Version identification: a literature review

Louise Allsop, Anya Somerville and Frances Shipsey
(additional editing Dave Puplett)

November 2007

Document history	
Authors	Louise Allsop, Anya Somerville and Frances Shipsey with Dave Puplett
Version	1a
Date	30 November 2007
Circulation	Project partners; submitted to JISC

Contents

Introduction 3

1. Multiple versions in a literary context 5

2. Multiple versions online and digital repositories 10

3. Concepts of bibliographic organisation 16

4. Software and versioning: systems, tools and solutions 20

   4.1 Version control systems (VCS) 20
   4.2 Content management systems and wikis 21
   4.3 Word processors 21
   4.4 Tools and protocols relating to version identification 22
   4.5 Proposed software solutions for versioning problems 22

5. Open access and digital repositories 28

6. Conclusions 30

7. References 32

Introduction

"Photographs, data and key documents are difficult to find, and when a document… is located, forensic work is needed to determine if the document is the 'right' version. Each project triggers another expensive investigation."

Arnold (2002) encapsulates the issues underpinning the Joint Information Systems Committee (JISC)-funded VERSIONS Project: the continual steep increase in online digital content and the consequent difficulties of locating specific versions on the web. As technological developments have made both the editing of existing versions and the dissemination of a variety of versions less complex, the issues associated with the management of multiple versions have been exacerbated. Conceptual issues, relating to how works encompassing numerous manifestations should be viewed, and more practical issues, including the selection, management and identification of versions for printing, dissemination, and preservation have long been considered and discussed. This literature review introduces a selection of papers that touch on the issues associated with version identification, and have in some cases proposed solutions to those issues.

The problems of identifying and preserving multiple versions of electronic documents are often considered in the light of technical issues such as media and format. Buckland (1997) feels that this technical focus has restricted our understanding of documentation. Taking this assertion as a starting point, Basden and Burke (2004) utilise the philosophy of Herman Dooyeweerd in an attempt to frame a common sense approach to understanding the nature of documents, and so shed light on tackling issues of identity and change. Rather than recount in detail the Dooyeweerdian philosophy regarding the meaning of things - as explained by Basden and Burke - we will attempt to describe only those aspects specifically pertinent to the reach of the VERSIONS project.

Basden and Burke describe Dooyeweerd's philosophy as 'meaning' rather than 'being' oriented. Applied in the context of understanding documents and versions of a work, they say the philosophy directs us to consider "what a document means (in a human context) rather than what documentation is itself". This philosophical approach aims to establish the "ontic boundary of documentation" and necessarily involves acknowledging that documents are subject to change, both during their creation and as a consequence of their use. It also necessitates understanding the context of the documents, and their links with the culture in which they are produced and used: "In veritable naïve experience things are not experienced as completely separate entities" (Dooyeweerd, 1955, Vol. III, p. 54).

The human context in which documents can find their various meanings is explored through the use of Dooyeweerd's "suite of aspects". The aspects describe the types of meaning a thing may have in different spheres and are used here to differentiate the ways in which a document functions, and illustrate the differing identities it can have when the roles of user, author and librarian are considered. They include, as outlined by Basden and Burke:

- the medium (physical aspectual structure);
- the shapes or signals that our senses can detect and our brains process (sensitive

aspectual structure);
- the raw symbols (analytical aspectual structure);
- the structure of the symbols (formative aspectual structure); and
- the content carried by the symbols, whatever it is (lingual aspectual structure).

The author's concern with content, the user's concern with their specific purpose in using the document and the librarian's concern with managing, storing and locating the document are all aspects of this structure and of the meaning of the document. Furthermore, just as a document can change over time so too will these concerns. As Basden and Burke summarise: "a document has a range of aspects that give it meaning, [and] change can occur in one or more of those aspects." Changes in the sense of versions of a work held electronically can for example occur linguistically, in terms of form, as well as in aspects relating to content.

Basden and Burke do not make any specific pragmatic recommendations regarding the management of versions of a work, rather they offer a useful framework for understanding the nature of documentation in terms of function, identity and change. The usefulness of understanding the functioning of a document, and its versions across a range of aspects, is that it allows us to recognise the versions of a work as individual entities which make up the whole of the work, and also to distinguish evolution and variability as well as that which remains constant. Acknowledgement of the inherent value of versions of a work in this way can be brought to bear upon the issues of responsibility, identity and change in managing, storing and locating versions in academic publishing. In the next section we review some investigations that have been carried out on this subject in the field of Literature, which offer interesting insights that can be usefully applied to similar concerns in academic publishing.

This literature review was conducting during 2006-2007 by the JISC-funded VERSIONS Project, as part of the investigation into the extent of the issues relating to version identification and of the need for standards in this area. Searches were made in the following databases and resources: *Library and information science abstracts (LISA), LISTA, ACM Digital Library, Wikipedia*. Search terms used included 'version identification', versioning, versions, 'version control'. In addition to systematic searches, citations were followed up and search engines were also used to uncover relevant material. The review is selective and represents a snapshot at a particular point in time. Conscious efforts were made also to review the literature beyond the fields of open access and digital repositories, in order to learn lessons from other disciplines such as literary criticism, traditional cataloguing and bibliographic organisation, and from software development.

1. Multiple versions in a literary context

The difficulties associated with version management existed long before the digital age. The work of the Romantic poets offers a case in point. Certain Romantic poets revised their work repeatedly, leading not only to a plethora of publicly available versions, but also to a series of problems in dealing with them. Stillinger (1994) considers the work of Coleridge in this regard, identifying ninety-four extant versions of the seven poems analysed. Stillinger notes that multiple versions raise questions about the ontological identity of works. The example he gives is whether the title *The Ancient Mariner* should refer to a single version of the poem or to all versions taken together? If a work is to be recognised as all of its versions together; should it be constituted by the process of its revisions, one by one, or by all versions existing simultaneously? Stillinger notes that, historically, editorial theory has assigned value to a particular version, that which is perceived as the best, most correct or most authoritative version. This approach is dependent on the existence of that single, valuable version. In common with Basden & Burke's recommendation that versions should be understood as having coherence within the whole of a work, Stillinger's paper suggests a pluralistic approach to versions; whereby there is no designated best, correct, or most authoritative text. Stillinger terms this approach 'textual instability'. This describes the idea that all versions have a value in their own right, each exhibiting the character and authorial intention of the time at which they were written, with no one version being more valuable than any other. This, according to Stillinger, is a more appropriate way to view the poetry of Coleridge:

> *"With writers like Wordsworth and Coleridge, obsessive revisers who lived for three or four decades after first drafting their most admired poems, numerous versions compete for attention; and a theory of textual pluralism – allowing that every separate version has its separate legitimacy and that all authoritative versions are equally authoritative – seems a much more appropriate way to regard their writing than an approach that ignores its extended textual history."*

Coleridge wrote eighteen versions of *The Ancient Mariner*, so how can one be viewed as the 'real' or 'most authoritative' version? This issue is particularly pertinent to literary writing, where the content is not strictly factual; authorial intention is not usually made explicit and thus cannot reliably be known. This issue is also relevant in the domain of academic writing. Allowing that every 'authoritative' version is equal, or at least significant in its own right, encourages consideration of the various stages in the life cycle of an academic research paper as a set of versions, all of which may be valuable to the end user. While the peer-reviewed, published version may be recognised as the final authoritative version, large amounts of informative content, data for example, are often excluded from published journals, as strict word limits and space constraints are imposed upon the work. Stillinger's theory of textual instability seems worthy of evaluation in the academic, as well as in the literary, sphere.

Stillinger proposes a framework by which to consider Coleridge's multiple versions, and thus encourages separation of the concepts of version and work. Version and work may have been interchangeable terms in the past, but to attempt to recognise all versions as part of the same work necessitates making some distinction between the two. Stillinger suggests that a work consists of all known versions, and can be physically embodied only in the form of its versions; one version cannot be viewed as the work, as there are

no logical grounds for excluding others. Therefore versions are viewed as particular instances of a larger concept, the work. The notion of the work becomes an organisational tool; referring to a common core of intellectual matter, to which various versions can be related.

In the context of academic papers, the organisation of versions around the common core of the research upon which they are based could have valuable implications for those seeking information resources. To source one version and find directions to others would certainly assist with resource selection. As already noted, early versions of academic papers can be of real value to researchers, therefore it could be argued that to exclude these from a repository in favour of published articles needlessly limits the potential intellectual worth of that repository. Certainly, journal articles can be viewed as superior in the key respect of having undergone peer review. And yet, if we consider the material potentially excluded - background information, detailed results and so on - a strong case for the validity and value of pre-published versions can be made. Considering all versions as the true embodiment of the work and in doing so making the versions accessible is useful to those wishing to gain a fuller grasp of a strand of research, and valuable in terms of preserving for posterity the evolution of that work.

Staying with Romantic literature, Mary Shelley's *Frankenstein* is a specific example of the value to posterity of versions of a work. Marilyn Butler, editor of a 1993 edition of the 1818 text, describes the novel as "the most protean and disputable of even Romantic texts". *Frankenstein* has had a complex publishing history, published initially in 1818 and again in 1823, before a third edition with substantive revisions by the author was published in 1831. According to Butler the third edition has largely been preferred to earlier editions: "…its version of the plot has been the standard one since 1831". Butler summarises the thinking behind the preference for the 1831 edition as an acknowledgement of authorial control over a text, that the changes made have literary value in terms of developing characterisation and improving the prose, and most interestingly, that the changes reflect what the story had come to mean to its readership.

The intervening years between the first and third editions saw debate intensify around the issues raised by Shelley in *Frankenstein*. Butler describes how the middle classes in particular were preoccupied with the threat radical science posed to the moral, religious heart of society and the "campaign to regulate 'family reading' […] was running full steam". Contemporary stage versions of the novel imbued the story with a religious morality that was not present in the original text. As Butler describes it, in making changes Shelley "…submitted to middle class opinion", and "...her cuts and rewriting were acts of damage limitation rather than a reassertion of authority". This evaluation of the significance of the changes differs markedly from Shelley's own assessment. Writing in the Introduction to the third edition, Shelley states that the changes are:

> *"[…] principally those of style. I have changed no portion of the story nor introduced any new ideas or circumstances. I have mended the language where it was so bald as to interfere with the interest of the narrative; and these changes occur almost exclusively in the beginning of the first volume. Throughout they are entirely confined to such parts as are mere adjuncts to the story, leaving the core and substance of it untouched."*
> Shelley 1831, in Butler 1993

Butler makes a strong case for preferring the 1818 edition, one that supports the validity of retaining earlier versions of works in general, and has implications which relate to the differing ways in which works function. For literary and sociological historians for example, the publication of the earliest version of a work accompanied by an appendix detailing subsequent additions and omissions better enables an understanding the evolution of the work. In the case of *Frankenstein,* Butler says*:*

> *"…[it] is only when the relations between the two versions are expressed this way round that readers can see how much turns on the addition of religious attitudes and judgements, and the cancellation or reinterpretation of the science".* Butler 1993

It is possible then to discern not only the changes but also to relate them to the context in which they were made. Readers unaware of the earlier edition miss the opportunity to take advantage of what Butler feels is the "more important and serious" version. The differing assessments of the significance of the changes as expressed here by Shelley and Butler highlight the impossibility of predicting the future value of a work. Application of this lesson to academic publishing makes it clear that while referees and publishers may feel they have made the correct critical judgement in cutting, editing or refining a work, posterity may take a different view. To best serve the needs of future scholars, and given the fact that research is often publicly funded, it is in our interest to preserve full collections of earlier versions of academic works, at least in cases where authors have disseminated such early versions.

However, even confidently identifying what constitutes a new version of a work is an issue of complexity, and this is also discussed by Stillinger. For example, when does a version become a new version, and when does a version become a new work? Stillinger considers Coleridge's *Dejection: An Ode*, versions of which differ substantially in theme, structure and style, with early versions around two and a half times the length of later ones. Strictly speaking, two documents with a single punctuation variant could be considered as separate versions, although the majority of readers may not be interested in alternative versions until a change in meaning or aesthetic effect has occurred. Stillinger suggests that such decisions are arbitrary.

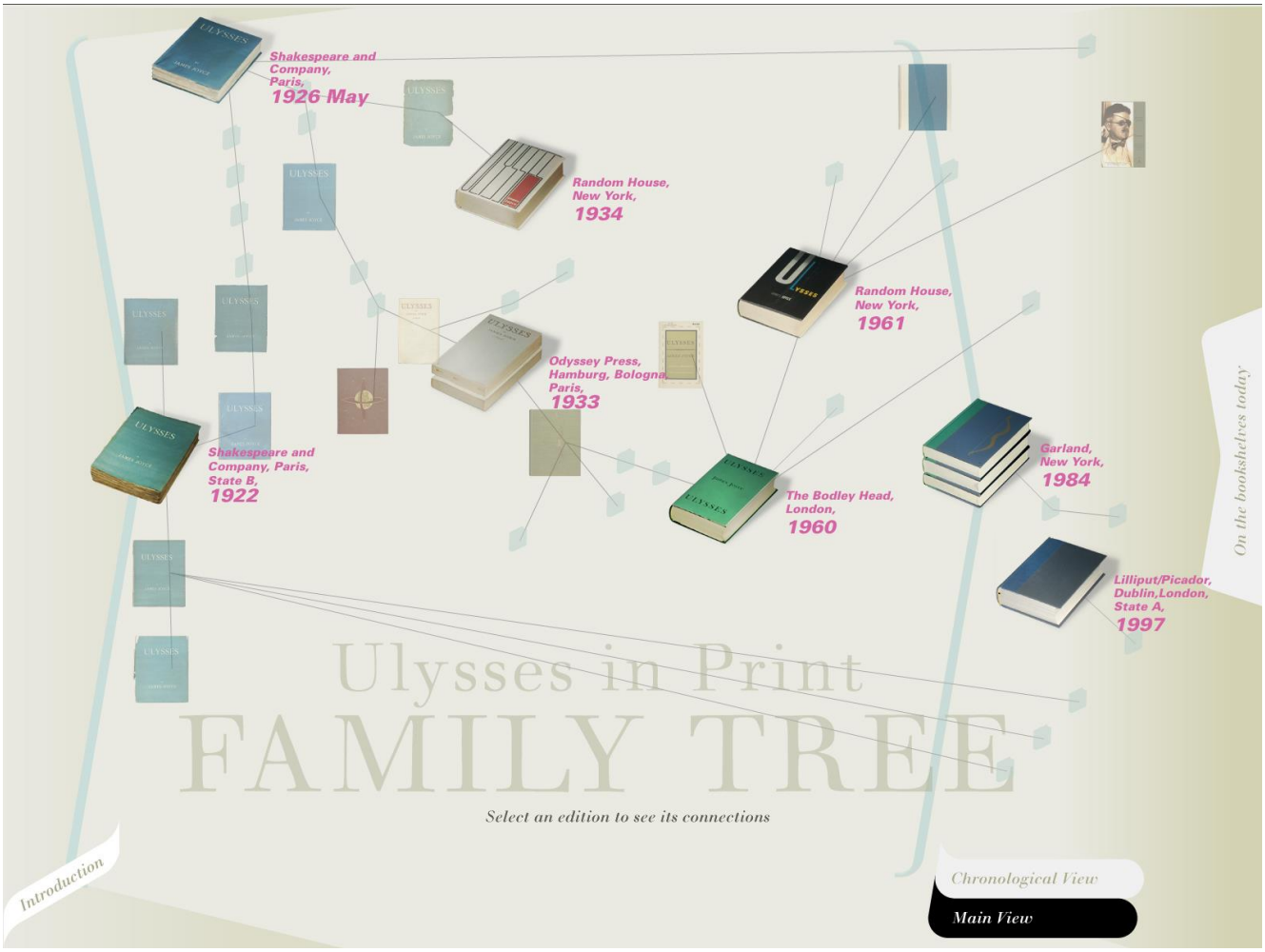
A further complication is that different user communities may use different parameters in determining whether one document is a different version of, or is in effect identical to, another. Stillinger proposed that implementing a set of definitions relating work to version could solve the problem of defining the version. The important factor seems to be that users are at least made aware of these definitions, in relation to the particular resources with which they are working. For an example of such a definition in the context of storing works in an institutional repository, the administrator could define documents which differ by more than a single line of text (or other threshold) as different versions. In this way users would then be able to make an informed decision about whether to read one, some or all of the existing versions.

Providing multiple versions of a work alongside one another electronically is of course fairly straightforward. However in the realm of traditional publishing the practical issue remains that editors must ultimately select a single version for printing. This problem was noted by Stillinger and also recognised by Tetreault (1997). Tetreault focuses on Wordsworth, as some of his poems were modified repeatedly over a 50-year period, resulting in markedly different works from those that had preceded them. In accordance

with Stillinger's findings, Tetreault's paper recognises that to select, use, or cite one version and thus exclude all other versions, is to deem that version more important than all prior or subsequent versions, perhaps inappropriately. This is a corollary of the medium of print for the dissemination of work, and consequently Tetreault views print as an insufficient medium by which to represent the evolutionary process of text development: "establishing a text in print has always meant that we must privilege one version over others, and settle for a static representation of what might be better understood as a dynamic process" (Tetreault, 1997). He views texts as transient and evolutionary; characteristics that can be seen to transcend the divide between literary and academic writing. As Stillinger and Tetreault explain, literary texts can evolve over many years. Likewise academic papers will likely progress through a series of versions before appearing in a peer-reviewed journal. The author's continuing research on a topic may also lead to further development of the intellectual content after publication in a journal. Tetreault notes that the electronic medium can reflect the changing nature of text, and so progression through versions, more effectively than print can.

Figure 1 below provides a practical illustration of how multiple versions can appear in part owing to actions on the part of the publishers. In common with *Frankenstein*, James Joyce's *Ulysses* has had a complex publication history. In this case, the proliferation of published incidences of *Ulysses* was largely due to publisher activity. The 1922 edition, (far left), was printed with numerous mistakes, as Joyce was writing and revising until the very last minute. Consequently, a series of re-prints were issued, as represented by the small volumes surrounding the 1922 edition. These re-prints contained errors and errata slips, and so a series of different versions were in circulation relatively soon after the publication of the first edition. A 'definitive' edition was published in 1926 (top left); subsequent editions should, therefore, flow from this one. Bodley Head's edition of 1960 and the edition published by Random House in 1961 do show a clear lineage. The Random House 1934 is an interesting example of an unofficial version entering the published canon. The small volume which separates the 1926 and 1934 Random House editions in the diagram represents a pirate edition issued in the US during the period when *Ulysses* was banned. Following the lifting of the ban Random House, wishing to publish an official version, inadvertently came into possession of such a pirate version. The pirate version was then republished 'officially', replete with all the mistakes of the pirate version. This example illustrates two important issues: that even in print, multiple versions come into existence relatively quickly and perhaps inevitably; and that tracking or labelling versions correctly is therefore necessary in order to prevent confusion regarding the status of versions.

Figure 1: The complicated publishing history of James Joyce's Ulysses



Reproduced with permission from University College Dublin and XCommunications

The value of multiple versions has been demonstrated, both in terms of their significance as an integral part of a work and in terms of the contribution a version makes in understanding the evolution of a work. Discarding or privileging particular versions of works is a function of editorship. The editorial role clearly has a value in discriminating between versions and steering works towards quality. Authors themselves have rights in determining what is disseminated in the first place. Readers, while perhaps not having rights, do at least have an interest in accessing different versions of works, particularly where these have been disseminated. Assessing the potential value of different versions to scholarship in the future is a difficult task.

Digital repositories offer an unparalleled opportunity to improve on the limitations of print in organising, tracking and preserving versions. As development of the electronic medium for document preparation continues to facilitate the production and circulation of multiple versions, the inevitable proliferation of versions underlines the pressing need to make identification of versions possible.

2. Multiple versions online and digital repositories

As we have seen, multiple versions and versioning issues have existed for centuries. We will now focus on versions of academic papers online and their discoverability in open access institutional repositories.

Relatively recent technological developments have opened up many possibilities for the dissemination of academic research papers. While authors still publish their findings via traditional peer-reviewed journals, they now also have the option of using multiple, alternative dissemination channels, many of which are electronic. Web-based options include the posting of material on personal or departmental websites, or in subject or institutional repositories. For the author, these channels can offer opportunities for both broader and faster dissemination of their research. For the researcher as reader, there is the benefit of increasing amounts of freely available material in addition to the material published in peer-reviewed journals.

One major development in online publishing is the Open Access movement, where articles, at almost any conceivable stage of their development and creation are posted online without restrictions on access from the point of view of the reader. While this has undoubtedly increased opportunities to access some material, it has also brought more versions of academic work to the public eye. Richardson (2005), writing from the publisher's perspective, urges caution here, describing how institutional repositories generally aim to include post-prints (author created post-refereeing but pre-copyediting and proof corrected versions). The proof corrected version will usually differ, albeit in minor detail, from the published journal article. Following the addition of a post-print version to a repository, and the work's publication in a journal, two versions of that work are publicly available. The author may have deposited a pre peer-review version in a preprint server, or elsewhere. When we consider that the same research may have been presented at conferences, or released as discussion or working papers at different stages in the writing process, and the numerous other dissemination channels through which any one of these versions could have been released, it is possible to see the potential for the online availability of many different versions.

Richardson considers how far this multiplicity of versions really improves accessibility of information, asking "will it simply add to the burden faced by readers looking for ways to avoid drowning in information?" The response to this question could be that for readers with no or little access to subscription journals, open access versions might represent not so much a danger of drowning as a salvation from drought. Quint (2006) echoes Richardson's concern, focusing particularly on how readers searching for information will be able to discern the quality of the versions that they locate online. Quint's concern, that less evaluated content may reach more users than expertly reviewed content, is especially valid if users are working outside a university or research environment in which subscription journals are readily available. That users may be unaware of the existence of such 'official' content, or that what they are reading may differ from it, shifts the focus of the issue slightly, from the availability of numerous online versions to the organisation, collocation and labelling of these versions.

Rumsey (2006) describes how institutional repositories can organise information; for dealing with varied publication types, multiple versions, and for expressing relationships

between them. With this in mind, in contrast to exacerbating the version identification problem, repositories could be viewed as a solution to navigating through the mass of online information at an institutional level. The importance of relationships between versions, and of expressing these coherently, was addressed earlier in this review. Rumsey highlights the potential for repositories to fulfil this task, for example by acting as a pointer to publisher versions.

Johnson (2001) makes the point that in academic publishing, the 'official' version of articles remains the peer-reviewed version, links to which will be embedded in repository versions. Although, as Richardson points out, repositories encourage the existence of parallel public versions, in a more positive light they can also be seen to address the version identification issue, by linking versions together. Furthermore, the link is likely to be highly visible on search engines due to compliance with international metadata standards, and probable registration with search mechanisms.

Markland (2006) investigated the visibility to researchers of material held by twenty-six institutional repositories in the UK. According to Markland, in order to achieve the aims of Open Access, depositing material in repositories is only the first step: "…[it] must also be possible to find, retrieve and use the scholarly content". Markland examines the effectiveness of repository search interfaces along with Google and Google Scholar, at retrieving items from the sample repositories. One item held by each repository was selected and searches for these were conducted using the complete title and keywords or a phrase from the title. The results of the study highlight issues pertinent to version identification, particularly that of multiple versions existing online.

In her discussion of the versions of the item yielded by each search, Markland often uses the phrase "appears to be" in evaluating what each version might be, indicating that there is a consistent failure to adequately, visibly identify the version. Markland explains how Google Scholar for example displays different versions collected together in 'one hit'. This means that users have less to trawl through, but she observes that "the 'collection' of different versions still requires careful evaluation".

Markland also notes the issue of users 'satisficing', settling for a quickly found or higher ranked item which is not actually the most appropriate or best for them to use. It may also be that a user finds a copy available for purchase and pays for access where they could have found a freely available copy with further searching. As Markland remarks, "a repository may showcase the intellectual output of the institution, but [that] does not of itself mean that anyone 'on the outside' will come looking to see what is there". A high number of bad or unhelpful descriptors were found such as 'e-print type-article'. Special mention is made of a document retrieved from the University of Edinburgh's Edinburgh Research Archive where the following information has been included on the title page: "This is a pre-copy-editing, author-produced PDF of an article accepted for inclusion in…published by John Wiley and Sons Ltd following peer review". Both the detail of the information and its visibility, as opposed to having such information simply contained in the metadata accompanying the document, are of considerable value to the user. Markland also suggests that given the relative novelty of pre-prints in some disciplines, citation information might also be usefully included.

Markland concludes her study with a number of comments on how repositories can improve the process by which users retrieve and assess versions. As already discussed, clearly indicating the status of a document is key. Finally she puts forward the possibility

that search engine and repository interfaces might develop in the future to incorporate the functionality to identify versions as the range and nature of available versions increases.

Illustrating the continued regard for peer review, Swan (2003) quotes that 94% of almost 1250 authors who responded to a survey on behalf of ALPSP see the process as 'very important' or 'important'. Thus, users require an effective method of clearly identifying version status in relation to *refereed* journal article versions. Current terms for describing stages in the publication process are, McCulloch (2006) suggests, creating confusion rather than clarifying the identification of versions. McCulloch cites the problem highlighted by SHERPA that the common terminology is not standardised in meaning or interpretation; authors and publishers seem to hold different views about what constitutes a preprint. What the digital repository community refers to as the 'post-print' (the version containing revisions following peer-review) may still undergo updates at proof stage, hence for publishers it remains a preprint.

The lack of a standardised vocabulary to identify versions, in relation to the peer reviewed and published versions, will contribute to difficulties for users when it comes to selecting which versions they wish to read. Effective terms to express relationships between versions could benefit all stakeholders. The final report of the RIVER study (Rightscom Ltd and partners, 2006) identifies a useful although tentative terminology for the identification of versions that acknowledges the need to reflect the lifecycle of a journal article in revealing the relationship between versions to users. The RIVER report also refers to the work of the joint NISO/ALPSP Working Group on Versions of Journal Articles, which has proposed (NISO/ALPSP 2006) possible terms and definitions for metadata used in association with document objects to help differentiate between versions. The terms are 'Author's Original', 'Accepted Manuscript', 'Proof', 'Version of Record', and 'Updated Version of Record'. The working group focussed on journal articles rather than other related document types such as working or conference papers. Nevertheless, their proposals are useful for illustrating the relationships between journal articles and instances of other document types.

The status of online material as an equal version to the print copy of academic research is also not fully established. Lynch (2003) highlights a historical tendency for journal publishers to view non-textual materials as add-ons to the published journal article. Whilst accepting that scientific journals are beginning to recognise additional materials, source data for example, as an integral part of the publication, Lynch states that there is still uncertainty over the commitment that publishers are prepared to make to incorporate such material into the permanent scholarly record. The article suggests that, due to the massive diversity of scholarly material, disciplinary repositories will find it difficult to accommodate supplementary materials for all research undertaken. Such a task "must be worked out also in the evaluation, tenure, and promotion practices in place at an institutional level" (Lynch 2003), where faculty can be involved in the process. Used in this way, institutional repositories can be seen as complementary to traditional publication channels, rather than a substitute for them. Institutional collections can manage, disseminate and preserve resources which other dissemination channels may not be able to cope with. As a result users are able to view these resources in conjunction with traditionally published material.

Van de Sompel et al (2004) take Lynch's notion of institutional repositories supplementing the functions of traditional journal publishing in scholarly communication

a step further. They consider some of the issues relating to the variety of work online, the distributed nature of the storage of that work, and the possibilities for improved scholarly communication. Beginning with the premise that the established system has not kept pace with changes in research practice and dissemination, their paper outlines proposals for an improved technical functionality to meet the demands of the changing nature of scholarly research and scholarly communication.

They suggest how versions of a work can be better integrated into a scholarly communication system. Initially, in order to facilitate such integration, a change of perspective is needed in how what they term a 'unit of communication' is perceived by systems. The system should, according to the authors, consider datasets, simulations, software etc. representations as units of communication in their own right. It ought to be able to cope with complex documents regardless of their format, constitution or location. Finally, a system should allow all units in the system, no matter what their stage of development to be registered quickly into the system.. This would facilitate collaborative network-based endeavours and increase the speed of discovery, just as text documents are currently treated.

Van de Sompel et al see the institutional repository as one type of 'hub' for scholarly communication, existing alongside other 'hubs' such as ArXiv and CiteBase. These 'hubs' implement the required functions of scholarly communications such as registration, certification, awareness, archiving and rewarding. Their vision is of a "distributed service approach" to scholarly communication, in which the required functions are implemented in particular ways by many parties, as opposed to the more vertical, fixed system of traditional journal publishing. The traditional system lacks the flexibility to trace the evolution of a work through the scholarly communication system. Van de Sompel et al assert that this is a significant failure, one that:

> *"needs to be remedied in a future scholarly communication system by natively embedding the capability to record and expose such dynamics, relationships, and interactions in the scholarly communication infrastructure".*
> Van de Sompel et al (2004)

As others have done, they note the intrinsic value of versions of works from the point of view of the researcher. They also note that potential for recording and identifying the evolution of work has value to the author in terms of evaluating their performance in the scholarly system.

Van de Sompel et al (2006) revisit the topic in a later article, proposing that a necessary "…technical step is the development of information models, process models, and related protocols to enable interoperability among existing repositories, information stores, and services". Van de Sompel et al re-assert their conviction that improving inter-operability in this way will "leverage the value" of the material hosted by institutional repositories by improving accessibility.

The article outlines a prototype "interoperability fabric", designed to allow the cross-repository use of digital objects and introduces the Object Reuse and Exchange (ORE) Initiative begun in October 2006 with Mellon Foundation funding. Digital objects are defined as being a data structure made up of digital data and key-metadata, which will include an identifier for the digital object.

Van de Sompel et al take the view that "digital objects are compound in nature". This can be related to Stillinger's, amongst others, assertion that versions should be viewed as particular instances of a larger concept, the work. Van de Sompel et al develop this idea, incorporating differing types of intellectual output such as media types as well as the variety of intellectual content types such as papers, datasets, simulations etc. They then take the pragmatic step of suggesting how this variety of types of work can be identified, tracked and shared, thus "leveraging the intrinsic value of scholarly digital objects beyond the borders of the hosting repository."

At the time of writing JISC is funding work on improving interoperability across digital repositories. The Scholarly Works Application Profile (SWAP) (formerly known as the Eprints Application Profile) has been developed as a Dublin Core Application Profile for describing scholarly publications held in institutional repositories, acknowledging the need to describe research papers using metadata that is more detailed and richly structured than at present. The SWAP is based on the DCMI Abstract Model and on the Functional Requirements for Bibliographic Records (FRBR). FRBR is well suited to handling versioning issues and more will be said about it in the following section.

Pinfield and James (2003) recognise that e-prints regularly contain more information than published versions. They cite the example of a piece by Steve Lawrence, an e-print written on open access, which contains more detail and data than the published version in *Nature*. Their paper suggests that the additional information contained in the e-print makes an important and different contribution to scholarly understanding; as such both versions should be preserved in their own right. This indicates that the availability of multiple versions can be advantageous to researchers as authors and readers alike.

A further benefit is the opportunity to gain early access to research that would otherwise be subject to a discipline-wide time lag between paper submission and publication in refereed journals. McKiernan (2001) also views expedited formal publication as a benefit of e-print servers. Pinfield and James' observation that there may be local priorities for preservation, relating to version type, is relevant here. If authors know that a submitted paper will not be published for some time, they are likely to be keen to disseminate conference presentations, working papers, discussion papers, pre-prints etc, in order to stake their claim on the research. Dallman et al (1994) demonstrate a clear subject-specific preference for particular versions in their paper on high-energy physics. They note that pre-prints are by far the most important source of information transfer in high-energy physics, with hard copy preprint flow comparing quantitatively to the distribution of established journals, up until the development of electronic preprint exchange.

Pinfield and James view the selection of versions for preservation as a key practical issue. It appears to be a similar dilemma to that experienced by the editors of Coleridge and Wordsworth in coming to a decision regarding which version of those works to print. The paper calls for further work in the area, and advocates the possible development of a set of criteria for preservation; preserving post-prints rather than pre-prints, or preserving e-prints where they are fuller versions of conventionally published papers, for example. The possibility of local priorities for preservation does make all-encompassing criteria more difficult. Differing publisher positions on author self-archiving could also make blanket criteria for version selection somewhat challenging. Furthermore, as Greig (2005) points out, universal requests for post-print versions may not be entirely fruitful. Discussions with potential depositors, carried out by the JISC-funded

DAEDALUS project, indicated that many authors do not retain a suitable version for deposit, and do not have time to create one.

Rusbridge (2006) looks into a storage strategy for digital preservation. There is no way of knowing what functionality future users will require, leading to a pressure to preserve documents with full functionality, which is expensive (though perhaps not by comparison with preservation of print). In the print world different user communities are served by different preservation activities. For example, the public will utilise widely available multiple editions of texts, whereas a scholar may require access to a rare early edition of a work, held in a special collection. Similarly, in the digital domain, the general public are likely to be content with fewer properties than the scholar, who may require full functionality in a digital document.

Rusbridge cites Kunze's proposal for storing versions in 'desiccated formats': documents with reduced functionality which are easier to preserve. Original data files will be retained, but will require the user to perform transformations in order to extract a fully functioning version. The extra work involved is likened to the work that is required of the scholar in order to access rarer print copies – travel to different libraries, learning languages, deciphering faded documents etc. Storing desiccated files could assist with ensuring lasting user access to a variety of versions.

Retention and preservation of suitable versions are not the only issues. Cummings (2006) writes, "the advent of digital technology is troubling scholars' ability to identify originals". Bradley (2005) elaborates, citing Ross (2005), who highlights the comparative ease with which digital objects can be altered, in comparison to print documents. As digital documents are so easily manipulated, the proliferation of new and unofficial versions is possible. Users of digital material are thus confronted with the additional task of ensuring that research sourced on the Internet is exactly what it purports to be.

Authenticity and integrity, while a necessary consideration for Internet users generally, could be considered less of an issue for users of institutional repositories, collections in which significant control is exerted over content. Bradley's study, addressing methods of securing digital content, suggests otherwise. At the time of writing, technology for securing repository content had not been implemented at the majority of institutions surveyed by Bradley; some indicated plans to address the security issue in the future, with the matter being viewed as low priority in comparison to increasing access and preservation. When considering digitisation of artefacts, several institutions agreed "that researchers should confirm the authenticity and integrity of digital material against the original." (Bradley 2005)

If users cannot be assured that repository versions are as described, than a further identification problem will present itself. Götze (1995) also voiced concerns over the vulnerability of online material, given the possibility of interventions such as making additions, modifications, deletions and the re-dissemination of modified, extended or pirated versions. This view illustrates how authors, as well as readers, are affected by the ease with which content can be manipulated.

There is a clear need for the end user to be able to trust the document(s) that they discover in their research, and to be able to understand which version it is that they are using. The literature reviewed here shows this to be a complex and multi-faceted problem.

3. Concepts of bibliographic organisation

As discussed in Section 2, hierarchical presentation of documents can be beneficial in illustrating relationships between entities, and thus could be of assistance in understanding the nature of differences between versions. This section looks at the role bibliographic organisation can play in making the organisation of versions transparent to the researcher.

Graham (1990), writing almost two decades ago and therefore before the 1998 FRBR Study mentioned below, recognises a proliferation of versions due, in part, to technological developments allowing the same work to be made available in a variety of formats. The paper identifies a need to exert bibliographic control over the increasing number of manifestations of works, and refers to an uncertainty over terminology and bibliographic description within library science:

> *"Library science lacks standard terminology for the discussion of works, versions, copies, and pieces. We have hopelessly intertwined the terminology for the physical pieces and the ideas that they contain."* (Graham, 1990)

Noting a tension between placing emphasis on either physical object or intellectual content in catalogue records, Graham cites Wilson (1989), who urged librarians to reconsider the basic unit of record, making it an abstract intellectual concept rather than a particular item. Wilson felt the catalogue record should represent a 'family of texts'; that is a text, all related versions, and other items such as translations, abridgments and so on, in a similar manner to Stillinger's conception of how a work should be viewed. The envisioned record is likened to "a perfect swarm of parasites of different sorts" (Wilson, 1989), organised around a single intellectual concept. The focus of the catalogue record moves away from physically describing the item in hand, towards describing the intellectual concept that binds a set of linked entities together.

Graham questions the necessity of treating different versions as entirely separate works, with discrete bibliographic records. A reduced number of records around which the physical descriptions are grouped should reduce the time spent locating the desired version.

Smiraglia (2002) notes that locating works is essentially the goal of a catalogue search. He observes, however, that the relationship between works and documents is not one of strict equivalence; rather, a given work can exist in numerous forms, and can appear as numerous documents. His paper reviews a number of theories relating to the concept of 'work' before noting three common strands; that a work is conceptual, that changes lead to new related works, and that a set of related works can therefore be associated with the original work's citation. This echoes the ideas of Wilson, outlined above.

Standard library catalogues, whilst effective inventories, have not commonly prioritised collocation of linked or related items. Smiraglia cites Green (2001), who succinctly describes the difference between an entity and any relationships associated with it:

> *"Whatever we consider to be the most basic elements or reality, we deem to be things, or more formally, entities. After that it's all relationships. Relationships are*

> *involved as we combine simple entities to form more complex entities, as we group entities, as one performs a process on an another entity and so forth."*
> (Green, 2001)

To maximise effective information retrieval, Smiraglia suggests, entity relationship phenomena must be identified, defined and mapped. This is necessary to "further advance the utility of the catalog in the digital age." (Smiraglia, 2002)

Tillett (2001) defines a series of relationships that can occur between bibliographic entities. Examples given by Tillett are 'equivalence' and 'derivation'. Tillett uses 'equivalence' to describe the relationship between exact copies, or between originals and their reproductions, whilst 'derivative' relates to relationships between a bibliographic work, and modifications based on it: editions, revisions, translations, adaptations, genre changes, free translations and so on. Tillett calls for the expression of relationships in library catalogues, stating that cataloguing rules should be specific about which relationships should be included.

Tillett also notes the existence of content relationships. Such relationships, pertaining to the extent of intellectual or artistic similarity between related documents, are given as the primary focus of most catalogues. Tillett says content relationships "…can be viewed as a continuum starting from an original work…". Closely related pieces such as copies, versions, and slight modifications, are often viewed as the same work from a cataloguing point of view. More distantly related resources, appearing thereafter on the continuum, such as adaptations, genre changes, and commentaries, are catalogued as different works.

Copeland (2001) examines some of the practicalities of collocating related material in the library catalogue, having found similar issues with current cataloguing procedures to those discussed in this review. Copeland cites Svenonius (2000), who suggests that researchers need bibliographic systems which can bring together 'almost the same' information, as well as 'exactly the same' information. Users need related resources to be linked, with coherent expression of the relationships between documents. In keeping with Smiraglia's observation that works and documents are not equivalent, Copeland notes that works are published as editions, and yet catalogue searches yield works, not editions.

Copeland, noting that works are essentially abstract concepts, looks at the ways in which specific types of digital materials are catalogued to illustrate how access is aided or hindered by cataloguing rules. Copeland also investigates the use of uniform titles, a common title assigned across a number of related entities, which could demonstrate the advantages of using an organising agent for different entities. Copeland also highlights the potential problem of inconsistencies resulting from a lack of detailed procedures in the application of the uniform title. Using OCLC's WorldCat, Copeland writes that a search for 'chaucer' and the uniform title 'works' yields almost 200 results. Dates are then necessary to distinguish between the editions. The uniform title is demonstrated here as a powerful tool for collocation, although the way in which the results of the search are displayed is not entirely satisfactory. Some digitisations are listed as reproductions, while others are listed as new editions; in one case the same edition is dated as 1896 and 1985 in separate records. These findings, the paper suggests, highlight the need for clarification of cataloguing rules in relation to digital material.

When employing a uniform title, additional organising information such as manuscript date will also be required, and clear guidance in relation to this is essential.

The issues relating to traditional catalogue organisation were the subject of the 1998 IFLA Functional Requirements for Bibliographic Records (FRBR) study, which defines a set of entities to express intellectual or artistic content, organised in a hierarchical structure with 'work' at the top. 'Work', for example Austen's *Pride and Prejudice,* is defined as a distinct intellectual or artistic concept, and realisations of that concept appear further down the hierarchy. 'Expression' relates to the intellectual or artistic realisation of a work, a particular translation of *Pride and Prejudice* for example. 'Manifestation' is defined as the physical embodiment of an expression; a particular paperback edition of a particular translation of *Pride and Prejudice.* Finally, 'item', is understood as an exemplar of a manifestation; a particular copy with notes in the margin, for example. Organising documents in this way enables collocation of related material under the umbrella of the 'work' concept, to some extent fulfilling the calls discussed earlier for catalogues to present a cluster of linked texts in search results.

Mimno et al (2005) assess the implications of a FRBR-based hierarchical catalogue record system for searching and browsing. They comment that removing the need for repetition of information is given as an advantage of this system. Records of 'expressions' do not need to repeat all information (original author for example), only information specific to the 'expression'. As a result, the cataloguing process is expedited, and records can be stored more efficiently. The major benefit for users is that complex works, with numerous manifestations, are automatically organised into a simpler, more understandable format. In addition, searches for title and author will bring up all available versions, including those that would not, of themselves, match the query.

Mimno et al identify some of the difficulties with the hierarchical organisation of intellectual / artistic entities. They also report that The Australian Literature Gateway found distinguishing between manifestations and expressions challenging, showing that clear designation of entity level may be difficult. Beall (2006) also notes this problem, seeing the FRBR terms as "vague and overlapping". Relationships between shorter works in larger volumes can be problematic. For example, a play which is a distinct intellectual concept, or 'work' in FRBR terms, will be identical to plays in a collected volume at the 'manifestation' level.

Beall urges caution in relation to several aspects of FRBR. Both Mimno and Beall raise the issue of potentially complex and ambiguous terms. In addition, Beall expresses a feeling that most libraries will not have more than one manifestation of a particular work and therefore considers the framework irrelevant for many catalogues.

While Ercegovac (2006) considers library catalogue organisation to be superior to the organisation of online search engine results, she suggests that there is room for improvement in terms of linking entities together, and displaying them in a manner that is useful to users. The article highlights the prevalence of derivative bibliographic relationships, citing as examples Tillett's report that 11.2% to 19.4% of Library of Congress catalogue records exhibit at least one and Smiraglia's finding that 49.9% of a sample of Georgetown University Library records also possess at least one. Ercegovac implements the FRBR hierarchy in relation to Abbot's *Flatland*, noting that the original work has parented four versions, which in turn have led to seventy 'offspring', and thus a potentially complex bibliographic network. The intellectual creation, or 'work' level entity,

is *Flatland*. All related texts are subsequently divided into textual and non-textual entities, and then sub-divided into textual English language and textual non-English language entities and so on. The paper notes that such a mode of organisation partitions larger results into smaller, more meaningful, clusters. This is viewed as practical due to the high incidence of bibliographic relationships, and the fact that results are displayed in a way that is natural for human information processing capacity.

Ercegovac cites Carlyle and Summerlin (2002), who suggest that organising retrieved records into intelligible categories may communicate search results more quickly and effectively. The results listing is not based on the entities themselves, rather on library systems or procedures. It is concluded therefore that in the case of multiple online versions, it would be more practical to view files organised in a coherent, partitioned manner than scattered randomly, or sorted based on criteria other than the entities themselves. Having highlighted the need to link entities intelligibly, Ercegovac provides an extensive set of terms with which to describe relationships between entities. Descriptions such as 'X is annotated Y', and 'X is published edition of Y' serve to clearly define relationships between two entities. These descriptions, in addition to a well-ordered display, should assist the user attempting to navigate multiple versions.

The current organisation of bibliographic records does little to ensure that all related items are retrieved and less still to ensure that the relationships between these items are described. The articles discussed suggest that catalogues organising entities around their central intellectual or artistic concept could be beneficial to users. A central organising principle will, they suggest, enable the presentation of items as a network in relation to the common core from which they stem. Versions form part of this network, along with more distantly related items such as criticisms, commentaries etc. Collocating versions and related items in this way offers benefits to the user concerned with the existence of multiple versions of academic papers. Organisation along with clear signposting about how versions relate, would alleviate the version identification issue for users of library catalogues and repositories alike.

4. Software and versioning: systems, tools and solutions

This review has so far noted a series of conceptual and practical issues relating to versions of documents in the print and digital environments, along with possible solutions. The software development sector is extremely important when considering version control and version management techniques. The process of software development necessitates the creation and management of many versions of source code, configuration files and documentation. It is therefore natural that the sector would have developed systems and tools to help manage these digital files.

At this point in the review it is worth declaring that the authors of this review have a background in library and information science and publishing, rather than in computing science. Owing to these admitted limitations on the part of the authors, the present section can be no more than a dip of the toe in the water of an extensive and technical literature on software and versioning. However, it is hoped that even if the selection of articles overlooks more important or relevant work, at least what follows may provide an overview of the systems and tools available for versioning digital objects as well as reviewing several research articles on possible solutions to aspects of versioning.

4.1 Version control systems (VCS)

Practical approaches to versioning are necessarily employed in the disciplines of software development and engineering, which are characterised by teamwork, stepping back to earlier iterations of plans or code and, in the case of software, support and development of different versions or releases of software. It is not surprising therefore that sophisticated standalone applications exist to manage and control versioning in these sectors (Wikipedia, 'Revision control' article, 2007).

The first type of software system to consider is version control systems or source control or code management systems, of which some well known open source examples are CVS and Subversion. Designed to manage the collaborative development of software source code, these systems typically assign a new version number to any new version committed to the repository. Some of the systems have the ability to lock code for editing so that only one developer at a time can make changes. This works on the idea of checking out code from a central repository and checking it in again when finished. While the code is checked out, other developers will only have read access. Another method of ensuring consistent version control used by such systems is version merging. A typical VCS will allow a developer to view different versions of code side by side with the changes highlighted, before checking in the new code. It is also possible to use version control systems in a distributed peer-to-peer environment, rather than a centralised client-server model. One further advantage of such systems is that storage space is reduced because only the differences between versions are stored (as deltas), rather than complete version of the file being made with each revision.

Although such version control systems were designed to manage files such as source code, it is now becoming more common to think of them also in terms of managing documents and other digital objects. Nagel (2006) suggests using Subversion to help with personal information management of documents stored on different computers and storage media.

4.2 Content management systems and wikis

Content management systems (CMS) are used to help manage content where there may be a large number of creators, for example on an institutional website with extensive web content. Typically a CMS will include some version management and tracking functionality.

The literature on content management systems frequently makes comment about versioning, indicating that this is an important aspect of systems dealing with large amounts of fast-changing content. Pitkow and Jones (1995) present a prototype web publishing environment, emphasising that version control should be supported, permitting multiple revision of documents, and graceful recovery from system and user errors. Browning and Lowndes (2001), along with Pyles (2001), also reference the importance of being able to retrieve previous versions, which may have been lost or overwritten. Guenther (2001) notes that the roll-back facility, enabling access to previous versions, is particularly useful in an environment where it may be necessary to show the content of a web page prior to changes being made, for example in a healthcare organisation. The above articles all point to the value of retaining previous versions rather than simply retaining a final copy.

Wikis can be viewed as a particular type of CMS and they share many of the same features including revision control functionality. One feature of great interest in the context of version identification and management is in the ability to display previous versions of wiki pages. Another key feature is the ability to compare two versions side by side and to display the differences between them using a *diff* feature. Thirdly, contributors have the option to mark any edits as minor. Taken together these form a powerful set of version management features leaving the reader with significant control over what they wish to view. Most wikis provide this functionality and *Wikipedia* which uses the Mediawiki software, is perhaps the best known example. Criticism of wikis focuses on the ease with which they can be vandalised. As the article on 'Wiki' accessed on 18 November 2007 puts it:

> *"Wikis are generally designed with the philosophy of making it easy to correct mistakes, rather than making it difficult to make them. Thus, while wikis are very open, they provide a means to verify the validity of recent additions to the body of pages. […] Using the Revision History, an editor can view and restore a previous version of the article. The diff feature can be used to decide whether or not this is necessary."*
> (Wikipedia, 'Wiki' article, 2007)

4.3 Word processors

Most word processors include some kind of version control functionality. Byfield (2006) outlines the three revision control features in OO Writer: Changes, Compare Documents and Versions and comments on how these three can be used together to make for reasonable version control. He also notes their limitations compared with a full version control system, for example the lack of automatic version numbering.

MS Word has two main version control features, Track Changes and Versions. Track Changes allows for review and suggested edits to be made to a document by collaborators. Versions allows a creator to save snapshot versions of the work, to add comments about the nature of the changes and to revert to earlier versions again if wished.
Google Docs provides an interesting approach to collaborative working on the web and includes an element of revision control so that changes can be made by collaborators in real time.

4.4 Tools and protocols relating to version identification

A range of tools frequently used as elements in larger systems assists with version control and identification. An overview of some of these is given here.

Timestamp. A sequence of characters representing the date and time of a specific event, for example the date a file was created or modified. These are stored in the properties metadata in office documents and are used in file systems to sort and differentiate between different versions of a document.

Checksum. A value generated by a hash function to check that a digital data object is authentic and unchanged. It can be useful for checking that one document is an exact copy of another. However it cannot help with detecting the nature and extent of differences between versions.

Diff. A file comparison utility which compares the differences between two files. The output of a diff is called a patch. Early versions of diff utilities worked well only with text files such as source code files or plain text documents, but not so with binary files such as word processed documents. However, later versions of the utility it appears can support diffing of binary files. Wikis typically use diff to present the differences between older and newer versions of wiki pages. There are many file comparison utilities available, some free or open source and others proprietary.

WebDAV and Delta-V. WebDAV (Web-based Distributed Authoring and Versioning) is a proposed open standard, an extension to HTTP, which allows for creation and editing of documents on a remote server. WebDAV is published as RFC 4918, Proposed Standard, June 2007, supersedes RFC 2518. Delta-V or Versioning Extensions to WebDAV is itself an extension to WebDAV and is published as RFC 3253. DeltaV aims to support better document authoring via the web by promoting internet standards for versioning.
(Wikipedia, 2007), (Hunt and Reuter, 2001)

4.5 Proposed software solutions for versioning problems

Khaddaj et al (2004) look at how object oriented techniques can be used in information management to help with object versioning. The paper looks at two case studies: a geographical information system (GIS) model and a document management model. The argument made for considering object oriented databases for object versioning is in its superior (to relational databases) performance at “handling complex relationships among objects”.

> *"For example, if a real life object is represented in object oriented form, rather than as an entry in a database table, associations with other objects (to which it is linked) can automatically 'inherit' any changes made, making it easier to track later."*
> Khaddaj et al (2004)

The authors discuss different approaches to storing versions in an object oriented model:

- Complete: each version is stored as a complete snapshot. Khaddaj et al posit that this could be costly in terms of storage space.
- Linear: one complete version is stored, and then other versions are "presented as differences between the versions." Khaddaj et al (2004). Each parent object has only one 'child', meaning that each version only related to the one that came before it.
- Branching: A branching technique is more complex and allows for more complex relationships between versions. It can be applied either forward or backwards – ie a 'child' document can have many parents, or a parent document can have many children.
- These could also be combined, depending on the application they were needed for.

For document management, Khaddaj et all suggest an object oriented model made up of the following composite classes:

- Document Object Class
- Version Event and Manager Classes
- Document Type Class
- Location Class
- Temporal Class

Each of these classes (which of course could have subclasses) would be able to handle events and features of the particular version of the object being described, such as object type (text or multimedia etc), time attributes (eg when the object was created).

Maliappis and Sideridis (2004) view knowledge as evolutionary; as such, representations of knowledge should be capable of illustrating this evolution whilst allowing for the simultaneous existence of various versions. They propose a framework of knowledge version handling which should include the following criteria:

1) Provide an unambiguous reference to what is being identified.
2) Make obvious the relationship between one version and other versions.
3) Offer a language of representation and provide suitable tools for its implementation.
4)Provide tools for easy management – change, control, and search.
5) Be secure.
6) Provide proper interfaces for knowledge exchange and distribution via the Internet.

Maliappis and Sideridis propose a framework where knowledge is assembled into independent portions (modules) relating to a narrow cognitive field. These modules use an inheritance mechanism, which leads to the assembly of bigger units of knowledge, and ultimately knowledge bases. They suggest that versioning systems based on inheritance could assist with mapping relationships between current and previous versions. When a new version of an object is created in their proposed system, a new

identifier is allocated, typically an incremental number, which links the object to all previous versions, as well as to the original version to which it belongs.

Such a system allows users to follow the evolutionary history of various objects, and to identify where the version at which they are looking fits into the collective work. Automatic creation of unique identifiers and the ability to track the evolution of a paper could thus assist users with document selection. Furthermore, the links expressing relationships between versions proposed by Maliappis and Sideridis could answer the calls previously noted for clear expression of bibliographic relationships in information retrieval systems.

Nørvåg (2004) presents 'granularity reduction', a proposed system for removing intermediate versions from a temporal document database. Nørvåg suggests that it is often more appropriate to delete intermediate versions than it is to remove the oldest ones in a temporal document database. It is suggested that it is improbable that authors will wish to store multiple versions reflecting only minimal changes, or that readers will want to be confronted with a large set of almost identical documents when sourcing information.

Nørvåg details strategies for determining which versions should be deleted from systems. Documents are logged with a 'transaction time' when entered; this information can then be used to select documents for deletion, when required:

1) Naïve granularity reduction
Every nth document is deleted. If n=1, all candidate versions are removed, if n=2, every other version is removed etc. The disadvantage of this strategy presents if a large number of versions are created in close succession over a short period of time - many of these very similar versions will be retained, whereas other versions, spanning longer time ranges, will be removed.
2) Time-based granularity reduction
Versions created within t (a specified period) will be deleted. If t=1 week, all versions separated by less than one week will be removed. The retained versions are not necessarily evenly spaced, time wise. Nørvåg warns that this may not always be an appropriate strategy, giving web versions of newspapers as an example. Here reducing granularity from one day to one week means only one in every seven versions is retained. In this case, the intermediate versions, selected for deletion, are just as important as those that remain.
3) Periodic-based granularity reduction
Versions created at a specified periodic interval are retained. For example, if the period is seven days then one version every Sunday could be kept. A maximum of one version can be retained per period, and the same version can be valid for longer than one period, if no changes have occurred during that time. The retained versions are evenly spaced time wise.
The above strategies utilise metadata only; the following consider document content:
4) Similarity-based granularity reduction
Compares similarity of document versions to decide which should be removed. If the similarity is greater than a given threshold, a version will be removed. This is based on the assumption that small changes are probably error corrections or similar. Naming a reasonable threshold value is difficult, however, since this will usually differ between documents.
5) Change-based granularity reduction

Considers words only. The difference between two versions is calculated using a difference function (d). If d is less than a specified threshold, the version can be removed. The function calculates the number of basic operations that will turn one version into the other. The result must then be normalised according to the number of text lines.

6) Relevancy-based granularity reduction

All document versions below a given relevancy rank threshold are removed. Relevance is calculated with respect to a query or set of queries. This approach is problematic because future queries cannot be predicted - a relevancy measure must be sought without this information.

Nørvåg suggests the possibility of using simpler low-cost strategies (naïve, time-based, period-based) as a filtering step, before applying the more costly content-based strategies. However, there is a risk of eliminating key versions during the filtering stage. The above-mentioned strategies, and the advantages and disadvantages of each are well worth considering. The systems based on metadata could be easily implemented, as they are based on file properties data (date) that can easily be accessed. Content-based systems would be more difficult to implement, but perhaps more appropriate.

Saccol et al (2007) propose a method of version identification to be used for XML documents which could overcome the problem of setting a threshold for similarity between documents noted above by Nørvåg. The method would use a combination of a similarity function and Naïve Bayesian classifiers to estimate whether a given document is a probable version of another. For the authors there are many applications of a solution to version detection, such as web page ranking, plagiarism detection and peer-to-peer searching. The authors state two important issues posed by the version detection problem:

> *"the first is how to measure similarity between files and the second is how to define the minimum degree of similarity required for a file to be considered a version of another."*
> Saccol et al

The proposed mechanism can be used both for linear and for branched versioning. It begins with a first task of conducting a similarity analysis. The second task described is the classification using Naïve Bayesian classifiers for which a small amount of training data is needed. The authors reported good results for their experimental results with the classifier detecting over 90% of versions in the sample data set.

Staying with XML and documents, an earlier paper by Rönnau et al (2005) describes the potential for XML version control of office documents. The purpose of their research is to explore how a specific XML diff algorithm could be integrated into existing mainstream version control systems as an API. Version control systems have traditionally not handled office documents well, since these are treated as binary files rather than text files.

Marshall (2006) reflects on the ways in which people meet the challenge of personal information management (PIM), and discusses the need for technological developments to support it. The discussion has broader implications, dealing as it does with not only the necessity of addressing the storage, preservation, and long-term access of digital materials, but also with the importance of "making these things intelligible to others".

Long-term storage issues of digital objects are affected by a number of issues reflecting the information needs of users. Marshall summarises these, including the difficulty of predicting the long-term value of a document, the digital context, distributed storage and format opacity, and "curatorial effort".

Marshall maintains that the urgency of these issues is perhaps not widely recognized as "we have been trained to approach technological progress with an air of optimism". There is a presumption that although digital formats may become obsolete, technology will somehow ensure that they always remain recoverable. And yet this optimism co-exists with a mistrust of technological storage – we expect computers to crash and be infected by viruses; there is an acknowledgement that documents we value are always subject to the risk of being totally lost. Marshall suggests that these polar positions have led to a "benign neglect" of the need to archive and organise digital material. Intentions with regard to the maintenance of some kind of quality control in selecting material for storage, for backing up material and refreshing storage media are good, but standards are not rigorously applied. This has a significant bearing on the advice that might be offered to authors making submission to digital repositories regarding the organisation of their own personal filing systems.

Marshall also looks at concerns affecting digital repositories and Internet archives, including canonicalisation and the implementation of policies and methods for preserving large-scale hyperlinked structures. She identifies a series of seven factors that should shape the development of archiving technologies:

> *(1) digital material accumulates quickly, obscuring those items that have long-term value;*
> *(2) digital material is distributed over many different off- and online stores, making it difficult to keep track of individual items;*
> *(3) digital material derives a substantial amount of meaning from context that may not be preserved with an item (for example, references to email attachments, links to Web pages, and application-specific metadata);*
> *(4) digital material is easily passed around and replicated, introducing a tension between protection (of privacy, copyright, and security) and future access;*
> *(5) digital formats are not only opaque to many users, but also may be incompatible with available applications or may become obsolete;*
> *(6) curating personal digital archives is time-consuming and requires specialized skills;*
> *(7) our current computing environments have yet to incorporate mechanisms and metaphors that support long term access.*
> (Marshall, 2006)

Marshall refers to Jones (2004) who points out, "the wrong information competes for attention and may obscure information more appropriate to the current task". The assertion here is that too much information can be as bad as too little information.

Identifying a need for "tools for managing distributed content and techniques for stabilizing digital references", Marshall calls for technological developments to assist in personal curation. She suggests the following strategies:

> *(1) Automatically generated visualizations that provide us with an overall gestalt of what we have;*

> *(2) Manually defined and circumscribed digital places and geographies that give us the digital equivalent of "the box under the bed" (for the most valued stuff) and "remote storage lockers" (for the things we aren't sure we'll continue to want);*
> *(3) Heuristics for detecting the relative value of individual items, because people demonstrate the worth of their belongings much more reliably than they declare it; and*
> *(4) Methods and tools for examining individual items that reveal an item's provenance and help increase its intelligibility (to oneself and possibly to others), for example allowing a user to distinguish among related copies of an item.*
> (Marshall, 2006)

Noy and Musen (2004), looking at the question of ontology management, examine a framework that allows developers to compare different ontologies and map similarities and differences among them. They highlight a need to support ontology versioning and to:

> *"provide mechanisms to store and identify various versions of the same ontology and to highlight differences between them." Noy and Musen (2004)*

Noy and Musen describe how ontology merging methods, developed to help users locate overlap between ontologies, can also be used to locate differences between versions of ontologies. In both instances, there are two overlapping ontologies and a mapping between them must be determined. In order to locate overlap, similarities would be identified. Where version comparison is required, the differences would be highlighted. Thus the processes can be seen as complementary.

As in the change-based system detailed by Nørvåg, the overlap can be expressed as a set of rules that would transform one ontology into the other. However, the authors point out that simple text file or data comparison is not helpful in identifying versions of ontologies, as ontology versions' can be exactly the same conceptually, but composed differently. They propose software solutions that could compare the structures of ontologies rather than their composition.

In this section, an overview of version control systems has been given, together with a brief look at specific tools and standards relating to versioning. Lastly several articles have been reviewed which look at specific problems in versioning.

5. Open access and digital repositories

The desire for timely exchange of research findings and the wish to ensure that the results of research supported by public funds are made openly accessible to the research community and others, have been the driving force behind the creation of institutional and subject repositories. There is extensive literature on the subject of open access and digital repositories which is not explored in depth here. Some of these reports and reviews highlight issues relevant to the scope of the VERSIONS project however.

The RCUK[1] draft Position Statement on Access to Research Outputs (2005) sets out the Research Councils' views on the issues surrounding the dissemination of research through open access models, amongst others. RCUK updated its position statement in June 2006 and the individual research councils have established their own policies on open access. The draft statement from 2005 acknowledges the role the Internet has had, and continues to have, in changing the scholarly communication landscape, and the possibility of a diminished role for traditional publication of journal articles as a means of communicating research outputs. This evolution is theorised as follows:

> *"'Publication' may thus be seen as a series or continuum of distinct activities (peer review, 'kite marking', editing, promotion, printing, distribution, placing in repositories, long-term curation...) undertaken according to circumstances by different players, thereby blurring the distinction between the roles of these players."*
> RCUK (2005)

This model implicitly illustrates the lifecycle of a work and the issues surrounding identification of the resulting versions. In the same statement, RCUK calls for clear identification of the position a version holds within its lifecycle:

> *"There must be an absolutely clear distinction, for users of such repositories, between articles that have not yet been peer-reviewed (pre-prints) and those that have (post-prints) – and also between different pre-print versions".*
> RCUK (2005)

Given the major developments taking place in scholarly publishing, the JISC Disciplinary Differences Report (Sparks, 2005) examines the needs of academic researchers across disciplines for information resources. The report also touches on researchers' attitudes to and awareness of repositories and the open access debate. In terms of the single most important resource, 65.9% of respondents to the survey nominated journal articles as being the resource they could not do without. Other versions of work such as e-prints (pre and post) were nominated by 2.4% of respondents. Among economics and econometrics researchers, 18.2% of respondents stated that pre-prints were the most important resource. The survey also identified problems researchers face in disseminating and publishing their work. Means of dissemination varies, with versions of the journal article (in all forms – pre-print, post-refereed version and final publication) commonly disseminated across fields. In common with the VERSIONS survey findings,

[1] Research Councils UK

respondents expressed the most frustration regarding publication with the pressure of space in highly rated journals and slow speed of reviewing and decision making by journals. In terms of gaining access to research resources, particularly journals, the majority did not report any problems, however the minority was large. The overwhelming majority of researchers in all disciplines do not know if their university has an institutional repository. Given the reported problems in dissemination and access, this is certainly a missed opportunity. However, there is a high level of awareness of debates about open access generally, and the majority of researchers in all disciplines favour research-funding bodies mandating self-archiving. Most scholars across the disciplines think journal articles will remain relevant to their discipline in the next ten years, but they also think new forms of dissemination will grow in importance. These findings suggest that scholarly publishing routes are undergoing changes in response to changing information needs, underlining the significance of understanding the needs of researchers in terms of versions of work and the importance of establishing standards for describing them.

Underpinning the need for standards for identification of versions to offer services such as searchability and interoperability is the need to assure users of repositories of the authenticity and integrity of information retrieved. Trusted Digital Repositories: Attributes and Responsibilities (2002), a report issued by RLG-OCLC defines the mission of trusted repositories as being the provision of "reliable, long-term access to managed digital resources to its designated community, now and in the future". With this goal in mind, the report makes a number of recommendations relating to the certification of repositories and the need for collaboration on issues such as technical strategies for ensuring access and systems for identifying the attributes of material that should be preserved.

Among these recommendations, and central to the versions issue, is the assertion that if a user is to trust the digital documents provided by digital repositories, it is essential that changes made to documents must be identifiable. Authentication techniques such as the use of checksum and public key encryption systems are already in use by some communities, for example in e-commerce. The report notes that although almost all of the metadata standards used in repositories include a placeholder for such authentication metadata it is not always required. The report recommends that the repository community should:

> *"Investigate and define the minimal-level metadata required to manage digital information for the long term. Develop tools to automatically generate and/or extract as much of the required metadata as possible."*
> Trusted Digital Repositories (2002)

6. Conclusions

This literature review has brought together resources on the subject of versions and version identification, a subject that has been mentioned from time to time in the literature on digital repositories, but usually at a cursory or 'in passing' level of detail.

By looking at the field of literary criticism it is possible to see that the question of multiple versions and their identification is by no means a new problem or particular to the digital environment. Indeed, though the present review focussed on a few selected works by the Romantics, it would have been possible to consider works in the field of textual criticism of biblical and classical works. This in itself is a major field of scholarship with its own methodologies for determining the authenticity and original texts of works when many variant versions are in existence often in manuscript form. Variorum editions are another technique of scholarship to collocate complete works of an author with editorial and commentator notes; variorum editions are also taken to mean editions of different versions of works by an author in a collected works edition, in order to allow for comparison of the variant texts.

The discussion of the Romantic poets and authors indicates how critical judgments about the value of one version of a text over another can change over time. Naturally authors themselves should and do have the right to control whether to disseminate versions of their work. The point raised by the Mary Shelley example though is that once versions of a work are in the public domain, the author is no longer the only party with an interest or a view about the relative value of the different versions. This is important in the context of open access digital repositories given that authors sometimes express a wish to remove early pre-print versions of their work from circulation once the revised and refereed version has been published.

The chapter looking at bibliographic organisation from the point of view of traditional library catalogues confirms that issues of versioning has been extensively explored by the library and information science community. Ideas about separating the abstract core of a work from its more physical manifestions and establishing relationships between versions of works have been circulating for some time. One outcome of this has been the IFLA FRBR Study, which has been taken up by the digital repository community through the work of the group implementing the Scholarly Works Application Profile. Practical implementations of FRBR do exist, for example in OCLC's xISBN web service which returns related ISBNs in response to a submitted ISBN. In the digital repositories world there is a working example in the institutional repository of the Science and Technologies Facilities Council (STFC).

Reviewing the literature on digital repositories has brought together some of the issues about version identification which have been discussed in the context of open access repositories in recent years.

An admittedly arbitrary selection of resources reviewed from the field of software development has brought together a few ideas about tools, systems and solutions from the software development sector. The sector has been dealing with large amounts of digital files for many years now and can bring a great deal of experience to bear on the question of version control. When thinking about the reader's experience of finding

multiple versions of academic papers online, it is clear that time is spent on establishing the nature of the differences between versions. Solutions to this are already available in systems such as wikis and version control systems. They are based partly on the use of the diff file comparison utility and are successful because the files compared are in open text formats rather than proprietary binary formats.

7. References

Arnold, SE. 'Feeling Content?', *Information World Review* (2002), 178, p30.

Basden, A. and Burke M.E. (2004) 'Towards a philosophical understanding of documentation : a Dooyeweerdian framework'. Journal of documentation 60 (4), 352-370. doi:10.1108/00220410410548135

Beall, J. 'Some reservations about FRBR', *Library Hi Tech News* (2006), 23 (2), pp15-16. doi: 10.1108/07419050610660744

Buckland, M.K. (1997), "What is a 'document'?", *Journal of the American Society of Information Science*, Vol. 48 No. 9, pp. 804-9. http://www3.interscience.wiley.com/cgi-bin/jhome/27981 [Accessed 07/09/07]

Bradley, R. 'Digital Authenticity and integrity: Digital Cultural Heritage Documents as Research Resources', *Libraries and the Academy* (2005), 5 (2), pp165-175. http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v005/5.2bradley.html [Accessed 07/09/2007]

Browning P. and Lowndes, M. 'Content Management Systems: Who Needs Them?', *Ariadne* (2001), 30. http://www.ariadne.ac.uk/issue30/techwatch/ [Accessed 07/09/2007]

Byfield, B. 'OOo Off the Wall: That's Your Version--Document Control in OOo Writer', *Linux Journal* (2006) (March 7, 2006). http://www.linuxjournal.com/article/8911 [Accessed 28/11/07]

Carlyle, A., & Summerlin, J.. 'Transforming catalog displays: Record clustering for works of fiction'. Cataloging & Classification Quarterly (2002), 33, 13-25.

Copeland, A. 'Works and Digital Resources in the Catalog: Electronic Versions of *Book of Urizen*, *The Kelmscott Chaucer* and *Robinson Crusoe*', *Cataloging and Classification Quarterly* (2001), 33 (3/4), pp161-180.

Cummings, S. 'As Versions Proliferate', *Library Journal* (2006), 131, p28. http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=22921994&site=ehost-live [Accessed 17/09/07]

Dallman, D., Draper, M. and Schwarz, S. 'Electronic Pre-publishing for Worldwide Access: The Case of High Energy Physics', *Interlending Document and Supply* (1994), 22 (2), pp3-7. http://www.ingentaconnect.com/content/mcb/122/1994/00000022/00000002/art00001 [Accessed 17/09/07]

Deadalus Project Website: http://www.lib.gla.ac.uk/daedalus/index.html [Accessed 17/09/07]

Dooyeweerd, H. (1955), A New Critique of Theoretical Thought, 1975 ed., Vols. I-IV, Paideia Press, Jordan Station, St Catherine's.

Ercegovac, Z. 'Multiple Version Resources in Digital Libraries: Towards User-Centered Displays', *Journal of the American Society for Information Science and Technology* (2006), 57 (8), pp1023-1032. http://www3.interscience.wiley.com/cgi-bin/fulltext/112593517/HTMLSTARTW [Accessed 17/09/07]

Götze, D. 'Electronic journals – Market and Technology', *Publishing Research Quarterly* (1995), 11 (1), pp3-20.

Google Scholar http://scholar.google.com/ [Accessed 17/09/07]

Graham, C. 'Definition and Scope of Multiple Versions', Cataloging and Classification Quarterly (1990), 11 (2), pp5-32.

Green, R. 'Relationships in the Organization of knowledge: An overview'. In Relationships in the organization of knowledge, ed. C A Bean and R Green. Dordrecht: Kluwer Academic, pp.3-18

Greig, M. and Nixon, WJ. 'DAEDALUS: Delivering the Glasgow E-prints Service', *Ariadne* (2005), 45 http://www.ariadne.ac.uk/issue45/greig-nixon/ [Accessed 17/09/07]

Guenther, K. 'What is a Web Content Management Solution?', *Online* (2001),25 (4), p81. http://search.ebscohost.com/login.aspx?direct=true&db=buh&AN=5010918&site=ehost-live [Accessed 17/09/07]

Hunt, JJ. and Reuter, J. 'Using the Web for Document Versioning: an Implementation Report for DeltaV', *icse*, p. 0507, 23rd International Conference on Software Engineering (ICSE'01), (2001) http://doi.ieeecomputersociety.org/10.1109/ICSE.2001.919123 [Accessed 15/11/07]

'IFLA Study Group on the Functional Requirements for Bibliographic Records: Functional Requirements for Bibliographic Records: Final Report. Approved by the Standing Committee of the IFLA Section on Cataloguing' (1998). http://www.ifla.org/VII/s13/frbr/frbr.htm [Accessed 17/09/07]

Johnson, RK. 'Effecting Change Through Competition: The Evolving Scholarly Communications Market', *Logos* (2001), 12 (3), pp166-170.

Jones, W. (2004, March 3). 'Finders, keepers? The present and future perfect in support of personal information management', First Monday 9 (3). http://www.firstmonday.org/issues/issue9_3/jones/ [Accessed 17/09/07]

Khaddaj, S., Adamu, A. and Morad, M. 'Object Versioning and Information Management', *Information and Software Technology* (2004), 46 (7), pp491-498. doi:10.1016/j.infsof.2003.10.002 [Accessed 17/09/07]

Lynch, CA. 'Institutional Repositories: Essential Infrastructure for Scholarship in the Digital Age', *Libraries and the Academy* (2003), 3 (2), pp327-336. http://muse.jhu.edu/journals/portal_libraries_and_the_academy/v003/3.2lynch.html [Accessed 17/09/07]

Maliappis, MT. and Sideridis, AB. 'A Framework of Knowledge Versioning Management', *Expert Systems* (2004), 21 (3), pp149-156. http://search.epnet.com/login.aspx?direct=true&db=buh&an=13537948 [Accessed 17/09/07]

Markland, M 'Institutional repositories in the UK: what can the Google user find there?' *Journal of librarianship and information science* 38 (4), Dec 2006, 221-228

Marshall, Catherine C. 'How people manage information over a lifetime' To appear in *Personal Information Management: Challenges and Opportunities* (Jones and Teevan, eds.), University of Washington Press, Seattle, Washington, 2006. (in press) www.csdl.tamu.edu/~marshall/PIM%20Chapter-Marshall.pdf [Accessed 17/09/07]

McCulloch, E. 'Taking Stock of Open Access: Progress and Issues', *Library Review* (2006), 55 (6), pp337-343.

McKiernan, G. 'E-print Servers', *Science and Technology Libraries* (2001), 20 (2/3), pp149-158.

Mimno, D., Crane, G. and Jones A. 'Hierarchical Catalog Records', *D-Lib Magazine* (2005), 11 (10) http://www.dlib.org/dlib/october05/crane/10crane.html [Accessed 17/09/07]

Nagel, W. Subversion: Not Just for Code Anymore', *Linux Journal* (2006), (January 26th, 2006). http://www.linuxjournal.com/article/8596 [Accessed 27/11/07]

NISO/ALPSP Working Group on Versions of Journal Articles (2006). Recommendations of the NISO/ALPSP Working Group on Versions of Journal Articles. http://www.niso.org/committees/Journal_versioning/JournalVer_comm.html [Accessed 17/09/07]

Nørvåg, K. 'Granularity Reduction in Temporal Document Databases', *Information Systems* (2006), 31 (2), pp134-147.
Granularity reduction in temporal document databases
doi:10.1016/j.is.2004.10.002 [Accessed 17/09/07]

Noy, NF. And Musen, MA. 'Ontology Versioning in an Ontology Management Framework', *IEEE Intelligent Systems* (2004), 19 (4), pp6-13. http://doi.ieeecomputersociety.org/10.1109/MIS.2004.33 [Accessed 17/09/07]

OCLC - Worldcat
http://www.oclc.org/worldcat/ [Accessed 17/09/07]

Pinfield, S. and James, H. 'The Digital Preservation of E-prints', *D-Lib Magazine* (2003), 9 (9).

http://www.dlib.org/dlib/september03/pinfield/09pinfield.html [Accessed 12/09/06]

Pitkow, JE. And Jones, RK. 'Towards an Intelligent Publishing Environment', *Computer Networks and ISDN Systems* (1995), 27 (6), pp729-737. doi:10.1016/0169-7552(95)00030-B [Accessed 17/09/07]

Pyles, M. 'The Three Steps to Content Management', *Imaging and Document Solutions* (2001), 10 (1), pp41-43.

Quint, B. 'Good, Better, Best: A Mission Statement', *Information Today* (2006), p7. http://www.infotoday.com/it/itnew.htm [Accessed 17/09/07]

RCUK position statement on access to research outputs (2005) http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2005statement.pdf [Accessed 17/09/07]

RCUK updated position statement on access to research outputs (2006) http://www.rcuk.ac.uk/cmsweb/downloads/rcuk/documents/2006statement.pdf [Accessed 28 November 2007]

Richardson, M. 'Post-print Archives: Parasite or Symbiont?', *Learned Publishing* (2005), 18 (3), pp221-223. http://www.ingentaconnect.com/content/alpsp/lp/2005/00000018/00000003/art00008 [Accessed 17/09/07]

Rightscom Ltd. and partners London School of Economics and Political Science Library, University of Oxford Computing Services. (2006). Scoping Study on Repository Version Identification (RIVER): Final Report: http://www.jisc.ac.uk/uploaded_documents/RIVER%20Final%20Report.pdf [Accessed 17/09/07]

Rönnau, S. et al. 'Towards XML Version Control of Office Documents', *Document Engineering: Proceedings of the 2005 ACM Symposium on Document Engineering* , 2-4 November 2005, Bristol, United Kingdom. pp10-19. ISBN:1-59593-240-2. http://doi.acm.org/10.1145/1096601.1096606. [Accessed 15/11/07]

Ross, S. 'Position Paper on Integrity and Authenticity of Digital Cultural Heritage Objects', thematic issue 1, DigiCULT: Integrity and Authenticity of Digital Cultural Heritage Objects (August 2002): 7. http://www.digicult.info/downloads/thematic_issue_1_final.pdf [Accessed 17/19/07].

Rumsey, S. 'The Purpose of Institutional Repositories in UK Higher Education: A Repository Manager's Point of View', *International Journal of Information Management* (2006), 26 (3), pp181-186. doi:10.1016/j.ijinfomgt.2006.01.002 [Accessed 13/11/06]

Rusbridge, C. 'Excuse Me… Some Digital Preservation Fallacies?', *Ariadne* (2006), 46. http://www.ariadne.ac.uk/issue46/rusbridge/ [Accessed 17/09/07]

Saccol, D. et al. 'XML Version Detection', *Document Engineering: Proceedings of the 2007 ACM Symposium on Document Engineering,* 28-31 August 2007, Winnipeg,

Manitoba, Canada. pp 79–88. ISBN:978-1-59593-776-6.
http://doi.acm.org/10.1145/1284420.1284441 [Accessed 15/11/07]

Scholarly Works Application Profile (formerly Eprints Application Profile)
http://www.ukoln.ac.uk/repositories/digirep/index/E-prints_Application_Profile [Accessed 01/05/07]

Shelley, M. *Frankenstein or the Modern Prometheus. The 1818 Text (1993)*, edited with an introduction and notes by Butler, M., Oxford University Press, Oxford.

SHERPA/RoMEO
http://www.sherpa.ac.uk/romeoinfo.html#colours [Accessed 17/09/07]

Smiraglia, RP. 'Further Reflections on the Nature of a Work: An Introduction', *Cataloging and Classification Quarterly* (2002), 33 (3/4), pp1-11.

Sparks, S. JISC Disciplinary Differences Report. Rightscom Ltd., August 2005.
http://www.jisc.ac.uk/uploaded_documents/Disciplinary%20Differences%20and%20Needs.doc [Accessed 02/05/07]

Stillinger, J. 'A Practical Theory of Versions', *Coleridge and Textual Instability: The Multiple Versions of the Major Poems* (1994), Oxford University Press, Inc., New York.

Stillinger, J. 'Preface', *Coleridge and Textual Instability: The Multiple Versions of the Major Poems* (1994), Oxford University Press, Inc., New York.

Swan, A. and Brown, S. 'Authors and Electronic Publishing: What Authors Want From the New Technology', *Learned Publishing* (2003), 16, pp28-33.

Tetreault, R. 'Versioning Wordsworth: Dynamic Collation in the New Medium', *Electronic Publishing '97 – New Models and Opportunities: Proceedings of an ICCC/IFIP Conference held at the University of Kent, Canterbury* (1997).
http://elpub.scix.net/cgi-bin/works/Show?97131 [Accessed 17/09/07]

Tillett, B.B. 'Bibliographic Relationships', in Bean, CA. and Green, R. (eds.), *Relationships in the Organization of Knowledge* (2001), Kluwer Academic Publishers, Dordrecht.

'Trusted Digital Repositories : attributes and responsibilities an RLG-OCLC Report' RLG, May 2002 http://www.oclc.org/programs/ourwork/past/trustedrep/repositories.pdf [Accessed 28/11/07]

Van de Sompel, H. (et al) (2004) Rethinking scholarly communication : building the system that scholars deserve, D-Lib Magazine 10 (9)
http://www.dlib.org//dlib/september04/vandesompel/09vandesompel.html [Accessed 17/09/07]

Van de Sompel, H. (et al) (2006) An interoperable fabric for scholarly value chains, D-Lib Magazine 12 (10)
http://www.dlib.org/dlib/october06/vandesompel/10vandesompel.html [Accessed 17/09/07]

Wikipedia. http://en.wikipedia.org. Specific Wikipedia articles consulted:
'Checksum', http://en.wikipedia.org/wiki/Checksum [Accessed 28/11/07]
'Content management system',
http://en.wikipedia.org/wiki/Content_management_system [Accessed 18/11/07]
'Diff', http://en.wikipedia.org/wiki/Diff [Accessed 18/11/07]
'File comparison', http://en.wikipedia.org/wiki/File_comparison [Accessed 18/11/07]
'Revision control', http://en.wikipedia.org/wiki/Version_control [Accessed 18/11/07]
'Subversion (software)', http://en.wikipedia.org/wiki/Subversion_%28software%29
[Accessed 18/11/07]
'Textual criticism', http://en.wikipedia.org/wiki/Textual_criticism [Accessed 18/11/07]
'Timestamp', http://en.wikipedia.org/wiki/Timestamp [Accessed 18/11/07]
'WebDAV', http://en.wikipedia.org/wiki/Webdav [Accessed 18/11/07]
'Wiki', http://en.wikipedia.org/wiki/Wiki [Accessed 18/11/07]