An Investigation into Free eBooks

Final Report
March 2004

**Ylva Berglund, Alan Morrison,
Rowan Wilson and Martin Wynne**

ahds literature, languages and linguistics

Acknowledgements

The authors would like to thank Louise Edwards and Hazel Woodward of the JISC e-Books Working Group who commissioned this report, and who assisted in the development of the workplan. A huge debt is due to Chris Armstrong and Ray Lonsdale, who ran the focus groups with great skill. We are very grateful to the participants in the focus groups who gave willingly of their time and gave us the benefit of their experience, but who must remain anonymous. We are very grateful to Paul Davey who contacted the representatives of FE institutions who responded to the questionnaire.

Michael Popham's work on the project was invaluable, as Head of the Oxford Text Archive at the time of the commissioning of the report. Michael wrote the workplan and led the early stages of the project. Annamaria Carusi conducted part of the survey of HE. James Cummings, Gregory Simpson and Stuart Yeates provided very useful input in editing and proofreading this report, made valuable suggestions for improvements and helped with the production of the document. We would like to thank the many people who volunteered to take part in the project advisory committee. RM Ltd kindly loaned the Tablet PCs which were demonstrated to the focus group participants.

Table of Contents

Executive Summary

The aim of this project was to inform the Joint Information Systems (JISC) e-Books Working Group on the availability of free e-books for teaching and learning in the UK Further and Higher Education sectors, and to evaluate user needs in this area. The main focus of the study was on resources for the arts and humanities subject area, to reflect the expertise and knowledge of the authors, and to complement the related projects. This project involved the following investigations:

- a survey of available free e-books
- a survey of the different formats in which free e-books are available
- a survey of current levels of usage in HE
- a survey of user needs to representatives of FE institutions
- an in-depth examination of attitudes to free e-books with representatives from FE and HE.

A combination of desk-based research, a questionnaire and focus groups were the instruments used for the different investigations. The results of these investigations are summarised below.

The survey of free e-books showed that there are a vast number of free e-books available to the arts and humanities scholar. The problem this community faces is not one of quantity, but of quality. It is questionable whether many existing free e-books are of any practical use to the academic because of the lack of quality assurance regarding text integrity, conformance to a single printed edition and adequate metadata. In terms of quality and supporting bibliographic information, e-books created by academic departments provide the best option to the academic user (although quality here is by no means always assured), while the multitude of ‘enthusiast’ sites provide little that can be used or trusted. Almost all arts and humanities subject areas have free e-books at their disposal, and the scope of material available ranges from the ancient classics, to the early twentieth century. Although there are a multitude of web sites which host or point to free e-books there is no single trusted repository where this material can be deposited for others to use. For the arts and humanities community, initiatives such as the AHDS, which maintains a nationally funded free archival and distribution service, are providing a template for possible solutions to the problem of locating and accessing high quality digital resources.

The survey of formats found that free e-books are available in a variety of different formats, some of which require specialised software and some of which are designed for particular hardware, such as Personal Digital Assistants (PDAs). A large proportion of the free e-books currently available are in plain text, with no structural markup at all. While it is unproblematic to print these texts, and to a certain extent to paste their contents into Virtual Learning Environments, they are not ideal. Their lack of formally marked-up structure is a barrier to their being reformatted for different devices, aggregated into collections, searched meaningfully or preserved. Texts which have been marked up in complex structural tagging schemes like the Text Encoding Initiative are readily reformatted, aggregated and searched. They can also be transformed easily into other structured formats for presentation, such as HTML, printed and incorporated into

Virtual Learning Environments. There is however, an overhead of expertise and time in producing such texts.

Users and potential users of free e-books found the following barriers to uptake:

- Lack of availability of a complete range of titles for any given course
- Doubts about quality assurance
- Lack of confidence in the persistence of availability of resources
- Costs involved in the cataloguing, archiving, management of resources
- Costs involved in computing support for users
- Poor design of free e-books and poor ergonomics of reading on screen

The following were identified as potential opportunities to promote the uptake of free e-books:

- ‘Freeness’ could be vital to Further Education institutions
- Existing digital repositories and resources do exist
- Free e-books tend to come in open formats and free of IPR restrictions, which means that they can be more easily repurposed, integrated into institutional systems and preserved
- Free e-books may be more useful for the humanities than other disciplines, because there is more use of ‘old’ texts
- VLEs represent an opportunity for the delivery of free e-books to the student.

This project was able to make the following recommendations to the JISC:

- Take measures to offer more comprehensive ranges of titles in specific areas
- Support efforts to migrate existing collections to common formats
- Institute a system of quality assurance (of text integrity and metadata)
- Ensure the permanence of collections
- Support the professional, standardised cataloguing of electronic resources
- Offer support for users in the basic ICT dimension of the use of e-books
- Offer help with integration into VLEs.

1. Introduction

There are hundreds of thousands of electronic texts freely available on the internet, and many of them are texts which are studied in colleges and universities in the UK. The following questions are therefore starting to be asked: are free e-books currently being used in teaching and learning in Further and Higher Education in the UK? Exactly what relevant free e-books are available? Are they of sufficient quality for use in teaching and learning? What is required to deploy them in the curriculum? Can we save money by using free e-books instead of buying print books or commercial e-books?

In late 2002, the Arts and Humanities Data Service (AHDS) was approached by the JISC/DNER e-Books Working Group[1] to undertake an investigation into free e-books and their potential use within the Further and Higher Education communities. This investigation was carried out by *AHDS Literature, Languages, and Linguistics*, based at the Oxford Text Archive (OTA), the AHDS centre with the most relevant expertise and experience. The job of *AHDS Literature, Languages, and Linguistics* is to collect, catalogue, and preserve high-quality electronic texts for research and teaching, and to give advice on best practice in resource creation. The OTA is part of the Research Technologies Service of Oxford University Computing Services, and has been operating as an archive of electronic texts since 1976. This report is the outcome of this investigation.

The following key questions are addressed by this study:

Availability: what free e-books are freely available with the minimum of intellectual property rights constraints?

User needs: Who are the actual and potential users of free e-books, and what are the possible uses?

Repurposing: To what extent can existing freely available e-books be repurposed, converted to other delivery formats, and assimilated into other activities or collections?

The focus of the study is on materials that are freely available to support work within the Arts and Humanities disciplines of the HE and FE community – but any significant differences from, or similarities to, materials available in other discipline areas were highlighted. This subject focus was chosen to reflect the experience and expertise of the AHDS, and to complement other studies commissioned by the JISC/DNER e-Books Working Group, which had concentrated more on science disciplines. It was also acknowledged that other projects were investigating textbooks and e-prints, and so these types of text were not considered.

But first, another question: what is a free e-book? It is not the intention of this report to explore this question in any depth. Much has been written on the topic, including a useful discussion in the report from the EBONI project[2]. In order to set

[1] JISC E-Books Working Group: http://www.jisc.ac.uk/index.cfm?name=wg_e-books_home

[2] EBONI's definitions of an e-book: http://e-books.strath.ac.uk/eboni/documents/definition.html

the parameters for the study, it was however necessary to have a working definition for the project:

> *An e-book comprises a text in electronic form, coupled with software and hardware in order to read it. Prototypically e-books are electronic editions of material published in print, and which attempt to emulate 'book-like' characteristics. "By a free e-book we mean one that involves no direct costs to acquire, access, read, copy, or use.*

Exactly what sorts of e-books this definition will include is explored in the following chapter.

2. Survey of Free e-Book Collections

Before proceeding onto the main part of this investigation into free e-books, one of the main remits of this report was to produce a survey of the scope and quantity of free e-books that are currently available online. It is certainly not difficult to find free e-books on the internet, but rather than attempt an exhaustive list of sites – an almost impossible task – this chapter attempts to look at different kinds of free e-book resources and tries to evaluate their characteristics. In line with the general scope of this report, the focus will be on e-books which are of use to the arts and humanities communities.

Tools to find free e-books

There are numerous resources available online which will help in locating free e-books. General search engines such as Google (http://www.google.com/) will probably be most people's first choice, resulting in the familiar avalanche of random hits. Entering the term 'free e-book' in Google, at the time of writing, produces a return hit rate of over two and a half million sites, most originating from commercial and marketing sites. Adding the search term 'humanities' to this reduces this number considerably, to around ten thousand, but still the sites returned have little to offer the academic visitor. Lastly, adding 'arts and humanities' to the search results reduced the hit rate by half again, the bulk of which remain from the non-academic sphere. This quick example perhaps demonstrates that while few web sites consciously identify themselves are primarily sites with free e-books for the academic community, there is a vast volume of information out there that potentially could be of use to the arts and humanities communities.

Inevitably, the blame for the volume and general nature of these hit rates is due to the misuse of metadata by individual web sites and the harvesting methods of the general search engines. Most academics who are familiar with the internet will most probably have knowledge of sites which group together links to arts and humanities only resources into e-book portals or hubs. Again, there are a large number of these types of sites, ranging from the amateur to sites which attempt to keep up with the ever increasing number of e-books being produced. One of the most impressive compendiums of links to external resources is The Voice of the Shuttle (http://vos.ucsb.edu/), established in 1994 and housed at the University of California at Santa Barbara. It sets out its missions as:

> *"to provide a structured and briefly annotated guide to online resources that at once respects the established humanities disciplines in their professional organization and points toward the transformation of those disciplines as they interact with the sciences and social sciences and with new digital media."*

While VOS has complete control over what is represented on the site, it does not have any influence over the sites it links to, and inevitably one finds that many of the links point to sites which have either moved address or have shut down all together. Sites such as VOS provide no reviews or comments on the web sites listed in their catalogue, however, some kind of peer review by subject specialists on individual sites can be extremely useful to users who have no knowledge of a site or resource. The Humbul Humanities Hub (http://www.humbul.ac.uk/), part of the Resource Discovery Network, catalogues, describes and evaluates online resources, including a description of a resource and its usefulness. Currently few actual free e-book sources are catalogued there, but it would be useful extension of their catalogue if such sources were added.

Other resources for e-book discovery are run by amateur enthusiasts rather than by academic institutions or projects. Links to resources are maintained by individuals or small groups simply out of interest in their subject. Almost as soon as access to the internet became commonplace, those interested in the possibilities of e-books started making collections of their own from what was freely available online. Online catalogues such as Alex (http://www.infomotions.com/alex/) started to appear, publicizing what was already in digital form for the benefit of others. As these sites had no institutional backing or direct funding, the style and efficiency of these sites were completely in the hands of those who ran them. New sites would (and still do) regularly come and go, and the community who used them just had to get used to their transience. While such sites are not resources of free e-books themselves, but rather simply collections of links to external resources, they nevertheless form an important link between those looking for free e-books and those who, for whatever reason, wish to catalogue the existence and location of free e-books. Finally, in this round-up of resource discovery tools, there are a few e-book-specific sites which offer specialised searching tools, such as e-book Locator (http://www.e-booklocator.com/) which will search only for e-books, although their usefulness is limited, as much of what they find are simply links to commercial e-book web pages.

A survey of free e-book sites

Regrettably, there is no such thing as a central national or international digital repository for e-books, free or otherwise. While there has recently been some progress in this area in the UK, there is no legal requirement yet to store published e-books in a legal deposit library such as the British Library. Perhaps because of their very nature, free e-books are regarded in the same way that printed ephemera has been in the past, and their curation is largely left to enthusiasts. For the rest of this chapter we will examine representative exemplars of the most important types of free e-book sites, which we consider represent the bulk of free e-book resources available to the arts and humanities scholar studying in the UK today.

Academic repositories

Within the UK there are virtually no digital repositories which offer a free, permanent archival storage space for arts and humanities digital resources, and a free distribution service. There are numerous reasons for this, which will be investigated further in this report, but one of the main causes is the lack of ongoing funding to support such an initiative. However, steps are being taken to remedy this situation, with the establishment of nationally funded services such as the Arts and Humanities Data Service (AHDS) (http://ahds.ac.uk/). The creation of the AHDS resulted from the increasing awareness by funding bodies and users of digital resources that some kind of central repository for electronic resources, created and used by UK academics, was crucial to ensure that valuable electronic resources did not disappear due to lack of professional curation. The AHDS was established in 1995 to collect, catalogue, preserve and distribute the digital resources created by UK academics for the benefit of the arts and humanities community as a whole. One of the five subject centres which comprise the AHDS is the Oxford Text Archive (OTA) (http://ota.ahds.ac.uk) which hosts AHDS Languages, Literature, and Linguistics.

The OTA pre-dates the formation of the AHDS, and was originally established in 1976 at Oxford University Computing Services, where it continues to be based. From the beginning, the remit of the OTA was simple: to provide free storage space to any electronic text, in any format and in any language. By publishing what had been collected (in paper catalogues, mainly at conferences) the OTA alerted other interested parties to what had already been created – thereby saving them the effort of re-creating it – and offered these texts free of charge for others to re-use in their research. One activity that the OTA specifically does not engage in is the creation of e-books themselves. Given such an open accessions policy, the OTA's collection of e-books has grown into an eclectic mixture of resources, covering classical ancient texts, medieval and Renaissance literature and the literary cannon (Shakespeare, Austen, Dickens, etc.), as well as a wealth of reference and linguistic materials. Because the OTA collected many of these resources in the time before standards for markup had evolved, the e-books in the collection span a wide variety of formats and markup schemes. The e-books are not restricted to the English language. In fact, the deposit of non-English texts is actively encouraged and around twenty-five different languages are currently represented. The OTA was quick to recognise the importance of the emerging standard for encoding e-books, and remains a strong supporter of the Text Encoding Initiative (TEI) (http://www.tei-c.org/). Many texts held by the OTA were encoded in TEI to test out the encoding scheme's strengths and weaknesses, and while it is not in the scope of the OTA to encode every resource in TEI, the OTA has adopted the TEI header as its preferred method of documenting and administrating its collections. So while many of the texts held by the OTA are less than perfect, it is hoped that the standard application of the TEI header will ensure that all are catalogued to the same high standard.

With the formation of the AHDS, the OTA has focused much of its energies on promoting good practice in the creation of e-books to the UK arts and humanities community. The AHDS is partly funded by the Arts and Humanities Research Board (http://www.ahrb.ac.uk/), which was set up in 1998 to support quality research and postgraduate training in the arts and humanities in the UK. One crucial requirement for any projects funded by the AHRB which create a significant electronic resource as a result of its funding, is that this electronic resource should be offered for deposit with the AHDS once the project is complete. Formalizing the relationship between resource creator and digital repository is the first step to ensuring that digital resources do not disappear once a research project has finished. The resources which are now being deposited by this method, are of a much higher standard than previous e-books, and tend to be large scale research materials, rather than simple individual texts. The AHDS offers depositors a non-exclusive deposit licence, which means the creators of the resource retain all intellectual property rights to their material, and are free to publish the material elsewhere, in addition to making a preservation copy available to the AHDS.

Access to the resources held by the OTA and AHDS is completely free of charge, via a web catalogue, although some resources do require users to request additional permissions to use certain restricted e-books. The only restrictions on the user, is that they do not re-use the resources for any commercial gain without the appropriate permissions.

Academic e-book creators

While many large universities do not have a recognizable e-book repository themselves, it is quite common for a university library to list electronic resources created within its own institution for the use of its own staff and students. These resources are often small and directly related to courses taught on a particular campus. Perhaps the most successful example of this is the Electronic Text Center at the University of Virginia Library, Charlottesville USA (http://etext.lib.virginia.edu/). Established in 1992, the E-text Center quickly established itself as one of the best sources of free e-books for the arts and humanities academic community worldwide. Initially e-books from other collections (including the OTA) were collected together and encoded in TEI in order that the standardised texts could be viewed and delivered via a common interface. As the collection started to attract interest, the Center encouraged faculty members at the University of Virginia to get involved in creating digital resource which would be of use in their teaching and to their students. This partnership between the digital library and its users proved to be very successful, and soon the holdings of the Center were complemented by numerous new collections, many relating to American history and literature. Like the OTA, the E-text Center remains committed to promoting the use of standards in textual markup, specifically the TEI. The TEI encoding scheme was applied systematically to all texts held by the Center, fulfilling many of the ambitions that were hoped for, but practically impossible to implement, at the OTA. The consistent application of the TEI guidelines meant that the e-texts could be viewed or downloaded in a variety of formats, and with the increase in popularity of hand-held e-book readers in th USA in the late 1990s, the Center began offering free e-books in Microsoft Reader and Palm formats. Again, this service proved to be successful, in the first year and a half of this service, the Center had shipped more than eight and a half million titles.

The collection at the E-Text Center at Virginia covers a wide range of arts and humanities subjects, with an emphasis towards American history and literature. Like the OTA its collection is multilingual, and the majority of its resources are available free of charge via a simple on-line catalogue and download feature. At present the Center boasts some impressive statistics, including a catalogue of seventy thousand texts and three hundred and fifty thousand related images, although there does not appear to be many new accessions being created or added to the catalogue at present.

"Enthusiast" e-book collections

As a publishing medium the internet revolutionised the means by which books could be publicised and distributed across the globe. A connection to the internet and a minimum knowledge of HTML was all that was required to set up individual web sites, the content of which could be entirely dictated by the owner. Naturally this medium provided the opportunity "vanity publishing", and also paved the way for the creation of numerous websites reflecting the interests of individuals. Many of the free e-book websites surveyed for this report fall into the category of 'built by enthusiasts'. Anyone can create a website, there are no barriers or specific requirements that need to be adhered to, there is no quality control or level of professionalism that needs to be attained. Therefore the internet provided a perfect forum for individuals to pursue their passions, and from the earliest stages e-books were a popular subject for such sites.

The earliest and perhaps the most popular site for free e-books that exists is Project Gutenberg (http://www.promo.net/pg/). Established in 1971, Project Gutenberg was founded by Professor Michael Hart, who is credited with sending the first ever free e-book over a computer network, a copy of the American Declaration of Independence. Michael Hart has remained as the driving force behind project Gutenberg ever since, and the site itself has gone from strength to strength, spawning mirror sites all over the World. It has also established satellite projects in different countries, such as the “Project Gutenberg of Australia” (http://gutenberg.net.au/) which specialises in the culture and literature of Australia, and whose growth is helped by the more liberal copyright laws afforded there. At the time of writing this report, Project Gutenberg were celebrating the posting of their ten thousandth free e-book, and were producing, on average, three hundred free e-books a month. An impressive statistic by any measurement, the production of these free e-books is even more remarkable because many are created by a volunteer effort.

From its conception, the basic principles behind Project Gutenberg have remained the same, to make as many e-books available to as many people as possible, in the simplest format that can be read by any computer. To begin with the volunteer effort was relatively small, and free e-books were produced at a slow rate, however, as the popularity of the website grew, so did the number of people who offered their services. As Project Gutenberg themselves put it:

> *It took 30 years to do the first 5,000, only 18 months for the next 5,000*
>
> *We Have Already Done Over 3,263 e-books In 2003 !!!*

The reason for this rapid escalation in e-book production is due to a worldwide army of ‘distributed proofreaders’ and, one assumes, of people who digitise. From its outset Project Gutenberg made it clear that its target audience was the ‘general reader’ and this assumption governs much of the selection and principles behind the choice and format of its free e-books. Michael Hart is often portrayed as the ‘maverick’ of the e-book community, and some of the pronouncements of Project Gutenberg perhaps bear this out, for example when justifying the selection of titles he says,

> *“We have also been told that nearly every Star Trek movie has quoted current Project Gutenberg etext releases (from Moby Dick in The Wrath of Khan; a Peter Pan quote finishing up the most recent, etc.) not to mention a reference to Through the Looking-Glass in JFK. This was a primary concern when we chose the books for our libraries.”*

The project includes what it regards as ‘light literature’ (Alice in Wonderland, Through the Looking-Glass, Peter Pan, Aesop's Fables) and ‘heavy literature’ (the Bible, Shakespeare, Moby Dick, Paradise Lost). More problematic is the fact that the construction of the e-books themselves is unorthodox. Sections or chapters of books are farmed out to various individuals across the world, each of who digitise or proof-read their own individual section of text. These sections are then returned to Project Gutenberg where the sections are edited together to produce a ‘Project Gutenberg Etext’:

> *"Project Gutenberg Etexts are usually created from multiple editions, all of which are in the Public Domain in the United States, unless a copyright notice is included. Therefore, we usually do NOT keep any of these books in compliance with any particular paper edition."*

The efforts of the Distributed Proofreaders group have been addressing this problem somewhat, where complete editions are now scanned and uploaded for everyone to use. Distributed Proofreaders also encourages use of Unicode for accented characters missing from older Project Gutenberg texts, but still transliterates the Greek alphabet.

Overall, however, the policy of creating non-specific editions aimed at the general reader makes the use of Project Gutenberg texts unreliable to the serious arts and humanities scholar. The dominance of Project Gutenberg as a bulk producer of free e-books is disappointing from the academic point of view, but even more alarming is that many Project Gutenberg titles populate a large number of other free e-book sites. For example the holdings of websites such as Blackmask Online (http://www.blackmask.com/) are simply offering multiple pre-formatted versions of original 'plain ascii' Project Gutenberg titles, thereby proliferating their editorial policies through such duplication.

It is unfortunate, from an academic perspective, that such a productive resource provides so few resources that are of use in teaching and research. Project Gutenberg holds a large quantity of useful titles, and they are currently branching out into non-English texts and even in to audio versions of classic texts. That so many other free e-book sites replicate the Project Gutenberg editions rather than promoting quality editions with known provenance or bibliographic details does not help the promotion of the use of free e-books in the academic world.

Single author/genre sites

One area of free e-books often over-looked, is the creation of web resources dedicated to a single author or genre of literature. These resources can be a valuable resource to the academic user, as they are often themselves the product of an academic department or research project. As these resources are also usually the creation of subject specialist they have much to offer, above and beyond the texts themselves, incorporating secondary resources such as commentaries and related online resources. The Darwin Correspondence Project (http://www.lib.cam.ac.uk/Departments/Darwin/) is based at the University of Cambridge, and was originally set up in 1974 as a print-only endeavour, to collect and publish the definitive edition of letters to and from Charles Darwin. The letters were transcribed into electronic form from the beginning of the project, but it was not until years into the project that the potential for making them available electronically was realised. The project now intends to publish the letters in print and digital form, with the probability that the final digital resource will be deposited with the Oxford Text Archive.

Increasingly, the output of an academic project comprises some kind of digital resource, which is usually made available from the projects departmental server (and hopefully a back-up copy deposited with the Arts and Humanities Data Service). However, unless the project makes intelligent use of its metadata, or is linked to one of the major e-book portals, it can often go unnoticed by exactly the community it is trying to reach. Central repositories such as the AHDS will help to publicise the existence of such rich resources, as they tend to contain resources of a very high quality and cover a wide range of arts and humanities disciplines. Notable examples include sites such as The Perseus Digital Library (http://www.perseus.tufts.edu/) based at Tufts University, a specialist in ancient Greek culture and texts, but expanding to include a range of quality digital cultural resources. The Blake Archive (http://www.blakearchive.org/) is an astonishing collection of the works of William Blake, which utilises to the full the tools and software available on the web to re-examine this author's work in a way impossible on paper.

Just as there are academic and 'enthusiast' repositories, so there proliferate a number of 'enthusiast' sites for authors and literary genres. Sites such as Austen.com (http://www.austen.com/) are typical of what you might find in these resources. Often the owner of the site has simply collected together all the free resources they can find on the internet and made them available on a single site. This might mean that the site includes texts from Project Gutenberg, links to external resources, discussion forums and the like. While these sites may provide the academic with some useful resources, they do not inspire the same scholarly trust as academic based project sites, nor are they any less susceptible to the transitory nature of many non-funded sites and can simply disappear without notice.

Reference, miscellany and foreign language resources

In our broad definition of what comprises an e-book in the introduction to this report, we include 'book-like' documents which are freely available at the point of use, and are readable electronically. One genre often over-looked in e-book surveys is the large number of reference works that are available online. The reference and linguistics resources held by the Oxford Text Archive are possibly the most requested items, and in many ways these resources are expertly suited to be used in digital form rather than in print. Dictionaries, old and new, thesauri and other reference materials offer the academic quick and easy access to practical information once in their digital desktop.

Sites such as yourDictionary.com (http://www.yourdictionary.com/) make good use of available technology to retrieve information quickly, and also offer a wide range of different kinds of dictionaries, in a variety of languages. The theology community were one of the first the realise the potential of e-books, and a thriving collection of religious and sacred texts are now available. The Bible Gateway (http://bible.gospelcom.net/) collects together many of the relevant available e-books and makes them available and searchable from a single point.

While this survey has inevitably focused on the availability of English language texts, it should be noted that there are many corresponding text collections for the study of foreign language texts. Not surprisingly these non-English resources tend to be housed in their native country, but can easily be accessed by UK academics. Sites such as the Spanish based Biblioteca Virtual Miguel Cervantes (http://cervantesvirtual.com/) offer large collections free of charge. While European languages are generally well represented, there is still a problem of easily displaying non-latin based languages online, and unless software is configured appropriately it can be difficult to read resources in languages such as Chinese, Japanese and Arabic. However, with the introduction of character sets such as Unicode and an increasing demand for these non-English texts, it should not be long before these become more widespread. More information on the impact Unicode is making can be found in the next chapter of this report in our survey of formats.

Comments and conclusions

It is evident from this brief survey of web sites that there are a vast number of free e-books available to the arts and humanities scholar. The problem this community faces is not one of quantity, but of quality. It is debatable what percentage of existing free e-books are of any practical use to the academic at present, but we would estimate that the numbers are low. In terms of quality and supporting bibliographic information, e-books created in academic departments or by research projects provide the best option to the academic user, while the multitude of 'enthusiast' sites provide little that can be used or trusted. Almost all arts and humanities subject areas have free e-books at their disposal, and the scope of material available ranges from the ancient classics to the early twentieth century. One obvious absence is the availability of digital reproductions of still in copyright and later twentieth-century material, as publishers are naturally reluctant to allow freely available versions of their titles to be downloaded. Such restrictions often force scholars to use earlier, out of copyright, editions, rather than have access to the latest edited versions. The Faber & Faber 1999 edition of the Anglo-Saxon epic Beowulf, edited by Seamus Heaney, illustrates the problem of gaining access to the latest interpretations of widely available texts, as this edition will remain in copyright for decades to come, and so users who want a free e-book version will be reliant on earlier editions.

Although there are a multitude of web sites which host or point to free e-books there is no single trusted repository where this material can be deposited for others to use. For the arts and humanities community, initiatives such as the AHDS, which maintains a nationally funded free archival and distribution service, are providing a template for possible solutions to the problem of locating and accessing high quality digital resources. Many factors contribute to the current state of affairs, perhaps the main one being that the e-book is still in its infancy as an educational tool and is only now being investigated seriously for its potential as a research and teaching resource.

3. Survey of text encoding formats

This chapter describes the file formats in which e-books are most commonly available, and attempts to sketch something of the development of digital text encoding. Appendix A contains some earlier history of text encoding, for those interested in the origins of the current formats. Readers unsure about the meaning of terms such as 'ASCII' and 'Unicode' can find explanations in Appendix A.

Reviewing these technical details will, it is hoped, facilitate an understanding of the issues surrounding the repurposing of e-books, as well as providing a context for the investigation of possible future developments in this area.

Plain Text

Description: ASCII (and Unicode) files are both commonly known as ‘plain text’. This means that they are streams of bytes that encode characters in a specified manner. Although ASCII does contain a sequence of 'control codes' (characters with no visual representation whose insertion affects visual text-flow) text styling metadata such as italicisation or underlining cannot be stored directly in the stream without markup. Clearly the absence of styling in plain text files means that they are not ideal for creating faithful digital representations of print books. Plain text's strength is its simplicity and portability. The structural descriptors that it lacks can be encoded into plain text files using markup.

Creation Software: Any modern personal computer operating system will come bundled with a basic ASCII text editor, for example ‘Notepad’ under Windows. Unix-based operating systems, particularly, place much of their configuration information within plain text files, thus making a text editor an essential system administration tool.

Viewing Software: As above, text editors, either supplied with operating systems or acquired separately, are the main means of viewing plain text files.

Portability: Unfortunately, despite its simplicity, ASCII has a serious portability issue. UNIX-based systems encode a line-break with a single ASCII linefeed character (code 10), DOS systems encode line breaks by the combination of the carriage return and the linefeed character (code 13 followed by code 10), while Apple Macintosh computers encode line breaks with just the carriage return character (code 10). While the majority of text editors on these systems attempt to identify this incompatibility and convert files automatically where appropriate, this difference in approach can still lead to problems. For example, even in the current version of Notepad (included in Windows XP) single linefeed characters are not displayed as breaks, and Unix-originating text runs together. Unicode provides a single line break character and a separate paragraph character, as well as containing the legacy control codes it has inherited from ASCII.

Further complication is added by the variations upon the 7 bit ASCII format introduced by various software firms to accommodate the use of characters outside the standard ASCII range. For example, Microsoft created various 8-bit (256 character) 'code pages' to accompany versions of Windows sold outside the US. Code pages for languages with more than 256 characters were encoded using a double byte coding scheme, where two bytes encode one character. The problem with these 'code pages' is that they are specific to the platform they were designed for, and conform to no other general scheme. Text editors can attempt to discern which code page a text is encoded in, and display it correctly, but this is a complex and inelegant solution. Later versions of both Windows and Mac OS use unicode to encode characters for all regions, thus eliminating the problem for newly generated texts.

File Structure: Standard ASCII is a stream of bytes with values ranging from 0-127. Thus ASCII streams will always encode one character per byte. Unicode is more complex, in that a single character can be represented by anything up to five bytes. The Unicode consortium provides a series of specifications for how the encoding of unicode characters should be undertaken by programmers, with schemes that involve the use of single byte sequences (UTF-8), paired byte sequences (UTF-16) and four byte sequences (UTF-32). The first of these, UTF-8, has the advantage that it is byte-identical with simple ASCII, as long as only simple ASCII characters are being encoded. Different programming languages use differing encoding schemes for their storage of textual data; for example the multi-platform development language Java uses UTF-16 as its native textual storage format.

Repurposing: Due to its simplicity ASCII text is easily converted into other formats, as long as attention is paid to the mapping of line breaks as mentioned above. Care must also be taken to check for proprietary extensions: in the past text encoders frequently tackle the problem of ASCII's limited character set by devising their own rough and ready character mappings, for example encoding an **é** as **(e/)**. Clearly all such mappings need to be identified and correctly mapped.

Unicode texts are readily portable into most modern text-handling software, particularly as the text-handling in the current versions of both Windows and MacOS is unicode-based.

Because document structure is not encoded in any systematic way (for example a paragraph might be indicated by many different combinations of carriage returns, line feeds and tabs, with variations even within one document) it is not easy to fully convert plain text documents into more structured formats such as XML without dedicating a lot of time to hand-specifying structure.

SGML

Description: SGML has its roots in research done at IBM by Charles Goldfarb, Ed Mosher and Ray Lorie. In the late 1960s, IBM were designing a document management system for use by lawyers which would facilitate creation, storage and retrieval. The work required the integration of several existing software solutions, each of which implemented their own control codes within the document. In one program, the control codes would encode presentation information such as which text was a heading and which words should be italicised. In another, they would mark areas of the text which represented topics and areas that were appropriate for indexing. Goldfarb, Mosher and Lurie decided that what was needed was a generalised method of inserting control codes. In this way, any application could read the document without the risk that one set of codes would conflict with another. The application itself could decide which codes were appropriate to it, and act upon them. This generalised method was dubbed GML, standing for Generalised Markup Language (as well as, conveniently, standing for Goldfarb, Mosher and Lurie). Later, the concept of a DTD, or Document Type

Definition, was added to GML. A DTD was a separate document which contained a template for the control codes (or tags) of a particular application. Using a program called a validating parser, a document could be checked against its declared DTD and certified to be structurally correct, without the overhead of having to process it fully.

IBM had soon realised that the concepts developed in GML were a good basis for a general document management system. The software they developed to exploit these inventions, Document Composition Facility (more generally known as Script), soon became the technological backbone of publishing and business document preparation the world over. Generalised Markup Language became Standard Generalised Markup Language (SGML) in 1986, when the International Organisation for Standardisation accredited ISO8879, which describes the information processing methods behind GML.

Creation Software: SGML is essentially a plain text stream with inline metadata in the form of tags, and so in principle any text editor is suitable for the creation of SGML. The ubiquitous text editing application Emacs has an SGML-editing extension known as PSGML, which will indent SGML documents to show their hierarchical structure, and help with identifying structural errors during their composition, although it does not include a validating parser. For that functionality, there is the SP toolset written by James Clark, which in recent years has been developed under an Open Source license as OpenSP. In the commercial arena there are many products which offer SGML and XML editing facilities, although it would seem that increasingly the SGML processing functionality is being edged out by the developers' concentration on XML technologies.

Viewing Software: Any of the editors mentioned above is also suitable for viewing SGML. Standalone viewers have become rare since the advent of XML. SoftQuad's Panorama, a windows application for displaying SGML, has been discontinued. Rather than viewing the SGML file itself, it is common for a user to transform the SGML file into a more presentational form, then view it. For printing this could be Postscript or TEX, while for on-screen viewing it is likely to be HTML.

Portability: SGML encoding is entirely platform independent. However, as actual SGML files are composed of plain text, they do inherit the portability problems mentioned in the 'Plain Text' section above. Any character set can be used to create SGML files, provided it is identified in the file itself.

File Structure: As mentioned above SGML is a plain text stream. The document structure must be validatable against the appropriate DTD in order to be valid SGML.

Repurposing: The concept of validation makes checking of the integrity of SGML files easy. This considerably eases the process of conversion, in that it ensures that there is an easily performable data integrity check that can be performed before conversion. Transforming valid SGML from compliance to one DTD to another is trivial, and transforming SGML into any another schematised structure is possible as long as a mapping can be established between the respective schemas.

HTML

Description: HyperText Markup Language is an SGML-compliant tagset designed by Tim Berners-Lee for his WorldWide Web project. Berners-Lee was working at CERN, the Swiss particle physics facility. CERN had an in-house SGML DTD for designing documentation, called SGMLguid, and Berners-Lee derived much of HTML from this tagset. The incredible success of the web has meant that there has never really been a design stage for HTML – it went from a working model to production language extremely quickly, and ever since then its extension and development has been driven by browser manufacturers and the desires of internet users. This history has lead to a focus on presentational markup and the embedding of graphical elements.

Creation Software: As with SGML, HTML can be created in any text editor. The W3 Consortium, who administrate the development of HTML, make an open source browser and HTML composing tool available, called Amaya. Commercial packages for designing HTML documents are widely available, with Macromedia's Dreamweaver being the current market leader.

Viewing Software: Web browsers have become a standard component of any modern operating system installation, so viewing rendered HTML is not a challenge. Any standard text editor can be used to view the HTML source directly. Mobipocket, a reader program for portable devices, can display XHTML.

Portability: The market-driven evolution of HTML has lead to some problems in the uniform rendering of HTML across different browsers. Microsoft and other browser developers took a decision a long time ago to create browsers that are tolerant of malformed HTML, and which attempts to render it as best they can, without reporting errors. As a result, there is a lot of broken HTML on the web, which nevertheless looks fine in the chosen browser of the person who developed it. Another major portability issue for HTML is the implementation of complementary technologies in the different browsers. This problem is most evident in the client-side scripting language javascript, and in the accompanying API for addressing elements in a loaded web page, known as the Document Object Model. Unfortunately these technologies are implemented differently in different browsers, meaning that a page which wishes to take advantage of these features must either exist in variant versions, one for each browser, or alternatively avoid areas of the implementation where differences exist. These issues can make HTML problematic as a truly portable document format.

File Structure: Like SGML, HTML is a stream of textual data with incorporated markup.

Repurposing: HTML markup began life as a subset of a fully-featured presentational tagset, and has developed new features as a result of the browser developers desire to extend their market share. HTML documents which stick to employing the core presentational markup are easily repurposed. Those which employ complex scripting or proprietary tags will require a considerable amount of work to standardise and export. It should also be noted that HTML is purely a presentational tagset, concerned with getting text onto a visual display. It lacks true structural markup, and as a result it can be difficult to automatically translate into forms that assume a much smaller display, such a WAP, or no display at all, such as an audio stream.

XML

Description: The astonishing success of HTML can be attributed in large part to its extreme simplicity. People without technical expertise could nevertheless author a simple HTML page after five minutes instruction – and naturally this helped enormously with uptake. The abandonment of a compulsory validation stage, and the willingness of browser developers to render even severely broken HTML meant that web-authoring became a widespread pastime. It seems extremely unlikely that this would have happened so completely if fully validated SGML had been chosen as the document format for the web. Having recognised these facts, the World Wide Web Consortium (W3C), who coordinate the development of web technologies, felt that there was a requirement for a new version of SGML that was tailored for web use. If it were pitched correctly, this new markup language could be used both as a more rigid presentational language to replace the annoyingly non-standard and broken HTML, while also being usable as a more friendly alternative to SGML for applications requiring structural markup.

XML is a derivative of SGML. In many ways it enforces a more rigid structure than its parent language, for example, requiring that all tags be explicitly terminated. It does not, however, require that an XML document be validated against a DTD. This is a significant potential simplification, in that it allows developers to leap right in and create XML documents, that, as long as they are legal in terms of their general structure, are compliant with the standard. And, of course, the mechanism of DTD validation is still there for situations in which it is desirable. The W3C is currently in the process of encouraging web developers to only produce HTML pages which validate against an XML DTD. This validatable HTML is known as XHTML.

An XML alternative to the DTD template has also been developed, the scema. An XML schema is an XML document which fulfills the same role as a DTD – providing a template against which to validate compliant XML documents.

Creation Software: XML can be created with any text editor which supports unicode. XML parsers are only required to be able to read unicode streams, although as noted above, it is possible to generate a UTF-8 compliant unicode stream using standard ASCII tools, as long as non-ASCII characters are either avoided or encoded as entity references (for example **ä** as `&#xe4;`). Some form of automatic testing of well-formedness and validity is extremely desirable in creating compliant XML documents, and thus a more specialised XML application is very much the preferred option. The Emacs extension PSGML can be used as a composition and validation tool for XML, just as it can for SGML. James Clark has developed an Emacs extension exclusively for XML, called nXML. Commercial validating XML editors exist for all major platforms - (not all of the following are available on all platforms) – oXygen, XMetaL, XML Spy, Emile, Framemaker.

Viewing Software: Given that XML is general structural markup, which does not necessarily imply any presentational format, viewing the content of a document may often involve a transformation into a more presentational format, such as XHTML. XML transformation can be achieved using an XSL stylesheet, which is itself an XML document. The XSL stylesheet contains information about how a document should be mapped into a different structure. Internet Explorer, in versions 5 and above, does an automatic transformation of XML documents into an indented tree structure where it is possible to 'open and close' tags to display or hide their contents. For untransformed XML, a text editor or and XML editor is an effective way of viewing the document's structure.

Portability: The enforced use of Unicode as a character set ensures that XML is viewable and processable in any Unicode-supporting environment. The identity of ASCII and UTF-8 when encoding only ASCII characters ensures that most XML is usable even on platforms which do not directly support Unicode.

File Structure: XML is a stream of characters incorporating markup.

Repurposing: Like SGML, XML is readily transformed between DTDs. In fact, because some of the structural constraints introduced in XML to eliminate ambiguous tagging, it is easier to transform than SGML.

TEI

Description: SGML's ability to markup content in a structured way was useful in areas other than business system development. Academics were excited by the possibilities it offered particularly those engaged in textual studies. Previous plain text encoding of academic texts had resulted in the loss of much styling and formatting information. Some procedural control-coding schemes had been created, such as COCOA, but these were not as extensive or extensible as SGML promised to be. In 1989, the Text Encoding Initiative was launched, a major international academic project to develop an SGML tagset for textual studies. The TEI DTD has become the standard for academic text encoding in the humanities and language studies, and continues to actively extend its range and usability. In its most recent revision, the TEI has been modified to also become an XML-compliant tagset.

Creation Software: As the TEI is an SGML/XML tagset, a TEI document can be created in any SGML or XML editor. Specific support for the TEI is built into a standard installation of the XML editor Oxygen, and there is an adapted version on Emacs available (TEI-Emacs) for TEI-compliant document creation. Sun's open source application suite Open Office uses an XML-based file format for its data storage, and it is capable of performing an XSL transformations upon these files before loading or saving them. XSLT filters exist to allow Open Office to load and save TEI XML documents.

Viewing Software: See the sections for SGML and XML (above) for software that facilitates the viewing of TEI-SGML and TEI-XML respectively.

Portability: TEI documents are subject to the same portability issues as their host languages SGML and XML – they are easily transportable between systems with some small provisos, described in the sections on SGML and XML above.

File Structure: TEI documents are instances of SGML or XML and have the same format as any file of these types (see respective sections above)

Repurposing: TEI is the most fully-featured schematisation of the content and structure of textual objects in existence. In both its SGML and XML forms, it is readily transformable into any of the other formats detailed here, and given its long history, many of these transformation processes already exist and are well-documented.

DocBook

Description: Docbook is an extensive SGML DTD for creation and interchange of book-like documents. It began life as an internal document format within the O'Reilly in 1991, and developed over the next decade to be an extremely fully-featured DTD. DocBook is currently administered by the OASIS group, who specialise in e-business standards promotion. DTDs exist for both SGML and XML.

Creation Software: Any of the editors mentioned in the SGML and XML sections above. Many commercial XML editors, for example Oxygen and XML Spy, support DocBook document creation out of the box.

Viewing Software: Raw DocBook files are viewable in any text editor. However, it is more likely that a reader will want to see the DocBook document transformed. Stylesheets exist to tranform both SGML and XML DocBook documents into a variety of formats such as HTML and PDF.

Portability: DocBook documents are subject to the same portability issues as their host languages SGML and XML – they are easily transportable between systems with some small provisos, described in the sections on SGML and XML above.

File Structure: DocBook documents are instances of SGML or XML and have the same format as any file of these types (see respective sections above)

Repurposing: The ready availability of transformation stylesheets for both the SGML and XML versions of DocBook mean that texts encoded in this format are easily translated to a number of other formats. For transformations that are not currently implemented, a stylesheet can be created, although this is not trivial.

Open e-book Publication Structure

Description: Open e-book is an XML-based document format, designed to provide a technology-neutral method of representing electronic texts. It is developed by the Open e-book Forum, who are a trade and standards body dedicated to the promotion of electronic publishing. The Forum is made up of many large players in the spheres of both publishing (Random House, McGraw-Hill, Harper Collins, Simon & Schuster) and technology (Microsoft, Adobe, Sony). The Open e-book tagset itself is a subset of XHTML. The Open e-book Forum was established in 1998.

Creation Software: Any software capable of creating XML or XHTML (see above).

Viewing Software: Any web browser will display an Open e-book compliant document. In addition to this, there is the eMonocle Open e-book Reader, and the Mobipocket reader for portable devices. Until recently Adobe made an Open e-book reader (called Adobe E-book Reader) available for free. However, with the release of Adobe Reader 6, this program has been discontinued, although its Open e-book display funtionality has not been bundled into Reader 6.

Portability: Open e-book documents are XML, thus they are extremely portable (see XML entry above).

File Structure: Open e-book documents are a variety of XML stream. Additional styling information can be stored in a linked Cascading Stylesheet file. Please see the section above on XML.

Repurposing: The Open e-book format is essentially a subset of XHTML, so documents in this format are extremely easy to republish on the web. The fact that this is an XML-based format also makes them easy to transform into other XML-based fornats, and relatively easy to programmatically convert into other non-XML formats.

Microsoft Reader

Description: Microsoft first released 'Reader', and it's accompanying 'lit' e-book format in August 2000. In an effort to capture a share of the Personal Digital Assistant market from Palm, Microsoft was promoting its PocketPC range of handheld computers. Palm PDAs had the popular Palm Reader as their built-in e-book display software. Microsoft wrote Reader to be its functional counterpart for the Windows CE operating system that powered their PocketPCs. In addition, they created a version that ran on full size PCs running Windows. More recently Microsoft have added a third version, designed to run on their Tablet PC (a laptop sized portable whose principle user interface tool is a touch-screen and pen rather than a keyboard).

The 'lit' file format is, in fact, the Open e-book format (see above) with a wrapper of encryption and application metadata. Microsoft make a range of digital rights management technologies available to creators of 'lit' e-books.

The Reader software allows the user to search the text, highlight passages and to add annotations to words or sentences. A free add-on downloadable from Microsoft enables text-to-speech functionality in the Windows version of the Reader.

Creation Software: A 'lit' creation plugin for the 'Microsoft Word' word processing package is available for free from the Microsoft website. It transcodes Microsoft Word documents into the 'lit' format for viewing in Microsoft Reader. For publishers with an interest in implemementing Digital Rights Management in their e-books, Microsoft has a series of third party developers who supply creation and encryption software. Microsoft Digital Asset Server can be used to broker the encryption of 'lit' e-books on the fly, for online booksellers – although again this software is only available via third parties.

Viewing Software: The Microsoft Reader is the only software capable of decrypting 'lit' files. It is only available for Windows, Windows for Tablet PCs and Windows CE. No other versions are planned.

Portability: Microsoft Reader 'lit' files are encrypted binaries which are designed to be viewed in just one application. This application is only available for Microsoft operating systems. There is, however, an open source tool (ConvertLit) which will export text from Microsoft Reader 'lit' files. It will even interact with the Digital Rights Management libraries installed with a copy of Microsoft Reader, allowing a decrypted version of a protected 'lit; text to be exported.

File Structure: A 'Lit' file is a single binary file containing an encrypted Open e-book-like textual section, some binary application metadata and a bundle of associated files, such as images. Annotations are stored in a separate binary file in unencrypted form. This is the lowest level of encryption available and is known as a 'sealed' book. In addition to 'sealing', 'lit' publishers can add further encryption to their works using third party solutions like OverDrive ReaderWorks. In order to read such files, a user must have 'activated' their copy of the Reader software, a process which requires registration with Microsoft over an internet connection. This associates a unique ID to the user, and allows publishers to mark content as having been acquired by that specific user. This can be done on two levels. The first ('inscribing') involves the 'lit' file being marked with the purchaser's name. The file can be read on any copy of Reader, but the 'inscription' discourages the owner from distributing copies. Finally there are 'owner exclusive' 'lit' files. These are 'inscribed' documents which are encrypted in such a way as to be only viewable in a copy of Reader that is registered to the purchasing user.

Repurposing: Once a document has been saved in a 'lit' file it can only be decrypted with the Reader software. It is to be regretted that even Microsoft's free Word plugin insists upon encrypting its 'lit' output.

Rich Text Format

Description: Rich Text Format is a text-based document format developed by Microsoft as an interchange standard. It consists of ASCII text with interpolated (non-SGML/XML compliant) application markup. Basic presentational markup is supported, as well as some more advanced features like tables.

Creation Software: Nearly all Word Processing packages will export into Rich Text Format including Microsoft Word, Wordpad, Open Office and Wordperfect Office.

Viewing Software: As mentioned above, RTF is widely supported as an interchange format. Nearly all text processing software will be capable of correctly displaying it.

Portability: As it is a plain text stream, RTF is readily portable. The RTF specification is freely available. RTF writer and reader software can be implemented on any platform, without licence restrictions. Commercial software is available that will convert RTF documents into various presentational XML formats.

File Structure: RTF files are plain text streams with presentational metadata interpolated. The grammar of the presentational metadata is freely available from Microsoft. However, RTF has a bad reputation for being complex to parse and difficult to validate. Assigning stylistic properties to text can be done in a number of ways, all of which must be understood by a compliant parser. Tables can also be defined in a number of ways, and it is easy to create table structures which will crash one or other RTF viewing application. In general it is fair to say that the RTF specification has a close relationship with the way that documents are modelled within certain Microsoft products, such as Microsoft Word. For this reason RTF is now considered to be an inferior interchange format to XML, which is entirely implementation-agnostic.

Repurposing: While RTF is a published standard, it is a complex one. Translating RTF files into other formats can be complex, given the plethora of different approaches to document structure that are possible using RTF.

Adobe PDF

Description: Adobe's Portable Document Format was developed in the early 90s to replace the aging Postscript page description language (also an Adobe standard). The PDF was designed to contain everything necessary to create print copies of electronically-designed pages, including vector images, bitmaps and fonts. PDF's uptake built steadily through the late 90s, and as it did so the emphasis on print reproduction was balanced by an equal emphasis on digital delivery and viewing of documents.

The initial PDF specification had not included much structural markup – it was intended to represent a page to be printed and nothing else. With the increasing emphasis on digital viewing, it became clear that more structure was needed in order to allow the viewing software to intelligently reformat the document for differing displays. For example, in PDF v1.0, a page with two columns of text might very well have been stored as a single column with a large gap in the middle of each 'line'. This would be irrelevant for the purposes of printing, but extremely problematic if the text was to be read aloud by a screen reader. Succeeding versions of the PDF specification added more structural tagging to avoid this kind of problem, but there still exists a body of early PDFs in circulation which are difficult to programmatically repurpose owing to this early omission.

PDF is now being marketed (successfully) by Adobe as both an e-book format and a print page description language. The most recent versions of Adobe Reader (formerly Acrobat Reader) have included mechanisms for securing the content of commercial e-books and limiting their use to a named purchaser, using the provision of unique user IDs and encryption.

Adobe publish the PDF specification, and freely allow others to implement writers and parsers, on the condition that the resulting software respects the intended access control mechanisms built into the format.

Creation Software: Acrobat is the official application for creating PDFs, and must be considered the most fully-featured. It is available for Windows and MacOS. As the PDF specification is freely available from Adobe, and permission is automatically given to developers wishing to write PDF parsers and writers (consumers or producers, in Adobe parlance), in theory any piece of document creation software could be written to generate PDFs. Both Open Office and Wordperfect Office support the exporting of documents in the PDF format.

Viewing Software: Adobe Reader is the official PDF viewing application (known until version 6 as Acrobat Reader). As mentioned above, the availability of the specification for PDF means that many applications are capable of opening and displaying PDFs.

Portability: Adobe Reader (or Acrobat Reader) is available of an extremely large range of platforms, including all versions of Windows back to Windows 95, MacOS X and earlier, PalmOS, Symbian, Linux, and many other Unix implementations. The availability of the specification means that potentially even unsupported platforms can run software which can read and write PDFs, given a willing developer.

File Structure: PDF files consist of a series of 'objects' such as fonts, images or text sections, along with a table of the locations of these objects within the file. Changes to the file are saved incrementally in the form of additional objects, with an additional lookup table. Thus, theoretically, changes in the PDF can be rolled back if necessary (in practise, an application can just rewrite the PDF as new with the changes in place, rather than appending them.) PDFs are frequently compressed.

Repurposing: Most PDFs are readily repurposed, given their open specification. The exceptions are PDFs which have access control implemented and, as noted above, early PDF files without structural tagging, which may require manual reformatting.

Palm PDB and PRC formats

Description: The Palm Operating System, which is used on a large number of portable computing devices such as Personal Digital Assistants (PDAs), has a pair of default data structures which are designed to be used by all Palm applications as templates for their own data storage formats. As a consequence, e-book formats for the Palm are all derived from these structures, which are essentially simple databases. PalmDoc, Palm Reader, iSilo, Mobipocket and Tome Raider are all PDB or PRC-based file formats, and all are capable of storing text, with varying degrees of styling and formatting information included. ISilo, Mobipocket and Tome Raider support encryption and varying degrees of digital rights management.

Creation Software: Many PalmDoc editors are available for a variety of operating systems including Windows, Linux and MacOS. Palm Reader documents can be generated using the free utility 'DropBook' provided by Palm Digital Media for Windows and MacOS. Dropbook will convert a standard ascii text document into a Palm Reader PDB file. In order to include formatting and styling information, proprietary markup (Palm Markup Language) must be used in the source text document. MobiPocket documents PDBs are in fact PalmDoc files with added HTML markup, and so can be generated with any PalmDoc editing software. Tome Raider files are created using a proprietary application which processes files marked up with a proprietary markup scheme.

Viewing Software: Many applications exist on the PalmOS platform to view PDB and PRC-based documents. Some reader applications, such as Mobipocket, are capable of opening and displaying a range of document formats. PDB and PRC readers also exist for desktop computers. PalmDoc and Palm Reader documents can also be read on PocketPCs with appropriate software. Certain viewers, such as Tome Raider, are available for some of the higher-specified mobile phones.

Portability: It is difficult to generalise about the portability of PDB and PRC files, as the availability of readers varies with each sub-variety of document. There are readers for PalmDoc documents available freely for almost every operating system.

File Structure: PDBs and PRCs are composed of a table of contents, a block of application metadata and a series of records. In a textual format these could represent pages, or entries in a reference work.

Repurposing: As these are binary files, in the main created out of text files with inserted proprietary markup, they are not particularly easy to repurpose. As they were designed to be used on devices with relatively little memory and low-processing power, their structure is utilitarian. For example, some employ proprietary compression techniques, to fit more text into the same space.

Conclusions

- Many of the document formats described above employ a markup scheme which is either a subset of, or a derivation from HTML.
- A document is easier to repurpose if it adheres to a well-documented standard.

- A complex document will be easier to repurpose if it contains explicit structural markup.
- More complex e-book functionality, such as user annotation and the incorporation of specific fonts with an e-book, tend to be present only in formats which have commercial backing, such as Microsoft's lit and Adobe's pdf.

4. User Survey

4.1 Abstract

It was found that in all sectors and subject areas surveyed, current levels of usage of free e-books are low.

There are some common and consistent concerns about quality assurance of texts and metadata, about the persistence of availability of electronic resources, about various types of hidden and indirect costs associated with the use of free e-books and about technical barriers to their use. There was also concern about the poor design of free e-books and the poor ergonomics of reading text on screen.

There was a general observation that the full range of titles for any given course were never available as free e-books, and so it was not possible to deploy them as a complete solution to providing primary texts.

The potential for making available books which are no longer viable for commercial publication was noted, especially in smaller subject communities.

There was some optimism that Virtual Learning Environments (VLEs)[1] will be a driver for the uptake of electronic resources in general, and therefore possibly for free e-books in particular.

Potential users of free e-books expressed interest in the following functions: copy & paste, automated searching, bookmarking, highlighting, annotation and printing of sections.

In the FE sector, very few of those questioned were interested in the potential use of handheld readers or mobile phones. Desktop computers were thought to be the most likely hardware that would be used to read free e-books. Also in the FE sector, it was thought that the availability of free e-books could make the difference between access and no access to the text given the financial constraints, and it was noted that many students are resistant to using library facilities, so the direct delivery of free e-books to students could be useful.

4.2 Introduction to the user survey

The preceding chapters have focussed on the availability and usability of free e-books for teaching and learning. This chapter shifts attention to the needs of current and potential users of free e-books in the UK FE and HE sectors.

The aims of this part of the investigation were to gauge current levels of usage, identify barriers to uptake and explore the potential for improved uptake and more effective use of free e-books.

Three main survey instruments were employed: in order to gather information about the requirements of the UK Further Education sector with regard to their use of free e-books, an email survey was employed. The HE sector was examined by more desk-based research, examining existing reports, and surveying resources and studies on the internet. A more in-depth investigation of both sectors was then

[1] JISC Requirements for a Virtual Learning Environment: http://www.jisc.ac.uk/index.cfm?name=mle_related_vle

carried out by conducting two focus group sessions. The results of these activities are now examined in the following sections.

4.3 Survey of Further Education

An online questionnaire that was constructed and advertised to selected respondents via email. We are grateful to Paul Davey of JISC for distributing the questionnaire through the Regional Support Centres. The questionnaire was divided into two parts – one for people who had already used e-books and one that everyone could fill out irrespective of their previous experience of e-books. We received 47 replies in all, only 14 of which were from people who said they had used e-books before. Though the replies in no way can be seen as representing the UK FE sector as a whole, they give some insight into issues which should be consider in future work in this area. The responses and comments to the questionnaire tally well with the results of the Focus group sessions (see below). The section below provides a summary of the replies. References to a particular question is given within parentheses (Q1-Q20). A copy of the questionnaire can be found in Appendix D.

Respondents

Of the people who replied to the questionnaire, about two thirds were librarians. The others were teachers (25%) and some held ICT-related posts (VLE administrator, web manager, etc). All but two were from the FE sector.

Previous use of e-books

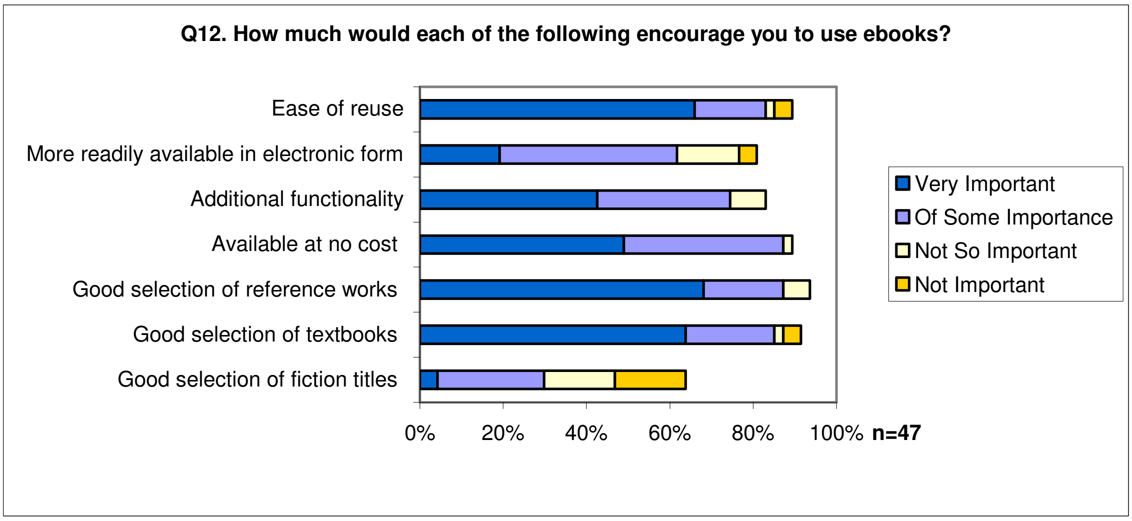
Over 90% of the participants had heard of e-books before. Less than a third of the people who submitted the questionnaire said they had used e-books. Of these, the majority had used reference works for research, teaching or study (Q4)[1]. Some had used fiction e-books for leisure. People had mostly found the e-books through a web search or via a web reference and did not generally considered it difficult to locate the titles they wanted (Q6, Q7). We only had 14 replies to whether the e-books people had used were free or not (Q5). Of those only one person said they had not used free e-books at all while nine said all e-books they used were free.

All but one of the people who replied to the question say they accessed the e-books on desktop personal computers (Q9). It appears most of the e-book users in our survey read their books on the screen although a few printed whole books or parts of them and some re-purposed the material (Q8). They used the word/phrase search more than other functions such as book-marking, annotation/notes, or copy text for quotations (Q10).

Potential use of e-books

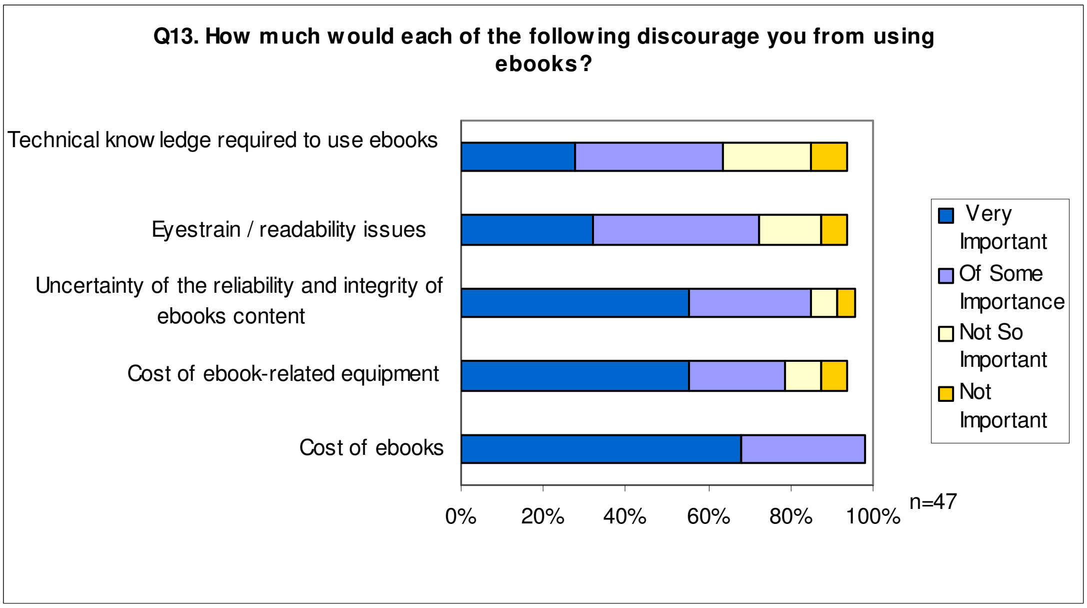
To get some idea of the potential use of e-books, what people would like to do or what they see as useful or deferring factors, we asked a number of questions about this. Figure Q12 illustrates the answers to our question about factors that would encourage the use of e-books (Q12).

[1] NB all questions are listed in Appendix D to this report.



As shown in the figure, factors that would encourage the use of e-books amongst the people completing this survey seem to be related to the selection of titles as well as the ease of re-use. Over 80% said a good selection of textbooks and reference works is very important or of some importance. Additional functionality and, in particular, ease of re-use were also important factors. Only one person said cost was 'not so important' (5 selected the 'no opinion' option). A good selection of fiction titles was not considered very important by the participants in this survey – only two people said that was very important.

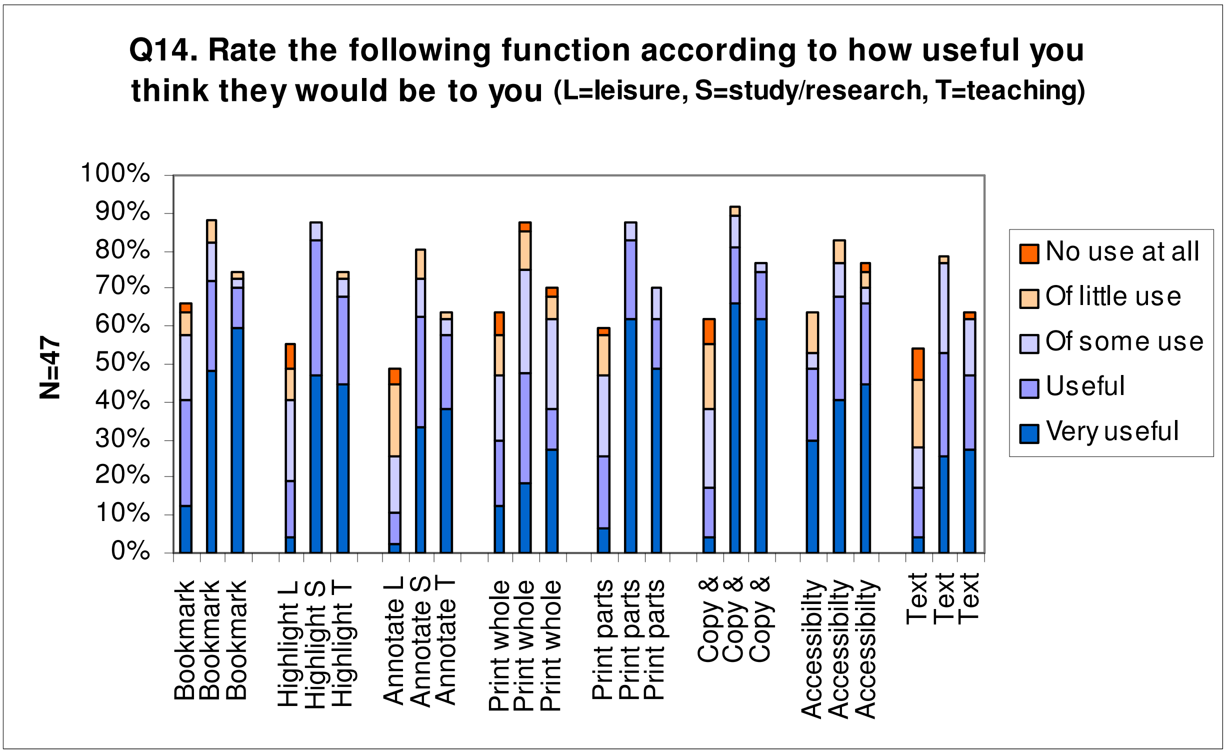
Turning to factors that might discourage use of e-books, the respondents suggest one major factor is the cost, both of the e-books themselves and the equipment needed to access them (see Figure Q13).



Although the majority of those who replied to the question said issues related to the technology were of some importance or very important (technical knowledge needed as well as readability issues/eyestrain), more people were concerned about quality issues. 40 of our 47 respondents said uncertainty of the reliability and integrity of e-books content was an important or very important factor which would discourage them from using e-books. It thus appears that cost and quality of e-books is more important than technical issues.

Functions

The recipients were asked to state how useful certain functions would be to them for leisure (L), study/research (S), and teaching (T) and the result is illustrated in Figure Q14.



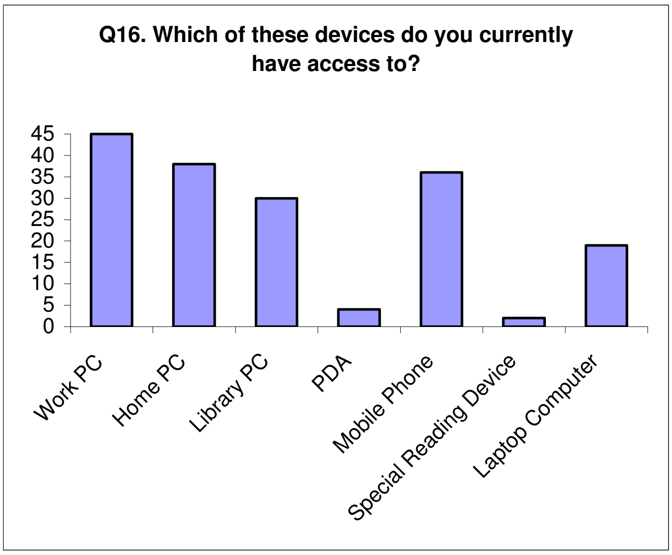
Q14. Rate the following function according to how useful you think they would be to you (L=leisure, S=study/research, T=teaching)

Generally, all functions were less important for leisure and fewer people offered an opinion on that. Where study/research and teaching are concerned, most of the people who replied said they would find all the functions of use. Printing parts or all the book scored high, which is an interesting observation in the context of this being about e-books and that few people said reading on the screen or technology would deter them from using e-books. Copying parts of the books for pasting into other applications was also a popular function. Many said they would find it useful to be able to bookmark, highlight and annotate the text. Text analysis was the least popular function of this selection.

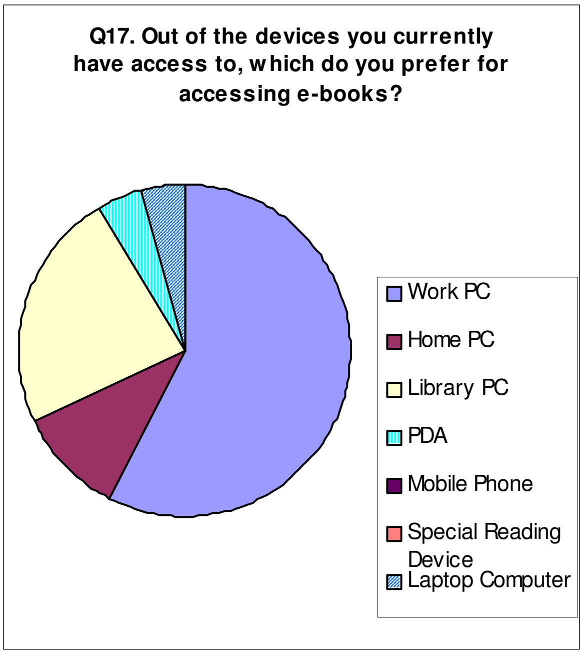
Most comments offered in response to the question of what else people would like to do with an e-book (Q15) related to repurposing. Some wanted to be able to add related material (glossaries, images, author biographies, multimedia) or links and there was also interest in making e-books interactive, linking to a VLE (or vice verse) and being able to compile a course book from sections of different e-books. Further suggestions were having more bibliographic information available, particularly with respect to the edition, and also to be able to link to the e-book from a library catalogue.

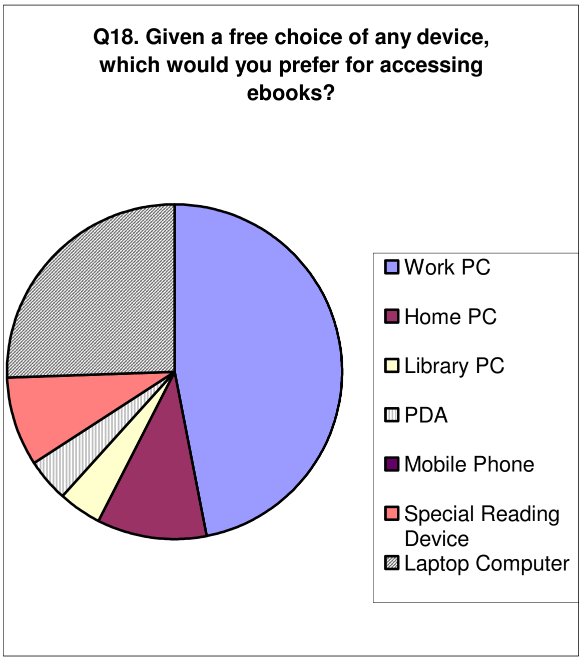
Accessing e-books

Three questions in the survey related to technology used for accessing e-books (Q16-18). It appears the respondents have access to PCs or laptops at work or at home to a large extent and also access to library computers. The majority have a mobile phone (36/47) but few have PDAs or special reading devices.



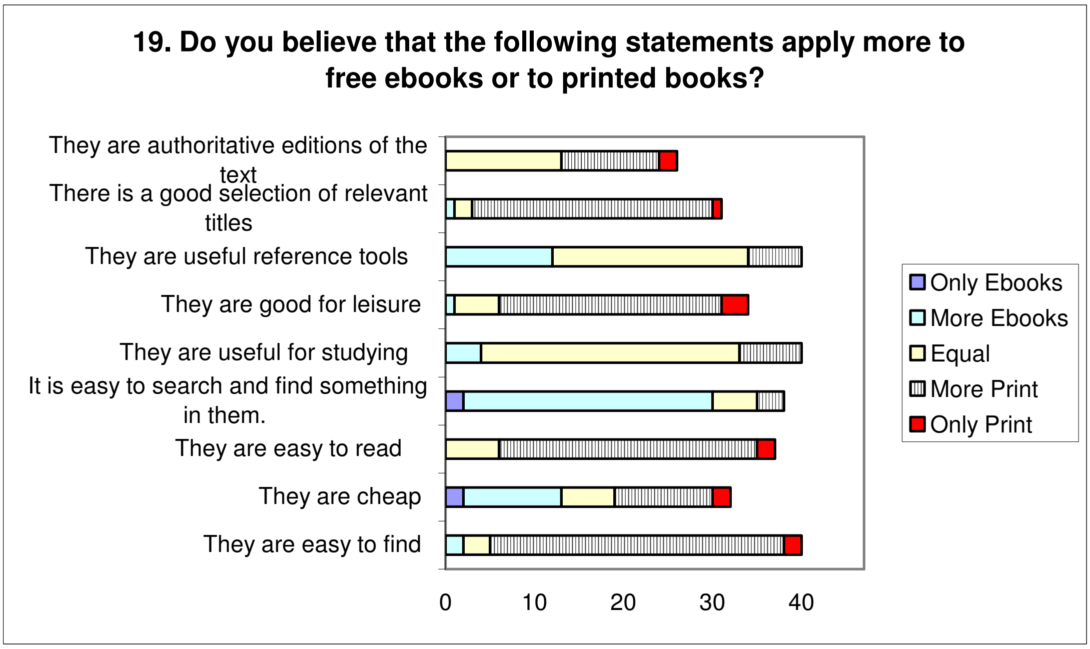
Amongst the devices available to the respondents, nobody would prefer to use their mobile phone or a special device for accessing e-books and only four said they would prefer a PDA or laptop (Q17). It thus appears a stationary PC is the preferred available means for accessing e-books, either one at work or one in a library. Given a free choice of device, more people would prefer to use a laptop or special reading device and fewer would choose to use library or work PCs (Q18).





Attitudes

The respondents were asked to compare e-books and print publications and say whether certain statements applied equally to the two, more to one of them or only to one of them. There was also an option ‘No Opinion’ (Q19). The statements and responses are given in Figure Q19. The proportion of ‘No Opinion’ replies can be inferred by the length of the bars in the figure.



As shown in the figure, the only statement which the majority of the respondents felt referred more to e-books is the one that relates to searching. Over 60% replied that "It is easy to search and find something in them" applied more or only to e-books. Usefulness for studying or as reference tools was something many felt refers equally to e-books and print publications. Nobody suggested e-books are easier to read and only one person felt "good for leisure" applies more to e-books.

Question 13 revealed that uncertainty about the reliability or integrity of e-books was a factor that would defer people from using them (see above). This sentiment is also reflected in the reply to Question 19 where nobody suggests "They are authoritative editions of the text" refers more to e-books. Interesting to note in this context is that this particular statement was the one with the highest number of "No Opinion" replies (21). This would suggest that this is an area where people might need further guidance and advice.

4.4 HE Survey

No formal questionnaire was sent to Higher Education Institutions (HEIs), but some considerable research was conducted in order to find examples of actual current usage. The JISCMail list **lis-e-books** was used to ask for examples of current usage. Major archives of electronic texts were asked if they knew of examples of deployment of their resources in teaching and learning. Internet searches were done, for existing reports, studies and surveys, and for instances of use. Distance and continuing education sites were examined, as was the VLE at Oxford University.

The general conclusion was that very few examples of actual current use of free e-books in teaching and learning were found. From this it may be concluded either that the current usage is hidden from these methods of investigation, or that usage is really very low. There are undoubtedly cases where an individual lecturer has employed a free e-book in a particular course or class, and this is not visible on the web and has not come to the attention of the librarians or e-learning specialists in the institution who may have been able to pass on the information to this report.

One example which was found was the CitySites project, a collaborative effort to deliver an American Studies course, developed by the Universities of Birmingham and Nottingham and King Alfred's College, Winchester[1]. According to the CitySites own description:

> CitySites *is an innovative web-based multimedia research collaboration that explores the meanings and forms of American urbanism in New York and Chicago in the modern period.* City Sites *is at the centre of the* 3Cities *project; a six year AHRB funded research project, based at the Universities of Birmingham and Nottingham, which seeks to foster new modes of analysing American urban culture as well as developing a network of international scholars working on US urbanism.*

CitySites is described by the authors as an e-book, and the text was written by various authors especially for the project. This was made possible by the funding granted to the project. CitySites is an exciting and innovative example of e-learning, and is a type of e-book which may well become an important part of teaching and learning in the near future. However, as a free e-book it is perhaps not typical of the type of resource considered by this investigation. We are chiefly concerned to explore the potential for the use of existing free e-books.

Institutions engaged in large amounts of distance learning, such as the Open University are clearly doing a large amount of research in the area of e-learning and online delivery of texts, but, perhaps surprisingly, are not actually currently employing e-books in their teaching to any noticeable degree.

Several useful studies of the potential for the use of e-books in HE in the United States were considered, but the US experience differs in important respects from that in the UK, and they do not differentiate free e-books from commercial ones, so their findings are not particularly useful for this study[2].

[1] http://artsweb.bham.ac.uk/citysites/

[2] While not focussing on free e-books, and based on the US experience, the reports nevertheless make interesting reading. Some examples are Columbia Online Books Project at http://www.columbia.edu/cu/libraries/digital/texts/about.html has two useful reports; Susan Gibbons, 'E-books: Some Concerns and Surprises' at http://www.lib.rochester.edu/main/e-books/studies/1.1gibbons.pdf; 'E-books and their future in academic libraries', by Lucia Snowhill, University of California, Santa Barbara http://www.dlib.org/dlib/july01/snowhill/07snowhill.html.

4.5 Focus Groups

Methodology

The information on the attitudes of current and potential users is gleaned from two focus group sessions organised at the Oxford Text Archive in June 2003. The participants were drawn from a mixture of backgrounds – FE and HE, librarians, lecturers and “IT champions”. There were 12 participants in each session, and the sessions lasted approximately two hours. The focus groups were facilitated by Chris Armstrong of Information Automation Ltd. and Ray Lonsdale of the Department of Information Science, University of Wales, Aberystwyth. The focus group sessions were structured around a set of questions which is printed in Appendix E below. The sessions were recorded on audio cassette and transcribed, and the results were analysed by the authors of this report. All of the quotations below are from these focus group sessions.

Results

All of the focus group participants had at least used e-books for some purpose, although the level and intensity of use varied. The majority, but not quite all, had used free e-books. The types of titles of which they had experience was wide, and included fiction, reference works, poetry, textbooks and official government publications. The discussions are summarised below under topic headings with quotations to exemplify the key points which were made.

Quality

There was a general concern about the quality of available resources:

> *If you look at texts of Shakespeare on the internet the quality is appallingly low. It’s lower than that of any printed text of Shakespeare since probably 1670. Humans have spent centuries building up ways of ensuring textural transmission as reliable, that problem texts are documented and ways of ensuring that the words are text or not, and a lot of that has gone by the board with e-text.*[1]

And it is not just amateur enthusiasts who are held responsible for this:

> *Even academics are creating e-text and they suddenly throw away all their training and stick them up there with no indication of what edition they are following, or what editorial principles they are using. Not all of them but some.*

The potential for re-editing and repurposing electronic texts can mean that there is a lack of accepted standards for electronic editions:

> *I’ve got a concrete example of* [poor text quality]*, which is an edition of Catullus, when someone chose to renumber the lines in all the poems with numbers which seemed good to them.*

There is an issue of who is to offer quality assurance, and how:

[1] A small amount of editing has been applied to these quotations in order to remove some of the disfluencies of naturally occurring speech, such as repetitions, hesitations and false starts. This was considered helpful in order to make them more easily readable and understandable, but the meaning of the utterances should be unchanged.

> *Quality is a big issue for us. We are very wary in the library of putting resources on our web pages for students to use, for students assume there is a quality endorsement there. We haven't got time to look at all of those E-books have we – its difficult. Plus we are not subject specialist enough to know necessarily whether that's the best translation, or best edition, there's not a lot of data around on the computer to tell you exactly where that text is coming from.*

As a result, quality assurance tends to be left to the judgement of the end user:

> [as a librarian] *you are not guaranteeing any level of quality – it's a bit like using websites from Google, isn't it? Users have got to make their own minds up, evaluate the material themselves. You haven't got the guarantee you get to a certain extent with print.*

There was strong anecdotal evidence of a lack of text integrity in free e-book products:

> *Strangely enough I just used an e-book yesterday and it falls right into this category of E-books that are created badly. Someone created an E-book from existing text and left out all footnotes, all bibliography, everything. At the end of the day I would say that rendered it useless – you couldn't get any of the references, anything. I had to go to the library and take out a hard copy.*
>
> *I would like to mention that the first e-book I read was on a hand-held, was some sort of detective story and the last 5 pages or so were not there.*

This experience can be backed up by the informal surveys of available free web resources. Not surprisingly, pirate editions seem to be particularly prone to particularly poor quality editions.

Availability

As well as concern about quality, there was concern about the ongoing availability of resources. Both are expressed in the following quotation:

> *I think it would be useful but we need to have assurances of quality and the fact that they would be there. If we are going to put a lot of resources into cataloguing they have got to be there in 10 years.*

Librarians in particular are concerned about their ability to assure the ongoing availability of resources. Cataloguing links to resources held elsewhere is particularly risky:

> *I think the real thing is the persistence of it - if its going to be there one year, are you going to be able to rely on it being there six to ten years on? These people that put things on are enthusiasts; what happens if they...*
>
> *What libraries have traditionally done is bought a copy of a book which they keep in their library and it is there in perpetuity, provided other things don't go wrong, whereas the majority of the subscription arrangements are licensed for the period you are paying your subscription and when you change your mind you no longer have the material in your libraries, that's the fundamental difference of approach.*

And while this is a real and widely felt concern, it is something which applies more to e-books bought on subscription rather than free e-books. There is the possibility of permanent accession of free resources, providing that rights in the electronic edition are observed. A similar concern about electronic subscription services, reproduced below, should also be mitigated by free availability:

> *Most of the time they ask you for the money before you see what you are buying. If I go to a bookshop I can look through the book before I say yes that's what I am looking for and then go and buy it. Whereas it says click on this link and then it says fill in your credit card details and then you'll get through to the site.*

There were concerns about accessibility for visually-impaired readers. It was noted that while electronic text opens up many opportunities and possibilities for enhanced access, it is also all too easy to make things less accessible if care is not taken.

The lack of availability of a complete range of titles in any particular area was noted. In some cases the lack of availability of relevant titles or editions may be caused by a US bias:

> *I think what's available in e-books, practically every content you get is US-based. Even fiction, new, breaking fiction tends to be US fiction. I think that is why we have not seen as much take up on e-books as we would with British fiction.*

On the other hand, one participant had noted the advantages of the availability of US material:

> *The other potential for E-books as well, one thing I like to get across to students is that they can go and look at British work, but then they can go and look at say American work and see how the Americans deal with it. Obviously you can't jump on a plane and go into an American library to get some information and ideas for e-books. The biggest potential advantage is that you can different views.*

The lack of availability of a complete range of materials in free e-book form, allied with concerns about quality and US bias, means that their use is not viable as a complete solution in most circumstances:

> *As head of IT … despite being an IT freak, I would still want my students to go to the libraries and look at books to get different perspectives on the same thing. So if one of my members of staff came to me and said I'm going to do the entire course by free e-books I'd say no. I don't mind them using them, using them extensively, but I would never like them to use them 100%. I would still like them to use traditional methods. That's my opinion at this time, it may change in 2 years time when we have got a decent collection but at the moment that's it.*

The limited range of available titles also means that books at an appropriate level are often not available:

> *Then I would go back to my point for further education colleges the levels aren't always appropriate for those colleges, they are just too specialised for my level.*

But 'freeness' can be a crucial factor:

> *The cost, and we are actually literally on extremely tight budgets. I am at a very small and very broke college … so cost to us is nearly everything. Well it's the difference of having it or not having it.*

Portability

The advantages which e-books offer of reduced storage space, online access and unlimited copying were cited by many participants as useful and important:

> *They don't deteriorate or get stolen or damaged.*
>
> *Especially with the fiction I am looking at, e-books with 10 novels on. In other words I would never carry 10 books around*
>
> *Saving shelf space. It blows my mind downloading government books. I am saying books instead of documents because they are thick and big some of them and they have got them on the shelf and I am thinking to myself there must be some way of storing them.*
>
> *One of the biggest effects we see is the secure means of making text available to a large number of people increasing the number of students chasing courses where you can keep them secure and hope that the licence allows you to make them available to a lot of people at once.*

This last point was brought up by several participants. Increasing student numbers mean that libraries cannot provide copies of popular books for all students, and so the fact that there are not the same restrictions on multiple access to free e-books is increasingly important:

> *That's the main thing for me is access. I was saying earlier in 1995 for the first year course that I teach now there were 30 students there are now 110 and the libraries plus the librarians find it impossible to keep up with the demand for books. They can't do it even if they are buying 3, 4, or 5 copies its just not enough and now it's become the major excuse as to why people can't do essays or assignments they can't find the books.*
>
> *For very large classes I'ts fantastic, they can never use the excuse" I couldn't get the book". As long as they can find a PC, if they don't own one, they can use the material.*
>
> *…what I find more and more often now is students in a big class saying "why aren't there enough copies of the secondary text for the whole class". In the past you might put in three or four primary texts of the library copy, but the book about Jane Austen you may have one copy, but our students nowadays in the age of mass higher education think that the library should provide them all with a copy of both the primary and the secondary text they need. A great deal offer to do it - virtual learning environments and all these other things. History will view us as trying to cope in different ways with the transition to a mass higher education system.*

Online access has other advantages:

> *It's very convenient you don't have to go to the library and get the book out, you just sit at home.*
>
> *It's not dependent on a building being open for them to have access.*
>
> *I'm at one of those universities which has invested a lot in wireless networking, and we find that the students' favourite place for reading is the*

> *car park. I'm not quite sure why yet, but I am working on it. We are thinking of extending our work to the car park. I think part of the reason is that we kick them out of the suites at 10 o'clock or whatever and they can carry on working for any time out there – working within range of the network.*

Some disadvantages of portability and ease of access were also pointed to:

> *I think reading an E-book is a different experience, reading a printed book you don't get a sense of context of what you are reading in the same way you don't know whether you are reading the first bit of a very long book or something that will finish on the next page and you don't have the same sense of whether you are in the middle of it or – you can find that out - but you don't that sense in your hand as you are reading it.*
>
> *One problem in practical terms of that is there anything to do with the browser you can jump out of the document into something completely different this is a problem and students hit the back button and they are into* [another website] *and when you confront them with the fact that they have misquoted something completely absurd they say it's on the site and you say it's not actually but that distinction is very hard to enforce so you can try to design around that.*

And not everything may always be easily accessible online:

> *I have encountered technical problems to do with the way the system works because of the firewalls. There are practical problems to do with institutions.*
>
> *One of the problems we have had relates to the sort of proprietary things we were doing a study with 10 e-books and when we download a book into the reader (a) that reader has to be registered and (b) the book gets tagged specifically to the hardware and if I wanted to lend somebody my book I have to lend them my reader as well, and if you are trying to do research study when you want as many people reading as you can its horrendous and I think again that's one of the things that is slowing down the uptake while people are locking it into proprietary models. Its just going to slow it down.*

Although in the latter case, this type of rights management software is less likely to be implemented in the case of free e-books.

The reading experience

Some librarians and users are addicted to printing:

> *Actually even for me for a research piece, unless its just one page or something, if it's a section I will print it out, but if I find a book I will print the whole thing out and I will bind it.*
>
> *I find it very difficult to read great long things on screen, its much easier to print it out and then read it.*
>
> *I have only started using* [free e-books] *recently, I have mainly used government and shall we say public bodies downloadable PDF files, say something like the Public Health Laboratory Service. Latest statistics on alcohol use in teenagers, which I actually bind up and put on the library shelf so I am using them as a free resource that is then borrowable so that students can either find them of their own accord on the internet or they can*

> *actually take it out as a hardcopy, and that to me is a very good free resource.*

But others are prepared to read on screen:

> *I would have agreed to that 10 years ago, but I think that in the last 10 years I have spent so much time reading things on screen, and different screens and reading different stuff that I think you do get into a different way.*
>
> *Yes I find I don't spot mistakes.*

So, while some may be comfortable with reading on screen, they are aware that they read in a different way.

Various problems with the ergonomics of e-books were explored:

> *The screen is only one part of the ergonomics you have got to get the whole of the ergonomics right. You have got to get the position right. If you watch students focus on screens they slouch down like this they have the key board across their kneecaps. If you can find somewhere that's comfortable for you that's the secret. You just become used to being comfortable with books and that's our intention. It's all down to the ergonomics, a lot of people put up with bad ergonomic conditions in the work place.*
>
> *Clearly ergonomics is important but the fact is that screens are still rather poor, the screen actually displays between about one and two million bits of information, whereas in colour slides there's fifty million bits of information per square inch, and you can carry on extracting information to that depth if you want to. So we are talking about several orders of magnitude from our screen and we are used to with our naked eyes looking at good quality images. Until screen technology is going to be better reading is always going to be more attractive in other ways.*
>
> *That's an interesting point because I think in order to do text well on screen you must do it differently to on paper, as you said on paper the printing is, say on a laser printer, some 300 dots per inch, whereas your best screen is going to be 75 dots per inch, which means that you can't be putting italics, serif fonts on the screen and expect it to be legible. You need sans serif fonts. Most of us probably work in Times 12 point when we are on the screen and that's the worst thing you can do. … if we want to optimise the screen, put it in Helvetica or something, put it in 14 point, and then format it at the end. So I think it is all tied in with the experience around e-books and if we are going to try to do too much book-like stuff in them, its going to look horrible.*
>
> *It's a deep physiological question as to whether the attempt to imitate to produce the facsimile of a professional book is always the best thing to be doing. It gives us sufficient clues which you rightly say we want, but probably in 50 years time we will think that this was like the early motor cars imitating horseless carriages or railways being four and a half feet wide for no better reason than that was the width of the Roman carts. We don't yet know which bits of those printed books will be beneficial to carry over and which we can junk.*

Several participants expressed a feeling, for themselves and for others, that the physical properties of books are attractive to the user:

> *There seems to be a sort of psychological…psychologically people are against it. They want to be physically holding the book, the paper…*

> *It's nicer, much nicer. They feel nice. It smells nicer.*

Note however, that it was also remarked upon that some groups of users find electronic resources attractive, and increasingly expect them to be available.

Resource discovery

> *I think the crux is how you can discover they exist, in other words how do you provide access through your catalogue of your normal sources, how you can integrate that record for an electronic book with the other stuff that you put before people. There are a lot of questions there.*

A variety of means are used to attempt to locate relevant free e-books. These include using websites specialising in free e-books and electronic texts, search engines and lists of links on university library websites.

> *I find other university library web sites very useful. For instance Warwick has a lovely page on e-books with it. E-books that it subscribes to as well as links to free e-book collections. And I use things like that quite unashamedly.*

Formats

There was a limited amount of discussion on the merits and problems of different formats, reflecting the limited experience that the participants had with actually using free e-books, particularly in proprietary e-book reader formats. There was however an interesting discussion of the merits of user-friendly interfaces:

> *The fact that all the proprietary software have non-standard navigation is a step backwards, There are more or less two standard models to navigate our way around the interface, but when we look at these formats we find a lot of icons we have never seen before unless we have used that particular e-book reader before.*

And of the merits of open standards:

> *Is this not a terrible contradiction that we are talking about free resources and then putting them in a proprietary format, which is owned commercially, controlled commercially could be changed at any time. The final format is one thing if it is going to be delivered to your screen or a printer that's fine, but surely we should be arguing for it to be originated and kept in a open standards format and then transformed from that to the device as necessary. So let's stick with open standards, please. Or we will make just another mess of proprietary stuff which breaks within three years and is unusable after 10 years.*

> *…on formats, we were speaking as if the whole world read and spoke English that's a great mistake. There are works in ASCII and we shouldn't be doing that any more really. Texts should be created in Unicode which covers all the world's languages and that's another advantage of XML because if you use XML you get Unicode.*

What you can do with them

The potential for annotating, reformatting and repurposing free e-books was noted:

> *The fact that they may be out of copyright you have more flexibility to do things to mess them around.*

However, it appeared that among the focus group participants, no-one was actually doing any annotation of free e-books. Technical difficulties with this were noted:

> *I think its because of the platforms because most of us at college have got PCs. You can't scribble on it you can't say "I don't agree with this, this is rubbish" You can't highlight it so you need to print it.*

The focus group participants had been shown some Tablet PCs, and were interested in the potential to use these for annotating electronic texts:

> *This is the first time that I've had a chance to look at these Tablets. I thought that the highlighting and note taking facility on those was quite good.*
>
> *I have only spent 5 minutes looking at these things and thought they looked absolutely wonderful, they were really good. But you know there are lots of things which look really good at first sight and then you don't find, I think as you have described, your students, who've said, well, in reality they won't and don't do it, so I don't know.*

Hidden costs

There was concern about hidden costs. As noted above, some of these were associated with the time and effort involved in quality assurance:

> *Quality is a big issue for we are very wary in the library of putting resources on our web pages for students to use, for students assume there is a quality endorsement there. We haven't got time to look at all of those E-books, have we? It's difficult.*

There was also concern about hidden costs associated with intellectual property rights. Even though resources may be made available on the web or elsewhere and advertised as free and freely available, there would still be a responsibility on whoever copied and used them to ensure that the practice was legal, and this could be time-consuming and costly:

> *If on top of that you have in addition to get clearance from different publishers, different rules, you can see it is a bit complicated, and as a lecturer I wouldn't go into that myself I would need an institution to do that for me and for someone to help on the design and so on.*

It was noted that there are costs involved in the accession and cataloguing of free e-books:

> *Well at the moment its certainly a problem, its mirrored for us and I think a lot of libraries, with economic journals when they come in bulk in big deals and most libraries find they cannot cope with cataloguing individual titles in the way that they would normally catalogue and so I think many places have had, perhaps, to resort to separate web pages directing people to electronic journals whereas we still expect people to use our catalogues to find the other journals and this is a major two way split which is very unsatisfactory and the same is likely to be true in the e-book field.*

Other hidden costs which concerned the participants were the costs associated with the use of ICT. The infrastructure and support necessary to allow students to use electronic texts could be very high in comparison to the costs involved in the use of print media.

> *Well the institution that provides them has got to have the hardware to enable them to start, or sufficient hardware to provide the equipment, freedom of access.*
>
> *And the reader needs to have access to quite sophisticated equipment which half of the world doesn't have. You say they are free but they're only free if you've got an enormous amount of kit and the bandwidth in order to download it*
>
> *As soon as you start making requirements you say to the students this is required reading for the course you have to support it and historically one of the lessons for supporting computing is that the people cost at least 10 times as much as the machines and perhaps more like 100 times as much as the machines as you can buy a machine for £300 a year once you write off its cost, but to buy a person it cost £30,000 a year. So you say to the kids here's your course reading, I want 10 e-books and I will loan you a laptop and they come back and say how do I do this why has this gone wrong and all that sort of thing and you end up paying a lot of support staff. So I agree nothing is free.*

'Business' models for free e-books

Some interesting perspectives were offered on the possibilities and problems relevant to specific subject areas.

> *I also think different subject areas have different problems with this. For my particular area it's great for old out-of-print texts, but I couldn't use it as a primary reader because it's not up to date enough for it. You give a student a textbook and it has to be the latest in that particular subject.*

In particular there may be opportunities where there is no commercial possibility of publication, but electronic versions of a text may be available. (It should be noted that the rights in an edition may however still be held by a publisher, effectively preventing the publication of a work.)

> *I use e-books sometimes and it tends to be text books that are nearly 150 years old I am involved with people who teach very oddball subjects, like Sanskrit, which no text book has been written for in 20 years … so in the terms of distribution no print publisher is going to touch this kind of thing, but … electronic ones seem to be available.*
>
> *I was just thinking of the way I use a free e-book in teaching actually and I am not too sure of the definition. I published a book in 1990 which went out of print and I still refer to it in my lectures. I've updated things as well and it's still a usable text and because the publishers didn't pick up the option to reprint, under the terms of the contact with them, the copyright reverted to me. So I now have the entire book on my web site and make it available to the students on my website so that I can still refer to the text even though the library copy perished ages ago.*

It was also noted that a free e-book can be accessioned by a library, whereas commercial works can only be licensed for the short term. This would help address the problem of persistence of availability (although there are risks and costs associated with the preservation of electronic data which would need to be addressed).

> *There have been lots of cases where the sensible thing to do is to pay 50p or even a couple of pounds to get a commercial product, but I think there has to be a major role for free academic sites. If only for the reasons you mentioned because if we could accept a guarantee that it will be around. You can own it outright, whereas a commercial company may take it away from you in a year's time.*

There was a feeling that the types of work that are available in free e-book form are more likely to be of use to humanities disciplines, where texts have a longer ‘shelf life’.

> *In humanities you would expect a publication to have a more or less indefinite lifespan, if you were in medicine or engineering then there is less of a difference because the content is likely to have obsolescence within a reasonably short space of time and so the difference between a subscription model to a purchase model is much less different than in humanities as is in social sciences. You have got a major contrast.*

VLEs – the future?

There was optimism that the increasing use of Managed and Virtual Learning Environments would make the introduction of electronic resources into the curriculum easier and cheaper:

> *It seems to me that there isn't much uptake on free e-books. VLEs may be the way of getting electronic resources to the students.*
>
> *I wouldn’t use it as a primary reading tool, it’s quite useful for taking big chunks out and putting it into a module just to save you typing it up especially if it’s in translation and also I do use some to some degree in the VLE. It’s there and it makes it easier for students to get hold of it.*

Indeed, students may soon expect electronic editions of texts:

> *Students' expectations are rising all the time. They are going to be using things like Blackboard, WebCT or whatever and they are going to be used to things that are more attractive, hopefully.*
>
> *Another thing is that students expect things to be on the internet. If you can push a good courses on the internet they’re happy – customer satisfaction.*

5. Conclusions

The findings of the research detailed in the preceding chapters are summarised here. These are first classified in terms of the questions asked in the introduction to this report.

Availability

What free e-books are freely available with the minimum of intellectual property rights constraints?

While there is a wealth of electronic text freely available on the internet, there are serious problems of quality assurance in respect of text and metadata. Furthermore, the kaleidoscope of different formats in which the texts are available is at best confusing, and at worst an insurmountable barrier to their use.

It is likely that even if a full range of the titles needed for a particular course or other teaching purpose is available for free in electronic form, then they will not all be available in a common format and of sufficient quality.

Furthermore, the restriction of the scope of this study to *free* e-books further reduces the available range of titles. In some subject areas, virtually nothing freely available is useful; in others, particularly in humanities disciplines which make extensive use of old texts, a large amount of the reading lists is potentially available, though not usually in the best critical editions.

User needs

Who are the actual and potential users of free e-books, and what are the possible uses?

The level of current usage is difficult to calculate, but is almost certainly very low, but may expand rapidly with the twin drivers of (i) the availability of vast amounts of resources on the internet and (ii) the introduction of an ubiquitous delivery mechanism for electronic course materials in the shape the of the VLE. The possibilities are limited mainly by the availability of free resources in a given subject area, and this will vary widely.

Possible uses are for reading primary course materials, search and retrieval, quotation, annotation and for more complex linguistic and content analysis using software tools.

Repurposing

To what extent can existing freely available e-books be repurposed, converted to other delivery formats, and assimilated into other activities or collections?

A large proportion of the free e-books currently available are in plain text, with no structural markup at all. While it is unproblematic to print these texts, and to a certain extent to paste their contents into Virtual Learning Environments, they are not ideal. Their lack of formally marked-up structure is a barrier to their being reformatted for differing devices, aggregated into collections, searched meaningfully and preserved in the long term. Texts which have been marked up in complex structural tagging schemes like the TEI are readily reformatted, aggregated and searched. They can also be transformed easily into other structured formats for presentation, such as HTML, printed and incorporated into

Virtual Learning Environments. There is however, an overhead of expertise and time in producing such texts.

Until the needs of students using VLEs become clearer it is difficult to make judgements on which e-book formats will be of the most use for teaching and learning. Storing texts in formats using open standards such as XML, with structural markup provided by a tagset like the TEI, ensures that the content will be able to be transformed and delivered in its most useful form in the future.

Many of the advantages of VLEs stem from their nature as networked applications – the linking of material with related material is a key feature. The increasing power of mobile devices, and their increasing networking capabilities, may well mean that in the medium term there will be a convergence of delivery formats between desktop computers and mobile devices. This progress will also facilitate the direct use of VLEs via the internet on mobile devices, and this may well be the model for academic use of e-books in the future – digital texts accessible anywhere while retaining all the advantages of placement within context in a VLE.

The findings of this investigation can also be classified differently, according barriers to and opportunities for the uptake of the use of free e-books in teaching and learning in FE and HE in the UK.

Barriers to uptake:

- Lack of availability of a complete range of titles for any given course
- Doubts about quality assurance
- Lack of confidence in the persistence of availability of resources
- Costs involved in the cataloguing, archiving, management of resources
- Costs involved in computing support for users
- Poor design of free e-books and poor ergonomics of reading on screen

Potential opportunities for uptake:

- ‘Freeness’ could be vital to FEs
- Existing digital repositories and resources do exist
- The fact that free e-books tend to come in open formats and free of IPR restrictions means that they can be more easily repurposed, integrated into institutional systems and preserved
- Free e-books may be more useful for the humanities than other disciplines, because there is more use of ‘old’ texts
- VLEs represent an opportunity for the delivery of free e-books to the student.

Drawing on these findings, the following recommendations are made to the JISC to help shape future policy in relation to free e-books in teaching and learning.

Recommendations

- Take measures to offer more comprehensive ranges of titles in specific areas
- Support efforts to migrate existing collections to common formats
- Institute a system of quality assurance (of text integrity and metadata)
- Ensure the permanence of collections
- Support the professional, standardised cataloguing of electronic resources
- Offer support for users in the basic ICT dimension of the use of e-books
- Offer help with integration into VLEs.

The ways in which these recommendations can be addressed are three-fold: Recommendations 3, 4 and 5 can be addressed by activities centred on a national archive of electronic texts, such as AHDS Literature, Languages and Linguistics. A checklist for evaluation and validation of electronic texts which has been proposed by the Oxford Text Archive is included in an appendix below to give a concrete idea of what this validation could involve. Such a central archive could also develop a collections development policy to address Recommendations 1 and 2. Recommendations 5 and 6 would be best addressed by drawing on and documenting existing expertise and experience which is currently being developed in the adoption of VLEs across the country.

Appendix A: Character encoding history

Character encoding before ASCII

By the late nineteenth century, early telecommunications firms were looking for a way to mechanise the process of encoding and decoding the text that they sent. From the invention of the telegraph, in 1832, Morse code had been used to transmit messages, with a human operator at each end performing the encoding and decoding. Morse uses variable-length character encoding - the codes for common characters are shorter than the codes for uncommon ones. This made Morse less tiring on the fingers of the manual telegraph operators. It also made it extremely difficult to mechanise. As telecommunications became more widespread towards the end of the 19th century, it became clear that it would soon be impractical to rely upon legions of trained Morse operators. A standard, easily machine-transmissible encoding scheme had to be created.

In 1874, Emile Baudot of the French Telegraphic Service developed a scheme based upon a five ‘bit’ encoding. Each character was represented by five binary values, giving a total of 32 possible characters (2x2x2x2x2). This covered the 26-letter Latin alphabet, in addition to the whitespace, line-feed and carriage-return. Two non-printing ‘shift’ characters were also assigned, allowing numerals and punctuation to be transmitted.

Baudot’s constant-length 5-bit encoding scheme was extremely successful, and a variation of it was adopted by the British Post Office as its telecommunications standard. By the 1930s, there were several differing 5-bit schemes in use around the world, all based upon Baudot’s. Finally in 1932 the Comité Consultatif International Telegraphique et Telephonique (CCITT), a cross-industry body, introduced a synthesis of the current schemes called the International Telegraph Alphabet number 2, or ITA2. This successfully became the standard, although as always in these matters, some variations were developed.

In parallel with the creation of a telecommunications scheme, a clerical system of encoding was being developed. In 1879, Herman Hollerith, a graduate statistician, started work with the US Census Bureau. The job of processing all the data gathered by the census was gigantic – the results of the 1880 census were tabulated by hand and were not finally published until 1888. Hollerith decided that it ought to be possible to use punch-cards - which up until that time had been used only in industrial machinery such as powered looms – to store the census data and make it subject to mechanical tabulation. Hollerith’s system encoded characters using twelve possible punch positions. Clearly this could potentially have encoded up to 4,096 different characters. However, in reality, the twelve ‘bits’ were only called upon to represent one of 69 potential characters – Hollerith was concerned with the amount of actual hole-punching data-entry staff would have to do, and tried to limit the encoding of any one character to a maximum of two punches. By 1890 when the next census was due to be taken, Hollerith’s scheme was ready. Where the last census had taken eight years to process, the 1890 census took just six weeks, thanks to Hollerith’s encoding scheme. Following this great success Hollerith quit his Census Bureau job set up in business on his own. He called this venture the Computing-Tabulating-Recording Company, and it soon dominated the market for mechanised clerical systems. Later it would become the International Business Machines Corporation, or IBM.

The birth of ASCII

By the late 1950s the ITA2 encoding standard and its American sister-standard Teletypewriter code were beginning to show their age. As the use of typewriters became widespread, the disparity between the characters that could be typed on a QWERTY keyboard and those that could be sent via teletype began to become irksome. The American Standards Agency decided to define a new encoding scheme that would allow the transmission of this larger range of characters. Starting in 1963, the ASA published a series of standards that they called ASCII, or American Standard Code for Information Interchange. ASCII was a 7 bit scheme, allowing for 128 different characters, and by the 1967 revision it included all 96 characters found on a standard QWERTY keyboard, in addition to some control codes representing such things as line feeds and carriage returns. Also in 1967 the International Organization for Standardization (ISO) recommended that a slight variation on the 1967 ASCII standard be adopted worldwide (the only difference was that ISO called for 10 of the unused character positions to be ring-fenced as space for adaptations of ASCII for various international character sets). At the same time, most computer manufacturers standardised upon ASCII as their character encoding solution. The only exception was IBM.

In order to retain backward compatibility with their own Hollerith-derived encoding schemes, IBM had invented an entirely different non-ASCII compatible 8-bit set, which they called Extended Binary Coded Decimal Interchange Code (EBCDIC pronounced 'Eb-See-Dic'). The success of IBM's mainframes during the 1960s and 1970s has ensured that this encoding scheme is still in existence today, in legacy data. However it was never popular, as it caused interchange problems with the spreading ASCII scheme. In addition to this, IBM had chosen to solve the issue of non-Latin character encoding by creating 57 different regional versions of EBCDIC, which they marketed with their mainframes around the world. Mappings between these versions were difficult to obtain, and this further contributed to the unpopularity of EBCDIC with large multinational customers.

Internationalisation has also been an issue for ASCII. It soon became clear that the 10 character 'internationalisation area' within the 7-bit ASCII scheme was inadequate. ISO recommendation 2022 describes how the 7-bit ASCII Latin set can be extended into an 8-bit international character, with more room for accented and variant characters than the original 10 spaces. ISO 2022 provides a template for people who wish to create an ASCII variant. It does so by defining where these additional characters should be placed within an extended 8-bit ASCII scheme. ISO-8859 is an example of an ASCII variant created using the ISO 2022 template – in fact it is a set of 14 different ASCII-variant sets which cover languages and characters used in Eastern and Western Europe and the US. The first of these 14 (ISO-8859-1) is the set that contains Western European (and thus US) characters, and is consequently the most widely used of the variants, both on the internet and elsewhere.

Unicode

The advent of the World Wide Web in the mid-1990s began to show the inherent deficiencies in the ISO 2022 approach to international character encoding. A web browser allows a user to view pages written in many languages, simply by clicking a link. Unfortunately this ease of use leads to many difficulties. Web pages are easy to create, but not easy to create well. Frequently authors will fail to specify the character set that the page should be viewed with. Their page will appear correct to them, and to others whose computers use the same character set by default. However, non-local browsers will see the page in their own default character set, which will almost certainly mean a garbled array of letters, numbers and ‘special’ characters. The idea of internationalisation through many ASCII-variant character sets is thus extremely problematic.

What was needed was a universal character set, that could not only encompass the contents of all the existing international and historic code pages, but ideally allow room for future expansion. Developers at Apple and Xerox (and later Microsoft) had independently found themselves contemplating this requirement and through their cooperation, the Unicode consortium was formed in 1991.

The Unicode standard allows for the encoding of up to 1,114,112 different characters, although only around 100,000 are currently assigned.

Appendix B: Sources of free e-books

Oxford Text Archive	http://ota.ahds.ac.uk/
Arts and Humanities Data Service	http://ahds.ac.uk/
Arts and Humanities Research Board	http://www.ahrb.ac.uk/
Humbul Humanities Hub	http://www.humbul.ac.uk/
Text Encoding Initiative	http://www.tei-c.org/
Google	http://www.google.com/
Voice of the Shuttle	http://vos.ucsb.edu/
Alex	http://www.infomotions.com/alex/
E-book Locator	http://www.e-booklocator.com/
E-Text Center University of Virginia	http://etext.lib.virginia.edu/
Project Gutenberg	http://www.promo.net/pg/
Project Gutenberg of Australia	http://gutenberg.net.au/
Blackmask Online	http://www.blackmask.com/
The Darwin Correspondence Project	http://www.lib.cam.ac.uk/Departments/Darwin/
The Perseus Digital Library	http://www.perseus.tufts.edu/
The Blake Archive	http://www.blakearchive.org/
Austen.com	http://www.austen.com/
YourDictionary	http://www.yourdictionary.com/
The Bible Gateway	http://bible.gospelcom.net/
Biblioteca Virtual Miguel Cervantes	http://cervantesvirtual.com/

These websites were all verified as available online on 21st November 2003.

Appendix C: Technical Links

Unicode Home	http://www.unicode.org/
Alan Wood's Unicode Resources	http://www.alanwood.net/unicode/index.html
Purchase the SGML Specification	http://www.iso.org/iso/en/CatalogueDetailPage.CatalogueDetail?CSNUMBER=16387
SGML Syntax	http://xml.coverpages.org/sgmlsyn/contents.htm
HTML Home Page	http://www.w3.org/MarkUp/
XHTML Specification	http://www.w3.org/TR/xhtml1/
XML Specification	http://www.w3.org/TR/REC-xml
The TEI Consortium	http://www.tei-c.org/
O'Reilly Docbook site	http://www.docbook.org/
The Open e-book Forum	http://www.opene-book.org/
Microsoft Reader Home	http://www.microsoft.com/reader/default.asp
Rich Text Format 1.6 Specification	http://msdn.microsoft.com/library/default.asp?url=/library/en-us/dnrtfspec/html/rtfspec.asp
Adobe PDF Specification 1.5 Fourth Edition	http://partners.adobe.com/asn/acrobat/sdk/public/docs/PDFReference15_v6.pdf
Palm Developers Forum	http://www.palmone.com/us/developers/
PocketPC Developer home	http://www.microsoft.com/windowsmobile/information/devprograms/default.mspx

These websites were all verified as available online on 21st November 2003.

Appendix D: Free e-Book Questionnaire

N.B.: The original online version of this form is still available at http://ota.ahds.ac.uk/e-books/JISC/form.html, where it is possible to view the range of possible answers to the questions.

The Oxford Text Archive

An Investigation into
Free eBooks

JISC

E-book survey

The Oxford Text Archive http://ota.ahds.ac.uk is undertaking an investigation into the free e-books and their potential use within the HE and FE communities on behalf of the JISC/DNER E-Book Working Group. You can see more about this project at http://ota.ahds.ac.uk/e-books/JISC/.

This survey is now closed. Our thanks to all who completed it.

Part One - Who are you?

What do you do?

□ Teacher

□ Librarian

□ Researcher

□ Student

□ Other …… (Please specify below)

Where do you do it?

□ School

□ Further Education

□ Higher Education

□ Other …… (Please specify below)

What is your email address?

We are asking for your email address in case we need to authenticate repsonses to this questionnaire and so that you can be entered in the prize draw. Your email address will not be passed to any third parties, and will not be added to any mailing lists by us without your further permission.

Part Two - Questions

1. Have you heard of e-books?

□ Yes

□ No

The term "e-book" is often used to refer to various kinds of electronic texts.
For the purposes of this survey, our definition of "free e-books" is as follows:

"A free e-book comprises a document in electronic form, coupled with software and hardware in order to read it. The e-book must be free at the point of use, where the user is not required to make any kind of payment or subscription in order to access or download the e-book. In almost all cases we expect the free e-books in this survey to be electronic editions of material published in print, and which attempt to emulate 'book-like' characteristics"

2. Have you used e-books?

If you answer no to this question, you will automatically be taken to question 12

□ Yes

□ No

3. Approximately how many e-book titles have you used?

□ One

□ 2-5

○ 6-10

○ More Than 10

4. What kind of e-books have you used and what did you use them for?

	Teaching	Study	Research	Leisure	Other
Fiction	□	□	□	□	□
Reference Work	□	□	□	□	□
Academic Textbook	□	□	□	□	□
Other Non-Fiction	□	□	□	□	□

Additional comments

5. Were the e-books you used free or not?

(**NB: Free** here is intended to signify that the e-book in question was provided without charge, either to you or your institution).

○ All Free

○ Most Free

○ Some Free

○ None Free

If you have used free e-books, please include some details of them in the space below:

6. How did you find the e-books you used? (tick one or more)

□ Through a colleague / peer

□ Through a web search

□ Through a reference on the web

□ Through a reference in a print publication

□ Through a print advert

□ Through email / electronic advert

□ Other (Please specify below)

7. In general, how easy have you found it to locate e-book titles that you want?

OK ▾

Additional comments

8. How did you use the e-book? (tick one or more)

□ Read on screen

□ Printed whole

□ Printed part

□ Reused (eg in web page or teaching materials)

□ Other (Please specify below)

9. On what kind of device did you use them? (tick one or more)

□ On a Desktop Personal Computer

□ On a Laptop Personal Computer

□ On a Personal Digital Assistant (PDA)

□ On another mobile device (WAP)

□ Other (Please specify below)

▭

10. Did you use any of the following functions? (tick one or more)

□ Word/phrase Search

□ Copy text for quotation

□ User-defined bookmarks

□ User annotation

□ Other (Please specify below)

▭

11. Did you encounter any problems with any of the following? (tick one or more)

□ Finding the e-book you were looking for

□ Finding your e-book in a suitable format

□ Accessing or downloading your chosen e-book

□ Using functionality such as bookmarks or annotation

□ Finding a free edition of your chosen e-book

□ Confirming you had the right to use your chosen e-book in the way you wanted.

□ Reliability or integrity of the text

□ Character encoding (eg uncommon characters displaying incorrectly)

□ Other (Please specify below)

12. How much would each of the following encourage you to use e-books?

Good selection of fiction titles	No Opinion ▾
Good selection of textbooks	No Opinion ▾
Good selection of reference works	No Opinion ▾
Available at no cost	No Opinion ▾
Additional functionality over print books (bookmarking, searching, etc)	No Opinion ▾
More readily available in electronic form than in print form	No Opinion ▾
Ease of reuse (eg in Virtual Learning Environments, web pages etc)	No Opinion ▾

13. How much would each of the following discourage you from using e-books?

Cost of e-books	No Opinion ▾
Cost of e-book-related equipment	No Opinion ▾
Uncertainty of the reliability and integrity of e-books content	No Opinion ▾
Eyestrain / readability issues	No Opinion ▾
Technical knowledge required to use e-books	No Opinion ▾

14. Rate the following functions according to how useful you think they would be to you

	Study/research	Leisure	Teaching
Bookmark	No opinion ▾	No opinion ▾	No opinion ▾
Highlight	No opinion ▾	No opinion ▾	No opinion ▾
Annotate	No opinion ▾	No opinion ▾	No opinion ▾
Print whole	No opinion ▾	No opinion ▾	No opinion ▾
Print parts	No opinion ▾	No opinion ▾	No opinion ▾
Copy & paste text into other documents	No opinion ▾	No opinion ▾	No opinion ▾
Accessibilty (eg text to speech)	No opinion ▾	No opinion ▾	No opinion ▾
Text analysis	No opinion ▾	No opinion ▾	No opinion ▾

15. What else, if anything, would you want to be able to do with an e-book?

16. Which of these devices do you currently have access to? (tick none or more)

□ Work Personal Computer

□ Home Personal Computer

□ Library Personal Computer

□ PDA (Personal Digital Assisitant)

□ Mobile Phone

□ Special Reading Device

□ Laptop Computer

17. Out of the devices you currently have access to, which do you prefer for accessing e-books?

Work PC

18. Given a free choice of any device, which would you prefer for accessing e-books?

Work PC

19. Do you believe that the following statements apply more to free e-books or to printed books?

They are easy to find	Don't Know
They are cheap	Don't Know
They are easy to read	Don't Know
It is easy to search and find something in them.	Don't Know
They are useful for studying	Don't Know
They are good for leisure	Don't Know
They are useful reference tools	Don't Know
There is a good selection of relevant titles	Don't Know
They are authoritative editions of the text	Don't Know

20. Any other comments

Workshop 'Investigating Free E-books'
Would you be interested in attending the workshop 'Investigating Free E-books' in Oxford on June 13th, or do you know someone who may like to come?

☐ Check here and we will contact you (using the email address you provided above) with further details.

Appendix E: Free e-Book Focus Group Instrument

Preamble

The objective of the afternoon part of the day is to gather some views both about e-books and – especially – about free e-books. For the focus groups, you will be divided into two groups with 12 participants in each. We will go through a list of questions and are looking forward to your responses. We will be recording the sessions and taking notes. No organisation or individual will be identified in our report – we are taking note of who is speaking simply so that we know that different quotes come from different people and so that we can cite a range of views; also it is useful to know whether the speaker comes from a large University or a small College of FE, etc. The recordings will only be used as backup for the note takers and will be erased afterwards.

Since the time for the focus group session is limited, we would appreciate if you would consider this preliminary set of questions in preparation for the discussion. As a precursor to exploring issues surrounding free electronic books, the first area (A) concerns your perceptions of e-books in general, while the second area (B) deals specifically with free e-books.

If you have any questions about the focus group, the preliminary questions or anything else, please do not hesitate to contact the organisers at e-books@ota.ahds.ac.uk.

Here is our provisional, working definition of e-books:

"An e-book comprises a document in electronic form, coupled with software and hardware in order to read it. In almost all cases we expect the e-books discussed in this survey to be electronic editions of material published in print, and which attempt to emulate 'book-like' characteristics."

By a FREE e-book we mean one that costs nothing to acquire, access, read, copy, or use. Do not confuse this with free at point of access (because your library has already paid for a licence).

A e-Books

Introductory questions

A1 Have you used e-books?

A2 What kind of e-books have you used and what did you use them for?

We mean both:

(A) Fiction, Dictionary/encyclopaedia, Text book, Exercise book, Other non-fiction, Other

(B) Teaching, Study, Research, Leisure, Other

Advantages of e-books

A3 What do you see as being the major advantages of e-books?

You might think of, e.g.:

Good selection of titles, Additional functionality, etc

E-books and print

A4 How would you compare e-books with conventional (printed) books?

For example:

Are they easier to find, cheaper, etc?

Access

A5 How did you find (locate) the e-books you used?

You may have:

Searched for them on WWW, Read about them on WWW, Read about in print publication, Print advert, E-mail advert, Other

Use

A6 How would you prefer to read e-books (and why)?

For example:

On screen; Printed whole; Printed part; Searched; Skimmed; None of the above

A7 On what kind of machine would you prefer/find easiest to use them?

Prompts:

On PC; PDA; Mobile; Special reading device; Other

A8 Where do you prefer (or are you most likely) to read e-books?

Prompts:

At work; Home; Library; Other

A9 How effective were the following functions?

Prompts:

Search; Copy whole/part; Bookmark; Annotate/write notes; Other

A10 Did you encounter any problems/find something problematic when using the e-book?

For example:

Finding or choosing right format

The following questions relate to _free_ e-books.

B Free e-books

B1 What experience have you had of free e-books?

Prompt: Those of you who have had experience of free e-books, do these account for the majority of the titles you have used?

B2 Would it make a difference to you/your use whether the e-book is free or you have to pay for it?

B3 What do you see as the potential advantages of free e-books?

B4 What do you see as the potential disadvantages of free e-books?

B5 Do you see or are you aware of any content problems?

B6 Do you have a preferred format for free-e-books?

For example:

PDF, etc, Implications of accessibility

Is this different from e-books in general?

B7 This is a question for the teaching staff:

How do you think you might use free e-books in teaching and learning in the future?

A particular book as primary reading on a course; to integrate quotations from a book into teaching materials; in a VLE?

B8 This is a question for the library/information services staff:

What do you see are the main issues in the selection, acquisition and use of free e-books?

For example:

Locating them…

Appendix F: Checklist for the Evaluation of Free e-Books

In view of the importance given to quality assurance by potential users, we have considered it useful to develop some ad hoc procedures for evaluation of free e-books. Here is the draft checklist developed for our in-house quality evaluation:

Is it what it says it is?

1. Is the text really available and free to the user?
2. Check for existence of metadata;
3. If the metadata claims to conform to an external standard (e.g. Dublin Core) check its grammaticality, completeness and relevance;
4. Is the relevant information about the particular edition of the text present and accurate;
5. Are intellectual property issues covered in the metadata or text? Is the treatment accurate? Does the resource provider have the right to distribute the resource, and are the creators of the resource credited in the documentation or metadata or text (as appropriate)?
6. Check the accuracy and completeness of the metadata for individual texts, where the resource is a collection of texts or samples;
7. Where there is no metadata covering these issues, check in particular whether the following particulars are as expected: text, language, file type, text encoding format.
8. Where there is more than one file, check that all relevant resource files are present in the correct file structure (i.e. as documented), and that file naming conventions are suitable;
9. Assess the file format: is it as documented, is it valid according to the normal standards for that format and is this a suitable format for interchange, storage, use and preservation?

Fitness for purpose

10. Assess the appropriateness of the format for the intended purpose (e.g. quality of design, representativeness, sampling etc);
11. Duplication: is the text available elsewhere in a usable way, or even in a more useful form?
12. Is this text likely to be of use in HE and/or FE? If so, where and how?

Text Integrity

13. Check integrity of textual material (Are bits missing? Have some elements been silently omitted?);
14. Check for erroneously repeated textual material;
15. Are footnotes, endnotes, other editorial interventions encoded, and if so are they done correctly;
16. Are front matter such as foreword, preface, introduction etc. correctly encoded;
17. Are appendices such as afterword, endnotes, bibliography present and correctly encoded?

Text format and encoding

18. Assess the character sets which are used: is the character set as per the documentation, if this exists? Is it suitable? Are there any invalid characters or entities?
19. Assess the choice of textual markup scheme: is it suitable for interchange, use and migration?
20. Validate the textual markup and evaluate the semantic accuracy and appropriateness (e.g. are chapter or paragraph tags correctly used?);
21. Validate the design, markup and annotation against external criteria; check that it actually works with software for the processing of the format, e.g. check XML is valid and parses, check Acrobat Reader can read PDF files;

Factors external to the text

22. Search for and follow up documented bug reports, comments and reviews which may be available at the repository or elsewhere;

23. Contact, and maintain ongoing dialogue with, the resource provider (where this is possible) to ensure the accuracy and completeness of metadata, and to manage enhancement of the resource where necessary.