"'EEBO and the politics of open standards': A paper for the conference '(De)materialising the early modern text: Early English Books Online inTeaching and Research' at Bath Spa University College, 8-9 September" by Gabriel Egan, Loughborough University <mail@GabrielEgan.com>

Abstract EEBO is an unmitigated good; I take that as an agreed starting point. This paper is concerned with the technological particularities by which such goods are disseminated, with special concern for two things: the on-demand delivery of materials over the Internet (as opposed to materials downloaded and kept locally) and the use of proprietary standards for formatting (such as Adobe Acrobat's Portable Document Format). EEBO users are at the end of a supply-chain of almost unimaginable technical complexity, for the maintenance of which they are dependent upon privately and publicly owned agencies, comprising (at the least) the content provider ProQuest, its Internet Service Provider, Teleglobe International (owner of the Atlantic undersea cables), JANET (the United Kingdom's academic Internet Service Provider), and the user's computer services department. Most users have heard of only the first and last of those four agents. Likewise, unseen agency (that becomes apparent only when it goes wrong) obtains in the proprietary format by which images are delivered to users. This paper will survey how these systems bear upon academics' use of EEBO and their implications for the power relations between publicly-funded library staff and academics and private content publishers. In particular, certain means by which the power relation can be adjusted in favour of the public side of the equation will be outlined.

I'll start with an anecdote. In March of this year, the library of my university, Loughborough, found that its access to the *Journal of Construction Engineering and Management* published by The American Society of Civil Engineers ceased to work, and academics and graduate students were for that reason unable to do their research. Investigation showed that the problem was at the supplier's end: the electronic publishing-distributor, Scitation Online, had blocked all requests originating from within a whole range of Internet Protocol (IP) addresses on the campus. Asked to explain why, the publishing-distributor explained that in one hour, one IP address on campus--that is, one user--had downloaded 265 articles from the journal, mostly in issue-number sequence, and that this violated the licence agreement between the university and the publishing-distributor. The university library apologized to the publisher, promised to try to prevent the same thing happening again (essentially, this is limited to stressing even more strongly the library's Acceptable Use Policy regarding online resouces), and after a day of so's interruption the library's access to the database was restored.

I tell this story because it illustrates a shift in power in the relations between academics, libraries, and publishers. It so happens that at the Library Users' committee where I heard of this, the main users of the database, scientists and technicians, thought that the publisher had behaved reasonably and, far from objecting to the publisher's high-handed behaviour, these users backed the university library's craven attitude towards the publisher. Downloading article upon article in sequential order from an online journal was, the users agreed, clearly an act preparatory to pirating the contents, because in science and technology one would never want to look at a journal in this way. As the lone arts-and-humanities representative on the committee, I was only person who thought that there might be legitimate reasons for this user's behaviour. A couple of years ago I spent a few weeks in the University of London Library, requesting from the stack every volume of the journal *The Library* in chronological order, starting with volume

1 from the late nineteenth-century, because I was researching how the style and coverage of the journal changed from a gentlemanly book-collectors' journal to a serious academic journal once A. W. Pollard took over as editor in the first decade of the twentieth century. I'm not suggesting that such an interest in the journal itself *was* the reason that someone at Loughborough was downloading article after article from an online science journal--I suspect that this was indeed action preparatory to piracy--but I want to point out that across the disciplines we might disagree strongly about what constitutes reasonable behaviour. In practice, of course, the terms of a product's licence are supposed to draw clear lines about what is allowed, but in this case the licence gave the publisher the right to terminate access if an "unreasonable" number of articles were downloaded at one IP address. What is "unreasonable"? 100 articles? 10? 5? My university library had signed up to a deal that left this crucial term undefined, indeed left it up to the publisher's discretion, and having seen the publisher was indeed quick to excercize its right to cut us off, I find worrying.

Before online digital media appeared in libraries, the rules about usage of the materials were largely imposed by the libraries themselves. Certainly, copyright has always been fitfully imposed at the photocopier, but it is worth noting that the analogy with the case I'm considering is not with photocopying but with fetching the material from the stack. That is to say, the act of reading an online article necessarily puts a copy of it into the personal computer being used in the library, so the old-fashioned difference between merely 'seeing' a work and 'taking of copy' of a work has, with this new technology, disappeared: merely reading necessitates taking a copy. Nowadays the rules about usage are imposed by publishers and, in the case I've described, the users were so desperate to have their connection restored that they would have agreed to almost anything to get it back. I was alone on the committee in holding that the university library should renegotiate its deal with the publisher so as to at least quantify the reasonable number of articles that could be downloaded in one sitting at one IP address.

What I draw from this is that what I used to think was only a potential shift of power concomitant with the new technologies has now become real. With paper materials and with CD and DVD materials, power rests with the possessor of the physical media, whereas by contrast the Internet-delivered media have shifted the power to the provider. Whether or not we trust particular publishers--and I've no reason to doubt ProQuest's probity in such matters--we must respond to these changing power relations so at the get the best deal for academic users and to ensure the longevity of materials. Complicating the situation is that the fact that with Internet-delivered materials the publisher is only one link in the long chain of supply that brings the media's to our personal computers. Academic users are at the end of a supply-chain of almost unimaginable technical complexity, for the maintenance of which they are dependent upon privately and publicly owned agencies, comprising in the case of EEBO the content provider ProQuest itself, then ProQuest's Internet Service Provider, then Teleglobe International (the company who own of the transatlantic undersea cables for Internet traffic), then the Joint Academic Network (JANET) which is the Internet Service Provider for the UK's universities, then our own university computer services department would run our campus networks, and finally our own computers, which for those of us in the arts and humanities are probably the most complex and fragile machines we will ever use. A technical problem affecting any one of those links in the chain is likely to prevent us reading the early modern books that we've called up from EEBO, and our chances of fixing the problem ourselves are virtually zero. Indeed, few users are even aware of the existence of most of these links in the chain.

What should we do about this? My answer is that, as professionals morally charged with the maintenance and dissemination of the literary part of our cultural heritage, we should pirate as much as we can. That is, we should wherever possible use online resources to download what we need to use and then store local copies of the materials so that when the supply chain breaks we are not cut off. It is no exaggeration to say that the new media are fundamentally altering the nature of property within late industrial capitalism, and that old notions of ownership simply do not apply in the new situations. There is already a reality of mass violation of old copyright laws in the form of users sharing music, films, and software over peer-to-peer (P2P) networks on the Internet and by copying and swapping their CDs and DVDs. This shows how the technology of almost instantaneous and absolutely perfect digital reproduction makes a mockery of laws written in the days when copying was painfully slow and never perfect. Moreover, the new technologies are throwing up their own new models of knowledge creation and dissemination, shown best in such phenomena as the Open Source software projects by which we get miracles like the Linux operating system[1] and the collaborative-writing *wiki* movement that produces such beauties as the WikiPedia online encyclopaedia.[2] New right-managements frameworks such at the Creative Commons (CC) licence[3] might bring a little order to these processes, but the important point is that the old licences just won't do and we should not consider ourselves bound by them.

If this sounds like reckless talk, it is worth noting that no-one in academia has ever been prosecuted for breaking the old licensing rules using the new media, and I suggest that we ought not allow ourselves to be cowed by legal opinions (for which our employers pay a lot of money) that inhibit our copying of the materials that we use in teaching and research. In practice, publishers such as ProQuest often allow us to download unlimited amounts from their products and this is just what we should do. A few years ago ProQuest dropped their 50-page limit per download on EEBO, recognizing, I suspect, that anyone with a little technical knowledge easily join together a collection of 50-page downloads, and that the limit was only serving to frustrate ordinary users. This relaxation is to be applauded, and other publishers should be encouraged to do the same. The Eighteenth Century Collections Online (ECCO) product from Thomson Gale can be seen as the logical continuation of EEBO, for it provides digitizations of 125,000 key texts from 1700 to 1799, and it is now also available at relatively low cost to university users via a Joint Information Systems Committee (JISC) licence. Less enlightened than ProQuest, however, Thomas Gale still imposes a page limit on downloading from ECCO and the publisher seems impervious to sensible arguments against it. Of course, publishers such as ProQuest and Thomson Gale will point out that if we give our students and colleages a locally-stored copy of a book from EEBO or ECCO rather than pointing them to the version on the publisher's servers, the users will be missing out on any improvements that the publisher makes to it products. This is true, but it is no different from the familiar situation when a library declines to buy the second edition of a book of which its has the first edition: the sum total of the new edition's improvements has to be great enough to give the user reason to discard the first in favour of the second. I think this is a useful incentive to encourage publishers to improve their wares, and we should not relinquish it.

Moreover, even without this reason, the very impermanence of online resources puts us under a moral obligation to pirate as much as possible, because we cannot rely on the materials surviving any other way. To see why not, take the example of the BBC's splendid LaserDisc project to create a new digital Domesday book recording life in the United Kingdom 900 years after the first Domesday Book. The resources assembled for this project are effectively lost to us all because as a standard for dissemination the LaserDisc and its associated home computer, the Acorn/BBC micro, are incompatible with the standard computer systems in use today.[4]. If piracy of materials from the project had been widespread--that is, if users had possessed the technical means to violate their licence conditions by copying what they wanted--most or all of the raw material of the project would be available to us in some form. This is not wishful thinking on my part: we have a clear precedent for it. As is well known, the BBC routinely wiped and reused tapes of radio and television programmes from the 1950s and 1960s, and in many cases the only surviving copies are illegal pirated recordings made off-the-air by listeners and viewers and stored at home. The BBC is now grateful to receive copies of these illegal recordings to fill the extensive gaps in its broadcasting archive. On a personal level, I'm sure I'm not the only person here whose list of publications includes an article commissioned for an academic website that no longer exists. In my case, the I only hope that (contrary to the terms of use published on the site) people did copy material from the Arden Shakespeare's now defunct ArdenNet website, else I'm the sole possessor of an text that was once widely available and that has been cited in more than one printed book.[5] I'm aware that new technologies such as the Digital Object Identifier (DOI) scheme are supposed to save us from some if not all these problems of impermanence in the future, but I remain sceptical.[6]

The BBC Domesday LaserDisc project, of course, pre-WorldwideWeb and it relates to the preceding argument about the important of piracy only by analogy. The obsolescence of formats is merely another way, apart from the break in the supply chain, by which might easily lose access to essential digital materials, and it should teach us the same lesson: don't accept the formats and rules dictated by publishers, rather make whatever uses you want of the material in order to preserve it. Personally, I have hundreds of books I've downloaded from EEBO, and these came to me in the Adobe Corporation's Portable Document Format (PDF), which most people read using Adobe's freely-available Acrobat reader. These PDFs I've turned into thousands of individual images (one per book opening) in the Tagged Image File Format (TIFF), which is an open standard. People treat PDF as though it were an open standard, and strictly speaking it is (it is the subject of an International Standards Organization definition), but for practical purposes it belongs to the Adobe Corporation and they can what they want with it. For example, Adobe could at any issue a new specification of the format, incompatible with the old, and release a new version of the Acrobat reader to allow users to read it. Almost all users of PDFs look at them using Adobe's Acrobat reader, and these users would respond by simply updating their reader, so in practice the definition of the PDF format, and hence the power, remains in Adobe's hands.

Generally, large software corporations such as Adobe and Microsoft make new digital formats and software backwardly compatible with the old ones, so that (for example) if you buy the latest version of the Word word-processing program you can read documents made in any of the previous versions. Of course, if you stick with your old version of Word, you'll increasingly find that other people are making documents in the new format and you cannot read them; this incentive to buy the latest version of its products is central to Microsoft's sales strategy. Were it not for this strategy, we'd all be using Word version 2 because it has virtually all the functions we ever need, and Bill Gates would not be the richest man in the world. Microsoft and Adobe are sufficiently large that they must take care to ensure at least backward compatibility in their products (that is, the new software can still read the old data): they do not want to be seen to hold to ransom the users of their formats. Smaller companies, however, have more incentive to be sharp in their practices, as one can see from the BBC's experience with the Real Audio format. The BBC was persuaded to convert thousands of hours of radio broadcast content into the proprietary Real Audio format rather than use open-standard MP3 audio, and it had assurances from the supplier, Real Networks Incorporated, that listeners would always be able to download a free copy of the Real Audio player in order to receive this content. Now, it is still possible to get from Real Audio a free copy of their player, but the company's website is so constructed as to make it difficult: almost all the links take you to an offer to buy the latest version of the player using your credit card, or a free version of it that expires in 14 days.

Essentially the same situation obtains with PDF format that EEBO uses to supply downloaded texts to users, for which format we are at the mercy of the Adobe Corporation. Adobe has made public the PDF standard and there are products for reading and creating PDFs produced by companies other than Adobe, but nonetheless, Adobe owns the standard. So many people use PDF that I imagine Adobe did change the format, there would be what is known as a forking of the format. A sufficiently large group of programmers would continue with the old standard, releasing new tools for working with it, while the Adobe company moved forward with the new standard, and these two standards would effectively become rivals for the same market. The Betamax versus VHS war of videocassette technology in the 1980s shows that technological superiority is no guarantor of success in such a battle between closely-related formats, and indeed the history of EEBO shows this too. Early users will recall that EEBO images were delivered to one's web-browser using the DjVu format from the company LizardTech, which format employs fractal compression software to squeeze large books into a very small file sizes. For example, a full-colour DjVu digitization of the 700-page Records of Early English Drama (REED) volume for Coventry comes out at around 8MB, which is only 10,000 bytes per page.[7] I haven't asked them, but my guess is that ProQuest went over to the current system of sending the images as GIFs files (apparently created on-the-fly from TIFF files in their database) because standard web-browsers can read these without modification, whereas the DjVu format required the user to install a plug-in reader provided by the company LizardTech. If so, ProQuest's was a wise decision: widespread compatible is much more important than technological superiority.

In conclusion, then, I urge academic users of new media such as EEBO to be as daring as their universities will let them get away with in their use of technologies of dissemination, thinking always not what is strictly within the terms of the licence but what is most likely to perpetuate these intellectual and artistic goods long after the current generation of lawyers (who write the end-user licences) are dead. It is important that we do not repeat the fiasco of the BBC Domesday project, in which what we might call 'edition one', the 950-year old paper version, turned out to have a longevity 100 times as great as that of 'edition two', the digital version, which was unusable within a decade of its creation. If we stick the letter of the law as laid down in the end-user licences, the new technologies represent a massive shift of power towards publishers and away from readers. Fortunately, by the familiar dialectic of technological progress, the new media also give us the means by which to frustrate the terms of these licences. I would encourage users of EEBO to grasp these means and exploit them to the full.

[1] See www.opensource.org and sourceforge.net for more on the Open Source movement.

[2] See wikimediafoundation.org for more on the *wiki* movement.

[3] See creativecommons.org

[4] There have been heroic attempts to 'reverse engineer' the Domesday Project in order to recover the materials. The work of the CAMiLEON project at University of Leeds and University of Michigan showed that the original hardware and software could be emulated in modern personal computers, and although it produced a working system that can read the original LaserDiscs the raw materials have not been made publicly available; see www.si.umich.edu/CAMILEON. Another team of engineers working in collaboration with the National Archive has pulled out the digital data from the project, but not the moving video and sound, and their results are available on the web at www.domesday1986.com. For an account of the technical projects to recover all the material on the BBC Domesday disks, including archiving the video and sound streams, see the article at www.ariadne.ac.uk/issue36/tna/

[5] Because ArdenNet foolishly demanded that users register for a free userid and password to access the contents of the site, automated WWW archiving engines such as the The Wayback Machine <www.waybackmachine.org>, which cannot make an application for a free userid, were kept out of most of the site and captured only the introductory pages.

[6]See www.doi.org for an account of this scheme.

[7] For reasons that I cannot fathom, this expensive book--indeed all the published REED volumes--are available for free in this DjVu format from the Million Books Project of the Internet Archive; see www.archive.org/details/millionbooks