DRM as outlined in 8.1.

To quote from the report again:

> "Digital products are very different because they allow the user to access information at much lower level of granularity. In some cases this can be as little as a paragraph, in others subsections, or even whole chapters. Users of digital information therefore need to have content delivered to them in a form that allows them to find and aggregate units of information into structures they define whether for reading, reuse or re-purposing. This forces publishers to take a 'content centric' approach to managing data and building products.
>
> To be able to adopt this 'content centric' view, publishers must have technologies that support content centric workflows which aim to create, manage and store data, with its associated metadata, in granular form in databases. With these technologies in place traditional and digital products can be easily and quickly assembled in direct response to customers needs.
>
> These technologies allow publishers to describe the low level granularity in streams of text —the units of information — and databases in which they are stored must understand the granularity. This can be termed structured content management or asset management. XML must be adopted because it is the only technology available that allows the data to be described with sufficient robustness to support the digital publishing model and is small enough, cheap enough, and quick enough to enable the development of products from which business and social benefits can be generated".

The message here is clear. XML is good news because it enables easy alteration of the message and a loss of integrity as standard practice. We will look into the implications below.

8.3.3 XML files do not protect authenticity

We have a different picture here, which we need to look at exclusively in terms of encryption and watermarking in the next two sections. XML is not intended to protect authenticity as it is. Plain text formats such as HTML are not encrypted formats and so if they can be read, they can be altered. Encryption can be used to protect plain text documents in transit, but not in use. Digital signatures are the only option available for protecting such documents in use, and only as a means of verifying whether the document has been tampered with or not, not as a means of preventing it from being tampered with.

8.4 ENCRYPTION, WATERMARKING, AND DIGITAL SIGNATURES

In the original scoping of this section, the intention was to cover the technologies involved in some detail. However, in view of the fact that the general theme emerging from our consideration of the context of the technologies is that they are in fact not of much relevance to the study, what follows will be something of a simplistic treatment of a complex range of options. I have raised questions relating to the operation of their systems with a number of purveyors of intellectual property protection. Although most have admitted lack of interest among publishers in questions of authenticity, none have agreed to put anything in writing. The word indeed rarely appears in the sales literature. An exception is in the claims for the DocuRights technology (for example, see http://www.docurights.com/faqpub.html).

8.4.1 Some definitions of the technologies

There is good reason for some definitions, because the technologies are sometimes confused and usually not fully understood by the lay person. When considering encryption and the literature relating to it, one has to be carefully to distinguish between temporary encryption of data in transit and permanent encryption to prevent unauthorized use. Most of what is written has been concerned with the first purpose, whereas for the purposes of this study, the second purpose is more important. It is also important to recognize that protection of data from being altered can be achieved, or at least aided, in two different ways. Encryption can be used to protect data in that only a computer program that has access to the decryption key can read the data. Digital signatures do not themselves prevent alteration, but they do make it impossible to make a change that is not detected by any software that understands the digital signature. Encryption prevents, signatures deter but do not prevent.

One can summarize as follows. Encryption involves the enlisting of mechanical means to prevent the use of content without the knowledge of and agreement with the content owner. Watermarking prevents the paternity of the content being ignored or its assertion obliterated. Digital signatures validate the paternity and presumably the integrity. We will return to digital signatures in the next section when we consider archiving. In our view and for practical purposes in this context, digital signatures are less relevant to content protection than they are to financial transactions, but for another view see Bide (page 8).

8.4.2 The circumvention of encryption.

There is however, considerable interest in encryption and its circumvention at the present time. The debate centres on the Digital Millennium Copyright Act in the USA (**DCMA**). Proponents see penalisation of circumvention as nothing more than the implementation of already agreed WIPO provisions (page 3). It is concerned both with unauthorized access and unauthorized copying (page 4) Opponents, very vocal among the library community, consider that the Act, by seeking to prevent all circumvention, makes those allowed to access and/or copy under fair use provisions liable to prosecution. The debate continues. The point to be made here is that during this debate the question of assaults on integrity and paternity has been noticeable by its absence.

8.4.3 The relevance of the technologies

One again we return to the commercial justification for spending sometimes considerable sums of money on protection using these technologies. Certainly such money is spent for some high value information in, for example, the area of B2B publishing. In scholarly communication, the revenue to be extracted from control of onward use does not (in general) currently warrant the spend. There is a second point to make about the technologies, which is naturally somewhat glossed over by the companies that sell their own implementations. Encryption does make it difficult to access content. That is the purpose. We have touched on one response to this in the previous subsection. It is also very relevant to the content owner who wants to encourage use of a publication that any protection discouraging use by making it more difficult, is a mechanism that needs serious justification. Users are less and less intolerant of passwords or any barriers (multiple clicking, for example) that cut them off from the immediate gratification of what they seek. We have seen earlier that the prevention by the use of appropriate settings of PDF files of cutting and pasting does arouse some hostility. It is another example of the way in which the scholar as author may react differently as a user or reader. One partial source (**Sealed Media**) puts it neatly:

> "The common objection to the majority of current DRM systems is that they are far too inconvenient and invasive to the content owners and their users" (page 5).

8.5 PUBLISHING POLICIES

Kircz (6), whom we have quoted before, issued a challenge in 1997:

> "Since the scientific integrity and certification of the original (and each updated) version must be uniquely defined in an electronic archive (or library), standards for dating and electronic watermarking must emerge. This will enable future generations to follow trails in scientific discussions even if the documents evolve dynamically and more authors change and improve an electronically available text".

This is a tough agenda, which, as we have seen, does not seem to have attracted much attention as yet. Even where watermarking or encryption are made use of in scholarly communication it is essentially for a static and not a dynamic digital object. DRM is currently concerned with the print equivalent not digital entities of a dynamic type — but there is of course no reason why this should remain the case.

We have recognized that XML files do not lend themselves to protection but that money can be spent to make sure that any tampering with them can be recognized. We have already looked at what publishers are doing and are likely to do with their PDF files to protect authenticity. Our conclusion is that they are likely to pay to protect at any higher level than is currently intrinsic in the creation of such files. One informant pointed to an example where money is spent on watermarking of PDF files for a specific purpose, special to the complex relationship between medical publisher and the pharmaceutical industry. This publisher writes concerning the practice of a particular large company:

> "The forthcoming implementation of watermarks on our PDFs is as a result of concerns of our special sales people about the illicit copying of articles. They have noticed reprints of articles on the stands of drug companies where we have not sold reprints or rights. We believe that some drug company staff persuade themselves that, because they have access to a PDF via their legitimate subscription they can do what they like with it. We believe that if we watermark the PDFs, thus making it plain on the printout that this particular copy has not been licensed for multiple copying, they will go through the proper channels".

How do these discoveries relate to the <indecs> property (**3**) quoted at the beginning of this section? The short answer is that there is as yet no pressure from the scholarly community to demand this security, except in the sense that the representatives of academe as users do begin to demand that what they pay for is what they get. These representatives are the librarians, and it is interesting to observe that the recognition, particularly relating to the maintenance of the integrity, is just beginning to become a topic on library list-serves. The context is usually related to the question of whether it is appropriate for publishers or intermediaries to remove publications from web-sites — you can in the digital environment but cannot effectively in the analogue. Nevertheless, this could be just the start of a wider interest in the topics discussed in this study. For the most current, active consideration, see the threads available through http://www.library.yale.edu/~llicense/.

9. Archiving and preservation

The main concern of this study has been to establish the place of and threats to authenticity in current scholarly communication. Nevertheless (for example, in subsection 1.4.4) it has already been established that much of such discussion of the authenticity of these 'messages' has in fact been prompted as part of the considerations raised in contemplating the archiving and preserving of digital entities. We write of contemplating advisedly. In this section, we are not really writing of the practical application of the principles elicited in the long discussion of section 5, because these seem to have little relationship with the way actual decisions are being taken by those responsible. Obviously, however, there is and has to be some sort of connection.

It is worth pointing out here that when we write to archiving digital entities we are well aware that archiving, for example by publishers exploiting content by repurposing, does not need to be associated with preservation 'in perpetuity'. In this section, unless stated otherwise, we are always associating the two concepts.

In the first subsection, we will look at the context of decision-making, and in particular at the drivers involved in money being spent on the archiving and preservation of digital publications. We approach the extensive literature on archiving a tangent. Authenticity is a concept at the margins of both the policies adopted and the technical methods and considerations proposed and beginning to be adopted.

The substance of this section can be divided under two questions. In the first place there is the question of selection. Do questions of authenticity have any reference to what those concerned with archiving and preservation are seeking to archive and preserve? The second broad question relates specifically to preservation. The quotation in 5.1.3 is taken up again. What is essential to the intellectual message encapsulated in the digital entity? Taking vocabulary from the world of licensing we could denominate these questions as 'upstream' — what is it we are taking on board? — and as 'downstream' — what of it can we save? In these subsections our study might be described as the anatomization of the processes and procedures involved by looking at the what, the why, the how and the where of archiving and preservation in that order.

The theoretical discussion in section 5 has already been mentioned. Other relevant issues have also already been raised in sections 3 and 5. The section on submission metadata in 9.2.2 is obviously related to the discussion of metadata in general in section 7.

There is an extensive literature on digital preservation. The best entry to the subject is the bibliography at http://www.loc.gov/preserv/digital/dp-news.html. Unfortunately, much of what is written is not strictly relevant to scholarly communication in general and certainly does not take into account the concerns of this study. From the point of you of this author, the IFLA Guidelines are particularly disappointing and narrow in its concerns (**Lariviere**)

9.1 THE CONTEXT OF ARCHIVING AND PRESERVATION

We have suggested throughout this study that the consideration of questions of authenticity, within any particular framework of decision-making, cannot be separated from the realities of the context in which decisions are taken on why and how to spend money. Why is archiving and preservation undertaken and by whom?

In section 5 it became clear that we start essentially with models derived from the print environment. Actually, implementation in the digital environment is only just beginning and, because it is just beginning, we do not know if what is being attempted will work in practice. It also has to be admitted that at present there is very little content of importance to scholarly communication that is available only in digital form, but this is changing. However, it is not just because 'born-digital' entities are now evidently of importance that there are pressures to archive and preserve digital entities. Our examination of the drivers involved will show a more complex situation, in which various movements towards archiving bring different baggage that might impact on whether or not concerns about authenticity have more than a nominal place.

9.1.1 Old concepts adjusted to meet new circumstances?

If we look at the practicalities, how far are the old concepts derived from the archiving and preservation of print translated into the digital environment? In section 5 there is some emphasis on the way in which the principles of selection and preservation can be carried over and should be carried over, because they represent guidance not otherwise to be found. As we will see, the theoretical assertions reported in that section are relevant to selection but much less so to preservation. In the literature a whole new vocabulary has emerged which is different from that relating to the preservation of paper items. There are a lot of differences in the context.

There is, for example, the question of urgency. Books produced using inappropriate paper take time to yellow, crackle and fall apart. Digital entities placing content on platforms that become obsolescent or relying on proprietary software with what turns out to be a short shelf-life can quickly become either irretrievably lost or only available after a highly expensive rescue process. **Feeney** encapsulates much of the discussion as follows:

> "Traditional library and archive materials...and the materials used... present many preservation problems, with which conservators and preservation administrators have been wrestling for years, but as far as paper-based information resources are concerned the solutions are generally well understood. As we move towards ever greater dependence on electronic sources of information, however we encounter preservation problems of completely different order of magnitude and a completely different type".

As we have seen (section 5) and shall see in the practical context, much of the vocabulary remains but the meanings are different. Not only that, and more importantly, there is the matter of new costs, and who pays them. Feeney again in her chapter five summarizes them rather well. In a sense, her summary is too

clear, because in practice, making a commitment to preservation 'in perpetuity' is like writing an open cheque. We really do not know how expensive the commitment will be. I have, myself, been closely involved with the estimates of cost, demanded by government, that are part of the preparations for the Legal Deposit Bill. The documentation provided in the Impact Assessment and in other discussions within the committee 'sponsoring' the Bill demonstrate that there has been much uncertainty about how to provide the statistics asked for.

The role of the Andrew W. Mellon Foundation's e-archiving program and its programme director Don **Waters (2)** will be mentioned again. It is significant. His comments on costs on page 3 are worth quoting here:

> "In addition to flexibility and functionality, e-journals have promised lower costs, but this goal remains elusive. Major journals are rarely published [only] in e-format, and the costs of archiving are unknown. Without trusted electronic archives, it is unlikely that e-journals can substitute for print and serve as the copy of record, and so we have a duplicative and even more costly system — a system we all hope is transitional'.

As to who pays the costs, although archiving is still seen by most to be a job for librarians (see 9.4) there are some who argue that in the digital environment publishers should bear some of the costs involved. **Waters (2)**, writing about the US situation and a particular project concerned with e-journals, concludes that the 'basic value proposition' is as follows (page 8):

> "Publishers would bear the costs of transferring their content in an archivable for to a trusted archive and allow a limited but significant form of access or secondary use as part of the archiving process. Given this form of participation by publishers and universities, e-journal archives would maintain the content over time".

Who pays is important because he who pays the piper, calls the tune. Whatever publishers may think of the proposition by Waters, no-one would deny that there are changed responsibilities.

The Mellon project that Waters refers to is concerned with finding solutions agreed by all stakeholders in "broad interest of the scholarly community", but as we have seen and will see, there are different emphases that divide publishers and librarians in where this interest lies. We will see that most library thinking in this area is conditioned by experience of preparing digital content for archiving and preservation (see 9.3.1). The rather different circumstances that obtain when a national archive of publications is being organized will be examined in the subsequent subsection 9.1.4. Authenticity considerations do not come high on the hypothetical checklists of either intermediary group.

There are new pressures at work. In the print environment, libraries are characteristically pledged to preserve what they are given. Selection raises its head as a concept in the sense that libraries decide what to take in and what eventually (if at all) to discard. In the new environment, libraries might only be able to accept from what is made ready for them and then only to preserve what they regard as the essence of what they have selected. There are all sorts of opportunities for different understandings of authenticity here, and these are

understandings that are not necessarily related to the needs of scholars. Technical convenience may be more important.

9.1.2 The status of e-only communication

It is almost a truism, frequently repeated in the literature about electronic publication and mentioned earlier in this study (**KEY PERSPECTIVES**), that a main reason for the relative lack of acceptance of electronic-only publication as a medium for scholarly communication is the concern by scholars about archiving. Will their articles be preserved for access by future scholars? There is no need to rehearse the evidence here. In parenthesis, it should be reiterated that we are using the word 'article' advisedly. E-journals are where the current e-action is.

Less frequently stressed by some of those who count e-journals and demonstrate an increase in their number, is the fact (perhaps a consequence) that very little scholarly communication is still as yet in an e-only form. Causal relationships are not clear here. Evidence in presentations from individual libraries suggests that in 2003 librarians are beginning to implement at last the ditching of print (**Anderson**), but among traditional publishers there is still no enthusiasm for e-only journals or e-only books. Even within the general range of initiatives aimed to dethrone traditional publishing in science and medicine (create change) there is no concentration on e-only. The strong advocates of open access (see http://www.biomedcentral.com) are balanced by the sponsorship of competitive journals in both media. For the range of SPARC programmes, including the sponsorship of new journals with print versions see http://www.arl.org/sparc/. There is also a general recognition in library and funding circles that forward planning should be based on the concept of the hybrid library (**RSLG**). The strong statements made in the report referred to in the previous sentence provide a salutary counterpoint to some of the sentiments that are expressed below.

In a previous section (6) on digital informational entities, we have examined the theoretical position of Kircz and his group, and in 6.2 the realities of some serious existing e-journals. As we have seen, however, there is very little work on how such journals can be archived (section 9.3). In practice and in a sense rather surprisingly, what work on archiving and preservation of a practical and enduring sort, that has been done so far, has concentrated on e-versions rather than e-only.

Boyce has an admirable record of proposals being implemented and prophecies coming true — as his archive at http://www.aas.org/~pboyce and his publishing plan for the IEEE (**Boyce**) demonstrates. He wrote back on 1st February 2000 as part of a discussion of how to handle electronic versions of print publications on the ListServ (liblicense-l@lists.yale.edu) as follows:

> "The real evolution [in scholarly communication] will be away from simple words or pages on the screen and into interactive information, live math, equations into which a user can plug her date, 3-D visuals with which the user can interact, up to the minute databases of small chunks of information from a whole variety of web sources. This is the world we should be preparing for — and it is a lot harder to understand what to do. But let us not take up too much time describing how to handle a situation, which will not be relevant in five years".

This quotation is relevant here because it demonstrates, explicitly and implicitly, hopes and projections that those involved in what one might call the e-enterprise entertain, and how such hopes and projections tend to conflate. Boyce himself is an astrophysicist. In this discipline there has been a serious move into an e-only environment, but maybe not with all the consequences he suggests. Two questions raised by the quotation are appropriate to consider now. There is the take-up of the functionality on offer and there is the continuance of the print.

For many years the flexibility and functionality that is mentioned by Waters in the previous subsection has been on offer by publishers of paper-based journals to a greater or lesser extent. The history (**Pullinger**) of the SuperJournal project (still after five years a prime source for real user preferences) explains that, during the project, there was a serious attempt to persuade authors to offer papers including non-print elements, but that there was absolutely no take-up of the offer. In retrospect, the reason is obvious. Scholars were not then accustomed to presenting their 'messages' in anything other than print form. There is some evidence now that scholars in certain fields (not just in astrophysics) are beginning to submit papers to journals that do contain dynamic elements (mostly video clips), and they have been linking to databases etc, for some time. The evidence, as far as this author knows, is anecdotal but no doubt surveys are being done. Not all the non-print elements in such articles is regarded as part of the normative or definitive version — the situation is confused — but, even though e-only journals with a serious role in scientific communication may be thin on the ground, normative e-versions are becoming worth serious thinking and serious investment. We shall look at this question again in 9.3.3.

In the previous subsection, the quotation from Waters implies another question already touched on in 5.5.2. The cost of printing and distributing a journal is a cost both publishers and librarians would like to do away with, but there is no agreement on what costs still remain as integral to the publishing process. Scholars, however, obstinately insist on printing out and filing. Recent research (**Boyce** in **Casalini**) demonstrates that the predictions of **Butterworth** do not seem to be, or are not yet, coming good. Young scientists still print out and study off-line, even where their searching behaviour has changed — even in astrophysics. There is no agreement about whether the behaviour is just slower to change than was anticipated, or whether the behaviour will never change. Gurus such as Peter **Mayer** seem to think that the change will occur when the generation, who has always been on-line, becomes active scholars. There is some confusion about what this means in practice, but I think that we cannot test this hypothesis until about 2013. However, there are those who argue that the paperless future will never come about (**Sellen**), for what are basically mechanical reasons. If in practical terms we have to retain PDF alongside XML, there are consequences. If a print form (whether print as traditionally disseminated or print downloaded) is likely to be the one actually used for study, for the absorption of knowledge, what is the status of the principle that underlies much of the argument of this section — that there is a definitive e-version, the authentic version, that has to be archived and preserved? If there has to be an adequate print version that is at least intellectually adequate for transmitting the 'message', what, therefore, is the status of the enhanced and 'fuller' version? Are we in the realm of unnecessary 'bells and whistles'? We will leave this as a question hanging over the whole e-enterprise that we are discussing.

9.1.3 Digitization and born-digital

A study by the Research Libraries Group (**Waters 1**) gives a useful overview of archiving and preservation problems from a library viewpoint:

> "Digital materials for libraries and archives range from simple (e.g. text-based) digital files to complex multimedia and database resources ... For materials that have a physical counterpart, preservation decisions take into account considerations such as the condition of the original materials and the reasons for digitizing (e.g. for increased access to the materials). Materials that are 'born digital can present more challenging problems because their 'being digital' is not only a method of access, it represents their value as an information artefact. For many born-digital resources, effective preservation will rely as much on preservation of its basic intellectual content. More importantly, when a library of archive digitises its own collections, it can control decisions about standards, formats, quality control and documentation".

There are a lot of themes brought up in the above paragraph; a series of statements which seem to me to be full of sense in an area not always marked by clarity of thought. The challenge of 'born-digital' will be dealt with in 9.1.5, and one aspect of the question of 'access' in 9.1.4. In this current subsection, the quotation will be used as a peg on which to hang a short exploration of library attitudes to archiving. The sort of thinking that is made explicit here can have a serious impact on the way some library thinkers view questions of authenticity, or one might say, do not view seriously enough.

It is indeed arguable that librarians have on the whole looked at problems associated with archiving and preservation in the light of digital entities that they themselves have created. These entities can be characterized reasonably accurately as resources. They are digitized both to make them accessible and to preserve them. There are various types of resources, such as manuscript collections and out of print or out of copyright publications that are not directly part of current scholarly communication except to perform a function that is analogous to databanks. They do not on the whole raise any problems of authenticity, at least with respect to the questions that this study is concerned with. Other digitized resources include electronic course-packs. These are created, usually by librarians for academics, out of a range of content, some of which is licensed. An example of this category of resource is the materials produced under the auspices of the Heron project (http://www.heron.ac.uk) now a commercial service. In this case there are significant problems relating to authenticity but (thankfully) we are here in the teaching/learning environment not the environment of scholarly communication. They do not in any case come under the remit of this section because they are not produced with archiving and preservation in mind. Most digitized resources are.

Just because librarians are practically concerned with digitization and its consequences, rather than the preservation of born-digital items (though see the next subsection), it does not of course follow that they think in terms of digitized materials in their approach to questions of archiving and preservation. The literature, however, seems to confirm that they do. The UK **CEDARS (1)** project

is an excellent demonstration of this tendency. The excellent procedures intend to secure digital content for posterity assume digitization as the starting point, which is not say that the painstaking work or the evolved standards is wasted or irrelevant when born-digital content is what needs to be handled. Other important studies of digital preservation from the USA include that of Hedstrom and Montgomery (**Hedstrom 2**)**,** where the concern extends specifically to those born-digital items, for which libraries take responsibility, but is primarily concerned that libraries should co-ordinate their digitization and holding policies. There is also the study by Janet **Geertz**, which, in spite of being titled *Selection Guidelines for Preservation* is entirely concerned with policies for digitization. Geertz, in her survey, notes that some libraries see digitization as a preservation process in the same way as microfilming is (or was).

It is reasonable to project a library viewpoint based on a common concern to preserve resources including scholarly communications needed by their patrons in the future, but there other drivers in play in this environment. The enthusiasm in certain circles a digital scenario that mandates full use of the functionality available has been mentioned earlier and such an enthusiasm can override concerns about preservation. Faculty in a Texas university, encouraged by its library support, is said to have insisted that its students, in preparing dissertations, must include multimedia components without making any provision for the continued access to these components. Unfortunately we have no reference to this claim.

It could also be argued that the availability of a digitized version will lead to the preservation of that version being preferred rather than the preservation of a born-digital version that is authoritative and definitive, but which has been produced without preservation in mind. My own experience within the JISC e-books Working Group (http://jisc.ac.uk) indicates to me that there is a healthy wish within the library community to give access to the best versions available of, for example, a classic work of fiction, even when free versions are more easily available. One can hope that this sort of attitude will generally be evident in other contexts directly relevant to this study.

A contrary indication comes from the movement towards institutional repositories where some (though not all) of those who espouse the movement (**Crow 2**) propose as an argument for such repositories that librarians can point their patrons to a free version of a scholarly article. This preference is because it is free rather than to the version on the publisher's site (which has to be paid for). The word 'version' is used deliberately. There is of course no guarantee that the author will not have altered the definitive version under their sole control. The author of the publication quoted recognizes that this use of repositories should not be emphasised in explaining the advantages of repositories to scholars, therefore, one must assume, recognizing the possibility that such arguments will not necessarily find favour with those most immediately concerned with scholarly communication. There will be more on institutional repositories in the next subsection.

9.1.4 Archiving and access

This group of subsections is concerned with drivers that impact on paying for archiving and preservation. Questions of archiving and preservation, and access

to what has been archived and is to be preserved are tangled. There is a range of initiatives concerned, either directly or indirectly, with making sure that the scholar as a user has continued access to digital content. Some of them are discussed below.

Part of the context here is the change from the purchase of publications by a library in the print environment to the licensing of publications by the library from the publisher in the online environment. The outcome is that the content remains physically located on the site of the publisher. There is an active campaign by libraries to maintain access for their patrons to licensed content by holding it and preserving it themselves. Librarians rightly regard it as part of their professional duty. Much of the concern is with current access or access to recent content, but it could be argued that proposals or projections relating to current access cannot realistically be separated from considerations of long-term access. When does short term become long term? At any rate, the preservation of access is an important driver leading to investment in archiving and preservation. For example, in the UK this is particularly the case with investment by Higher Education Founding Council (HEFCE) through its appropriate arm, the Joint Information Services Committee (JISC) — see http://www.jisc.ac.uk. For them, continued access and (preferably) hosting under their own auspices is an essential part of any licensing agreement on a national basis for the use of digital content.

It is difficult to see how any funding of archiving and preservation can reasonably be sought, unless the material so preserved is accessible by the people for whom it is preserved, the scholars. Yet questions of access are highly contentious. The Council for Library and Information Resources (CLIR) seeks to find common ground between publishers and librarians (http://www.clir.org). To succeed in this quest they have tried to separate the two concerns of long-term archiving and continuing access, with very mixed success. This is partly because publishers have, in general, real problems deciding when the life of a publication from the point of view of commercial exploitation is effectively over.

The so-called 'threshold' moment is difficult to define. It is the understanding of this author that the Mellon financed Harvard project (http://www.diglib.org/preserve/harvardfinal.html), otherwise rather successful, has not been able to find a consensus of when the threshold begins. Another Mellon financed project, which involved as principals Yale University and the leading STM publisher Elsevier, provides one of the best explanations of the publisher approach to archiving (see below 9.1.5). This is in the presentation by Karen Hunter of Elsevier available at http://www.niso.org/presentation/hunter_ppt_01_22_02/. Ms Hunter was one of the two project leaders. The leader from the Yale library was Scott Bennett and she quotes him with an excellent definition of the way many publishers look at access in terms of the journal business life cycle. He writes of an "information half-life, which is the point at which the commercial value of e-journal content to the publisher has declined to the point, where the publisher hands off preservation and access responsibilities to an archiving agent". In the UK, the Joint Committee on Voluntary Deposit did achieve a consensus and felt able to support legislation. At the time of writing, the apparent agreement within the publishing community does seem less secure. There are demands from some quarters, expressed in the debate on the second reading (see the official report

on this debate of 14 March 2003 at http://www.parliament.the-stationery-office.co.uk/pa/cm200202/cmhansard/cm030314/debtext/30314-15.htm), for greater definition and more safeguards against commercial loss. At least no-one in the UK supports the so-called 'dark vault' approach, which does seem at the least to handicap scholarly communication.

The movement towards institutional repositories likewise can be seen in a similar light, but the relationship between the protagonists of such repositories and the certified article is much more complex, and questions of authenticity do come into at least part of the picture in a more obvious way. From within the library community, there have been some pertinent comments on the use of the word 'archive' by the Open Archive Initiative or OAI (**Hirtle**). This is not just a semantic disagreement. Hirtle is concerned with the similarity of this abbreviation and that of the Open Archive Information System — a standard that is central to digital archiving and abbreviated to OAIS (and that is touched on in the next subsection). The OAI is concerned with interoperability protocols. There is little in the literature to demonstrate that there is any serious concern with archiving and preservation in the long term either insofar as the OAI is concerned or in connection with institutional repositories viewed as a subset of OAI. Institutional Repositories have evolved from the preprint/e-print movement, which is where the word 'archive' originally came into the picture. The situation as described is probably beginning to change. The DARE (Digital Academic Repositories), in the Netherlands, not only seeks to make the research results of all Dutch universities digitally accessible, but envisages "long-term storage" at the Koninklijke Bibliotheek (see http://www.surf.nl for further references). We will return to this particular question when looking into submission metadata below in 9.2.2.

As far as questions of authenticity is concerned, there is likewise as yet little discussion, though personal communication between the author and one of the leaders of the movement have revealed to him that there is at least a concern about issues of certification in some quarters. There is also an interesting FAQ on one of the Southampton web-sites of Stefan Harnad and his colleagues at www.eprints.org/self-faq/#2, where the questions indicate the worries of the academic community about the preservation of authenticity, and the site provides some trenchant answers, presumably by Harnad himself. We have here one of these questions of 'trust', which will be examined further in 9.4.

At the moment, it would nevertheless seem that the movement is primarily concerned with setting up 'archives' and only secondarily dealing with other matters relating to them. The remit of institutional repositories is not to publish as such but to facilitate, gather and make accessible any content that faculty might want them to handle (**Crow 1**). There is some simple metadata, indicating the type of content which is available, which is apparent when some sites are accessed, see for example, the Glasgow site referred to in **Nixon**. However, it does not seem that it is felt as part of the job of the repository to indicate what has been refereed, what is not refereed, and more important what has been refereed and has been subsequently altered. This role, concerned with such levels of discrimination, is in the hands of individual authors or, in some cases, with individual faculties who act as gateways. The interesting question is raised: is it always in the interest of academic authors to preserve integrity even though it must be in the interest of the users and the scholarly process as such? We return to this question in 9.4.

It is instructive that Dspace, which has set out a clear mission statement on its web-site (http://www.dspace.org) and which seems to be a model of organization, is starting out mainly as repository of grey literature, and it is not clear how it will handle in practice certified content. Content in a repository will be a mixture. In a sense, this does not matter too much. Scholars using web sources of uncertain authority check them out against the refereed literature. There is, however, a strong 'political context in the movement towards institutional repositories. This has already been touched on elsewhere in this study, especially in 9.1.3. In earlier versions of the scenario now more developed, there was a strong emphasis on building alternative sources of certification, perhaps based on individual universities or perhaps based on learned societies. Such alternative sources of certification are not now so obviously being developed but the movement aims to undermine current arrangements. We will turn to these questions under the headings of trust and responsibility in subsection 9.4.

Nevertheless, if access to what **Berry** has called the Global e-Archive is not made easier, and the arrangements are not put in place to secure access for posterity, the demand for alternative if inadequate access is going to grow. As Berry points out, the sort of access that is postulated by him and his group does not need to be free of charge.

9.1.5 Preserving the national digital heritage

Where do publishers come into the picture? As we have seen throughout this study, publishers as intermediaries for the author community have taken central stage. Publishers represent or should represent scholarly authors and indeed scholarship in the preservation of authenticity in digital entities. In principle, this representation should extend to preservation of the entity itself but here publishers are ambivalent. Publishers do have an interest in short-term archiving, as has already been mentioned, because they want to continue to exploit and re-purpose. The time-scale involved is, however, too short at present for publishers to have to face up to any of those questions of preservation that throw up those questions of authenticity, which we will discuss particularly in 9.3.2.

It has long been recognized that publishers cause problems. As early as 1995 a Canadian study cited by Lunau mentioned some of them that were already apparent in the national drive for a virtual library:

> "A number of issues in selecting, managing and preserving electronic publications were identified. Issues included the lack of standard formats for on-line electronic publishing, which makes it difficult to collect, provide access to and preserve these publications; multiple versions of the same publication; limitations on access created by copyright and licensing agreements; dealing with hypertext links in a document; storing and preserving electronic documents for long-term use when the necessary technology may no longer be available; difficulties training staff; and integrating the processing of electronic publications into existing workflow".

I am not aware of any treatment in the formal or informal literature of any questions relating to authenticity, which might have been brought up in the

digitization of back volumes by those companies and organizations that have undertaken this task — and there are quite a few of such. Not surprisingly, librarians do not trust publishers to archive and preserve, a lack of trust that is examined further in section 9.4. Not surprisingly, in previous subsections we have been concerned with library agendas, where at the best publishers are essentially passive players in terms of the actual archiving and preservation process.

We have already mentioned the work of national libraries and in particular the work of the Joint Committee for Voluntary Deposit (JCVD) in the UK. Because I have been personally involved in the work of the JCVD and that of the committees and reports preceding its institution, where statements are made about attitudes and processes they might not be referenced. This is because they cannot be substantiated in publicly available sources and in some cases because they result from confidential discussion between the committee and stakeholders. There is no site as such for the JCVD, but the best collection of documents relating to its work can be found at http://www.alpsp.org/arc.

In most developed countries the state takes an active interest in preserving the national published heritage in print. The driver is the idea that something important in the national interest will be lost. For example, see a short list provided to members of the UK parliament in a BL press release at http://www.bl.uk/news/letter.html. In many countries there is a growing movement towards legislation aimed at extending legal deposit to what is quaintly called in the UK 'non-print materials'. Of particular interest in this connection are Australia (http://www.nla.gov.au, Canada (http://www.ncl-bnc.ca) and the Nordic countries, as well as the UK. In the Netherlands, where the national library is among the front runners (http://www.kb.nl), voluntary deposit (but taking in almost all serious publications) has always been the chosen route to a national collection. The information available about policies and progress in the acquisition of digital publications is very hard to find from all the sites mentioned including that of the British Library. The exception is the site of the National Library of Australia (NLA). The policies of the NLA will be quoted later in this section. The situation in the USA is somewhat different; the Library of Congress, though now active in the field, is not a national library in the same sense that the British Library is. The archiving and preservation of digital publications is of course only part of the picture. There is a lot of interest in the harvesting of web-sites in most of the countries mentioned.

In the UK, at least some of co-operation to be mentioned below did start with a wish by publishers to work in the national interest. There is indeed another driver that is causing publishers to be a lot more interested in archiving and preservation than they were only a few years ago. Authors do not want to entrust e-only content or e-only components of a digital entity to a publishing process, which does not have any plans for saving what they have put together as their 'message'. Librarians and publishers alike agree that there is a definite pressure.

The main point of this subsection is, however, the following. Legal deposit involves not just a compulsion on publishers to deposit. It also involves a compulsion on libraries to ingest and preserve, and to ingest and preserve moreover digital entities that have not been put together with archiving and preservation in mind. This fact leads to collaboration between librarians and publishers rare in the current environment of scholarly publishing. We have

mentioned the delicately poised agreement over access arrangements, which do not need to be discussed in detail, but looking forward to the workings of the extended legal deposit system in the UK (assuming the Bill is passed) it is clear that to work there will need to be consensus in a whole range of different areas from selection, through deposit metadata and acceptable XML DTDs to decisions involving authenticity in migration of content. There are precedents in the print world of publishers taking into account preservation needs. There is the case of acid-free paper, but the take-up was greatly helped by a relatively small difference of cost. There were also of course no serious theoretical questions involved in a change in paper buying policies and the change did not involve the publications being more or less acceptable in the marketplace — almost always the deciding element for publishers. However, the working together for the benefit of the whole information chain, which seems likely to be involved, is new and of significance outside the margins of this particular study.

9.2 DECISIONS ON WHAT TO ARCHIVE AND PRESERVE

Decisions on why one should archive and preserve digital content are closely related to what content is regarded as appropriate for archiving for reasons, which we shall develop in the first subsection below. The why is more explicitly related to the what than it is in the print environment. There does not have to be a high level of theory involved, or perhaps one should more properly say that there does not seem to be much serious theoretical thought about selection of publications. Harvesting of the web is a different matter. For a librarian developing a practical policy in this area, it is obvious that digitized content is digitized with archiving and preservation in mind (**Jones 2**), as we have seen in 9.1.2. The same goes for content, where arrangements for continued access have been negotiated (9.1.4). General decisions, which we will describe in 9.2.1, will have been made before the decision to invest is made. As we will see, the preservation of the national electronic heritage makes necessary a more difficult decision-making process. After examining the background to general decisions, we will look at submission metadata schemes and how they relate to the purpose of this study (9.2.2), in particular we will consider attitudes (where they are expressed) within the library community to versions of record (9.2.3).

9.2.1 General decisions on what to keep

In this subsection we will tend to write of libraries and librarians as if these institutions and individuals operate independently of their role as intermediaries on behalf of their reader, patron or user community. There is a tendency for publishers to assume some independent action, which must be resisted. Peter Graham of Rutgers University (then) provides a corrective in a collection (still useful) of 1996 (**UKOLN** page 6):

> "Readers' needs will continue to be what they long have been. Readers will want information to be reliably locatable, so that when they do there (whether personally or on the net) they can expect to find what they are looking for. Readers will want information easily accessible: the cataloguing must be clear and accurate, and the information must be promptly retrievable. In the electronic environment the needs for access tools will be more evident, and readers will expect appropriate and standard software to be readily available. *Readers will expect information*

> *that was placed in the library's care a long time ago to be available; and they will expect the integrity of the information they get from the library to be assured."* [my italics]

Selection is more important in the digital environment than it is in the print environment. **Jenkins** quotes a definition of digital preservation presented by Margaret Hedstrom in 1999:

> "The planning, resource allocation and application of preservation methods and technologies necessary to ensure that digital information of continuing value remains accessible and usable".

The concept of 'continuing value' is central. It is not a new idea. The Research Libraries Group (**Waters 1** page 24) points out:

> "Selection processes for archives of all kinds — paper and digital — are matters of intellectual judgement about what to include and what to exclude. Criteria for such judgements are largely tied to the intrinsic qualities of the material and many of the criteria that have proven useful in the paper world will no doubt translate to and prove equally effective in the digital environment".

The investment in archiving and preservation is more obvious than it is in the print environment (see Mirjam Foot of the British Library in **UKOLN** page 29). Books and journals are put on shelves, the latter (in the past but not always now) after binding, and the decisions involved are concerned with disposal rather than keeping. The difference between how one makes a decision in the different environments of print and digital can of course be exaggerated. There will be disposals in the digital environment as was perceived in the excellent report of 1996 from the Research Libraries Group previously mentioned (**Waters 1*):***

> "Selection for digital archives must be a continuing process. Given the need to migrate digital information regularly from its hardware and software environment, the stimulus and occasion will recur to reappraise the value of the material being migrated".

The apparent need to build an extension to house more printed materials does lead to decisions relating to acquisition. For example, it can provide a stimulus to stop subscriptions to print serials as it has done recently at two British universities (Manchester and Leicester). Nevertheless, the bottom line for digital preservation is, more starkly, that it is not worth preserving information that it not worthwhile.

We have again to be clear that in this study we are concerned with only one part of the mission of the archiving and preservation of digital entities. This is not to suggest that the larger mission is irrelevant. **Waters (2)** already quoted extensively states that "we need a serious investment in archiving because we are in danger of losing our cultural memory". Waters lumps together schemes for harvesting the web with the preservation of publications, which in this study we have taken to be what is central to the transmission of scholarship. In this, most librarians follow him. In print some of the huge quantity of printed material of little or no contemporary scholarly value (ephemera for example) survives by chance and provides a quarry for cultural historians of the future. Without

intervention, nothing like this will survive in the digital environment because it will, even if it is not wiped from the record, become impossible to read.

For most libraries selection is primarily a matter of deciding whether the item selected fits in with collection policies. Weinberger writes in what is basically a primer on handling born-digital items **(Weinberger)** that:

> "The digital objects should be approved by the subject specialists, who also understand the scale of costs involved in preserving these objects".

It is notable that costs often play a part in selection policies in a way that, on the whole, they do not overtly play in decisions about print. **Gatenby** makes the same point in describing the policies of the National Library of Australia.

Even in print national libraries do not collect everything they are entitled to receive free. We have mentioned ephemera (not characterised as publications) above, but the British Library (BL), to give a good example, does not collect all print publications. It "practices selectivity". There is a whole list of published items not collected by the BL, including local transport timetables and wall charts, and, more relevant to what follows, new editions in paperback — see for example the review produced by the BL in 1997 (**British Library 1**). In the scenarios presented to government, as a preparation for legislation to extend legal deposit to non-print materials, these principles of selectivity and prioritization are applied in the digital environment. The favoured scenario assumes a collection strategy that involves "initial incremental progress to reach capture of 90 per cent of unique e-monographs, 80 per cent of unique e-serials titles, and 60 per cent of other e-serials by 2005" (**EPS 4**).

The central question for our purposes is whether the decision to classify a digital publication as one of 'continuing' value contains within it a recognition that the authenticity of this publication has to be determined as part of the selection. Or, to put it another way, is a digital publication chosen for preservation only when it can be sure that it is an authentic digital publication. From conversations with senior management at one major national library (BL) it was indicative to me that at present, the library essentially accepts as authentic whatever the publisher delivers to the library. There is little doubt that this is an area where a policy is needed across the world of archiving and preservation, and no doubt such a policy is currently being worked on.

Most of those questions raised earlier in this study about the nature of authenticity are raised when considering selection of digital entities. There is a useful entry on Authenticity in the extended glossary provided by the National Library of Australia on its PADI site. PADI stands for Preserving Access to Digital Information. The whole entry is relevant (http://www.nla.gov.au/padi/topics/4.html) but the following quotation particularly:

> "The authenticity of a digital object refers to the degree of confidence a user can have that the object can is that the same as that expected based on a prior reference".

If the prior reference is to a publication should one go to the publisher (or its sanctioned intermediary) or can one go elsewhere? The BL, and no doubt other national libraries, are encouraged, for example in the most recent report (**RSLG**) from library sources to favour what are often referred to as emerging models for digital publishing. It is relevant to refer back here to the short discussion of institutional repositories in 9.1.3. When a library, particularly the national library with a special responsibility for archiving and preservation, needs to select and there are several apparent sources for the same entity, where do they turn and how to they decide? It seems to me that there is no difficulty with other emerging or non-traditional models where a publisher is involved. The Open Access pioneer BioMed Central (http://www.biomedcentral.com) has strong and clear policies relating to authenticity (though not expressed as such) and have indeed established an agreement with the BL for archiving its material in the short term. In practice, it seems likely that librarians and publishers share a belief that what publishers sell is an authoritative version. It is a matter of provenance. We will examine further in 9.4 the supposition that whereas the authentic document can be obtained from the publisher, the publisher cannot be counted on to maintain it.

There is a related problem. Can the selector be sure that there has not been corruption of the text, a loss of integrity, before the digital entity is deposited? The problem of possible alteration before deposit is surely not of major consequence with formal publications, although the Vanishing Archive experienced (discussed further in 9.4.3) does give pause for thought, but is much more relevant to self-publication. The thinking of Lynch on canonicalization is, as usual, highly relevant (see section 3.3.3) but, equally as usual, it is not clear how far it has impacted on the way in which librarians are groping towards solutions. The mechanics of the preservation of intellectual integrity are discussed in 9.3.4. There are obviously some relationships with the considerations of protection of digital objects covered in section 8.

The BL is in theory not concerned with digital versions of print in developing its archiving and preservation policy. The decision was made at an early stage in the work of the JCVD (see 9.1.5) but there could be pressure to change the policy. The sort of developments sketched in 9.1.4 means pressure to preserve hosted resources that are currently to a large extent available in print. It is with publications of this type that serious interaction between publishers and national libraries seems to start. The activities of the Koninklijke Bibliotheek in the Netherlands are particularly significant. After some years of experimentation and negotiation, they have become the official archive for leading Dutch publishers of scientific and medical literature (see http://www.kb.nl/resources/frameset_kb.html?/kb/pr/pers/pers2002/elsevier-en.html for the press release, which is quite informative). It is the journals of Elsevier Science that the Royal Dutch Library is particularly concerned with and the electronic versions of few Elsevier Science journals differ from print except insofar as they enable linking.

This is rather important and represents another problem for selection. The press release just cited indicates that Elsevier are depositing the journals in both a PDF version and a structured version, soon no doubt to become an XML version. We will discuss 'versions' further below. It is not just a matter of two electronic formats. We have seen how the fact of two e-versions (at least) was understood

as a problem in 5.5 as long as two years ago. It is a matter also of the publishing platform or platforms. The Code of Practice (**JCVD**), relating admittedly to offline publications, states that they "should be deposited in the form in which they are made available to the public". However, a working group on online deposit within the context of the JCVD has established that for online publications, such a simple rule does not work.

Finally selection by a national library or indeed any library has to be determined to a large extent not by the continuing value of the publication and not by its provenance but by its nature. There are a number of reasons why a submitted digital entity cannot be preserved. Only some will be listed here.

As **Weinberger (2)** points out, the library faced with a born-digital publication will need to be sure that the licence to the software that enables the entity covers the purposes of preservation. The use of software is regarded as sufficiently important to be covered by one of the short Legal Deposit Bill in the UK, mentioned elsewhere in this section. Software worries a lot of CD Rom publishers, but it is not so much a concern to publishers of online content. We are not going to unpick the rather general term 'software' which, in the thinking of the Bill, relates primarily to the CD Rom and can be for searching purposes and much as presentation. In the online environment, there has been little investigation of the problems downstream with embedded software applications, which need to be picked up at ingestion. On the whole, these aspects of the complete product do not seem to be identified in as much detail as one would assume is necessary, but the picture is not clear. For example, in the **CEDARS (1)** Metadata there is a range of elements concerning "Change History Before Archiving" (1.1.3.1.4), Original Technical Environments (1.1.3.14), Prerequisites (1.1.3.1.4.1), Procedures (1.1.3.1.4.2) and Documentation (1.1.3.1.4.3), but will they pick up some of the meta-textual information picked out as problematic by Francis Cave (section 5 of **Bide 2)**? Cave's presentation to the seminar, reported on in this publication, contains the following warning:

> "Publishers manage much of their content in proprietary technological environments and in proprietary formats. These are hidden from users by middleware (which for example may create HTML 'on the fly' from content held in proprietary databases. Even where neutral content formats are adopted (SGML, XML) their effectiveness for long-term preservation will depend to some extent on the quality and availability of the documentation of their DTDs (which is frequently deficient today). This will present substantial difficulties for any third party attempting to manage different versions of a publisher's DTD".

Not all these illustrations impact on the authenticity of the 'message' of the author by encouraging distortions further on in the preservation process. Publishers do however need to consider the interests of their authors as well as the demands of the market (when they go for some bell or whistle). Librarians have to organize themselves to get all necessary information from the publisher at deposit and in the construction of suitably detailed metadata so that the right questions are asked. The attitude to proprietary formats, for example even PDF, within library circles can be a handicap but **CEDARS (2)** does propose a solution in open source rendering tools, which may enable a way of dealing with undocumented formats (3.5.2)

There is a useful taxonomy (dealing mostly with offline publications) in an appendix to the report that preceded the JCVD in the UK, the so-called **Kenny** Report. I was one of the authors of the taxonomy and I consider that it holds up well. There is also a hierarchy of digital objects produced independently by Lynch (see 3.3.2). It lists the different types of digital entity in what is in practice a range from possible to preserve to impossible to preserve. The words 'static' and 'dynamic' are often used, but can be misleading. An electronic file with a video or audio item clipped on can be preserved as it is, but a database that changes hourly as new information is added can only be harvested.

The word 'database' gives away the fact that most truly 'dynamic' entities are not part of primary scholarly communications, as we have known it in the past. Most scholarly messages are not 'dynamic' in themselves. Nevertheless, as we look ahead we have to take into account the prophecies of Boyce reported at 9.1.2. Essential reference to outside databases promises real problems of capture, but in many fields this is the way science works and the pressure (it is assumed) must be for formal presentation of their findings to reflect the way this work is actually done. Fortunately, it is now fashionable to think about downstream consequences of what is called e-Science (see the appendix to **RSLG**), which may well impact on ideas of archiving and preservation and (hopefully) on the maintenance of authenticity. Even when we examine the potential problems thoroughly, there is some hope for the future. Kircz (whose views are described in a previous section) described problems with different media, which are theoretically possible to solve but which present immense practical problems now.

For the moment it is not just web-sites but 'publications', as defined for this study, which must be handled in a way that can only be described as less than ideal. The IFLA Guidelines set out problem and solution in a way that is basically legalistic (**Lariviere**):

> "While some suggest that depository libraries should not be collecting dynamic electronic publications because their permanent updating implies they are not meant to be preserved, others say that it is the responsibility of a national deposit library to collect, preserve and make available the cultural and intellectual heritage of a country no matter how it is expressed. While it is almost impossible to keep a permanent deposit copy up-to-date unless a publisher agrees to maintain two parallel versions, the legislation could require that the publisher send a 'snapshot' of its dynamic publication on a regular basis as fixed by the law" (chapter 6).

The usefulness, not to mention the representational qualities, of such a sample is dubious if the concern is the transmission of knowledge rather than the preparation of a quarry for future cultural historians.

To be fair, **Lariviere** does make his recommendations clearer later in the chapter:

> "What should be deposited are the separate and complete 'intellectual units that are stored either separately or as part of a database. Whenever a database is made up of separate and complete units — such as a legal database that includes cases, journal articles, etc. — it should be an

> object of deposit. But when a database is made up of raw data (i.e. unorganized data that could be selected and gathered on order by an individual to create a separate and complete 'intellectual unit for his/her own private use), it should not be subject to legal deposit. While there is a need to preserve those raw data, it is not within the normal mandate of a national legal deposit institution to be responsible for collecting and preserving them. But that same institution could play a leadership role in convincing governments that such valuable information and/or material should be preserved for future generations. As Mackenzie Owen and Walle write, "The conclusion to be drawn is that publications which cannot be acquired as an independent, self-contained and coherent entity (in general documents, which cannot be downloaded from the network but only accessed) should not be selected for deposit. Providing access to such documents is not a task for the deposit function"."

Of course any 'message' that depends on any sort of linking presents problems. An early but important study by the National Library of **Canada** came to the considered conclusion that links would just have to be chopped off. However, work on identifiers points to some solutions. For some further comments on identifiers see subsection 9.2.3.

In fact there will not be for the moment much effort to solve these practical problems among libraries attempting to archive and preserve. We can safely say that in their selection procedures, libraries with a mission to archive and preserve will have to stick to those digital formats closest to print. It may be that the authentic message in its full integrity will not be chosen for archiving because it is too difficult. Authors will have to be self-limiting if they want to be available to posterity.

9.2.2 Submission metadata and what it covers.

Bide (2) and his colleagues, rather hopefully described the situation and prognosis in 1999:

> "Discussions are required to decide *what* metadata is required at the point of accession and both the format and the protocols by which it might be delivered. To these deliberations, the work of the <indecs> project, the International DOI Foundation, CEDARS and BIBLINK will be significant contributors. We recommend that deposit libraries begin work with publishers as soon as possible in developing appropriate metadata schemes for deposited content".

Bide (in his next sentence) linked these schemes directly to the definition of conditions of access. This is, as we have seen in 9.1.4, reasonable insofar as it goes, because the wish to maintain access is such an important driver for librarians. Nevertheless, as the list of organizations mentioned above also shows, the needs of the scholarly community could well be seem implicitly as taking second place to the needs of the intermediary functions within the information chain.

The work of Bide and his associates (**Bides 2** and **3**) was commissioned specifically to feed into the development of standards by the Joint Committee for Voluntary Deposit (JCVD), which represented just such a collaboration as he proposed. However, the documents connected with deposit are very minimalist in nature (**JCVD 1**, **2** and **3**). The guiding principle was to get publishers on board and, in spite of the meagre amount of metadata required, there was a limited amount of cooperation. We will return to cooperation by publishers in subsection 9.4. It is a little unfair on the BL to report their thinking in this rather constraining context. **Spivey**, recording the experience of the management and library science publisher Emerald explains how they actually work with publishers.

Bide does not mention the Open Archival Information System (OAIS) in this quotation, and it is not surprising. However, OAIS does almost always get mentioned in connection with archiving and preservation. There is a short but helpful overview of ISO standards in archiving at http://ssdoo.gsfc.nasa.gov/nost/isoas/. The procedures set out at length in **OAIS (1) —** the so-called Blue Book — are derived from detailed work on the preservation of space data. This is in itself not a problem. The World Wide Web itself emerged, as is frequently pointed out, from the world of high energy physicists, that is the world of science, but in practice OAIS does not seem to this reader to have a lot to offer to the questions concerned with in this study. There is actually a study on ingestion arrangements, in principle relevant from submission metadata schemes (**OAIS 2**). This so-called Red Book is not really relevant. It is interesting that the producer-archive relationship (the analogy one might assume of the publisher-library relationship) makes it very clear that it is the job of the producer to fit in with the structure of the archive, which is probably very reasonable in this particular sector.

We have mentioned in 9.1.4 that the Open Archives Initiative (OAI) is not to be confused with OAIS. We have also discussed OAI in 7.3.4. but it is worth underlining the fact that OAI is all about access through harvesting. All that the relevant metadata section (**OAI 2**) tells us is the date of the last modification (datestamp 2.4), but there is no attempt to tell/ask us anything about the modifier. In any case, in repositories such as the Glasgow University one, what is being submitted and accepted are individual articles from individual scholars, which can be characterized as from refereed journals. There is an excellent review by **Day** of the problems for e-print services and 'long-term access' (already quotes in section 5).

What can/do libraries attempt when taking responsibility for a whole publication, perhaps a serial, covering the work of many different authors? We have already looked at metadata schemes in section 7 and have examined them to see how far they can be concerned with questions of authenticity. How far do librarians aiming to preserve aim for anything more than the OAI in this regard? The answer is not much, at least at present.

The **OCLC**/RLG report on "preservation metadata" is a thorough investigation of the state of play (in 2002) of the subject. This report asserts that OAIS is the 'de facto standard' (page 4). OAIS is a complex construct and the report forms a useful guide to its complexities and, because of the status of OAIS, it is worth looking at what is said about submission metadata and where questions of authenticity come in.

Central to this is the concept of an 'information object'. In the report (page 6):

> "An 'information object' is defined as a Data Object combined with Representative Information. In a digital environment, this implies a sequence of bits, combined with all data necessary to make a bit stream viewable and understandable".

This seems to be the digital information entity of section 6 accompanied by such metadata as is necessary for preservation.

Reception or submission of this is object in such a way that there can be successful ingestion requires an Archive Information Package (AIP) that comprises four items. One of these is the Preservation Description Information (PID) element. The components of the PID are as follows.

The first component is Reference Information, the information that is needed for secure identification "such that it can be referred to unambiguously". The second is 'Provenance Information' which documents the history of the 'content information'. Thirdly, there is Context Information. Context Information documents relationships, which might include the reason why the object was created and its relationship with other objects. Finally, there is Fixity Information, which sets out authentication mechanisms like digital signatures. I do not have much faith or perhaps rather hope in authentication mechanisms for reasons expressed elsewhere in this study. They are too expensive for the originator or publisher to adopt. One can see all sorts of correspondences with the <indecs> schema investigated in the previous section.

There is plenty of scope for bringing considerations of authenticity within the terms described in the previous paragraph. Nevertheless, as is the case with digital rights management, the intention is clear and the devisers (and presumably adopters) have purposes which do not seem to include such concepts as whether we are examining a definitive version or not. The report is clear (page 35):

> "It is important to note that informational requirements associated with managing the preservation process not those aimed at facilitating understanding and interpretation of the intellectual content".

Naturally, the purpose for which the metadata is created is very important. You cannot afford to gather everything, but when you consider the sort of decisions to be explored in 9.3.2 relating to the 'essence' of a piece of scholarship the usefulness of this approach becomes a little less valid.

It should also be noted that OAIS, like <indecs> is one of those schema, which are characteristically implemented in a cut-down form. What remains is significant. CEDARS submission data is a slimmed down version. **Jenkins** points out that the provenance data (as slimmed down) yet requires all the information necessary for rights management. **Jenkins** (page 9) draws attention to the fact that the NEDLIB schema has an even smaller dataset with a concentration on problems of technical obsolescence.

There is a strong connection between what metadata is delivered or acquired with the informational object and those decisions required in migration or emulation. This is, as usual, well understood by **Lynch (2)**:

> "Key processes involved in the management of digital objects over time include the tracking of authenticity as part of provenance, maintaining the integrity of the digital object and ensuring the referential integrity of links to that object (from other objects or from metadata records) and understanding how reformatting damages the integrity of the object. These involve both the digital object and the metadata".

It seems likely that submission metadata has to be treated very seriously. You cannot go back to the creator or publisher later on when you need to be sure about what you have got, what its 'essence' is, because you need to migrate the entity. I believe, in the context of this study, that it can be argued that in the current environment there has been too much emphasis on procedures and not enough on the intellectual status of the content under consideration.

9.2.3 Versions and identifiers in archiving and preservation

Thinking about versions is not new to the planning for the archiving and preservation of digital entities. Back in 1997 the work package for the **BIBLINK** project set out the need and the problem (section 2.1):

> "While traditional publishers have developed an elaborated practice for revisions, reprints and new editions of print publication – which is reflected in the elaborated cataloguing riles for the recording the edition statement of a print publication – this practice has yet to evolve in the (networked) electronic environment. Publishers have underscored the problems of version control and the need to record content changes of dynamic online publications. There is however no "good practice" yet and this is reflected in the lack of any edition statement in metadata formats like the Dublin Core Metadata Set. It is therefore recommended that an authentication technique should be used for version control — awaiting for publishers and metadata creators to develop appropriate ways to control and record versioning practices in the electronic environment".

Two years earlier, the Canadian pilot project (**Canada**) uses the term, but clearly with a different definition in mind:

> "It is widely believed that there will be more versions of publications in the electronic environment than in print. Versions are documents with identical, or nearly identical content, but different physical forms" (page 18).

This quotation is concerned with choice of physical versions. The Canadian pilot chose a particular physical format by preference. However, our concern is with the word 'nearly' and with the intrinsic differences between (say) a 'manifestation' that is a print equivalent and one that has additional material made possible by online functionalities. We have mentioned the question of format chosen, when considering selection, and elsewhere have considered whether or not a PDF format (designed for printing) has any (or much) value from an archiving point of

view unless it represents the complete 'message'. PDF tends to be undervalued by archival theorists and suffers too, in some quarters, because it is not strictly an open format. We shall not explore these sort of different versions here. As to the other types of version differences, where the differences are those of content, there have been a lot of discussions of such distinctions in different contexts and in different sections of this study. Characteristically, the National Library of Australia in its Guidelines for Selection (**Pandora**) does provide tantalising evidence of thought about these questions:

> "The National Library will not attempt to preserve all versions/editions of a selected online title, just as we do not attempt to preserve all stages of a print loose leaf item. In the online environment, publications can and often do change frequently and it is not feasible to capture all instances of change" (3.5).

This quotation is tantalising, because almost certainly the policy is only concerned with "snapshots" of "dynamic" or "cumulative" (**Bide (2)'s** definition) and not with the questions raised in the section on the definition of a publication.

In this current discussion of archiving and preservation, there is not a lot more to say, mainly because, in spite of these two early quotations, the approach of those concerned with selection and submission/ingestion have shown little interest in versions.

Rothenberg (1) is an exception, but it is very unlikely that his ideas have filtered through into the various protocols and schema. He has a long subsection on strategies for defining authenticity, but when his thinking (pages 5-8 in particular) is examined closely, it seems to me to have very little to do in practice with the concerns of much of this study. In any case, it is his contention that the strategy for archiving he advocates "makes the details of how we define authenticity all but irrelevant from the perspective of preservation". A lot of his "strategies" seem to concern "provenance" (in its broadest sense), or, using his vocabulary "custodianship", which we can refer to again in section 9.4.

It is also not surprising that a pamphlet (**Meyers**), directed to an audience of scholarly publishers, as well as suggesting as a litmus test of what is worth archiving whether the work "contributes to knowledge development" cuts through much of the debate we have earlier documented with these pithy sentences:

> "Not everything warrants archiving. For example, a work may be a transitory one that is being replaced by a more complete version. If so, archiving the first version may have little value and providing access to it may actually be confusion. (Witness the multiple versions of articles that authors leave on the Web.)".

The Canadian study also touches of another related question:

> "The incorporation of standard identifier information in e-publications would facilitate the finding and retrieval of e-publications by the intended audience". (**Canada** page 16).

Identifiers are touched on in the library literature that relates to archiving and preservation. Unfortunately, the topic is usually handled in outline only and gingerly, probably because of concerns that still exist about the digital object identifier (DOI), a system devised by publishers There is an important proposal in **Bide (1)**:

> "We would strongly recommend that careful thought be given to the ways in which unique identifiers are used at the point of deposit, particularly to the extent that they may be used in the future as finding and location aids. The long-term value of the deposit archive will depend, to some extent at least, on the approach taken to identification".

As far as I know, this recommendation has not been taken up seriously as yet. However, it is finessed by **Paskin (2)** in a typically far-sighted article that deals with the question of digital identifiers in the context of digital preservation of the record of science. He points out that:

> "As yet (1999), the DOI system has not investigated or implemented a Repository approach to any of its applications, though it seems certain that there could be some useful application to be developed here; DOI has brought the indecs approach to metadata to digital object management, and the prospect of integrating this with the full digital object architecture is attractive".

In this short article Paskin throws out some hints on how it can be determined "what (precisely) is going to be preserved". He sees 'interoperability' as the connection between preservation and DOI work:

> "Interoperability in the face of *legitimate change* [my italics] has been the theme of the DOI work. The problem of preservation is when the dimension of change is time: "how do we interoperate with the future".

The phrase 'legitimate change' is italicized because of its relevance to the next subsection. As far as we know, these ideas have not yet been followed up, or if followed up in identifier circles, they have not yet found their place among those concerned with archiving and preservation. Their obvious relevance to questions of authenticity cannot be exaggerated.

9.3 DECISIONS ON HOW TO ARCHIVE AND PRESERVE

In this section 9, we have separated the consideration of submission metadata from metadata that is required for preservation, because it seemed necessary to do so in order to control the material and also for reasons explained in the next paragraph. However, those responsible for preservation, however it is done, can only work with the information they have received and in some of the discussions below, for example over canonicalization and over document type definitions, we are looking at the whole archival process. There is also the problem, which in practice we have to ignore, that "the situation is complicated by the perception that different kinds of metadata will be required to support different digital preservation strategies or digital information types" (**Jenkins** page 4).

In the previous subsection, we have been concerned primarily with one aspect of the search for authenticity, the determination of the authentic document. Sometimes this has been a matter of the authentic document or authoritative document as compared with similar but less authentic versions. Almost nowhere (as we have seen) is the contrast explicit, but any sort of selection demands such a determination and selection is necessary on economic grounds as well as for the needs of scholarship. Sometimes we have been (more explicitly) concerned with making sure that the preserving body receives what information is necessary for subsequent (downstream) preservation.

In this subsection we return again to integrity, and how we preserve it. **Bide (2)** points that:

"It becomes essential to disentangle questions relating to the preservation of *products* from those relating to the preservation of *content.* Different strategies are required for each". [Bide's italics].

We have chosen for the next subsection a title that names the two main strategies for preservation; but fairly recently it has become clear that the situation is both more complex and more encouraging. It is not just a battle between two schools of thought in spite of some of the writing, e.g. **Bearman (3)**. Of course almost all of this discussion is for the future. Publications are being ingested but in most cases there has not been much need as yet to refresh them. In the following two subsections we shall pick out two aspects of the process that seem specifically relevant to the theme of the study. Finally we shall in 9.3.4 touch on the mechanics

9.3.1 Different strategies for preservation: is it emulation or migration or both?

I have as been told that, in the first place, emulation and migration as strategies in the practical environment are coming together and secondly, from within the BL, that some types of entity are more appropriate for one approach to preservation and some are more appropriate for another. A similar perception can be found in the Preservation and Access International Newsletter No 10 at http://clir.org/pubs/pain/pain10.html and here it is specifically linked to the need to define authenticity in preservation.

The key text for what follows is the CLIR proceedings published in 2002, which demonstrates that this understanding is partly true but much more complicated. These proceedings are mainly concerned with establishing a theoretical base, or perhaps establishing (successfully) that there is a theoretical base. There is a lot of practical work going on. For example, there is a useful summary in ICSTI, but this is now outdated. No doubt there is a new conspectus of all the work going on but the author does not know of it. One hopes that the 'practical' work is sustainable and will lead to procedures being embedded in the daily life of the libraries concerned. One of the problems of the sort of work funded by such important bodies of the Mellon Foundation is that, when the funding runs out, so the work finishes. This is not as bad as it sounds, because a lot has been learnt but we badly need operating archives. One advantage of the ponderous move forward in national libraries is that there is little danger of the work being stopped. The problems have to be addressed. As Bide (1) points out:

> "We cannot back away from addressing these questions for the inevitable corollary would be that users of the future would not be able to access the resources of today *at all.*" (page 15, Bide's italics).

Granger makes a similar point, writing in late 2000:

> "These problems have, of course, been exercising the library and information communities for some time, but as yet no one solution or set of solutions has been reached. Solutions have to be found urgently if we are not to sink into what Rothenberg calls 'technological quicksand'.

There are positives and negatives about the increasing cooperation between librarians and computer scientists in this area. Librarians have to be practical but computer scientists do not. Libraries have an explicit mission to preserve authenticity.

Thibodeau describes the theoretical base, already mentioned, in his important and lengthy essay in **CLIR** (chapter one). His title is *Overview of Technological Approaches to Digital Preservation and Challenges in Coming Years,* which in itself provides a framework. He writes:

> "Discussions of digital preservation over the last several years have focused on two techniques: emulation and migration. Emulation strives to maintain the ability to execute the software needed to process data stored in its 'original' encodings, whereas migration changes the encodings over time so that we can access the preserved objects using state-of-the-art software in the future. Taking a broader perspective, IT and computer science are offering an increasing variety of methods that might be useful for long-term preservation. These possibilities do not fit nicely into the simple bifurcation of emulation versus migration".

Our concern is obviously not with how these preservation methods work, but how the workings of these preservation methods preserve the integrity of the digital objects preserved.

Thibodeau provides further help:

> "Four criteria apply in all cases: any method chosen for preservation must be feasible, sustainable, practicable, and appropriate. *Feasibility* requires hardware and software capable of implementing the method. *Sustainability* means either that the method can be applied indefinitely into the future or that there are credible grounds for asserting that another path will offer a logical sequel to the method, should it cease being sustainable. The sustainability of any given method has internal and external components: internally, the method must be immune or isolated from the effects of technological obsolescence; externally, it must be capable of interfacing with other methods, such as for discovery and delivery, which will continue to change. *Practicality* requires that implementation be within reasonable limits of difficulty and expense. *Appropriateness* depends on the types of objects to be preserved and on the specific objectives of preservation".

The first three criteria echo the comments made earlier in this subsection. Our concern is with appropriateness. The process of preservation involves loss whether the digital object is migrated or emulated. **Bearman (3)** writes:

> "Rothenberg's abstract argument, that translation always involves loss of information, is plausible, but not, as he imagines, very relevant. If it was true, his own case for emulation, which depends on a much more complex translation than that envisioned by those who would move each generation of records forward incrementally, would be fatally flawed".

Bearman, in this polemical piece, is writing about 'records' under the control of the librarian or archivist, and therefore, it seems likely that there is probably more 'loss' in the situation of scholarly communication.

We have to accept the idea of 'loss' which has added to the language a variety of forms concerned with qualification of the absolute. This is essentially a loss in interpretation. **Gilheany** has written:

> "We know we can preserve the bits. We must guess that we, or some other interpreters, will be able to make use of them" (page 16).

A decision has to be made what loss can be afforded if the authenticity of the object is to be maintained. Such a decision depends on the purpose of the object, or 'message' as we might say in the context of this study, and also the use. We will examine this further in the next subsection.

In this subsection it is worth drawing out a number of relevant points from the articles already mentioned. The statement of Bearman, quoted above, draws attention to the fact that in practical scenarios the migration or alteration of the digital entity will not be a one-off affair. Technology will change and with it accessibility and every generation or so one has to envisage a further 'translation' (to use his neutral term), which each time will bring about the possibility of further loss of authenticity. It is therefore important that the nature of authenticity is established on a consensus basis, acceptable to authors and readers as well as publishers and librarians. One can also envisage new procedures to deal with updatings and corrections becoming established. The procedures we currently have in place are not designed for the digital environment but reflect some very important principles (see 9.4.3).

Secondly, Thibodeau writes:

> "The variety and complexity of digital information objects engender a basic criterion for evaluating possible digital preservation methods, namely, they must address this variety and complexity. Does that necessarily mean that we must preserve the variety and complexity? It is tempting to respond that the variety and complexity must indeed be preserved because if we change the characteristics of digital objects we are obviously not preserving them. However, that response is simplistic".

We have something of a stake in the preservation of digital entities, whether or not they fit in with the demands of some imposed procedural solution to the problems of preservation, so there could be warning signs here. However, Thibodeau provides an example of what he means, which seems to show that

what is at stake is the question of significance, of the 'essence' we will discuss below.

In this subsection we will not return again to examine metadata. The developments in metadata sets, which are ongoing, seem essentially to be a matter of implementing the OAIS framework. We have already pointed out that the composition of submission metadata, where most of the work has reasonably been done, has downstream consequences. Preservation metadata (in the sense of metadata concerned with the management of preservation) cannot be separated from and is part of package with submission metadata. It comes along with the digital entity. This was understood early on in the 1996 report of CPA and RLG (**Waters 1** page 15):

> "A concern for provenance in preserving the integrity of digital information means that digital archives must document what happens to the information within their own organizations. They must keep a record of migration activity, and particularly of the transformations they make to keep information objects current with new technologies for use. Only be tracing such migratory activity, can digital archives establish its own chain of custody back to the original object".

In this section, and elsewhere in this study, we have quoted a number of baseline documents, where the assertion of principles does not have to be qualified by the pressure of practical grappling with the problems mentioned or rather referred to above. We do apologise for this. Anyone concerned with questions relating to authenticity comes to realize that only by constantly returning to first principles can we avoid the danger of such first principles being practically ignored or only partly observed.

9.3.2 What is the essence?

In an important article, already quoted in 3.3.3 **Lynch (2)**, in discussing methods of preservation, writes:

> "There are continual, imprecisely framed questions about how the capabilities of representations (formats) interact with the meaning of an object, and how translations from one format to another may alter that meaning and thus damage the object's integrity".

This observation (criticism) is a reasonable one, but it is probably inevitable when the considerations, properly worked out by those authors referred to in section 3, impact on the task of librarians in actually devising practical systems. The relationship between sections 3 and 9 in this study is not an exact one of theory and application (alas). In part it is because the community (including me) which is concerned with knowledge transmission has real problems with the theoretical computer science community, who are providing this philosophical framework, and partly because the philosophical concepts are irrelevant to the task we are concerned with. Lynch's work is an exception.

To the layman, canonicalization, as described at Lynch in the article quoted, as way of capturing and making possible the translation in time of the 'essence' is, like the emulation proposals of Rothenberg, difficult to understand in terms of the practical illustration produced in the article. Lynch writes of capturing the

essential characteristics of a type of object, while recognizing that it will be an idealized form of the object. This will be an idealized form of this type of object. Specific representations of the object may be richer. As we are concerned with specific projects, each scholarly message is specific, is there a problem here or is what Lynch proposes a starting point rather than an end in itself. Lynch does try out what he admits is a relatively simple case study, on image data. It seems to work. However he admits that canonicalization for other types of digital object are more of a problem, because they have less formal models, but he proposes lines of research. Among the advantages of this approach is that it makes precise what is important about a class of objects. Do we need therefore to identify a class of object of which a digital entity is a member?

There are a lot of questions to follow up but the questions, from the point of view of this study are the right ones. Is there follow-up research? It is not at all clear that there is. A Google search reveals that the 1999 paper has been listed in every sort of relevant bibliography, and no doubt files (in print or in digital form), but there is no evidence, from that source, of any real interaction. When we think of the potential (and in some cases actual) richness of the digital entity as described or revealed by Kircz, the amount of work in different media with different characteristics and different sophistication of standard regimes, it is obvious what a big task lies ahead. From the point of view of this study, the important element of such work as that of Lynch lies in the fact that he starts with the concept of authenticity at the centre of his thinking.

At a different level of conceptualization there is an important trend in some of the thinking about preservation, which needs to be countered. As we have seen, the work of Rothenberg has met with a mixed reaction, but it would be a pity if some of his insights were rejected along with the theory of emulation. Rothenberg (1) in his section on deriving authenticity principles from expected ranges of use does treat seriously the concept of 'look and feel'. He places 'look and feel' in a sequence of "decreasingly stringent principles ... in terms of the relationship between a preserved digital informational entity and its original instantiation". These principles "levy different demands against preservation". The context is the range of expected uses. In the view of Rothenberg:

> "Since an authenticity principle encapsulates the preservation implications of a range of expected uses, it should always be derived from a specific range of this sort".

He admits that future use is speculative, but it is an interesting principle with significant implications. He sees it as a call to the creator — to aid preservation. We will discuss the call to the creator and publisher both in the next subsection and in section 9.4.

What is always of concern in any discussion of preservation is the tendency in some quarters to dismiss 'look and feel' as irrelevant. We have discussed in an earlier section a tendency in some quarters, mostly from within information technology, to distinguish sharply 'look and feel' from 'essence' in determining what is authentic. It is good to see that, at least in theoretical circles, this is being resisted.

9.3.3 Breaking down the digital entity

It was the intention of the author to develop the ideas of Kircz as set out at length in section 6 above. It does me that when we look at the sort of scholarly communication that takes advantage of the functionality of the web to produce the message in a number of different media, we need, for preservation purposes, to take each component as a separate problem, a separate problem just to hold never mind preserve with due regard for authenticity. However, such an excursion into different types of media is not a realistic aim for a work even of this length. Research on some of these components is beginning not enough research is beginning to become available. This is particularly true of images. The big ICSTI study of 1999 records a workshop on metadata for image preservation and notes that:

> "There was significant discussion of issues relating to preservation, particularly the movement of an image from one collection to another, on its way to long-term permanence, and how to verify the authenticity of the image over time... As draft elements, guidelines and white papers are developed, they will be made available on the work groups Web site (http://www.niso.org/images.html)". (page 63)

Nevertheless, there is one example of example of research, springing from the Mellon Foundation Harvard University Library project (Inera), that it is worth doing more than just drawing attention to. This is because the mission, from which this E-Journal Archival DTD Feasibility Study, is so much in accordance with the purpose of this study and also because the research itself is concerned with immediate and practical issues. The mission statement of the archive project is interesting, because the wording could mean that a lot of the theoretical issues we have been mentioning and which are discussed in more detail in section 3, are at the best ignored:

> "The archive's purpose is to preserve the *significant intellectual content of journals independent of the form,* in which that content was originally delivered in order to assure that this content will be available to the scholarly community for the indefinite future. Functionally the archive is designed to render text and still images and other formats as practical *with no significant loss in intellectual content.* The archive reserves the right to freely manipulate the internal format of the manifestation over time as long as the *plain meaning of the intellectual content* is preserved".

The italics are my own and in my view this is a brave if not foolhardy mission statement, which is likely to worry co-operating publishers and the scholarly authors they might represent. In particular, the use of the word 'plain' alongside 'significant' might ring all sorts of warning bells. However, the commissioning of the report reflects great credit on the library because the subject, the basic SGML (or now probably XML) in which the files they might receive from journal publishers is structured is part of the essential building blocks of any archive of scholarly communication. Ten journal publishers took part in the project.

The conclusions of the report (page 62) are significant. The headline conclusions are as follows:

> "[We] believe a DTD or Schema can be developed that will allow successful conversion of significant intellectual content from publisher SGML and XML files into a common format for archival purposes... While we are confident that the design and development of an archive DTD can be successfully completed, we believe there are significant challenges to be faced with its deployment and use".

There is a lot of subtext here. From subsequent presentations it is clear to me that some of the DTDs surveyed are better than others for any purpose and that there are very good reasons for some publishers to improve their DTDs. Anyone with any knowledge of how publishers arrive at their DTDs will not be surprised. In the listing of challenges mentioned in the second sentence of the conclusions there is a strong indication that publisher must include their quality control. This is not a surprise to any publisher. What is doubtful (see the next section) is the assumption, certainly held by the library and the Foundation, that publishers will provide the common archival DTD, although it seems that the conversion involved is not usually onerous.

The concern of this study must be that if the publisher chooses key structural elements not just because, as is always says, the publishers perceives a particular expression of the message as necessary because of their perception of the demands of the marketplace, but also because it is what they believe their authors want. For publishers of journals the marketplace is always the author or the representative of the author (the learned society for example), and not just the reader or user. The demands of different disciplines, what scholars in different disciplines want to convey, differ from one discipline to another. A loss of elements is, or could be, a loss of authenticity

9.3.4 The mechanics of intellectual preservation.

There is a mechanical side to the preservation of authenticity, which has exercised by librarians. It is instructive and probably realistic that the excellent topic site from PADI, the National Library of Australia, to which reference is given in 9.2.1, spends over a third of the one page on authenticity on the "range of strategies for asserting the authenticity of digital resources". They are all what one might call mechanical. The range of methods covers identifiers but it also includes hashing and digital time stamping. These are characterized as 'public methods'. The note continues:

> "Another class of methods for establishing authenticity includes encapsulation techniques and encryption strategies. A digital watermark can only be detected by appropriate software and is primarily used for protection against unauthorised copying. Digital signatures are used to record authorship and people who have played a role in the document".

There is a more detailed run through of many of these approaches in **Graham (2)**, where he advocates an algorithmic solution. However, his argument is a lot more wide-ranging that this statement might indicate and he tries (usefully) to look forward in time. It is my impression that, in practice, few of these mechanical approaches have been taken up by librarians engaged in developing strategies for now. Again the question of cost comes in. It is also not clear to us how mechanical strategies relate in practice to the general questions of trust, so

central to the thought of Lynch. Some of the more general and less theoretical questions of trust are touched on in the next subsection, but this is more about trust in institutions or organizations than trust mechanisms for evaluating the entities themselves.

Graham (2) also looks back to submission and ingestion. His note 10, quoting from Battin is particularly interesting:

> "For analog information, we must develop triage strategies for the past; for digital, prospective triage strategies at the point of acquisition or creation".

9.4 DECISIONS ON WHO ARCHIVES AND PRESERVES

This final main subsection of section 9 is built around a paradox. The scholarly author characteristically entrusts his or her message to the publisher, but the publisher is not seen as an archiving entity. Libraries often have difficulties understanding the behaviour, motivation and output of the scholar as author, but they are expected to preserve this output for posterity. The first subsection will cover an issue that has already been touched on in previous sections, but from a different angle. The second subsection will provide a concrete example of a general problem of preserving authenticity.

9.4.1 The role of trust and the question of mission

As we have seen in 3.4, Lynch and other writers have raised the question of who can be trusted as the origin of metadata in the digital environment. The question here in this subsection is a different one. Who is responsible for archiving the electronic publication or, more correctly, who should be responsible? This is at least as important a question.

The Research Libraries Group with OCLC produced a major report (**RLG 1**) in 2001 on the attributes of a trusted digital community. They produced the following definition:

> "Long-term preservation means two *distinct but equally important* functions: long-term maintenance of a bytestream and continued access to its contents through time and changing technology" [their italics].

Authenticity is not mentioned in this definition and indeed the emphasis throughout is elsewhere:

> "Materials that are 'born digital can present more challenging problems because their 'being digital' is not only a method of access, it represents their value as an information artefact. For many born-digital resources, effective preservation will rely as much on the object's digital characteristics or properties as on preservation of its basic intellectual content" (page 18).

From the point of view expressed in this study, the concept of "basic intellectual content" as somehow less difficult to preserve than physical characteristics is

rather surprising and the definition of 'basic' would be interesting. It is expected that at submission:

> "The repository must, in consultation with the depositor/rights owner and systems managers, assess the digital object and determine which of its properties are significant for preservation" (page 27).

The idea of a conference over each item is a little difficult to take seriously and it is a little surprising that the concept of the object's 'preservable essence' is passed over without explanation.

While there is some uncertainty here, it is clear that there can be little trust in publishers or for that matter authors:

> "While commercial publishers are beginning to provide some guarantee of continued access, most licensing agreements are perilously vague about how the digital library will be maintained and how long-term access will be ensured. Reliance solely on creators or producers of digital materials for long-term preservation of is potentially risky, not least because digital resources are not generally created or engineered with long-term preservation in mind".

This author would remove the word 'potentially'. In the UK, the same sort of concerns, as expressed in theory above, are expressed by the JISC committee of Electronic Information at August 2000 (no reference available) in a practical but more plaintive way:

> "The publisher should accept responsibility for ensuring that an archival copy of publications in digital format is maintained. In the event that the publisher is not in a position to take direct responsibility for maintenance of such archival copies, he should ensure that satisfactory alternative arrangements have been made. The licensee should have the right to preserve one copy of the files for archiving and for use in perpetuity".

In a sense it is difficult to understand why publishers should ever have been seen, even by themselves, as trusted to archive and preserve. Publishers publish, and libraries preserve. However, in the digital environment, or perhaps because of the digital environment, traditional roles have been called in question.

While librarians embarked on sponsoring competition and even running university presses, electronic or otherwise, some larger publishers seem to have at least considered taking on the library role of preservation. The fact of licensing rather than purchase, where the digital object (characteristically the journal) remains on the server of the publisher, naturally engendered ideas of this sort. In general however, ideas did not lead to a viable strategy that was worth the investment. The Elsevier arrangement with the Royal Dutch Library has already been mentioned and surveys have shown that most publishers in countries where a national deposit will or has become available look to their archive there. Nevertheless, short-term archiving for business reasons is a serious investment for publishers. There was an important explanation of how at least one publisher is going about creating their archive in the presentation by Geeti Granger of John Wiley & Sons at the final CEDARS Workshop in 2002. There are two reports, both

by Michael Day (2 and 3). He highlights the fact that the publisher is creating their archive for "business motives", to facilitate on-demand printing and to provide "material to support new business ventures". At the same time, the creation of this archive forces the publisher to adopt metadata standards, which can become convertible submission metadata for archives for the longer term.

Granger (2) (this is Stewart Granger not his namesake) made two important points to supplement and expand on the Research Libraries Report already quoted from (RLG 1). In the first place, it is his view (and my own) that:

> "Even a cursory examination of the problems of digital examination of the problems of digital preservation indicates the positive need for collaboration amongst interested parties and institutions".

He points out that such a collaborative mechanism does not currently exist, and goes on to the second related point - "Data creators are a different set of people from potential users of data" - and argues for a "plethora of differing motives and cultures".

The lesson in the context of this study is that it is probable that librarians should be trusted to archive and preserve. It is part of their traditional mission. They cannot, however, do so in isolation because they have to know the nature of the digital entities they are pledged to acquire. It is in the interests of both publishers and librarians, and the authors and users that they serve, that part of this necessary relationship should be the determination and then the maintenance of authenticity.

9.4.2 The vanishing archive

This final subsection of section 9 has been added in to the structure of this study because it relates well to the problems and hopes raised in the previous subsection.

In a submission to the International STM Association I wrote the following:

> "The medical librarian T.Scott Plutchak describes the area of contention, from a library standpoint but not unfairly, in an article entitled *Sands shifting beneath our feet* to be found in the Journal of the Medical Library Association for April 2002, 90 (2) 161-163. The complaint was against the removal by a leading STM publisher of a published article from all their databases on the grounds that it was 'entirely inappropriate'. The argument in the article, later supplemented by less reasoned complaints by professional publisher-bashers, was that just because we can now remove an article, as if it might have never existed, as an alternative to a published retraction linked to the article does not mean that it is an appropriate response. The publisher is Elsevier Science and the quoted policy on Article Withdrawal cited in the Plutchak article sets out circumstances under which "article is published that must later be withdrawn". The problem lies not so much in the policy, appropriate in print, but because of the finality of what can be done by the publisher in the digital environment, particularly

> where, as is often the case currently, there is no independent archive. Plutchak writes: "We must never forget that preservation of the historical record, with all its faults, mistakes, and corrections, is an essential part of the service that librarianship performs for society"."

As a result of these and other comments and criticisms, Elsevier has promulgated a new policy on Article Withdrawal that covers rules for article retraction, article removal and article replacement. This new policy has attracted general approval from the library community but not complete agreement. There has been a positive reaction to the fact that any article removed for whatever reason will remain available in the official archives of the publisher at the National Library of the Netherlands. There is also some concern about what is perceived as a "lack of transparency" in some of the language advocated to explain reasons for withdrawal. There is unfinished business. Plutchak, in a posting of 7th February, writes on this point:

> "Perhaps the most significant thing about the policy is Elsevier's commitment to be active in the development of international standards. The challenge to the rest of the publishing/editorial community is to develop similar policies and to make them public".

The new policy of the publisher is explained in an article in the Chronicle of Higher Education (http://chronicle.com/free/2003/02/2003021002t.htm.).

In preparing a submission to the STM Association, I advocated that the Association passed this up to the International Publishers Association as a problem to be solved in collaboration with the International Federation of Library Associations. We will see what happens. It is worth including this little story because it illustrates that we need new standards in the digital environment, that both publishers and librarians can become exercised by such standards given the right context, and that there may be mechanisms for solution. The question of the authenticity of a journal issue has not been raised in this study, but it is becoming of interest to some those concerned with these topics, for example Clifford Lynch. He considers that, for example, there is a fair case for attaching the names of the editorial structure, which accepted an article to the electronic files of that article.

10. Concluding comments

These concluding comments are neither a summary of conclusions distilled from a lengthy text, nor a pointer to work that can be done in the future. The comments have features of both approaches but there is no attempt to be comprehensive.

As is obvious from the study, my position is that in looking at an existing system one starts with the system, finds out how it works and what it does for whom. The analysis by the late Professor Sir Bryan **Coles** of the STM Information System, which was published as long ago as 1993, remains an important exemplar and, because of its approach, of serious use even a decade later. Coles was both a senior publisher and a distinguished physicist. Too few academics, who are not information scientists, have attempted anything like this work and it is certainly difficult to think of a publisher or a librarian who would be up to it. Of course he only deals with the scientific, technical and medical sphere of discourse, and only some of the conclusions are relevant to the arts and social sciences. Nevertheless, because this is where the research and researchers we have analysed for this study, it is germane to what in practice we have been trying to do for the practical reasons that we have expressed from time to time in the course of the work.

Many of those who have approached the questions dealt with in this study have approached it from a theoretical position. Scholarly communication has certain features. It should therefore work in a particular way. It does not work optimally in the current environment. Therefore, the way in which the pressures of the environment are expressed in institutional and other structures should be changed. This is rather a caricature of a perfectly reasonable way of working that comes naturally to those trained as a philosopher or at any rate in logistics. Academic computer scientists tend to have a good background in logic. I am trained as a historian.

Another position, which does seem to emerge from computer science or, perhaps more correctly, information technology, is, to our mind, less valid. There are too many commentators on the digital revolution who seem to be convinced that what can be done in the digital environment must not only be intrinsically a good thing but must lead to good results. This cannot be acceptable. It is the progress of knowledge that is the aim of scholarly communication, and the sort of causal approach, which leads to progress of this type, does not change just because the environment changes. The same, it might be argued, goes for the importance of authenticity as part of the way in which the whole process moves forward. Scientific communication is, to be worthwhile, about science not about the transient convenience of scientists. The advantages of the digital environment, especially the ease of searching, do not cut out the hard business of thinking, of testing hypotheses by experiments, which is central to much of the scholarly process and the scholarly communication underpinning it.

That being said, why is there so much ferment particularly among librarians? Why is there genuine concern that the system is not working properly and needs repair? **Coles** again outlined the drivers succinctly (page 1) in his introduction:

> "Difficulties in reconciling the increased supply of scientific information with the decreased ability of libraries to acquire it mean that the nature of the STM information system is altering, and may be expected to change further as time goes on. Financial and commercial pressures will soon become much stronger drivers of the change because of the shifts in the funding of the collection of and dissemination of STM information".

We will return to the theme of the second sentence later. For the moment, consortia deals notwithstanding, there are many who expect and welcome change for good reasons. It does not matter so much whether or not there is a crisis. The belief in a crisis has caused the system itself to be looked at more closely.

There is another way of distinguishing my position (and for that matter the position of Coles) from a number of those who write about the subject matter of the study (for example **Van de Sompel** or **Roosendaal**). It is the assumption that there is an information chain in the print and a similar chain in the digital environment. Sometimes this chain, crudely characterised as author–publisher–librarian–reader, is perceived as no longer relevant in a digital environment. There are diagrams showing the new structure or structures. How valid are these new structures?

Roles of intermediaries in scholarly transmission in the print environment have been worked out over a period of time, as both **Guedon** and **Mabe** point out starting from very different positions. It is interesting how different writers or presenters differently interpret the role of Henry Oldenberg. He was the founder/owner of the first scientific journal or almost the first and an inhabitant of the first slide in many a presentation. Everything now is moving much more quickly. It is not surprising that in the digital environment the role of the intermediary needs re-interpreting. It may even be that roles have changed but the functions assisting creation and distribution and dissemination and preservation (to mention some functions) are still there. Throughout this study we have emphasised that, in our view, if librarians take on the publishing function they have to take the responsibilities associated with that function and likewise, for publishers, the demands of archiving and preservation must be undertaken properly — 'in perpetuity' — or not undertaken at all.

It could of course be argued that neither publisher nor librarian is needed and that authors and readers, enabled by the Web, can interact more or less directly. **Ginsparg** at one time held a view not unlike this, and it was built into the range of concepts underlying the JISC Electronic Library (eLib) project in the UK — as I can personally confirm. This argument is made less frequently now. Throughout the study, we have had to accept publishers as intermediaries for authors and librarians as users, because that is how it happens on the whole. Both publishers and librarians are (fairly) keen on disintermediation in general, as long as they are not expelled from the chain, but in general it is perhaps significant that that this clumsy word is no longer used in the sort of analysis where it once was de rigeur.

There is change, because scholarly communication is happening online rather than delivered in print, but there is continuity, because (it is our contention) the

creation of knowledge is much the same. The central sections are so lengthy because these are the central questions, or rather statements.

The context, as set out above, is the context as I, the author, sees it, for considerations of authenticity. As we have asserted the handling of questions of most aspects of authenticity was in the hands of the publisher as the agent of the author. The publisher protected, or was expected to protect, integrity and paternity. The publisher was responsible for certification and therefore of the definitive version, the version that becomes part of the minutes of science. Indeed the publisher hands over the definitive version gift-wrapped, as it were, for the librarian to put into a drawer. This was true of the print environment and is true of the digital environment even if in some circumstances a publisher does not exercise the publishing function.

There are a number of rather sweeping statements made above. However, if any one message is presented by this study it is that such sweeping statements about questions of authenticity are dangerous, misleading and incorrect. They are dangerous because they lead to rules being made that are premature. Publishing online is in its early days. There are so few e-only journals, for example. They are misleading because there is so little evidence. There is a real paucity of information about what is involved in running an e-only journal with non-print components as part of the essential message. Editors are finding their way. It is interesting, in spite of all that has been said above, that we are talking of editors and not much yet about publishers. Sweeping statements about authenticity are incorrect because they are incorrect. It is obvious from the research and experience quoted in this study that there are a huge range of attitudes by the author and user to scholarly communication online and such vital areas as peer review. The answer to each question involves discovering and presenting not one attitude or position but a spectrum across disciplines, within disciplines and perhaps, though this seems less likely than if frequently stated, across an age range. There are a number of paradoxes. A stronger belief among the scholarly community in peer review than obtained in the past seems to have been established, though comparability across time through surveys is lacking. It is a pity that **Tenopir** and King did not look at some of these questions of attitude. At the same time there is little doubt that peer review, as practised, does not in fact deliver what it claims to deliver — this fact is well understood. Associated with this understanding is the fact that the quality of peer review is only one of the factors leading to a journal being regarded as prestigious. Whereas it seems generally to be the case that most scholars see a clear-cut division between formal and informal communication, how formal communication comes about is not at all clear in all cases.

I began work on this study by believing that the maintenance of integrity and paternity was truly important to scholars, but, as we have seen, they are usually fairly relaxed about it. The protection of authenticity is one of the arguments publishers put forward when claiming copyright (see **Gadd**), but publishers, in practice, are not interested in maintaining integrity unless, as we will see below, there is money to be made. Librarians acting for users often do not see the point in preventing cutting and pasting that is an obvious attack on integrity. We are now writing about the digital environment, where integrity can be threatened in a more obvious way, but in general issues, that must surely be important, are not faced up to. Is there a need for an education in authenticity? There is some

progress in archiving and preservation, but in that context, as we have seen, there is not much interest in the provenance of what is received to be archived, if the digital entity is born-digital and comes from outside as it were. At least there is a debate about the relationship between a string of bits and intellectual content, though it is not central to the overriding concerns of preservation metadata, which is more concerned with the technical issues than making sure that the message is secure. Once the term 'look and feel' comes into a statement, there must be concern about intellectual content.

It is quite clear that 'traditional publishers have not really thought about the issues raised in this study. The last sub-section in the penultimate section on the vanishing archives shows how the largest publisher could lose thirty or so articles without noticing it. I was at a publishing meeting a few weeks ago at which the controversy over these actions was mentioned and the arguments set out. The publishing lawyers present seemed to think it self-evident that if there was any chance of a legal threat, part of the scholarly literature could be deleted for ever, should be deleted for ever, without a second thought. We assigned most of a section to the legal issues surrounding authenticity, especially those concerned with moral rights. It was clear that moral rights mean very little and are unlikely to be protected by the publisher. There are currently debates within publishing houses about whether or not to put articles online before copyediting. These debates are not characterized by any soul searching about establishing and protecting a definitive version. The debate about the definitive version, discussed at length in this study, does not impinge more than peripherally in the decision making processes familiar to the author. It is a matter of cost, of market interest and procedures that will, on the whole, decide policies.

It is in the area of metadata that the lack of interest is most obvious. Metadata schemes have attracted some very serious thinking but identifying the definitive version is way down on the list of elements to be included, as we have seen. If the concept of the definitive version is as important a concept for the progress of knowledge as we have suggested, such an identification should be near the top of the list, even if it is a simple description that cannot be checked.

It is in the area of metadata also that the importance of economic drivers is most obvious. We have already provided a quotation that raises this point. Constructing metadata is expensive, and protecting documents is expensive. No one is going to pay out money without a return, though archiving and preservation does represent something of a qualification in this regard. Is there any reason to suppose that the situation will change, that the authentic document (the definitive or authoritative version) is an entity that can be sold because of its status? What are the circumstances that will make such a change possible?

In this study we have not proposed standards. We have reported. It seems to us that conventions have to be established before standards are worth constructing. Conventions thrive when there is a consensus. There is no consensus over issues of authenticity, not because the importance of authenticity is not recognized. The work of **Lynch**, who is highly regarded in library circles, has made it certain that the issues are well placed in the literature, but, as we discussed in the study, there is little evidence of implementation of his ideas. I have come to the conclusion that the SPARC approach, the attempt to overthrow the system, pervades so much of the literature that it makes the achievement of any

consensus on such issues as the definition of a publication very difficult to achieve. It is sad because that particular attempt of definition did have cross-sectoral approach in that publishers and librarians were included in the committee. Again all the work put into the proposal, explained in the study, seems to have been wasted. No-one is taking it further.

There is probably more hope of collaboration and cooperation in archiving and preservation than in any other area of endeavour connected with authenticity. I have been much involved in the JCVD (see references). The procedures have been built on cooperation. The attempt to extract submission metadata from publishers in a way that fits in with their workflow has motivated the research on ONIX for serials. We have mentioned ONIX in the text but not explained the work on using the standard for submission metadata. In another context, the Harvard project funded by Mellon, we have seen work, which will lead to publishers having to pay to convert. It is my understanding that such a heretical view was not entirely thrown out of court. It is also in the area of archiving and preservation that CLIR have been able to bring together the arch-enemies ARL and AAP. Again conversations are said to have been cordial and constructive. It is interesting that it is in this area that what interest in authenticity there is has come to the fore.

It seems me that the behaviour of scholars has always been pragmatic. The divorce in attitudes between the scholar as author and the scholar as user is often remarked upon with surprise especially where the same person fulfils both roles. For pragmatic one could substitute wayward. Perhaps it is part of the job of the publisher and librarian as intermediaries to enable an optimal working environment by exercising a controlling function, to provide order in an academic environment that needs it to work properly, but an order that the academic communities cannot achieve on their own. Of course, as is always pointed out, in the end the academic community will vote with their feet. They will, for example, leave those big expensive commercial journals and demand open access journals — or they will not. But is it arrogant to see the intermediary role as not being just passive but active on behalf of scholarship. If the intermediaries do not work together, the roles of both publishers and librarians are handicapped. There are a range of questions relating to authenticity examined at length in this study. Is it too much to hope that some entity like CLIR (http://www.clir.org) will bring librarians and publishers together to establish some ground rules? Or is it too soon. The opportunities presented by e-publishing have only just begun to be explored. There have been big investments in linking to enable searching to lead to the object desired. This sort of initiative is supported across the information chain. However, in practical terms, an investigation of how digital entities work when they are true digital entities has hardly been touched on by any company with the money to spend on such an enterprise. It is perhaps too soon for the sort of programme set out by **Kircz** but the sort of procedures laid out in his writings give some indication of the sort of problems that are going to exercise the minds of many. It is my impression that the time for multi-media messages is nearly with us, subject of course to the range of qualifications that have to be made to any generalization about scholarly communication

BIBLIOGRAPHY

The Preface contains comments on the difficulties of finding 'literature' on this topic.

A good example of the problem implied is what happens if 'authenticity', as a search term, is applied to the excellent bibliography by Charles W Bailey (see below). It gets no results though it must be admitted that 'authority' did bring up some useful references.

Another bibliography available online (Dworaczek – see below) does recognize the term, which it clusters as follows: Authentication/Authenticity/Authoritative Version. This cluster listed 43 items early in 2002, many of which were relevant. Liblicense (http://www.library.yale.edu/~llicense/) incidentally brought up three postings in early 2002, but one of them is by me. Liblicense also brought up seventy-nine postings if the search term used was 'authority' but almost all of these were concerned with 'authority to sign' in libraries.

Personal communications are not referenced in this bibliography but the nature of the communication is made clear in the text, though the author is not fully identified. In section 9 there are some statements based on my membership of the Joint Committee for the Voluntary Deposit of Non-Print Publications (JCVD) and the JISC e-Book Working Group in the UK, which cannot be referenced as they are based on discussions within the committees.

I have a final apology. I am not absolutely certain in all cases whether the references in the text refer to the right 'publication' by a particular prolific author pointed to as the source in this bibliography. The utterances of some of the more significant authors quoted, Kircz, Lynch and Paskin, are particularly certain to track down in the most authoritative version, and sometimes versions are different. The fact that I have had this problem does of course have obvious significance for the topic under discussion.

AAAS/UNESCO/ICSU Worksop on Developing Practices and Standards for Electronic Publishing in Science (1998) available at http://www.aaas.org/spp/dspp/sfrl/projects/epub/report.htm

ALPSP/ EASE Peer Review Survey October-November 2000 (Current Practice in Peer Review Report), available at http://www.alpsp.org.uk/pub4.htm.

Armstrong, Chris, *Metadata, PICS and Quality,* Ariadne issue 9 available at http://www.ariadne.ac.uk/issue9/pics

Bailey, Charles W, *Scholarly Electronic Publishing Bibliography.* The latest version is available [in HTML] at http://info.lib.uh.edu/sepb/sepb.html and other formats are available at related URLs.

Bearman (1), David and Jennifer Trant, *Authenticity of Digital Resources: Towards a Statement of Requirements in the Research Process.* D-Lib Magazine 4 (6) June 1998. Available at www.dlib.org/dlib/june98/o6bearman.html.

Bearman (2) et al, *A Common Model to Support Interoperable Metadata: Progress report on reconciling metadata requirements from the Dublin Core and INDECS/DOI Communities,* D-Lib Magazine 5:1 January 1999 available at http://www.dlib.org/dlib/january99/bearman/01bearman.html

Bearman (3), David, *Reality and Chimera in the Preservation of Electronic Records,* D-Lib Magazine April 1999 5 (4) available at http://www.dlib.org/dlib/april99/bearman/04bearman.html

Beit-Arie, Oren, *Linking to the Appropriate Copy: Report of a DOI-Based Prototype,* D-Lib Magazine 7:9 September 2001 available at www.dlib.org/dlib/september01/caplan/09caplan.html

Bently, Lionel and Brad Sherman, *Intellectual Property Law.* Oxford: Oxford University Press, 2001.

Berne *Convention for the Protection of Literary and Artistic Works*. Available inter alia at www.law.cornell.edu/treaties/berne/6bis.html

Berry, R.Stephen and Anne Simon Moffat**,** *The Transition from Paper*: *Where are we going and how will we get there?* American Academy of Arts and Sciences: Cambridge MA, published online only in 2001 and available at http://www.amacad.org/publications/trans.htm

BIBLINK, *Work Package 6 of Telematics for Libraries Projects BIBLINK* (LB 4034], 1997, available at http://hosted.ukoln.ac.uk/biblink/wp6/d6.1/toc.html.

Bide (1), Mark and Trevor Hing, *User Identification and Authentication: a brief introduction.* London: Book Industry Communication, 1998. Available at http://www.bic.org.uk/userid.pdf

Bide (2), Mark, Liz Potter and Anthony Watkinson, *Digital Preservation: an introduction to the standards issues surrounding the deposit of non-print publications,* London: Book Industry Communication, 1999. Available at http://www.big.org.uk/digpres.doc.

Bide (3), Mark, *Seminar Report: DIGITAL PRESERVATION – Standards issues surrounding the deposit of non-print publications,* London: Book Industry Communication 2000. Available at http://www.bic.org.uk.

Bosak, Jon, *XML, Java, and the future of the Web* in its final version 1997 available at http://metalab.unc.edu/pub/sun-info/standards/xml/why/xmlapps.htm

Boyce, Peter *et al. –Three Year Plan for Developing Electronic Publishing at the American Astronomical Society,* available at http://www.aas.org/Epubs/webinfo/Plan/epplan92.htm.

British Library, *A review of the policy and arrangements for the legal deposit of printed material in the British Library*. British Library: London, 1997.

Butterworth, I, *The Present Situation and the Likely Future: Introduction.* In *The Impact of Electronic Publishing on the Academic Community* (I. Butterworth, ed). London: Portland Press 1998. Available online at tiepac.portlandpress.co.uk/tiepac.htm.

Canada (National Library of Canada) *Report of the Electronic Publications Pilot Project (EPPP) of the National Library of Canada,* National Library of Canada: Ottawa, 1996

CEDARS (1) Project Team and UKOLN, *Metadata for digital preservation,* available at http://www.leeds.ac.uk/cedars/MD-STR~5.pdf. [The version used in this study is March 2000]

CEDARS (2) *The Cedars Project Report (April 1998 – March 2001),* June 2001

CLIR, *The State of Digital Preservation: An International Perspective,* Proceedings of a conference held in Washington in April 2002, Council on Library and Information Science: Washington DC, June 2002, available at http://www.clir.org/pubs/reports/pub107/contents.html

Cockerill, Matt, *Distributed and centralized technologies: complementary tools to build a permanent digital archive* available online in Nature webdebates at http://www.nature.com/nature/debates/e-access/Articles/cockerill.html

Coles, B.R, *The Scientific, Technical and Medical Information System in the UK* (A study on behalf of the Royal Society, The British Library and The Association of Learned and Professional Society Publishers), British Library: London 1993. This study is British Library R&D Report No. 6123

Copyright, *Copyright, Designs and Patents Act 1988 chapter 48,* HMSO: London, 1988, available at http://www.hmso.gov.uk/acts/acts1988/Ukpga_19880048_en_1.htm

Crow (1), Raym, *The Case for Institutional Repositories: A SPARC Position Paper,* The Scholarly Publishing and Academic Resources Coalition: Washington DC, 2002**,** available at http://www.arl.org/sparc/IR/ir.html.

Crow (2), Raym, *SPARC Institutional Repository Checklist & Resource Guide*, The Scholarly Publishing and Academic Resources Coalition: Washington DC, 2002, available at http://www.arl.org/sparc/IR/IR_Guide.html

Cullen, Charles T, *Authentication of Digital Objects; Lessons from a Historian's Research.* In Smith (1) cited below and available at www.clir.org/pubs/reports/pub92/cullen.html.

Day, Michael, *E-print Services and Long-term Access to the Record of Scholarly and Scientific Research,* Ariadne issue 28 available at http://www.ariadne.ac.uk/issue28/metadata/

Day (2), Michael, *The Final CEDARS Workshop: a Report from Manchester UK,* RLG Diginews, April 15 2002, Volume 6, Number 1, available at http://www.rlg.org/preserv/diginews/v6_n2_conference.html

Day (3), Michael, *Cedars Final Workshop Manchester Conference Centre, UMIST, Manchester, 25-26 February 2002,* Workshop summary available at http://www.leeds.ac.uk/cedars/pubconf/umist/finalWorkshopRep.html

DCMA. *The Digital Millennium Copyright Act of 1998.* U.S. Copyright Office Summary, December 1998 available at http://www.loc.gov/copyright/legislation/dmca.pdf.

Dempsey, Lorcan and Stuart L. Weibel, *The Warwick Metadata Workshop: A Framework for the Deployment of Resource Description,* D-Lib Magazine July/August 1996 available at http://www.dlib.org/dlib/july96/07weibel.html

Denning, Peter J. and Bernard Rous, *The ACM Electronic Publishing Plan,* Communications of the ACM, April 1995 and available at http://www.acm.org/pubs/epub_plan.html.

DLF (Digital Library Federation), *The Andrew W. Mellon Foundation's e-Journal archiving program,* DLF/CLIR last updated 26 June 2001. Available at http://www.diglib.org/preserve/ejp.htm

Doctorow, Cory, *Metacrap: Putting the torch to seven straw-men of the meta-utopia*, Version 1.3 26 August 2001 available at http:well.com/~doctorow/metacrap.htm

Dorner, Jane, *The Internet: A Writer's Guide second edition.* London: A&C Black 2001.

Dworaczek, Marian, *Subject Index to Literature on Electronic Sources of Information.* November 15th 2001 Edition available at http://library.usask.ca/~dworacze/SUBJIN_A.HTM

Ekman, Richard and Richard E.Quandt, eds. *Technology and Scholarly Communication.* Berkeley: University of California Press, 1999.

EPS (1) Electronic Publishing Services, *XML Xplained*, EPS Monthly Briefing Paper: August 2000 available from http://www.epsltd.com/Order/Order_main.asp

EPS (2) Electronic Publishing Services, *Digital Rights Management: Unlocking the value of content,* EPS Monthly Briefing Paper: October 2000 available from http://www.epsltd.com/Order/Order_main.asp

EPS (3) Electronic Publishing Services, *Publishing after Copyright: maintaining control online,* EPS Monthly Briefing Paper: March 2001 available from www.epsltd.com/Order/Order_main.asp

EPS (4) Electronic Publishing Services, *The impact of the extension of legal deposit to non-print publications: Assessment of cost and other quantifiable*

aspects, Prepared for the Joint Committee on Voluntary Deposit by EPS Ltd, 2002, and available at http://www.epsltd.com/ExtensionOfLegalDeposit.htm.

EU-CD. DIRECTIVE 2001/29/EC OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 22 May 2001on the harmonisation of certain aspects of copyright and related rights in the information society. Available inter alia at http://www.alpsp.org.uk/EUDir.rtf

Feeney, Mary (editor), *Digital Culture: maximising the nation's investment,* National Preservation Office/British Library 1999.

Fleming, Janice L, *Maintaining the Integrity of Electronic Publications: Potential Problems and Possible Solutions II,* in **AAAS** at http://www.aaas.org/spp/dspp/sfrl/projects/epub/ses3/fleming.htm

Frankel, Mark and Roger Elliott, co-chairs of the 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) October, 251-8. Available from http://www.catchword.com/alpsp/09531513/v13n4/contp1-1.htm

Fredriksson, Einar H (ed.), *A Century of Science Publishing. A Collection of Essays.* Amsterdam: IOS Press 2001.

Gadd, Elizabeth, Charles Oppenheim and Steve Probets, *ROMEO Studies 1: The impact of copyright ownership on academic author self-archiving,* accepted for publication in the Journal of Documentation and available in self archived form at http://www.lboro.ac.uk/departments/ls/disresearch/romeo/RoMEO%20Studies%201.pdf

Geertz, Janet, *Selection Procedures for Preservation,* 1999, available at http:www.rlg.org/preserv/joint/geertz/confpapers.html

Gilheany, Steve, *Preserving Information Forever and a Call for Emulators,* presented at the Digital Libraries Conference Singapore 1998 available at http://www.archivebuilders.com/aba010.html

Gilliland-Swetland, Anne J and Philip B. Eppard, *Preserving the Authenticity of Contingent Digital Objects: The InterPARES Project.* D-Lib Magazine 6 (7-8) July 2000. Available at http://www.dlib.org/dlib/july00/eppard/07eppard.html

Ginsparg (1), Paul *Winners and Losers in the Global Research Village.* Joint ICSU Press/UNESCO Expert Conference on Electronic Publishing in Science UNESCO, Paris, 19-23 February 1996 available at http://associnst.ox.ac.uk/~icsuinfo/Ginsparg96.htm.

Ginsparg (2), Paul, *Creating a global knowledge network,* invited contribution for Conference held in Paris February 2001, Second Joint ICSU Press – UNESCO Expert Conference on *Electronic Publishing in Science* available at http://arxiv.org/blurb/pg01unesco.html.

Ginsparg (3), Paul, *Update, Sept '96.* Appeared in the APS (American Physical Society) Newsletter November 1996 and available at http://arxiv.org/blurb/sep96news.html.

Godlee, Fiona and Tom Jefferson (eds), *Peer Review in Health Sciences.* London: BMJ Books, 1999.

Graham (1), Peter S, *Intellectual Preservation: Electronic Preservation of the Third Kind,* Washington DC: Commission on Preservation and Access, 1994, available at http://web.syr.edu/~psgraham/pgsite/pglibwork/pgtexts/cpaintpres.HTML.

Graham (2), Peter S, *Issues in Digital Archiving,* a chapter in *Preservation* (edited by Paul Banks and Roberta Pillette), Chicago: American Library Association 2000 (pages 97-113) available in a submitted version at http://web.syr.edu/~psgraham/pgsite/pglibwork/pgtexts/digarchiv2000.html.

Gregory, Vicki L, *Scholarly Communication/Publishing.* Lecture Notes (1999) available at http://www.cas.usf.edu/lis/lis6260/lectures/schpub.htm

Granger, Stewart, *Emulation as a Digital Preservation Strategy,* D-Lib Magazine October 2000 (volume 5 number 10), available at http://www.dlib.org/dlib/october00/granger/10granger.html

Granger (2), Stewart, *Digital Preservation and Deep Infrastructure,* D-Lib Magazine February 2002 Volume 8 Number 2, available at http://www.dlib.org/dlib/february02/granger/02granger.html.

Guedon, Jean-Claude, *In Oldenburg's Long Shadow: Librarians, Research Scientists, Publishers and the Control of Scientific Publishing.* Washington: The Association of Research Libraries (2001), available online, chapter by chapter, at http://www.arl.org/proceedings/138guedon.html.

Hanna, Marsha, *Born Digital – have Digital? Maybe.* ICSTI Forum (38) June 2001. Available at http://www.icsti.org/forum/38/

Harnad, S and M. Hemus, *All or none: no stable hybrid or half-way solutions for launching the learned periodical literature into the post-Gutenberg galaxy,* in **Butterworth.**

Harmsze, F.A.P *et al.*, *A modular structure for electronic scientific articles*, available at http://www.science.uva.nl/projects/commphys/papers/infwet/infwet.html.

Hedstrom (1), Margaret and Clifford Lampe, *Emulation vs. Migration: Do Users Care?,* RLG DigiNews 5 (4). Available at http://www.rlg.org/preserv/diginews/diginews5-6.html

Hedstrom (2), Margaret and Shaun Montgomery, *Digital Preservation Needs and Requirements in RLG Member Institutions,* commissioned by the Research Libraries Group, 1998, available at http://www.rlg.org/preserv/digpres.html#toc. [The pdf version of this survey refuses to download]

Herman, Lee and Alan Mandell, *The Given and the Made: Authenticity and Nature in Virtual Education.* FirstMonday, 5(10) 2000. Available at www.firstmonday.org/issues/issue5_10/herman/index.html

Hirtle, Peter, *Editorial – OAI and OAIS: What's in a Name?,* D-Lib Magazine April 2001 7(4). Available at http://www.dlib.org/dlib/aapril01/04editorial.html

Hitchcock, Steve and Wendy Hall, *How Dynamic E-Journals can Interconnect Open Access Archives.* Pages 183-193 in Arved Hubler at al (editors) *Electronic Publishing '01:2001 in the Digital Publishing Odyssey.* Amsterdam: IOS Press 2001 and available at www.bib.ecs.soton.ac.uk/data/6878/html/elpub01-online.html

Hunter, Karen, *STM Members Viewpoints and Development*, a presentation to the ICSTI conference in Paris February 2002, available at www.niso.org/presentation/hunter_ppt_01_22_02/, and subsequently published inby IOS Press in Information Services & Use, Volume 22, Numbers 2 and 3, pages 83-88.

ICSTI/CENDI, *Digital Electronic Archiving: The State of the Art and the State of the Practice,* ICSTI 1999. Available at www.icsti.org/99ga/digarch99_MainP.pdf. [ICSTI is the International Council for Scientific and Technical Information and CENDI is the Federal Scientific and Technical Information Managers Group]

Inera Incorporated, *E-Journal Archiving DTD Feasibility Study,* Harvard University Library Office for Information Systems E-Journal Archiving Project 2001, available at http://www.diglib.org/preserve/hadtdfs.pdf

Jenkins, Clare (project director), *Cedars Guide to Preservation Metadata,* The Cedars Project 2002. Available at http://www.leeds.ac.uk/cedars/guideto/metadata/.

JCVD (1) *Code of Practice for the voluntary deposit of non-print publications: general information and explanatory notes,* [British Library: London], available at http://www.alpsp.org/codeexfn.pdf.

JCVD (2) *Code of practice for the voluntary deposit of non-print publications Form 1: Publisher Information,* [British Library: London], available at http://www.alpsp.org/volfrm1.pdf.

JCVD (3) *Code of practice for the voluntary deposit of non-print publications Form 2: Publisher Specific Information,* [British Library: London], available at http://www.alpsp.org/volfrm2.pdf

JISC: Site of E-BOOKS WORKING GROUP. To be found at http://www.jisc.ac.uk/dner/ebooks/index.html

Jones, Hugh, *Publishing Law.* London: Routledge, 1996. [It is this edition that is cited but there has this year (2002) been a new edition]

Jones (2), Maggie and Neil Beagrie, *Preservation Management of Digital Materials: A Handbook,* London: The British Library for Resource. 2001. There is an updated web version at http://www.jisc.ac.uk.

Key Perspectives Ltd, *Authors and Electronic Publishing (The ALPSP research study on authors' and readers' views of electronic research communication),* The Association of Learned and Professional Society Publishers: Worthing 2002.

Kenny, Sir Anthony, *Report of the Working Party on Legal Deposit,* [date], available at http://www.alspsp.org/kennyrep.htm.

Kiernan, Vincent, *Why do some electronic-only journals struggle, while others flourish,* "reprinted" in the Chronicle of Higher Education by permission of The Journal of Electronic Publishing. Available at http://www.press.umich.edu/jep/04-04/kiernan.html

Kircz (1), Joost and Anita de Waard [compilers], *Change and Continuity in Scientific Communication,* proceedings of a conference at KNAW Trippenhuis June 2001 and available at http://www.niwi.knaw.nl/ccsc/index.htm.

Kircz (2), Joost and Hans Roosendaal, *Understanding and Shaping Scientific Information Transfer* in Joint ICSU Press/UNESCO Expert Conference on Electronic Publishing in Science UNESCO, Paris, 19-23 February 1996 available at www.science.uva.nl/projects/commphys/papers/unesco.htm

Kircz (3), Joost G., *New Practices for electronic publishing: how to maintain quality and guarentee integrity,* in Proceedings of the Second ICSU-UNESCO International Conference held in Paris 20-23 February 2001 and available at http://associnst.ox.ac.uk/~icusinfo/kirczfin.htm

Kircz (4), Joost G, *New practices for electronic publishing 1: Will the scientific paper keep its form,* Learned Publishing 14:4 October 2001 pages 265-272, available at http://mustafa.catchword.com/vl=28504003/cl=11/nw=1/rpsv/cgi-bin/linker?ini=alpsp&reqidx=/catchword/alpsp/09531513/v14n4/s4/p265

Kircz (5), Joost G, *New practices for electronic publishing 2: New forms of the scientific paper,* Learned Publishing 15:1 January 2002 pages 27-32 available at http://elvira.catchword.com/vl=28645841/cl=24/nw=1/rpsv/cgi-bin/linker?ini=alpsp&reqidx=/catchword/alpsp/09531513/v15n1/s4/p27

Kircz (6), Joost (G), *Scientific Communication as an object of science [1],* a contribution to the Academia Europaea workshop *The impact of electronic publishing on the academic community, Stockholm April 16-20 1997* and now available at http://www.science.uva.nl/projects/commphys/papers/aceur.htm

Kling, Rob *et al.*, *Locally Controlled Scholarly Publishing via the Internet: The Guild Model,* CSI working paper available at http://www.slis.indiana.edu/csi/WP/WP02-01B.html.

Lariviere, Jules, *Guidelines for Legal Deposit Legislation,* International Federation of Library Associations, published by UNESCO: Paris, 2000 and available at http://www.unesco.org/webworld/index.shtml

Levy, David M, *Where's Waldo? Reflections on Copies and Authenticity in a Digital Environment.* In Smith (1) cited below and available at www.clir.org/pubs/reports/pub92/levy.html

Licensing Digital Information: A Resource for Librarians. Available at www.yale.edu/~llicense/index.shtml.

Lindquist, Mats G, *Citations in the Digital Space,* The Journal of Electronic Publishing March 1999 4:3 available at http://www.press.umich.edu/jep/o4-03/linquist.html

Literati: for the Literati Club see http://www.emeraldinsight.com/literaticlub/

Lunau, Carrol D, *The National Library of Canada;s Initiatives Towards Building a Canadian Virtual Library,* ICSTI Forum, no 34, May 2000, available at http://www.icsti.org/icsti/forum/fo0005.html.

Lynch (1), Clifford A, *Integrity Issues in Electronic Publishing.* Chapter 8 in Peek and Newby cited below.

Lynch (2), Clifford, *Canonicalization: A Fundamental Tool to Facilitate Preservation and Management of Digital Information.* D-Lib Magazine 5 (9) September 1999. Available at www.dlib.org/dlib/september99/09lynch.html.

Lynch (3), Clifford, *Authenticity and Integrity in the Digital Environment: An Exploratory Analysis of the Central Role of Trust.* In Smith (1) cited below and available at www.clir.org/pubs/reports/pub92/lynch.html.

Mabe (1), Michael, *What SuperJournal has taught us about authors.* From the SuperJournal Conference London 1999 and available at http://irwell.mimas.ac.uk/sj/confmabe.htm

Mabe (2), Michael, *Digital Dilemmas.* ASLIB Proceedings 53(3) March 2001 pp. 85-92.

Maddox, John, *Electronic journals are already here,* Nature 365 21 October 1993.

Martin, David, *Beyond Dublin Core – the need for high quality product information,* presented at BIC/BL Seminar on *Trading Electronic Content* held on 10 March 1998 and available at http://www.bic.org.uk/beyonddc.rtf.

Meadows, A.J, *Communicating Research.* San Diego: Academic Press, 1998.

Meyers, Barbara and Linda Beebe, *Archiving from a Publisher's Point of View,* The Sheridan Press: Hanover PA, 1997.

Morris, Sally, *Metadata – next steps,* presented at BIC/BL Seminar on *Trading Electronic Content* held on 10 March 1998 and available at http://www.bic.org.uk/nextstep.rtf.

Nature – *E-optimism on a tide of red ink* from Webdebates on *The future of the electronic scientific literature* available at http://www.nature.com/nature/debates/e-access/Articles/opinion2.html. ["This article is a slightly expanded version of that published in the print edition of Nature" but no reference to print nor any date of compilation is given in this web version]

Nixon, William J, *DAEADALUS: Freeing Scholarly Communication at the University of Glasgow,* Ariadne issue 34 December 2002/January 2003 available at http://www.ariadne.ac.uk/issue34/nixon.

OAI (1) = Open Archives Initiative *Frequently Asked Questions* [2002], available at http://www.openarchives.org/documents/FAQ.html

OAI (2) = Open Archives Initiative *Protocol for Metadata Harvesting* [= **Van de Sompel (3)]**

OAIS (1) [Consultative Committee for Space Data Systems], *Reference Model for an Open Archival Information System (OAIS), BLUE BOOK,* January 2002, available at http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf

OAIS (2) [Consultative Committee for Space Data Systems] *Producer-Archive Interface Methodology Abstract Standard, RED BOOK*, December 2002, available at http://wwwclassic.ccsds.org/documents/pdf/CCSDS-650.0-B-1.pdf

OCLC/RLG Working Group on Preservation Metadata, *Preservation Metadata: The OAIS Information Model,* OCLC 2002. Available at http://www.oclc.org/research/pmwg/pm_framework.pdf

Owen, Lynette, *Selling Rights 4e.* London: Routledge, 2001

Odlyzko, Andrew, *The Economics of Electronic Journals,* First Monday Vol.2 No.8 - August 4th. 1997 available at http://www.firstmonday.dk/issues/issue2_8/odlyzko/index.html

Pandora = National Library of Australia, *Guidelines for the Selection of Online Australian Publications Intended for Preservation by the National Library of Australia,* available at http://pandora.nla.gov.au/selectionguidelines.html#Introduction.

Paskin (1), Norman, *The DOI ® Handbook Version 2.0.0 March 2002,* available at http://www.doi.org/handbook_2000/index.html

Paskin (2), Norman, *Digital Object Identifiers,* a preprint from the author (n.paskin@doi.org) of an article in an ICSTI Seminar: *Digital Preservation of the Record of Science (February 14/15 2002)* subsequently published as Information Services and Use, Volume 22 numbers 2 and 3, 2002, pages 97-112

Paskin (3), Norman, *Digital Object Identifier: Implementing a standard digital identifier as the key to effective digital rights management,* The International DOI Foundation, 2000, available at http://www.doi.org/doi_presentations/aprilpaper.pdf

Peek (1), Robin P., and Gregory B. Newby, eds. *Scholarly Publishing: The Electronic Frontier*. Cambridge, MA: The MIT Press, 1996.

Peek (2), Robin P, *Where is Publishing Going? A Perspective on Change"*, Journal of the American Society for Information Science 45 December 1994)

Plutchak, T. Scott, *Sands shifting beneath our feet*", Journal of the Medical Library Association , April 2002, 90 (2) 161-163, available at http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=100760.

Pullinger, Pullinger D and C Baldwin, *Electronic Journals and User Behaviour* London: Deedot Press, 2002.

Recommendations of the Second ICSU-UNESCO International Conference on Electronic Publishing in Science Paris 20-23 February 2001. Available at associnst.ox.ac.uk/~icsuinfo/recom01.htm.

Regier, Willis G, *Electronic Publishing is Cheaper*. Chapter 9 in Ekman and Quandt cited above.

RLG (1) [Research Libraries Group]-OCLC Report, *Attributes of a Trusted Digital Repository: Meeting the Needs of Research Resources*

RLG (2). *Developing a Digital Preservation Strategy for JSTOR, an interview with Kevin Guthrie,* Editor's Interview in RLG DigiNews 4 (4) August 15 2000. Available at http://www.rlg.org/preserv/diginews/diginews4-4.html.

Roberts, Peter, *Scholarly Publishing, Peer Review and the Internet.* First Monday Vol. 4 No. 4 - April 5th. 1999, available at http://www.firstmonday.dk/issues/issue4_4/proberts/index.html

Roosendaal, Hans E. and Peter A. Th. M. Geurts, *Forces and functions in scientific communication: an analysis of their interplay,* contained in M. Karttunen et al (eds), *CRISP 97 Cooperative Research Information in Physics,* available at http:/www.physik.uni-oldenburg.de/conferences/crips97/roosendaal.htm.

Rothenberg (1), Jeff, *Preserving Authentic Digital Information.* Available at www.clir.org/pubs/reports/pub92/rothenberg.html and contained in **Smith (1)** cited below.

Rothenburg (2), Jeff, *Avoiding Technological Quicksand: Finding a Viable Technical Foundation for Digital Preservation.* Council on Library and Information Resources 1999.

RSLG [Research Support Libraries Group], *Final Report* [to the four UK higher education funding councils, the British Library and the national libraries of Scotland and Wales], 2003, available at http://www.rslg.ac.uk/

Rust (1), Godfrey and Mark Bide, *The <indecs> metadata framework: Principles, model and data dictionary,* 2000 available at http://www.indecs.org/pdf/schema.pdf

Rust (2), Godfrey, *The fire and the rose: An Integrated model for Descriptive and Rights medata,* 1998, available at www.bic.org.uk/rights.html

Rzepa, Henry and Peter Murray-Rust, *A new publishing paradigm: STM articles as part of the semantic web,* Learned Publishing 14:3 July 2001 available at http://antonio.catchword.com/vl=73071230/cl=14/nw=1/rpsv/cgi-bin/linker?ini=alpsp&reqidx=/catchword/alpsp/09531513/v14n3/s3/p177

Sandewall, Erik, *Open Reviewing, Closed Refereeing: Where's the Publication.* This is chapter 25 pages 285-303 of **Fredrikkson** see above.

Schutz, Bernard F and Theresa Velden, *Living Reviews in Relativity – A Living Electronic Journal,* a contribution in **Kircz (1)** and available at http://www.niwi.knaw.nl/ccsc/talks/schutztalk.htm.

Sealed Media: *Core Technology White Paper,* 2001 at www.sealedmedia.com

Sellen, Abigail.J and Richard H.R.Harper, *The Myth of the Paperless Office,* Cambridge MA: The MIT Press 2002

Shaw, Dennis and David Price (eds), *Economics, real costs and benefits of electronic publishing in science – a technical study,* ICSU Press,1998. Available online only at www.bodley.ox.ac.uk/~icsu.

Shoffner, Ralph M, *Appearance and Growth of Computer and Electronic Products in Libraries.* Chapter 12 pp 209-256 in Richard E. Abel and Lyman W. Newlin (eds), *Scholarly Publishing: Books, Journals, Publishers, and Libraries in the Twentieth Century.* New York: Wiley, 2002.

Silverman, Robert J, *The Impact of Electronic Publishing on the Academic Community.* Chapter 3 (pages 55-70) of **Peek** and Newby (see above)

Shum, Simon Buckingham, *JIME: an interactive journal for interactive media,* Learned Publishing 14:4 October 2001 pages 273 – 285 available at http://elvira.catchword.com/vl=34159161/cl=41/nw=1/rpsv/cgi-bin/linker?ini=alpsp&reqidx=/catchword/alpsp/09531513/v14n4/s5/p273

Smith (1), Abby (ed.), *Authenticity in a Digital Environment.* Council on Library and Information Resources, 2000. Available at www.clir.org/pubs/reports/pub92/contents.html.

Smith (2), Abby, *Authenticity in Perspective.* In Smith (1) cited above and available at www.clir.org/pubs/reports/pub92/smith.html.

Smith (3), Philip N *et al.*, *Journal Publishing with Acrobat: the CAJUN project.* Electronic Publishing 6(4),481-493 (December 1993). Available at http://cajun.cs.nott.ac.uk/cgi-bin/getpaper?paper=compsci/epo/papers/samples/ep6x4pns.pdf

Smith (4), Richard, *Maintaining the Integrity of Electronic Publications: Potential Problems and Possible Solutions I* in **AAAS** at http://www.aaas.org/spp/dspp/sfrl/projects/epub/ses3/smith.htm

Spivey, Catherine, *Online Archiving with the British Library – The Emerald Experience,* Serials, Volume 15, no 3, November 2002.

Thomas, Charles F and Linda S. Griffin, *Who will create the Metadata for the Internet,* First Monday 1988 3:12 at http://www.firstmonday.dk/issues/issue3_12/thomas/index.html

UKOLN (organiser), *Long Term Preservation of Electronic Materials,* A JISC/BRITISH LIBRARY WORKSHOP 27th and 28th November 1995 at the University of Warwick,(British Library R&D Report 6238). Available online at http://www.ukoln.ac.uk/services/elib/papers/other/preservation/intro.htm.

Van de Sompel (1), Herbert and Carl Lagoze, *The Sante Fe Convention of the Open Archives Initiative,* D-Lib Magazine February 2000 6:2 available at http://www.dlib.org/dlib/february00/vandesompel-oai/02vandesompel-oai.html

Van de Sompel (2), Closing Keynote Address at CNI Fall 2000, available at http://www.cni.org/tfms/2000b.fall/handout/HVDS-CNI-2000Ftf.pdf

Van de Sompel (3), and Carl Lagoze, *The Open Archives Initiative Protocol for Metadata Harvesting.* Version 1.1 2001. Available at http://www.openarchives.org/OAI_protocol/openarchivesprotocol.html.

Waters (1), Donald and John Garrett (Task Force Co-Chairs), *Preserving Digital Information: Report of the Task Force on Archiving of Digital Information,* Committee on Preservation and Access and Research Libraries Group 1996, available at http://www.rlg.org/ArchTF/

Waters (2), Donald, *Good Archives Make Good Scholars; Reflections on Recent Steps Toward the Archiving of Digital Information* in *The State of the Art of Digital Preservation: An International Perspective,* Conference Proceedings, Document Abstracts Inc (DAI) Institutes for Information Science published by Council on Library and Information Resources (CLIR) 2002. Available at http://www.clir.org/pubs/reports/pub107/waters.html

Watkinson (1), Anthony, *Electronic Solutions to the Problems of Monograph Publishing.* London: Resource: the Council for Museums Archives and Libraries, 2001, available from http://www.publishers.org.uk.

Watkinson (2), Anthony, *Developments in Global Academic Publishing,* pp. 30-36, in *Publishing 2001: attitudes to technological change.* London: Bookseller Publications, 2001.

Watkinson (3), Anthony, *The STM Information System: An analysis.* Learned Publishing (1999) 12 (1), 11-24. Available at http://gottardo.catchword.com/vl=11944279/cl=17/nw=1/rpsv/catchword/alpsp/09531513/v12n1/s3/p11

Watkinson (4), Anthony, *Journal Publishing: current issues.* Learned Publishing (1999) 12 (2), 93-96. Available at http://gottardo.catchword.com/vl=11944279/cl=17/nw=1/rpsv/catchword/alpsp/09531513/v12n1/s3/p11

Watkinson (5), Anthony, *Trends in Journal Subscriptions 1998.* London: The Publishers Association, 1999.

Watkinson (6), Anthony, *What's so special about not-for-profit publishers?* Learned Publishing (2001) 14 (4) 313-4. Available at http://elvira.catchword.com/vl=9960770/cl=18/nw=1/rpsv/cgi-bin/linker?ini=alpsp&reqidx=/catchword/alpsp/09531513/v14n4/s13/p313

Watkinson (7), Anthony, *The Role of the Publisher in Scholarly Communication,* a contribution to International Conference on Academic Presses and Scholarly Communication, Florence March 22nd 2001 published by Florence: Firenze University Press (2002) online only and available at http://biblio.unifi.it/documents/archivio1/00/00/01/37/unifi00000137-00/International_Conference_on_Scholarly_Comm.pdf

Weinberger, Ellis, *Digital objects as manuscripts. How to select material that is born digital for long time preservation,* 1999, available at http://www.cus.cam.ac.uk/~ew206/d-as-m-article/

Williams, Don, *Image Quality Metrics,* RLG Diginews 4(4) August 2000. Available at http://www.rlg.org/preserv/diginews/diginews4-4.html

Wittenberg, Kate, *Transformational Publishing: Threats and Opportunities in the Digital Word.* PSP Bulletin 2:3 Pages 2-4 Fall 2001 available at www.pspcentral.org.

XML Canonicalization Requirements. At the time of writing the latest version is available at **http://www.w3.org/TR/NOTE-xml-canonical-req**