UNIVERSITY OF CAMBRIDGE

ARCADIA

Digitisation-on-Demand in Academic Research Libraries

Edmund Chamberlain, 2010

arcadia@cambridge:
rethinking the role of the research library in a digital age

The Arcadia Programme is a three year programme funded by a grant from the Arcadia Fund

Table of Contents

With thanks to the following for help and advice:

- John Naughton
- Michelle Heydon
- Rufus Pollock
- Ted Krawec
- John Norman, Laura James and their colleagues at CARET
- Erik Baber and Caroline Murray at Cambridge University Press
- Patricia Killiard, Sue Meherer, Emma Coonan, Grant Young, Vanessa Lacey and Don Manning at Cambridge University Library
- Anyone at Cambridge who assisted in the survey, especially Libby Tilley, Linda Washington and Jenni Lecky-Thompson
- Rick Anderson and colleagues at the Marriott Library University of Utah
- Marcus Gipps and Leon Dufficy at Blackwells Charing Cross

Introduction

There is a growing perception in the academic community that research and teaching material offered by a library is or should be wholly available online. Whilst attempting to meet this need, libraries are coping with the financial and physical constraints of maintaining large print collections, for which demand often remains high. This is especially true for large legal deposit libraries such as Cambridge.

The wide range of material available in such a library and its correspondingly varied demand is often referred to as the 'long tail'. In business, long tale operations only become successful when the tail is well exposed to its potential market. It is argued that library services orientated around the physical object only go so far in exposing the value of the long tail.

Digitising existing physical stock would allow for the growing online demand to be met whilst also potentially limiting the costly circulation and storage of print material. It would also allow for the long tail to be effectively exposed. Mass digitisation of collections still remains expensive enough to be largely unworkable for a single institution with no external or commercial backing. Copyright legislation also prevents libraries from engaging in wholesale digital digitisation of their physical stock.

Technological developments, specifically the Kirtas automated book scanner and the Espresso book printing/binding machine have made fast digitisation and print-on-demand a potential option for libraries. Along with similar technology, they represent an opportunity to provide a self-sustaining digitisation-on-demand service for an academic research library.

As on-line interactions evolve, library users are increasingly used to quick or immediate responses, from online shopping for furniture to digital delivery of media. Mechanisms to gauge and assess user demand form the basis of modern retail logistics. Libraries are already themselves adopting more immediate models of demand-driven service, especially in eBook acquisition.

The following report investigates digitisation-on-demand as a potential service model for an academic library, with a focus on Cambridge University Library.

Brief summary

The investigation finds that digitisation-on-demand and print on-demand services have the potential to provide greater value access to libraries' collections and could help a library to realise its true potential as a 'long tail'. There are at present a number of practical and financial limitations that prevent this from being fully realised.

Whilst the concept remains a viable one and demand is noted, copyright legislation restricts the material available for full digitisation to a niche subset of a library's' whole collection.

For digitisation-on-demand, start-up costs remain high, which itself endangers a higher level of risk if a self-funding service is not used. Lease hire models for equipment could help mitigate this.

For print on demand, start-up costs are also relatively high. Third party solutions could provide an alternative. In both cases, users may object to additional costs.

Definitions

Digitisation-on-demand in the context of this report refers to an emerging library service. A library user places a request for a digital copy of a work previously only available in print. The library digitises the work and a usable digital copy dispatched to the requestor in good time.

Print-on-demand refers to print delivery of a work that may or may not have already been digitised. The print copy will be printed exclusively for that request, no mass-production or warehousing of stock is involved. Print-on–demand is widely used in the publishing and bookselling industries and in some libraries.

Scope

The report focuses upon libraries supporting higher education teaching and learning and academic research. There is a specific focus upon Cambridge University Library. The University Library is a national legal-deposit library in the UK. It forms the major research library for the University.

Methodologies and activity

- A brief survey was created with the SurveyMonkey online survey tool. This was circulated amongst academic and library staff to gauge expectations of and demand for a digitisation-on-demand service

- Circulation data from several libraries in Cambridge was examined, broken down by publication date. Data was extracted from the Voyager Library Management System used by the majority of libraries in Cambridge.

- Key stakeholders within Cambridge University were interviewed regarding their opinions on digital service developments

- Cambridge University Press book digitisation infrastructure, the University of Utah's' and Blackwells books print-on-demand services were examined as case studies in equipment operation

- Potential costs for digitisation of works were examined. These were compared to available figures from existing library digitisation and print-on-demand services

- A prototype extension for the Cambridge University Aquabrowser catalogue (branded LibrarySearch) was developed, tying its bibliographic data into the copyright calculator service provided by the Open Knowledge Foundation. This was used to test the calculator and highlight issues in copyright calculation from library metadata

- Policy and technical documentation from existing digital libraries was examined to assess suitability for reuse of digitised material. The examination focused on legal and technical restrictions around access to full text

- Theoretical workflows and use-case scenarios for digitisation on demand were devised

Conclusions

Potential demand:

1. 91% of Cambridge academics surveyed would be interested in a full text digital copy of an out-of copyright work. 65% would also be interested in a print facsimile. 62% would be interested in a partial digital copy of an in-copyright work if available. Some requested that access to existing physical stock should be preserved and that academic users should not be asked to meet additional costs
    - *More information in sections 3.6, 4.6 and Appendix #6*

2. Copyright legislation and the perceived fragility of pre-1850 material realistically limits full text digitisation to material published between 1850 and 1920. Library circulation data indicates demand here is 'niche', accounting for less that 2% of total transactions
    - *More information in section 3.6 and Appendices #5 and #8*

3. Striking the appropriate balance between quality, cost and speed of any digitisation output would be a critical factor in establishing a successful service. Doing so may require an initial test period and further market research
    - *More information in section 3.3*

4. 'Long tail' operations are only successful when the tail is fully exposed to its potential user base. Adequate promotion of an on-demand service to current library users and the wider academic community would be necessary to ensure success
    - *More information in section 1.1*

Copyright:

5. The complexities around copyright law remain a barrier to library digitisation activity in all respects. If sensible precautions and take-down policies are in place, copyright decisions for digitisation could be decided on a risk-basis
    - *More information in section 2.3 and 2.1*

6. Automated tools such as copyright calculators are beginning to emerge. Their use in aiding decisions over copyright should be encouraged and they could be usefully integrated into library catalogues and other online interfaces
    - *More information in section 2.3 and Appendix #3*
    - *Watch a short film about copyright calculators at http://vimeo.com/15678944*

Costs:

7. Kirtas book scanners represent a cost-effective way to digitise full texts quickly at an acceptable quality level, although the initial capital investment remains high. If a rapid digitisation capacity is not already present in a library, a sustainable digitisation-on-demand service will require large levels of capital investment
    - *More information in section 3.3 – 3.6*
    - *Watch a film about the Kirtas book scanner: http://www.youtube.com/watch?v=l2cP14mEQKI*

8. Survey information reveals that academic users would prefer to pay under £15 for a digitised copy. Achieving this at cost or with a small surplus would be a challenge. Attempting to recoup capital investment directly would push costs beyond this 'sweet-spot' price point

- *More information in section 3.6 and Appendix #6*

9. Alternative models for digitisation including lease/hire of equipment. For Cambridge University Library, complete out-sourcing of digitisation for material published between 1850 and 1920 currently remains unrealistic given the risks involved in transporting material
    - *More information in section 3.5*

10. For print-on-demand services, the Espresso Book machine was examined. Operating and capital costs are seen as lower than those for digitisation. The option to provide quick affordable self-publishing services may help recoup initial investment. Alternatively, the use of third party services for print-on-demand present a lower risk means to trial and scope a service equipment may prove more cost effective if demand is low
    - *More information in section 4*
    - *Watch a short film about the Espresso Book Machine at http://www.youtube.com/watch?v=OIq0VqF0MnA*

Related developments in digital library services:

11. For print-on-demand, the use of existing digital copies from online repositories such as Google Books and the Hathi Trust is limited by technical and legal restrictions. The EspressNet Books-On-Demand service represents a potential way to sidestep these barriers
    - *More information in section 4.5 and Appendix #4*

12. Library metadata requires consistent and accurate use of numerical identifiers to inter-operate effectively with external digital libraries. OCLC Worldcat identifiers are the most commonly used
    - *More information in section 4.5 and Appendix #4*

13. Developments in hand-held devices, specifically eReaders, Tablet PCs and smartphones are driving change in online text provision. Reflowable text-based output (as opposed to image only) is vital for viewing on handheld devices and for future potential re-use
    - *More information in section 3.9*

14. As public domain print material is digitised, there is a growing perception that the digital copy should retain the same legal status. Use of Creative Commons licensing would be the best means to achieve this
    - *More information in section 2.6*

Recommendations

Copyright and copyright calculators

- Libraries should work with the Open Knowledge Foundation to assist in the development of the calculator and incorporate it into library catalogues

- Once an assumption on copyright status is made, risk could be mitigated by transparency over how the assumption was made. Risk could be further mitigated by developing a take-down policy and feedback mechanism to be used alongside copyright calculation and its use to present any digitised version of a work

- Material in the public domain that is already digitised should be released under an appropriate explicit Creative Commons license. In some cases, very high-resolution versions of images for use in publications could be retained and sold under a commercial license

Digitisation-on-demand

- Digitisation on demand when done in-house would work best as part of a greater digitisation program, for preservation or other purposes. It could be seen as one 'intake stream' in a wider shared digitisation service, potentially across several institutions

- If this is not available, given the high capital cost and uncertain levels of demand, use of third party suppliers or leased equipment for initial scoping is recommended. As cost for digitisation and print on demand decreases, this situation could be reviewed over time

- Greater involvement in shared digitisation and digital library projects such as the Hathi Trust could allow for reliable access to existing digital surrogates for a print on demand service without large investment in digitisation infrastructure

Print-on-demand

- Print-on-demand services could supplement or enhance a number of traditional library services. Further investigation of the Espresso Book Machine service is suggested, with a possible trial of print-on-demand services from bookscan-bureau

Section one – project background

1.1 Library operations and the long tail

The term long tail is frequently used to describe online business models:

> *"The **Long Tail** or **long tail** refers to the statistical property that a larger share of population rests within the tail of a probability distribution than observed under a 'normal' or Gaussian distribution. The term has gained popularity in recent times as a retailing concept describing the niche strategy of selling a large number of unique items in relatively small quantities."* [1]

Chris Anderson used the concept in Wired magazine [2] to describe the success of online retail businesses such as Netflix and Amazon.

At first glance, the long tail could be easily applied to any major research or legal deposit library, given its depth of collection and diversity of user base. Anderson however believes that in the long tail, demand for any item never reaches zero. This contradicts the fact that in modern research libraries, a lot of stock is rarely used (i.e. circulated). Colorado State University recently found 40% of its stock never circulates [3]. One of the reasons behind this may be restrictions around circulating a physical item; another may be limited exposure of the material to its potential reader.

Lorcan Dempsey has further challenged the view that libraries are in the long tail [4]. He believes that libraries need to examine some particular aspects of how a 'long tail' is successful in practice as a business. He believes that libraries do not have many of the components important to online retail success, notably aggregation of supply and demand.

True long tail operations make use of well-organised and accessible data, aggregating supply by effectively exposing the whole long tail. Recommendation services also play a key role in online business success, tying in popular items with less popular ones, those further down the tail. As Anderson notes when examining the failed MP3.com, the long tail by itself is cannot succeed, the obscure must be related to the popular to make the potential customer aware of its existence.

Dempsey acknowledges that click-stream and 'also borrowed' information in particular is not generally available to libraries. This has been vital for Amazon in exposing its long tail. There has been some recent effort within the UK to address this obvious gap in online library services [5].

Dempsey also talks about a consolidated web presence being vital in aggregating demand [6]; the long tail only works when little used resources are adequately exposed to their customer base. To work on this scale, a library must take steps to improve discovery within their interfaces and push discovery into avenues beyond the catalogue. Dempsey also argues cost and time of library transactions need to be brought down [7].

---

[1] "Long Tail - Wikipedia, the free encyclopedia."
[2] Anderson, Chris, "Wired 12.10: The Long Tail."
[3] "Morgan Library makeover moves out miles of books | coloradoan.com | The Coloradoan."
[4] Dempsey, "Libraries and the Long Tail."
[5] "JISC MOSAIC."
[6] Dempsey, "Libraries and the Long Tail," 3.
[7] Dempsey, "Libraries and the Long Tail." 5.

1.2 The digital context

In spite of these shortcomings in resource discovery, the mass of unique material in many libraries represents a tempting potential digital resource, especially when framed in the context of full-text online services. Full text searching of a digitised collection could provide a much more effective means of exposing the long tail. The large body of information in libraries compared to that available to individual publishers or booksellers has been noted [8].

Initially, digitisation efforts within libraries were focused upon unique or rare material with the aim of exposing them to a greater audience. Only a small proportion of academic libraries have been involved in mass digitisation. This process is costly and usually requires commercial support from companies such as Google or Microsoft, or collective approaches.

When such large bodies of digitised public domain works are tied into print-on-demand services, the long tail is again an apt description, arguably more so than to libraries with purely physical collections where complete access to the final work is much harder to achieve.

1.3 User driven selection – a new ethos for library digitisation

Scan or digitisation-on-demand services has recently emerged as an alternative or supplement to mass digitisation, with the aim of fulfilling reader requirements for material on a short turn over, per-item basis. The City Archive of Amsterdam has placed a 'request scan' button on the online interface to its archive database. This helps the archive decide priority for its ongoing digitisation efforts [9].

The National Library of Australia has placed a 'copies direct' button on its online catalogue record pages. This service is underpinned by an automatic mechanism for copyright assessment. If a work is deemed to be out of copyright, the choice of a full scan and physical copy is offered. If it is copyright, the service can still be used to order a copy of the subset of the work as allowed under the legal entitlement [10].

Harvard University Libraries launched a Scan and Deliver service in 2009. The service is free to staff, students and faculty in Harvard. Users can request partial copies of works as a PDF file via the Hollis catalogue [11].

The McGill Library in Montreal Canada is trailing a digitisation-on-demand service alongside its print on demand services [12]. Their website describes the ethos behind the service:

> *"At McGill, Digitize on Demand refers to the fact that the Library will, on request and for a cost, scan rare books and other items from our collections and create electronic versions of them for download. The items must be indisputably free of copyright, and their physical condition and format must make them suitable for scanning.*

[8] Reddy et al., "A web service for long tail book publishing."
[9] "Archives Database, How does it work."
[10] "Copies Direct - Help | National Library of Australia."
[11] "Harvard Libraries Launch Scan and Deliver Service - HCL News - Harvard College Library."
[12] Schmidt and O'Neill, "The 'DOD' and 'POD' project in context at McGill: Part of digitizing collections to preserve content, provide access and enrich research."

Print on Demand is a service where the Library creates print reproductions of books which have been digitized, on request for a cost.

McGill Library is digitizing its collections to enhance universal access to knowledge, to facilitate discovery and delivery of its rare and unique treasures, and to make them openly accessible to researchers, students and all interested individuals in the McGill community and around the world."

Once digitised, McGill make copies of digitised works freely available on their library catalogue. In short, they provide a user-driven, user-funded library digitisation project.

Other libraries that have undertaken mass digitisation processes have recently begun to make use of print-on-demand services. Some are working directly with the commercial vendors supplying publishers, specifically the British Library and Library of Congress with the Amazon owned Createspace [13]. Due to copyright restrictions, they are invariably sourcing print-on-demand texts from pre-scanned collections of material already in the public domain.

1.4 User driven selection models in other areas of library activity

Libraries have always used reader recommendations to inform acquisitions processes and collection development strategies. As licensed online content overtakes traditional print intake, new approaches to handling recommendation are being adopted, involving direct user interaction and monitoring of user activity.

One specific application has involved eBook acquisition. Sales models are now offered that allow users to directly influence selection. eBook vendors pre-loading records for their entire eBook collections into a library catalogue. When a text receives a certain number of 'click-throughs', this is converted into a sale, allowing the library and its users perpetual access to a text.

Debate continues regarding the strength, balance and depth of collections that are developed in this fashion. A recent in-depth study from Claremont University Consortium comparing different models suggests that collections with user-centric development models have more unique users per item and less items with no use [14].

1.5 Advantages of on-demand digitisation services

- Costs could be met by the end user at the point of request, so a large private or corporate donor to kick-start mass digitisation would not be required

- Existing bodies of digitised material in the public domain could utilised, so as to avoid unnecessary duplication of work

- Existing workflows and practices could be tapped and potentially repurposed
- The library could contribute any newly scanned material to existing open access collections, thus increasing the size of the digitised public domain

[13] "The Library of Congress Revives Public Domain Works via CreateSpace Print on-Demand and Amazon Europe Print on-Demand - MarketWatch."
[14] "Ebook Library Blog » "Beguiled by Bananas" – A Statistical Analysis of Patron-selected vs.Upfront Acquistion presented at Charleston Conference 2009."

Digitisation-on-demand can perhaps also be tied to Ragnathans' five laws of library science [15]. These include statements that books are for use, every reader should be able to get access to his or her book and that readers' time should be saved whenever possible.

Placing choice of digitisation directly in the hands of the reader rather than a librarian, donor or publisher would arguably be an act in the spirit of these laws.

1.6 Barriers to success

The following major problem areas for digitisation-on-demand have been identified. These four themes will be explored throughout the remainder of the report, Whilst they will be discussed in the context of Cambridge University Library, the arguments and issues raised will be applicable to any research library considering digitisation and print-on-demand.

Copyright

Assessment of the copyright status of a work remains a major barrier to provide a quick, automated digitise and print-on-demand service. There are also many extra complications around unpublished works and continuing works. As such, the scope of this report will be largely limited to published monographs. Mechanisms are currently being developed to automatically assess the copyright status of a work. Assessment and prototyping of these mechanisms and their basis to drive a user forms a major component of this report.

Cost

Cost of digitisation and any potential print delivery raises several questions. How should scanning be funded, by donor or by the requester? Should profit be made? Raise the costs too high and choice of work becomes prohibitive. Setting them too low could affect the quality of digital and print outputs. This report makes an initial attempt to investigate costs and pricing options for both digitisation-on-demand and print-on-demand.

Quality

In order to keep costs low, both scanning and printing processes would need to be as automated as possible. In particular, post production work on scanned images to make them as fit as possible for reprint would be required. This could affect quality. Balancing the quality and cost of any product is a key factor in success.

Potential market

Copyright realistically restricts the material available to mid to nineteenth and early 20th century material. These collections are not blockbuster fiction or primary academic textbooks. They fall strictly within the long tail and potential demand is hard to gauge. As has been noted, the long tail only becomes effective when it is fully exposed through aggregated supply and demand. The potential market for digitisation-on-demand and print-on-demand is assessed by a survey of academics in Cambridge and analysis of circulation statistics across several Cambridge libraries.

[15] "Five laws of library science - Wikipedia, the free encyclopedia."

Section two – copyright, the public domain and copyright calculators

Introduction

> *"Information professionals have a duty to foster the fullest possible access to information for their users whilst also respecting the intellectual property protections afforded to people for their creativity and innovation"* [16]

Assessment of copyright is an important part of any digitisation workflow. This section highlights issues and potential problem areas in relation to digitisation-on-demand services.

Due to the complexities of copyright law the focus is upon copyright under UK jurisdiction regarding published texts, specifically monographs.

2.1 Copyright in the UK

Copyright forms part of a wider set of intellectual property legislation.

Duration of copyright for literary, musical, dramatic and artistic works is largely governed by the 1988 Copyright Designs and Patents Act, as amended by *The Duration of Copyright* and Related *Rights in Performance Regulations* 1995 (SI 1995 No. 3297) [17]. Further background information on UK Copyright legislation can be found in Appendix #2.

The following points are of relevance to the concept of digitisation-on-demand:

- Copyright for literary, musical, dramatic and artistic works generally exists for the life of the author or creator for 70 years (expiring on midnight 31st December for that year), published or not
- If the author is unknown, the copyright is 70 years from the date of creation, or if made publicly available, from the date of availability
- Copyright status of a work published outside of the UK is dependent upon a large number of reciprocal treaties between the UK
- Variations should be considered when assessing the copyright status of a printed work. In all cases, knowledge of the place and date of publication and death of author is required to give an accurate indication of copyright status
- Further complications apply to unpublished works
- In assessing copyright of a work not published in the United Kingdom, county of publication, its copyright relationship with the UK and the duration of copyright in that country should all be known
- Within a printed published work, the preface, an illustration or even a typeface may hold separate copyright status. This issue should be viewed from the point of view of potential risk rather than complete accuracy. Providing only the original text of the

[16] Pedley, *Essential law for information professionals / Paul Pedley.*, 19.
[17] Padfield, *Copyright for archivists and users of archives / Tim Padfield.*, 22.

work in OCR derived plain text form with no illustrations, in a public domain font would be one alternative

- Two so-called exemptions exist under UK legislation, one for 'fair-usage', the other for 'preservation'. Neither exemption would allow for a library to digitise and distribute full copies of any work within copyright

- Digitisation-on-demand workflows could potentially be used to deliver partial copies of work under the fair dealing exemption, in lieu of the full work. The science Libraries in the University of Cambridge are already fulfilling this function through article provision

- Any work where copyright is expired or not affected by the larger set of intellectual property laws can be judged to be within the public domain.

2.2 Cambridge University Library and the public domain

As part of a wider study of the size of the public domain, the economist Rufus Pollock has recently conducted an appraisal of the pre-1960 collections of Cambridge University Library. He has blogged extensively on the preliminary findings. Pollock notes that:

> '*Computing PD status is non-trivial largely because a) it is hard to match a given item to a work or person b) we lack data such as authorial death dates and dates of first publication that are required … As such we need to adopt approximate and probabilistic methods* [18].

Two separate initial approaches were taken, one a conservative estimate based upon publication date [19] and the other based on author death date information sourced from library bibliographic records [20]. In order to gain the best estimate, results from the two methodologies were combined [21]. The final table is reproduced below:

Pub. Date	Items	% PD	No. PD
1400-1850	304,587	100	304,587
1850-1860	40,970	100	40,970
1860-1870	43,734	100	43,734
1870-1880	50,564	95	48,035
1880-1890	66,857	90	60,171
1890-1900	66,883	85	56,850
1900-1910	70,360	65	45,734
1910-1920	60,489	40	24,195
1920-1930	78,670	25	19,667
1930-1940	90,576	10	9,057
1940-1950	72,692	6	4,361
1950-1960	118,251	0	0
1960-1970	262,974	0	0
1970-2009	2130,509	0	0
Total	3458,116	19	657,361

Figure 1. Estimations of University of Cambridge holdings within the public domain. R.Pollock 2009

[18] "miscellaneous factZ – The online home of Rufus Pollock » Blog Archive » The Size of the Public Domain."
[19] Ibid.
[20] "miscellaneous factZ – The online home of Rufus Pollock » Blog Archive » Size of the Public Domain II."
[21] "miscellaneous factZ – The online home of Rufus Pollock » Blog Archive » Size of the Public Domain III."

Despite the simplistic approach to some of the legal complexities surrounding copyright, estimation such as this are undeniably useful in making a case for greater digitisation of the public domain. From an institutional point of view, a sizable proportion of CUL holdings, between 15-20% of works held can be estimated to fall into the public domain.

Much of the material from within the public domain originates from the late Nineteenth century. Nineteenth century ephemera are now important source material for twenty-first century research, as well as being of potential interest to a wider audience. Much of this material has been catalogued through the University Library Tower project [22]. Catalogue data from the Tower project has itself recently been published for free use and reuse under Public Domain Dedication and License (PDDL) [23].

2.3 Copyright calculators

Copyright calculation software is being developed to automatically assess the copyright status of a work. The software takes information on a specific work, potentially sourced from a bibliographic record such as a MARC-encoded catalogue record and attempts to resolve the copyright status of a work by passing information through a series of algorithms that describe copyright legislation.

In order to provide the greatest flexibility, such calculators should ideally operate using application programme interfaces or APIs, allowing them to be queried automatically or en-mass by other software.

Within the UK, the Open Knowledge Foundation has driven developments in copyright calculators. They have produced a UK specific API-orientated calculator for the Public Domain Works project and are working as part of the Europeana, the European Digital Library to produce a series of web based calculators that operate on an international basis [24].

Copyright calculators could be used in a number of ways in a digitisation-on-demand workflow. They could indicate the status of a work in library catalogue and advise a reader on the potential copyright status of a work, or be used by library staff as a guide to advise on a final decision regarding copyright of a single work or collection-en-mass.

Copyright calculators are not necessarily intended as an outright replacement for a full assessment of a work, as Europeana project member Christina Angelopoulos notes:

> *"It should of course also be noted that there is a limit to the extent to which an electronic tool can replace a case-by-case assessment of the public domain status of a copyrighted work or other protected subject matter in complicated legal situations. The Tools are accordingly accompanied by a disclaimer indicating that they cannot offer an absolute guarantee of legal certainty.* [25]*"*

Any digitisation workflow, even one working on a 'per-case' basis will need to move quickly and effectively. The risk of stalling a digitisation workflow or project through indecision over the copyright status of a work may itself outweigh the risk of incorrect determination of status.

---

[22] "Tower Project - Introduction."
[23] "JISC OpenBibliography: CUL data release | Open Biblio (graphic) Projects."
[24] "PublicDomainCalculators/About - Open Knowledge Foundation Wiki."
[25] "Open Knowledge Foundation Blog » Blog Archive » Public Domain Calculators at Europeana."

Effectiveness of copyright calculators is largely dependent on the bibliographic information available to support them.

As noted by Pollock, the key factor in assessing copyright status of a work lies in readily available information on author death. However, this cannot be assumed. Libraries record birth and death dates of individuals as a means to distinguish between different individuals with the same name. His final paper notes that for the 1890-1900 period, only 43% of all records in CUL have an author death date record. For the next decade, it falls to 42% and continues to decline [26].

Beyond a library catalogue, there are a number of potential sources of individual death date information. The largest source in the UK is the index of Births, Deaths and Marriages derived from historical census information.

Two potential solutions are available here. Libraries can continue to publish bibliographic and name authority data in a linked fashion, so it could eventually be semantically associated with other sources of individual death information. They can also improve the quantity and quality of author death dates within existing records.

Other risks exist around the issue of underlying material in a work. Prefaces, illustrations and editorials are all treated as separate works under copyright legislation. A thorough assessment of a work's status would need to be aware of the separate underlying entitles and their creators' date of birth. Edited versions of a work, translation, annotations and abridgements could also have separate copyright status. Such material often accompanies a reprint of an out-of copyright work. Library catalogue data does contain some information on additional material, but this may not be written in a reliable codified form, or with accurate creator information. These issues are further highlighted in appendices two and three.

2.4 Managed risks and copyright calculation

Leaving aside availability of useful bibliographic data, the issue of copyright is still complex. Even when restricted to printed monographs, the exceptions and variations in copyright ownership outlined above clearly produce significant barriers to quickly identifying the copyright status of a work.

A workflow based around copyright calculators would need to take potential variations in term into account and provide fallback mechanisms for situations where data is not available for accurate assessment.

The following table simply presents major risk areas and some potential practices to help mitigate risk:

Issue	Risk	Mitigation
Bibliographic data is not complete, specifically death date not present in underlying name authority data	• Calculator is unable to guess copyright status correctly • Library is sued for breach of copyright	• Calculator could assume 100 years from birth date if available, or from publication date, in addition to standard term • If birth date is not available, a

[26] Pollock and Stepan, "The Size of the EU Public Domain," 14.

		reasonable safe cut off point publication (150 years) could be assumed
Calculator performs incorrect assessment due to configuration error	• Calculator incorrectly assesses status of a work • Library publishes incorrect status and digitises in error • Library is sued for breach of copyright	• Full transparency in interface over calculator operation in interface (and source code) • Clear mechanism for copyright owners to make a claim over incorrect status • Implementation of immediate takedown / blacklist mechanism
International copyright information not available	• Calculator incorrectly assesses copyright duration • Library is sued for breach of copyright	• Calculator only displays results for jurisdictions it can cover
Element of work covered by separate copyright (chapter, preface, typeface, image etc.)	• Calculator incorrectly assesses status of a work • Library publishes incorrect status and digitises in error • Library is sued for breach of copyright	• Information on subsections and illustrations is not generally recorded in a bibliographic record • Train those involved in digitisation workflow to spot potential problems

In all cases, the risk is largely the same, that of incorrect assessment via a calculator and subsequent breach of copyright. This may or may not result in legal action against the library by the copyright holder.

The implications of this risk could be assessed be examining the potential severity of any legal action and the chance of actual occurrence. In the UK, this could potentially be severe. Breach of copyright is a civil offence, so imprisonment is unlikely. Damages awarded in copyright cases are not fixed to any particular measure, but usually measured against any perceived commercial value of the infringement [27].

2.5 Current approaches to copyright in library digitisation

Large corpuses of digitised public domain works are now in existence, generated in the main through ongoing library-based digitisation.

What is considered out of copyright in the US may be in copyright within the UK. International sensitivity is clearly a major issue. The safest approach would be to reassess

[27] "Copyright Infringement | Damages claims | UK intellectual property lawyers law firm."

and potentially rescan work from a UK legislative point of view. Publishing a take-down policy and contact details may help to mitigate risk if work in-copyright is accidentally digitised and made freely available.

When considering risk, it is helpful to briefly examine the approaches taken towards copyright in previous and existing digitisation projects:

Open Content Alliance

The Open Content Alliance, a subsidiary of the Internet Archive has made over two and a half million texts freely available online through a number of different avenues and funding streams. It includes the Project Gutenburg collection of public domain works. The British Library, through sponsorship from Microsoft has contributed large amounts to this collection. The Biodiversity Heritage Library project has also contributed material[28].

An early mass digitisation effort led by Carnige Mellon University, the million books project claimed full compliancy with U.S. Copyright law. Digitisation was limited to material published before 1920. The project later recommended a proactive approach to copyright clearance for works within copyright:

> *The million book project will make a good faith effort to clear copyright on appropriate materials by sending the publisher of record a letter asking for permission. Replies will be recorded in the administrative metadata. If the publisher has returned the rights to the author, the author will be contacted. Subsequent copyright holders will be contacted as needed. If the permission letter receives no response, then materials will be digitized as a part of the project. If rights holders subsequently identify themselves and request that the material be removed from the project, that request will be complied with immediately* [29].

Output from the project was eventually deposited in the Internet Archive.

Biodiversity Heritage Library

The Biodiversity Heritage Library is an international project aiming to digitise material relating to the life sciences. Focusing initially on English language material, the project is expanding to cover collections from a wider range of European countries.

The collection is comprised largely of public domain material, but some material within copyright has been sourced under license from copyright holders. This has in turn been made available under a Creative Commons Attribution Non- Commercial 2.5 license. The BHL project operates in an international context, but largely bases its assumptions on choice of digitisation on material scanned around US copyright law.

The project acknowledges that material within its collections may still be in copyright in some territories outside the US but does not attempt to advise on specifics. The consortium makes no copyright claim itself over material digitised by the project [30].

[28] "Open Content Alliance (OCA) » Contributors."
[29] Reddy and St Clair, "The Million Book Digital Library Project."
[30] "Biodiversity Heritage Library - Licensing and Copyright."

Project Gutenberg

Project Gutenberg began in 1971 as an attempt to create simple text based digital versions of major works. It became the first collection of eBooks and is still widely circulated today, making over 33,000 eBooks available in a variety of formats. Whilst not a digitisation project in the normal sense, similar copyright issues apply. All works chosen by the project are held in the public domain according to U.S. legislation [31].

Hathi Trust libraries

The Hathi trust is a growing consortium of major US academic libraries who have pooled digitisation infrastructure and resources. They collectively hold over six million total digitised volumes of which over one million (22% of total) are in the public domain [32]. Much of the material was scanned under the Google books project.

The Hathi Trust holds both public domain and copyright material. Access to all material is governed by a rights database added to the trusts' repository. The trust will provide open access to public domain material but will not provide access to material in copyright unless specifically required as an alternative to print (i.e. for accessibility or preservation reasons).

A clear take-down policy and contact address is provided on the trusts' website [33].

Google books

The best-known collection of digitised full texts lies in Google books. Created with partner libraries and publishers, Google books is believed to hold over seven million digitised full texts, of which at least one million are in the public domain[34].

Since its conception, the Google books project has come under heavy criticism for its approach to intellectual property law. Most notoriously, this includes wholesale digitisation of works that are in copyright under a fair-usage for preservation clause [35]. Google has faced several lawsuits from publishers with regards to this practice.

Only some libraries in the Google books project are having in-copyright material, scanned, others are restricting digitisation activity to material within the public domain.

Works in the U.S. public domain (pre. 1923) are viewable from within the U.S. only. Google has taken a cautious approach to international copyright and only expose the full text of a work in a country where it has confirmed the copyright status of a work in that territory. This internationally sensitive approach is markedly different to that of the Bio-Diversity Heritage Library.

The difference in US and UK term durations presents an issue in potentially re-purposing scanned material from US based digitisation efforts. Assessment based on UK legislation is still required – hence Google banning US public domain works from being fully viewed in any territory.

---

[31] "No Sweat of the Brow Copyright - Gutenberg."
[32] "Welcome to the Shared Digital Future | www.hathitrust.org."
[33] "Rights Management | www.hathitrust.org."
[34] "In Google Book Settlement, Business Trumps Ideals - PCWorld Business Center."
[35] Grogg and Ashmore, "Google book search libraries and their digital copies," 133.

2.6 Copyright after digitisation - ownership and licensing of digitised public domain material

In some cases, a digitised surrogate of a work may be seen as an original work itself, which could thus gain its own copyright. This depends on how much perceived value is added through the act of digitisation. A large digitised collection would also collectively attract database rights [36].

Depending on the viewpoint of the digitising institution, this might be useful in protecting any revenue from charged access or print distribution it may want to collect. Doing so however would also place additional burdens and restrictions upon the re-use of the digital surrogate, thus limiting its true value to the reader.

A digital surrogate of a work has far greater potential flexibility and usage in an educational and research context than a print original. However, restrictions of copyright could limit the opportunities to fulfil this. Instead of asserting copyright over a digitised surrogate or even entire collection, a Creative Commons usage license [37] could instead be attached to public domain material produced by a digitisation project.

Various types of license exist, some specifically allowing (and encouraging) commercial re-use and re-purpose. Depending on the nature of the license, this could also include partial or complete copying or editing of a work for any purpose with no requirement to attribute the original work.

Making digital surrogates freely available under some form of Creative Commons license would arguably be the best means to ensure that greatest value is obtained from digital surrogates. This would not necessarily negate collection of profit. A print-on-demand version of the work could still sold at a profit. If in-copyright material was re-published incorrectly under a specific Creative Commons license, it may increase the risk of the digitiser being sued for breach of copyright.

2.7 Copyright as a barrier to academic and educational work

It is clear that the combination of complex copyright legislation and long terms of duration in the UK create major obstacles for library digitisation projects.

The effect of copyright restrictions on the wider education and research community is also worth considering, and is the subject of some debate. The British Library has recently published a collection of small essays from a range of academic practitioners. They speak of individual experiences and call for a rethink of copyright legislation in light of technological advances. In the prefix, Lynn Brindley, CEO of the British Library writes:

> *"There is a supreme irony that just as technology is allowing greater access to books and other creative works than ever before for education and research, new restrictions threaten to lock away digital content in a way we would never countenance for printed material"* [38].

The essays present a number of differing views about how to approach copyright change. Different issues are highlighted, but a clear consensus rests around the lack of fair dealing

[36] Oppenheim, "Legal issues for information professionals VI: copyright issues in digitisation and the hybrid library.."
[37] Schloman, "Creative Commons."
[38] Bentley et. al., "Driving UK Research – Is copyright a help or a hindrance?."

for music and sound recordings, issues of text-mining and the potential dangers of any further copyright term extension.

Vince Smith, a taxonomist at the Natural History Museum takes the argument a stage further and notes the restrictions that modern copyright legislation place upon modern scientific practice:

> *"We urgently need to separate cases where there is substantial loss of income to a content creator through content dissemination (e.g. a professional musician) from those that make no income from dissemination and rely on this as part of their scholarly activities (e.g. a professional scientist)."*

He also raises the challenges copyright raises to any potential scaling of digital library operations in the UK:

> *"Making copies for strictly archival purposes should not be subject to copyright control. Libraries in particular should be able to preserve digital copies in perpetuity, which technologically means regularly making copies"*

Copyright and copyright calculator: conclusions

- The complexities in determining copyright duration in the UK represent a barrier to any digitisation project**.** Copyright calculators are a valuable tool in helping to overcome this barrier
- Information on copyright status of a work has value outside of any digitisation service
- Due to differences in international copyright, many 'public domain' digital surrogates available on the Internet will still be in copyright under UK legislation
- For a library digitisation service, any decision on copyright should be taken on a risk-basis rather than attempting to seek outright and total assurance over the status of all elements of a work

Section three: Digitisation-on-demand

After a decade of digital library development, scanning a book generally presents little in the way of technical challenges. Commercial products and services to provide 'book-to-PDF' services are readily available and being actively marketed towards libraries.

This section of the report will briefly look at readily available technologies for quick, non-destructive scanning of material and examine key issues of conservation and post-production. The focus throughout is on 'good enough' delivery rather than archival level quality. The main issue raised is that of balancing quality and cost to provide a financially viable and useful on-demand service. Potential business models for digitisation and re-use of existing digital copies are also discussed.

3.1 Mass digitisation apparatus

Kirtas automated book scanner – Cambridge University Press
Cambridge University Press (CUP) use Kirtas automated book scanners to digitise selected University Library material for the Cambridge Libraries collection of important academic texts.

The device utilises two cameras to simultaneously image two open pages of a book. A vacuum equipped robot arm and a series of clamps automate page turning. CUP and other operators have found this to be less than reliable in practice, especially when using scanning older or fragile material. As a result staff still man the 'automated' terminals, monitoring the arm and correcting its behaviour as required.

Nonetheless, CUP have been able to achieve a relatively fast throughput on their digitisation service. For octavo-sized volumes, resolutions of up to 600 dpi are available.

Kirtas machines are available for hire/ purchase within the UK. The British Library has used them in the 25 million-page project. Several US libraries are making use of them for various digitisation projects.

Internet Archive Scribe scanner
The Internet Archive operates and leases scribe scanning devices to libraries in clusters referred to as pods.

The machines consist of two high quality commercial digital cameras and an adjustable cradle. Page turning is done manually. The rig is adjusted for each work digitised. A simple cloth back and photography studio and lighting is also provided. Internet archive typically trains and provide their own staff to operate a scanner, leased to partner institutions in digitisation projects.

Two attached network PC's deal with image capture and batch processing, synching finished works to Internet Archive servers on a nightly basis. Interestingly, much of the software used to manage the scribe has been open-sourced, allowing for the potential self-creation of a scribe-like machine.

Machines are leased to library digitisation partners either individually or in pods. Partners are charged a fixed fee per page scanned. This includes operating costs, consumerables, hire of equipment and staff time [39].

Google book scanner
To date, Google have not publicly revealed details of their book scanning process. They have patented an infra-red based digitisation method that compensates for the curvature of a page [40]. Googles' throughput is not known but suspected to be very high.

D.I.Y. Book scanner
A collective effort to design and build book-scanners has been underway on the Internet for some time [41]. Aiming to prove that digitisation need not require expensive equipment, plans and software to run dual camera scanners have been developed and shared by an international community of enthusiasts. Approaches vary, with changes and developments being shared and tested. Most designs focus around a wooden cradle with glass plates to hold pages down. Dual cameras are used to image pages. The process requires manual effort, but throughput of up to 1,200 pages an hour has reportedly been achieved. Some works digitised through DIY scanners have been submitted to Google books.

3.2 Post-production

Digitised images often retain speckles and distortions from digitisation. The curvature of the page will also need to be compensated for. The copy of a work being digitised may well have marks and blemishes that obscure the original work. In order to prepare digitised images for print, pages will need to be cropped and adjusted.

Post-production at Cambridge University Press
Cambridge University Press uses both automated and manual approaches in post - production.

The Kirtas workstations are used to scan and create RAW output. Batch image processing software supplied by Kirtas is used to de-speckle scans, crop and remove unwanted additions to an image and adjust to compensate for curvature on the page. Depending on server activity and the size of the work, this can take around three to four hours.

The scanner workstations include raided storage to hold RAW and TIFF output. A final PDF master is passed onto the Press's main server. This is the only archival output.

In order to achieve commercial quality output, PDF versions are then subject to lengthily additional manual editing using image editing software. Much of this work is outsourced to India, with a final editorial-check taking place in the UK. Marks and obstructions over the text are removed manually to bring the digital copy as close as possible to that of the original work as it went to press.

CUP staff source cover images for a work and produce additional editorial information, focusing on the importance of the work to modern scholarship.

[39] "The Internet Archive Keeps Book-Scanning Free."
[40] Lefevere and Saric, "United States Patent."
[41] "DIY Book Scanning | A forum dedicated to book scanning, open source, DIY digitization.."

The involved post-production and QA work leads to a longer much turnaround from scanning to print. Whilst vital for commercial level print products, this additional cost and time is clearly not suitable for an on-demand service.

3.3 Quality, cost and funding models

This case study raises the issue of quality versus speed and cost. The following rules can be assumed of any digitisation-on-demand service:

- The higher the quality, the more post-production will be required, the higher the cost of digitisation
- The higher the quality, the more post-production will be required, the greater the duration of digitisation
- The greater the cost of digitisation, the less likely a customer is to make an on-demand request
- The greater the duration of digitisation, the less likely a customer is to make an on-demand request

Given that the 'customer' would pay up-front for the cost of digitisation and delivery, the right balance must be struck between speed, quality and cost. The following factors are suggested as useful point aids in decision-making:

- *Quality:* If the role of a library is to get the information to a user as quickly and usefully as possible, useable content that is fit for purpose rather than commercial quality output should be sufficient, provided it meets expectations
- *Quality:* Managing the expectations of a potential customer is vital in ensuring satisfaction and repeat custom
- *Duration:* Work of a usable quality should be delivered as quickly as feasible. Users should be informed before committing to purchase of timescales
- *Costs*: Digital copies and print surrogates must be affordable to the average academic user. Given the 'on-demand' nature this should ideally be somewhere between the cost of an academic-text book and mass-market paperback. Costs could be greatly reduced for print surrogates by sourcing an existing PDF from a digitisation project
- *Costs*: If up-front costs remain to high to be tempting or viable to a reader, an alternative business model whereby not all costs of digitisation are incurred directly could be considered

3.4 Potential business models

1. User pays up front (for digitisation and optionally print-on-demand and delivery) and in doing so bequests a full text digital copy of a work to the public domain. In order to simplify costs, banding for the size of a work could be used.

2. Digitisation sourced from elsewhere – some model as above but cut costs for digitisation

3. Digitisation funded or part-funded by donation or institution, supplemented by POD costs or some other balance

4. Users express a preference. Items with the most preference are scanned with funds from a central pot. This could be done via a Facebook-style like button or by analysis of circulation statistics

A combination of these on-demand models, possibly alongside more traditional, donor-lead funding streams may be needed to provide sufficient return on any capital investment or to keep a service running in a sustainable fashion.

In order to further inform decision making, further information on costs and demand is required. The next two sections examine this.

3.5 Costs and throughput

A cost modeller for digitisation-on-demand using both a Kirtas book scanner and an Internet Archive Scribe service was developed. Based on the experiences of Cambridge University Press, it assumed a constant level of staff supervision during digitisation and drew information on wages and overheads from the University of Cambridge. The following was noted:

- An initial analysis of the operating cost of a Kirtas scanner suggests a 400-page work could be digitised at a cost of around £20. Cost could scale by size of a work, so smaller works could be digitised more cheaply.
- This does not include maintenance fees or any attempt to recoup capital costs for the scanner itself. Attempting to recoup this costs directly through would push prices above £40, depending on the volume of material passed through
- Neither does this does not include VAT or any surplus. In order to maintain flexibility of service, some surplus should ideally be retained to fund future development as needs and demands in digital delivery change
- In order to recoup staff and operational costs, an annual throughput of well over 1,500 would be required. With one operator on a single rota basis, limited to core office hours, this would be difficult to achieve.

3.6 Assessing Demand

In order to gauge demand for a full text digitisation-on-demand service in Cambridge two approaches were taken.

3.6.1 Survey

Two short surveys were circulated to academic and library staff. 61 academic and 16 library staff responded to their respective surveys. Full results are presented in Appendix #6.

Summary of academic responses:

- 91% of academic respondents would be interested in a full text digital copy in place of, or in addition to, requesting a book from the University Library stacks, assuming it was available under copyright restrictions

- 62% of academic respondents would also be interested in a partial digital copy, (i.e. one chapter) if a full copy was not available

- 66% of academic respondents would be prepared to pay between £10 and £15 for a full digital copy, with 35% willing to pay between £15 and £25

- 44% of academic respondents would be willing to wait a week or longer for a digital copy. 10% would be willing to wait for only 24 hours, indicating a time delay would not be so much of an issue

Summary of library responses:

- 93% of respondents would be interested in a digitize-on-demand service from the University Library. Reasons given included replacement of lost or damaged items, to add chapters to a VLE, or as an alternative to inter-library loan or a second hand purchase

- Librarians would generally be prepared to pay a bit more, with 21% only will to pay £5-£10, w8% willing to pay between £10-£15 and 28% willing to pay up to £25

- 42 % would be willing to wait over a week for delivery

Several academics provided additional responses, most voicing concern about a new service perceived as a replacement for access to physical materials. Several questioned the value of digitisation as a replacement for full text. One academic noted:

> *"I think it important to add that I think it is important for the soft copy to be additional to, and ***not*** in place of, hard copies. Nearly no-one likes reading lots of text on a screen, so for the university to hold only soft-copies of material simply off-loads the cost of printing onto individual academics, and we all end up with reams of A4 printed matter we don't need."*

One academic who did not respond to the first question was concerned that digital copies would replace access to the actual physical text, stating:

> *"I can't do this because we aren't given the option of either 'in place of' or 'in addition to'. I would be interested in various possible options for the latter but absolutely not in anything relating to the former: if I want something from the UL, I expect to be able to get it, not to have to wait to purchase a digital copy."*

Another respondent was unwilling to pay for additional services, commenting:

> *"I don't think the university should be charging academics for access to research materials. Seriously, the two questions about funds struck me as ridiculous".*

Quality was also a concern. One noted:

> *"Some of the books I have bought through Amazon have been so poorly scanned as to be useless, but it then becomes difficult to say who exactly is at fault, especially if the original book scanned is in poor condition, difficult to know what your rights are as a customer, and what the policy on returns might be. There are*

likely to be some difficulties with a percentage of scans and this ought to be considered an important problem for developing a reliable and useful service,"

3.6.2 Usage data

Circulation figures were also examined. The statistical analysis focused on the University Library and three other Cambridge libraries in the humanities. This was a deliberate choice given the perceived greater interest in late 19th century and early 20th century material from this period. The circulation data is presented in full in Appendix #5.

- Circulation transactions (loans) from non-periodical material published between 1850 and 1920 were examined
- The sample period of 48 months 1st 2008 to 1st November 2010
- All transactions rather than transaction for specific items were examined
- In addition, fetching requests for pre-1850, non circulated material from the University Library reading room are also presented

The following trends were noted:

- In all libraries, circulation transactions form 1850-1920 material accounted for less than 2% of all circulation transactions in the sample period

- In the case of the University Library, this was still a relatively large volume, accounting for 6918 transactions over the period, (an average of 150 a month)

- By contrast, the University Library Rare Books Reading room provides anywhere between 2,000 and 4,000 fetches a month for pre 1850 material. This indicates a somewhat greater interest in rare material in Cambridge than in published material, although the comparison is crude at best

3.7 Conservation issues

Both digitisation options above require considerable handling of a work. Whilst neither can be seen as destructive, there is a greater risk of damaging material. In particular, the automated page turning mechanism of the Kirtas machine poses a potential problem for conservation.

Staff at the Rare Books Reading Room in Cambridge University Library often provides conservation assessments for pre 1850-material undergoing partial digitisation. This can take upwards of 20 minutes and requires a trained observer.

For some post 1850 material such levels of input from trained conservators may not be required. However, level of preservation assessment should be considered in any digitisation workflow. This could potentially be a simple assessment of a volume by fetcher or digitisation staff.

In all cases, in-house digitisation is preferred by special collections staff within Cambridge University Library due to the potential risks around transporting material and having it digitised in an unsupervised way. This greatly limits the use of third party digitisation services.

3.8 Digitisation management and storage infrastructure

Cambridge University Press only hold PDF 'archival masters' of fully digitised works. This ties into their existing print and digital workflows.

A library may well want to store archival quality JPG or even TIFF output from a digitisation unit. This presents considerable storage problems in the context of any digital library development. Whilst full investigation of this issue is beyond the scope of this report, the following should be noted:

- Growth from an 'on-demand' service may be difficult to predict –this would affect any storage specification
- Assuming a digital surrogate is sourced from an external digital library, should it also be archived internally? This would also affect growth
- Any rights or ownership metadata attached to a digital object for the work may want to reflect the originator of the scan

3.9 Digital delivery

The commercial eBook market is finally beginning to mature. A recent survey on publishing in the digital era notes that:

> '*Dedicated eReaders and multipurpose tablets are finally becoming commonplace. A prerequisite to the digital publishing era, adoption rates are projected to reach 15 per cent to 20 per cent of the population in developed countries*. [42]'

The survey goes on to suggest that developments in both dedicated eBook reader and multi-purpose tablet computers are fuelling eBooks sales growth. It lists the unwillingness to 'abandon the paper experience', cost of device and the tiring effect of reading on a screen as the three most common barriers to takeup.

The authors of the report indicate that Moores' law [43] will take care of the final two, but the first will imply that the two mediums will co-exist for some time to come [44].

Within the UK, the BookSeller has conducted annual surveys of publishers and booksellers attitudes and opinions towards eBook vendors. Publishers and booksellers had differing ideas regarding how quickly the change would occur. By 2020, more than half of publishers believed digital sales would account for 20% of the market. Only one third of booksellers polled believed the same thing [45].

The growth of tablet and dedicated eReader devices also brings with it new conflicts in file format. ePub remains the most widely used format in text based viewing, but is not supported by Amazons' popular Kindle eBook platform.

Output from any modern digital library project will need to take this rapidly changing market into account. Most eReader devices rely on text-based file formats. The primary output from most library digitisation projects up till now have been image based, displayed on screen via PDF.

---

[42] "Bain & Company: "Publishing in the digital era" < Bain briefs < Publications."
[43] "Moore's law - Wikipedia, the free encyclopedia."
[44] "Ebooks Winners & Losers | Monday Note."
[45] "Booksellers at risk as digital growth accelerates, survey says | theBookseller.com."

OCR-derived text output with the correct format and mark-up has been relatively difficult to produce without expensive manual intervention. Going forward, textual output which can be repurposed to work on a variety of devices and in a variety of formats is arguably as more useful than a as a digital image. Search is arguably a major USP of the text -based eBook. As popular author Stephen King stated in the Wall Street Journal:

> *"I downloaded one 700-page book onto my Kindle that I was using for research. It didn't have an index, but I was able to search by key words. And that's something no physical book can do."* [46]

At the recent Charleston conference, Librarians were advised to build in capacity to manage changing outputs into a digitisation project. The consultant Joseph Esposito noted that flexibility in management, business plan and in funding sources is required for this to happen:

> *"A project also can't be run merely to recover costs, Esposito said. Instead, a project with long-term goals must aim for a surplus in order to fund ongoing enhancements. He urged the audience to think back a year and consider all of the digitization projects planned and funded before the release of the iPad, which has since altered the way many end-users wish to access collections and materials."*

This change is reflected in the publishing industry. Publishing systems have until recently revolved around producing a PDF document specifically formatted to act as a master for the printing process. Realising that true digital delivery occurs over a number of formats and devices, this approach is being seen as unsuitable. Instead, storage as a flexible marked up text, usually in XML provides a much more flexible approach. A parallel can be drawn with the use of TEI and other text-markup formats in library digitisation.

In contrast, some organisations are advocating the web browser as the primary means of making text available online. At the time of writing, Google eBooks has recently launched in the U.S. It takes a browser-based approach to viewing text, with and in-browser viewer and browser based applications available for a variety of devices. All text purchased is held in a cloud environment. HTML 5 technologies are used to display and reformat text and other document elements.

Whilst libraries are not unprepared for the digital preservation issues that will arise from these developments, their role in the post book word as enablers of access to information is certainly likely to change as the concept of the book is rethought. The traditional library practices of classification, storage and will be further eroded as digital distribution becomes more widespread.

Digitisation-on-demand: conclusions

- Kirtas book scanners represent a cost effective way to digitise full texts quickly at an acceptable quality level, although the initial capital investment for total ownership remains high

- Alternatives include leasing of equipment or use of third party services

[46] "Stephen King's "Full Dark, No Stars" - WSJ.com."

- Despite the high interest in a service, analysis of potential circulation data indicates that demand for 1850-1920 materials is likely to be limited to a niche audience

- Given the high throughput (over 1,500 requests p.a.) needed to meet staff costs, a self-funding digitisation service would be difficult to achieve

- Cost of digitisation cannot be fully met by the requestor at an affordable price point that the survey indicates potential users would be willing to pay, (between £5 and £20 for digitisation)

- There is greater interest in older rare material, that could not necessarily be digitised quickly using a Kirtas book scanner

- Re-purposable, re-flow-able text based output will be as important as image based output in the near future. Where available, manual transcription and OCR technology will add great value and ensure a digitised text is useable and re-purposed for use on handheld devices

Section four – Print-on-demand in libraries

Background

Print-on-demand is not a new concept in the book market, but one that is now being increasingly championed as an alternative to the traditional model of print-run based production. Many users are still unwilling to read material on a desktop, laptop or handheld device. Despite developments in screen reader technology, many readers still prefer a print copy of a work, especially when getting to grips with a text [47]. As noted in the previous section, print delivery will still have a role to play in libraries and the publishing industry for some time.

The traditional centralised publishing and book retail model is seen as being under threat, largely from changes in market conditions [48]. Print-on-demand is seen as a solution to these problems in publishing as it can be used to sidestep some of the burdens associated with the traditional printing model [49]. Publishers can themselves exploit their own 'long tail' of out of print content, often requested by readers. Frequently, print-on-demand operations are outsourced by publishers to specialist companies who are able to achieve greater economies of scale through serving the requirements of many publishers [50].

Print on demand already plays a role in library acquisition workflows. Companies such as Hollingworth and Moss provide print-on-demand services to libraries to replace damaged or lost items from stock. Many academic journal vendors now only offer print journals through print-on-demand.

One piece of technology, the Espresso book machine promises to de-centralise printing by providing a small, relatively cheap print and binding service through a single piece of equipment. This section focuses upon the Espresso book machine, its use within libraries and the potential impact upon traditional library services.

4.1 Case study #1 - Espresso book machine at the University of Utah Marriot Library

The Marriot Library in the University of Utah launched an Espresso based print-on-demand service on November 2009.

They are sourcing material from the EspressNet print-on-demand service, which contains over a million items from Google Books, and over two million from the Internet Archive, as well as an increasing amount of in-copyright material from publishers. EspressNEt is the primary means of getting content onto an Espresso Book Machine, in effect, the 'iTunes to its iPod'. In addition, they print material from the University of Utah Press and selected digitised material from their special collections. They aim to print anything that is 'open source or out of copyright'.

The pricing model meets allows the library to meet all costs associated with production. The Library also offers a 5% discount to members of the University. Utah note that interest

[47] "JISC national e-books observatory project » JISC national e-books observatory project: Key findings and recommendations."
[48] "Henry Porter | As I start to write my latest book, I fear for the future of publishing | Comment is free | The Observer."
[49] BARBER, "Meet publishers' enemy No. 1; Sci-fi novelist Cory Doctorow is shaking up the traditional book-selling model."
[50] Dean, "Books live forever in POD future.."

has been high, with self publishing becoming the unexpected focus of the device, taking five times the custom of print-on-demand titles from the public domain.

By contrast, readers found much of the material offered by ExpressNet to be rather old. Utah state that were digitised content offered by ExpressNet to overlap with their collections, they would expect Acquisitions and Inter-Library-Loan services to be affected. Utah would classify the EBM as disruptive technology:

> "*It undermines the need for traditional subject selection, disrupting a major sub-discipline of librarianship. By doing so, it also undermines the rationale for a large research collection—if the purpose of the collection is to meet patrons' information needs, and if they can now be met without buying and housing a large just-in-case collection, then how do we defend the unbelievably expensive and arguably quite wasteful practice of traditional collection building?"*

It has they argue a similar effect on publishing:

> *"It also undermines the need for publishers to print speculative runs of new books, thus potentially changing in a drastic way the logistics of the publishing world. In a rational marketplace, every bookstore would have an EBM or something that works on the same principle, and books would only be printed at the point of demand and purchase"*
>
> *"By making it possible to hold off on printing a book until the need for it has been demonstrated, and then to deliver the printed copy in virtual real time, the EBM essentially changes everything about the book business."*

The most notable disappointment from users relates to the limit amount of useful material available on the EspressNet service. Despite this Utah remain optimistic:

> *"Obviously, its full potential has yet to be realized—but the fundamental model is now in place. What are left to fix (bad metadata, incomplete catalog, rights issues, etc.) are the details. In most cases, fixing them will require only money and effort, and as roadblocks go those are relatively simple ones"*

This attitude ties into the wider issues of user selection and its effect on traditional library services.

4.2 Case study #2 – Blackwells Bookshop - London

Blackwells bookshop introduced the first Espresso book machine into a UK bookshop in 2009. The machine was placed in its Charing Cross Store in central London. The service offers both books from the EspressNet books-on-demand and a self-publishing service.

Print-on-demand sales using material on EspressNet are incorporated into the main Blackwells online bookstore. Staff noted that there was some discrepancy between prices for on-demand material and that in stock, but they hoped this would be resolved over time. The EspressNet service offers a different selection of material within the United Kingdom, including a different selection of material sourced from the public domain. Blackwells staff were aware of the discrepancy in copyright legislation, from the US. They expect more in-copyright material to become available as the service develops. In particular, material

offered through Lightning Source, a large print-on-demand supplier can also be potentially printed on the machine.

Blackwells have adopted a pricing model that also covers all costs, if not the initial capital expenditure for the machine. Placed centrally and visibly in the store, the device also serves to increase footfall, which is vital in attracting custom. This could help to justify some of the additional capital expenditure.

For self-publishing, an initial setup-fee is charged to format and enter a work into EspressNet, with each copy then being charged at the cost of printing. As with Utah, self-publishing was a lot more popular than initially realised, accounting for a large amount of transactions.

The machine is permanently staffed, with time split between desktop-publishing work to format print-on-demand material and operation and occasional maintenance of the machine.

Operationally, the machine has presented few problems. The ability of the main operator to quickly attend to problems and troubleshoot the odd mechanical problem is seen beneficial. Support engineers can access most functions of the machine externally. It is also equipped with a webcam to quickly diagnose mechanical faults.

Most staff in the store can easily perform basic functions, such as selecting and printing a work using the interface.

Blackwell's staff noted interest had been high and that the service had been developing well. They intend to roll out more Espresso book machines to other stores in the future.

Generally, customers are happy with the results, although managing user expectation, especially with regard to quality of works had been an issue. There was a noted difference in quality between works sourced from digital images (appearing closer to a bound photocopy in quality) and those from digitally typeset files. Some problems have occurred through mismatched cover and page size information.

4.3 Costs and throughput

As with digitisation, a cost modeller for print-on-demand services was developed. It was based around theoretical use of the Espresso book machine at Cambridge University Library.

- According to the modeller, a 400-page book could be printed at a cost of £8 and recoup operational costs. This does not include any attempt to recoup capital, surplus or maintenance costs.
- In terms of throughput, the Espresso book machine has a very high capacity. On-demand-books suggest a throughput of over 1,000 volumes per year would be required to meet operating costs.

Some real-world print-on-demand pricing models are presented in Appendix 8. Some sites such as Utah have chosen to pass the full cost onto the user on a per-page basis, others are charging standard fees based on size bands.

4.4 Perceived demand

The surveys of academic and library staff also touched on print-on-demand services. The following was noted:

- 65% of academic respondents would be interested in a bound print-on-demand version of a digitized request, In addition to, or in place of a digital copy
- 42% of academic respondents would be willing to pay £10-£15, 33% £15-£25
- 81% of Librarians surveyed would be interested in a print-on-demand version of a work
- As with digitisation-on-demand, Librarians would be willing to pay a more, with 15% willing to pay between £10-£15, 38% would pay up to £25 and £38% would pay over £25
- In both cases, there was a higher perceived value in a physical work, even though the actual act of digitisation is costlier than producing a print-on-demand copy

4.5 Existing Public Domain digital versions of a work

Sourcing a digitised surrogate from elsewhere could greatly reduce the cost of delivery for a total print-on-demand and digitise-on-demand service. The Hathi trust, Google books and the Internet Archive make public-domain material freely available via the Internet.

The full text could be accessed personally by a reader or Librarian, or potentially programmatically via an Application Programming Interface. All services have API access to material and rights information. Each API uses numerical identifiers' to match works.

In assessing each resource to source material for a digitisation or print-on-demand service, the following issues need to be considered:

1) Does the digital surrogate have any rights or usage restrictions in place?
2) Depending where the work was digitised, would it still be within the public domain within the UK?
3) Is the copy of acceptable quality and resolution for any physical reproduction?
4) Is it readily available in a re-purposable format?
5) Can a URL for full text be retrieved programmatically?
6) Can rights data be retrieved programmatically?

Each point is been examined in the context of the Hathi Trust, Internet Archive and Google books. Complete findings from this exercise can be found in Appendix #4.

For all three digital libraries examined, reliable access requires to existing digital objects will require accurate numerical identifiers for a bibliographic work. OCLC numbers are the most widely used identifier.

Programmatic access to source full text surrogates from anywhere other than the Internet Archive is limited. Google do not expose PDF file-paths directly via an API, and the Hathi trust restrict downloads of full PDF versions of works to authenticated users who are members of the Trust. Only texts from the Internet Archive can be easily accessed.

Creative-Commons licenses for material cannot be assumed. The Hathi Trust makes no mention of them on digitised text and both Google and the Internet Archive leave license assignment to the contributing party.

Sourcing a copy that is viable under UK copyright legislation (as opposed to US) also remains a challenge. Using a copyright calculator to assess a work' public domain status under UK jurisdiction before attempting to sourcing material may help to alleviate this. Both the Hathi Trust and Google books make some attempt to limit access based upon local copyright, by examining the IP address of the requestor.

According to OCLC figures supplied by Lorcan Dempsey, 20% of Cambridge University Library holdings as of July 2010 have been digitised and are available under the Hathi trust. This figure makes no attempt at assessing the UK public domain status of this proportion. Given the status of the University Library as a legal deposit Library, this is possibly comparable to other UK research libraries. Were Oxford University to join the Hathi Trust, it would be likely to increase substantially.

Given that the Hathi Trust is made up of some Google Books partners and represents about half of them, this percentage is itself indicative of the wider proportion of material already digitised in Google books.

BooksOnDemand and ExpressNet

BooksOnDemand, vendors of the Espresso book machine have tied their product to a print-on-demand 'catalogue of catalogues', allowing ready access to large amounts of public domain material. They also include in-copyright material made available under specific agreements with publishers. Much of the material available is sourced directly from Google books and the Internet archive. Integration options exist for catalogues and existing on-line book stores. This service could sidestep the legal and technical issues surrounding sourcing existing digital material for print-on-demand.

4.7 new forms of print output

Print-on-demand offers a number of intriguing options for library material.

- The Espresso book machine already offers large print versions of some works.

- Anecdotal evidence gained during the investigation suggests some researchers would like a separate copy of a work for 'treating badly'.

- Taking this a stage further, a cheap, notebook style version of a core text with every other page set aside for notes or comments could be useful.

- Blackwells academic publishing has recently launched a custom textbook publishing service. It allows course administrators to write and select articles, chapters and other information for a course-specific textbook. Blackwells then clear rights and can publish the book at a cost of under £30.

Print-on-demand: conclusions

- There is noted demand for print-on-demand in Cambridge from librarians and academics

- The Espresso Book machine was examined and seen as a potentially useful service for libraries. It can provide printed copies of works at a far cheaper price than any existing services at Cambridge University Library

- The same constraints over copyright that restrict digitisation-on-demand also apply, to print-on-demand. As print-on-demand services such as EspressNet start to offer material from publishers, this will hopefully change

- Operating and capital costs are seen as lower than those for digitisation, although capital remains high. The option to provide quick affordable self-publishing services may help recoup initial investment. Alternatively, the use of third party services for print-on-demand present a lower risk means to trial and scope a service equipment may prove more cost effective if demand is low

- Obtaining and re-presenting copies of digital works from other sources is technically possible but in practice currently difficult to achieve from any source other than the Internet Archive. The EspressNet service from books-on-demand could again sidestep many of the issues surrounding this

Appendix #1 – Theoretical use case scenarios for a library digitisation-on-demand service

These use-case scenarios assume a digitisation-on-demand service operating at Cambridge University Library.

The service would be run along these principles:

- Readers would initiate a request through a variety of online interfaces, principally the Library Catalogue
- Only work assumed to be in the public domain would be available for full digitisation
- Digitisation would be done to acceptable levels of quality rather than full preservation
- Scanning would be done at cost, paid in principle by the requester at the point of order
- All digitised material would then be made freely available to the public under a Creative Commons License
- PDF, plain text and potentially automated marked-up versions of works would be available
- The Library would offer print facsimiles at a reasonable profit

Teaching: Historical text in a course pack
An academic would like to use three chapters of a work in a course pack. They are aware that if a work as still in copyright this would be a breach of the fair-usage allowance for education purposes. Using the library catalogue, they can check for a viable estimate of the copyright status of the work and request a digital copy for distribution via a Virtual Learning Environment.

Post-graduate study: full text facsimile for 'getting to grips' with a work
A post-graduate researcher is investigating early photography techniques. She will be basing large amounts of research around a single 19th century photography manual. Only a few copies of this exist in legal deposit libraries, under closed access. Using the library catalogue, she can check the copyright status of a work and request a digital copy.

As this is a relatively cheap print-on-demand facsimile, she can feel free to write her own notes and amendments on the work, which should last for the period of her study.

The media: researcher wants copy of work to aid scriptwriter
A researcher for the BBC has been told to get multiple editions and variations of a classic work of literature in preparation for the next large costume drama. Using a digitisation on demand service, they can quickly source multiple PDF versions of the work for easy distribution across a large script-writing and editorial team.

For desktop and on-set reference, print-facsimiles of a couple of the versions of the work have been ordered.

The researcher: scripted analysis of historical texts
A post-doctorate researcher in gender studies and linguistics would like to analyse changes in language usage in literature over time. To do this, she is running a variety of analysis scripts over the full text of digitised works to spot language usage patterns.

She had been able to source many of the required texts already, but would like a further 20 from the University Library that have not yet been digitised anywhere. Furthermore,

some of the texts she has sourced are available in image-based PDF only, a format that makes scripted analysis next-to impossible.

Rather than wait to see if these texts become available elsewhere, she can check their copyright status and place digitisation requests directly via the library catalogue.

Within a few days, all works have been digitised with both image and OCR-derived text versions available for download.

Using extensions to the library APIs that provide direct URLs for full OCR derived text, she is able to harvest the texts automatically to run scripts for analysis of language patterns.

The international Creative Commons - open access publisher wants on increase its online portfolio
An online library of public domain works would like to increase the size and scope of its collections. The site relies on providing full text-based transcriptions of historical works, rather than scanned image based surrogates.

It has noted the digitisation-on-demand output produced by the University Library and would like to add the material it produces to its collections. Using API's from the digital library service, it can automatically access all new text-based material as it becomes available. The OCR derived text can then be forwarded onto the sites' team of volunteer transcribers, who can proof and edit the OCR text as required.

As the scans are provided under an unrestricted use Creative Commons License, clearly explained upon the website, this presents no problems.

The Public: a family gift
In researching his family tree, a retired accountant came across a series of short stories written by a distant relative. He is able to request digitised copies of the work from the University Library and can provide his family with copies as gifts.

Other libraries: copy for preservation purposes
A small specialist library and archive attached to a cathedral maintains the private collections and archives of famous figures from the history of the institution.

In order to preserve the existing stock, it can order print facsimiles, sourced from digital surrogates via the library catalogue. These can then be offered to readers and researchers within the library whilst the original items are preserved.

The cultural sector: repurposing of print based material
A museum curator would like to create an online database to catalogue a newly acquired collection. A print based catalogue of the material is already available, but digitising a copy is prohibitively expensive for an institution his size.

Requesting a digitised copy of the catalogue held within the University Library, they can take the OCR derived plain full text of the catalogue and pay a local software developer to parse it into a normalised form. This can then be entered into a database and used to develop a web-based catalogue.

Appendix #2 – Overview of UK copyright legislation

Copyright duration in the UK

Copyright is an automatic process; no registration or formal assertion of ownership is required. It is often said that copyright subsists rather than exists. In order to receive copyright, a work must first be 'recorded in a material form' [51].

The work itself must also be original. In the UK this does not itself mean that a work must be new, but that the author must have used 'skill, labour knowledge and judgement in its creation' [52]. Definitions of originality vary internationally.

The Copyright Designs and Patents Act 1988 (CPDA) is the most prominent piece of intellectual property legislation in the United Kingdom. It has been amended by the Copyright and Related Rights Regulations 2003, bringing into force key parts of the European Copyright Directive 2001/29/EC [53]. The Act seeks to protect the 'economic and moral' rights of creators by restricting the copying of material.

Economic rights include the right to:

- Copy the work (including digitisation)
- Issue copies of the work to the public
- Perform, show play or otherwise communicate the work to the public
- Adapt or translate the work [54]

When anyone other than the copyright holder or someone licensed on their behalf performs these actions, it is seen as a primary infringement of the creators' copyright. Secondary infringements also exist. These include possessing or dealing with an infringing copy and providing the means for making infringing copies. In the context of digitisation and print-on-demand services, these last two points could directly affect the library or digitisation service provider.

Copyright law varies in the UK depending on the nature of a work. Copyright duration of films and broadcast media is assessed through somewhat different criteria to literary, dramatic and musical works. There is also significant complication in the area of un-published works.

Additional specific provision is also made for databases as composite entities under the Copyright and Rights in Databases Regulations 1997 (SI 1997 No. 3032). Additional database rights can exists if there has been a substantial investment in assembling, verifying and presenting the contents of a database. Not restricted to electronic material, print based compilations of information can also qualify for copyright and database rights.

Duration

Copyright is not perpetual, but exists for a set duration. Any work where copyright is expired or not affected by the larger set of intellectual property laws can be judged to be within the public domain.

---

[51] Padfield, *Copyright for archivists and users of archives / Tim Padfield.*, 16.
[52] Padfield, *Copyright for archivists and users of archives / Tim Padfield.*??
[53] Pedley, *Essential law for information professionals / Paul Pedley.*, 22.
[54] Ibid., S.16.1 of the CPDA, in Pedley p.22.

For printed literary or musical works copyright term is generally for the life of the author/creator plus 70 years (expiring on midnight 31st December for that year), published or not. If the author is unknown, the copyright is 70 years from the date of creation, or if made publically available, from the date of availability.

Variations to duration
Any work that has gained a longer duration under a previous act still retains the original, longer duration as defined in that act (usually the 1988 un-amended or 1956 act). In addition, there are specific variations in term related to this issue that have arisen due to changes in term [55]:

1) If a work is not published or performed before 1 August 1989 and the author dies before 1969, copyright expires in 2039 (1988 act + 50 years)

2) If the author died more that 20 years *before* the date of original publication, assuming that publication date was before the 1st August 1989, the copyright expiry term is shortened to 50 years

3) With works of unknown authorship where the work was created/first published/ made available after 1969, copyright expires in 2039 (1988 act + 50 years).

Revived copyright is also a potential risk. Until 1996, the standard term in the UK was set at 50 years rather than 70. This has lead to copyright for literary, dramatic musical or artistic works to be extended or revived under specific circumstances.

The international context
Despite attempts to 'harmonize' copyright duration across the European Union to a standard 70 years after death, significant variation exists in approaches to determining originality of a work. There are also different variations in duration of copyright, most notably with regard to unpublished works [56].

Generally speaking, material held in a UK archive is protected by UK law, usually only variations in copyright duration apply. To be protected in the UK, the work must be published in a country that has signed one of a large number of reciprocal copyright treaties, the main one being the 1886 Berne Convention. There are very few countries with whom the UK does not share a treaty.

The biggest difference lies in the U.S. where literary works published before 1923 can be automatically considered to be in the public domain, (although additional complications for unpublished works apply). This difference in approach to the public domain can cause complications in joint U.S. and U.K. digitisation operations.

Underlying material, typeface and illustration copyright
Whilst copyright assessment is to some extent simplified by restriction to published monographs, there may be other variations in duration based around the nature of a work.

Even in light of restricting digitisation-on-demand services to monographs, many could be considered composite works. A published work may have elements contributed by several creators. If each has contributed individual chapters, then copyright would need to be assessed separately.

[55] Padfield, *Copyright for archivists and users of archives / Tim Padfield.*, 23.
[56] Ibid., 61.

A work may also be co-authored. Where multiple owners of copyright exist through co-authorship, duration is based upon the death of the last surviving author.

Furthermore, any typeface or illustration within the work may itself be subject to a separate copyright from the parent work.

Copyright exceptions for libraries

Various exemptions in copyright exist for libraries, allowing users to copy a work for specific purposes. The most well known is that of fair dealing. In the UK, this is restricted by convention to 5% or a single chapter or article from a work, whichever is greater, for research or private study. This exclusion provides no exemption for wholesale digitisation of a work in copyright, and relates only to literary dramatic and musical works. Sound and film recordings are specifically excluded.

The second is copying for preservation purposes, in which a library or archive can make a copy to preserve or replace the original work. Any library can make a copy, but only 'prescribed libraries' can receive them. The work itself must be out of print. The exemption again only applies to literary, dramatic or musical works and the illustrations within them[57].

[57] Oppenheim, "Legal issues for information professionals VI: copyright issues in digitisation and the hybrid library.," 204.

Appendix #3 – Technical issues encountered during calculator catalogue integration prototype

A simple php script was written to map bibliographic data taken from the University catalogue bibliographic record API. This was mapped against the JSON results from the open knowledge foundation copyright calculator.

100 sample results were viewed, chosen using the random search function built into the libraries' Aquabrowser resource discovery platform. A mixture of formats and material from a good international spread was investigated. In each case, the results displayed by the copyright calculator were assessed against the expected result.

Out of the 100 samples, 76 returned an expected result given the data available, 24 were judged as incorrect. Out of these, a further eight could have been correct if a publication date + 150 years safe cut-off date was assumed.

A3.1. Issues with MARC / AACR2

Many of the initial problems encountered in integration were due to problems with bibliographic data written to AACR2 standard in MARC21. Some also highlight issues with localised cataloguing practice in Cambridge.

A3.1.1 Death dates

AACR2 suggests that death dates be used to differentiate between two different individuals in a name authority file. This has lead to patchy and occasionally inaccurate death information. For the 1890-1900 period, only 43% of all records in CUL have an author death date record. Where no concrete death information is available, AACR2 also allows cataloguers to record when a writer or creator 'flourished'.

Indexes of Births, Deaths and Marriages and other registries of death information have not openly provided their data for re-use, or even provided programmatic interfaces. Information has to be sourced 'by hand' referencing printed indexes for birth and death information.

Given this patchy data and lack of open pragmatically available external resources, the calculator already assumes death as 100 years from birth. Assuming another safe cut-off point from year of publication of 150 years would have increased accuracy by 8%.

A3.1.2 Granularity of date information

MARC21 was designed to hold dates for display only, in a format designated by AACR2. All dates are held in a single subfield, with specific punctuation and annotation used to denote dates, e.g. a hyphen to separate birth and death dates and the use the prefix d. to denote a death date with no birth date.

Parsing dates into a normalised format useable by the copyright calculator requires manipulation of text through regular expressions. This in turn also requires that hand entered punctuation to be correct. Data stored or made available in the newer MODS format also lacks sufficient granularity, both only store a 'date' field. ONIX, a standard used heavily in the publishing industry has a more effective granular mechanism to hold dates.

A3.1.3 Identifiers
Lack of useful identifiers in Cambridge records; especially for historical material causes problems when attempting to identify previously digitised versions of a work, where ISBNs are not generally assigned.

It was not possible to usefully integrate API –derived data on full text availability from the Hathi Trust, Open Library and Google books due to this issue. The inclusion of OCLC identifiers in all Cambridge University Library MARC records would rectify this.

A3.1.4 Location codes and place of publication
MARC uses its own country codes. The Open Knowledge Foundation uses the more widely used ISO 3166 format. As such, mapping between the two codes had to be done by hand. This mapping does not currently extend to State specific US, Canadian and Australian codes. Some of the matches are imperfect.

A3.2 Issues with the calculator
The following section covers problems encountered primarily with the copyright calculator.

A3.2.1 Additional types
The copyright calculator currently accepts four different material types. 'text', 'composition', 'photograph', and 'law'. The following additional types are suggested:

Serial. To denote a work in continuing publication, possibly one with several authors. It may have a regular publishing date. In such cases – the calculator could simply return a warning requesting granular information on a specific article or part of the work

Recording. Provision for sound and video recordings should be made, given their separate nature. In particular, audio recordings of works out of copyright (i.e. a spoken word version of Pride and Prejudice) do not work well with the current calculator

A3.2.2 Non-UK Jurisdiction
The calculator currently has three separate areas of jurisdiction, UK, US and Canada. Expansion to other territories, especially those within the EU would be very useful. 13 records within the sample had a non US/ UK place of publication. Where the calculator does not support a jurisdiction, it should

A3.2.3 Crown copyright – how is this identified
Material published by the UK government is subject to different restrictions under copyright legislation, (variable term of 50-125 years from publication or creation). Identifying the crown as an author

A Library of congress name authority entry or identifier for the UK government as an author would be required (many currently exist, alternatively, identifying the publisher as her Majesty's' Stationary Office may be sufficient.

A3.2.4 Translations
MARC and AACR2 allow for translators to be entered as additional entries to a catalogue record. This information is normally stored in an additional entry field, although the role as the author as translator is not always recorded.

The copyright calculator as a 'person – type' attribute in its input data. This could potentially be used to denote a translator – identifying this work as a translation.

A3.2.5. Composite works and later editions – prefaces, editorial material, plates and illustrations

The biggest holdup in copyright clearance workflow can lie in tracing additional authors and illustrators in a work. According to local cataloguing practice, illustrators and additional contributors (preface etc.) creators could have be entered into the cataloguing. Whilst their names are often recorded, their role as editor, illustrator or additional contributor is often not. These are referred to as added entries.

Of the sample 100 records, 5 had added entries in place of as well as a given author. This could have affected copyright calculation. Another five had corporate authors as creators. In this case, the copyright would expire at the death of the employee.

Some notes referring to a plate, preface or illustration may have been entered into the physical description section of the catalogue record. This data could potentially be surfaced.

- *Library recommendation – parse additional entries into bibliographic record API – preferably with some means of qualifying role. Also attempt to pass details on plates and illustration.*

- *Calculator recommendation – expand the person type beyond author to deal with editors, illustrators and corporate authors.*

Appendix #4 – further details on API access to full text in major digital libraries

A4.1 Does the digital surrogate have any rights or usage restrictions in place?

Hathi trust:
For material in the public domain, Hathi trust states that:

> *"Works in the public domain in HathiTrust are open to all researchers—whoever and wherever they may be. Content in HathiTrust is discoverable through online search technologies within the repository and through Google, with no authentication, login, or password required* [58]*"*

No DRM is in place on public domain content. No Creative Commons license appears to be directly issued to public domain material provided by the trust. Material assessed to be in copyright under US law is restricted to search only access to external organisations:

> *"Access to materials in the repository is determined by 1) copyright law and 2) permissions granted by individual rights holders. Works that HathiTrust partners do not have rights to make available are not made available, or are made available under very limited circumstances (such as to certified users with disabilities who need to make use of a screen reader in order to access materials)"*

Google books:
Google state:

> "You can see books in Full View if the book is out of copyright, or if the publisher or author has asked to make the book fully viewable. The Full View allows you to view any page from the book, and if the book is in the public domain, you can download, save and print a PDF version to read at your own pace."

No DRM is in place on public domain content. Google has a programme in place where by partners placing material in Google books can specifically assign Creative Commons licenses [59].

Internet archive:
The Internet Archive text web pages state:

> "*This collection is open to the community for the contribution of any type of text, many licensed using Creative Commons licenses."* [60]

It places no restrictions on access. No site appears to give information on any technical restrictions (i.e. maximum download bandwidth restrictions).

[58] "Rights Management | www.hathitrust.org."
[59] "Google Books adds Creative Commons license options - Creative Commons."
[60] *2,559,807 itemsWelcome to Ebook and Texts Archive.*

A4.2 Depending where the work was digitised, would it still be within the public domain within the UK?

Hathi trust:
The Hathi trust describe in detail on their website the mechanism used to determine copyright status of a work. It uses bibliographic information to automatically determine if material is in the public domain. This includes:

- US federal government documents
- Published in the US prior to 1923
- Published outside of the US before 1870 [61]

The 1870 cut-off date would apply to UK publications in the Trusts' database. Material published in the US could not assumed to in the public domain with the UK, although US editions in the public domain may be an acceptable alternative. Further definitions, including negotiated Open Access and manually attributed public domain status can also be applied.

Google books:
Google books contains material both in copyright and out of copyright. As discussed earlier, it was sued in 2005 for breach of copyright by Author's Guild and the Association of American Publishers. According to an update on the Google books, material outside of the US is restricted to that in the Public Domain.

> *"Due to fundamental differences between U.S. and European copyright laws, Google opted to digitize in-copyright out-of-print and orphan works only from its library partners in the United States. It believed at the time, and maintains to this day, that it was allowed to digitized these categories of works under the United States' broad fair use exception to copyright law. Without an analogous exception for it to rely upon under European copyright law, Google chose from the start to restrict its European digitization projects solely to works in the public domain*[62]*"*

After the 2009 settlement, rights holders in the UK can voluntarily opt to have material in Google books preserved under US legislation. For the purposes of sourcing public domain content from Google, this will have little effect.

The Google books data API uses IP address to determine the location of a user and restricts the results returned accordingly.

Internet Archive:
Material held within the Internet Archive is deemed to be within the public domain under US legislation. The onus in determining copyright status of a work submitted to Internet archive is placed directly in the hands of the individual or organisation submitting it.

A4.3 Is the copy of acceptable quality and resolution for any physical reproduction?
In all cases, this will depend greatly on the project and mechanisms used to digitise a work in the first place. Many of the public domain items

[61] "HathiTrust Rights Database | www.hathitrust.org."
[62] "Update on Google Books Statement | IFLA."

In order to print effectively, a digital surrogate in PDF format will have to be in xxx DPI resolution in grayscale / CKMY / B/W format.

There are known quality issues with Google books scanning. In attempting to use a scan from a third party archive, quality assessment should be made

A4.4. Is it readily available in a re-purposable format?

Hathi Trust:
Hathi trust makes work available in a browser in PDF, text or image (JPG) based format on a per-page basis, without login. The trust also provides a REST-FUL data API to provide this information

To access full-text of a work from the trust a user must first authenticate via its Shibboleth derived federated access management system. This would make programmatic access to full text material from the Hathi trust potentially difficult.

Google books:
Formats available from Google books include image-derived PDF and OCR-derived ePub text based files and the open DAISy format. Availability of both formats is dependent on the supplier of material and cannot be assumed for all material within the public domain. Where they are available, no restriction is placed on download.

Internet archive:
The Internet archive makes texts available in the widest variety of text and image derived file formats, including PDF, text and ePub formats and the DJ-VU viewer and formats specific for e-reader devices such as the Kindle.

A4.5 Can a URL for full text be retrieved programmatically?

Hathi-trust:
The Hathi Trust provide a full data API to access granular (page and section) level OCR text, image and work level metadata. The API does not currently allow non-partner institutions to download full text in anything other than an atomised form, i.e. as a full PDF.

Google books:
Google has a variety of programmatic means to access works. Access to full text is based a round a client side Javascript API designed to embed a viewer in a web page. An additional data API based around the opensearch standard allows for the querying and manipulation of bibliographic information.

Google books has a simple URL structure is used for access to work level information:

http://books.google.com/books?vid=ISBN0451522907 or
http://books.google.com/books?vid=OCLCXXXXXXXX

PDF filepaths requires a Google books internal ID: x1Q9TxhYA and the known name of a work:

http://books.google.com/books/download/Little_Brother.pdf?id=x1Q9TxhYA3sC&output=pdf

There is no documented API to access full text in either ePub or PDF format directly. The client-side viewer API allows works to be surfaced and displayed in a web vlient only.

The data API returns bibliographic data only. Only the short description of a work in the dc:description field can currently be returned.

Internet Archive:
The Internet Archive allows full direct httpd access to all files in its text archive. It also exposes directory structures to remote browsing and access. Programmatic access is encouraged.

A4.6 Can rights data be retrieved programmatically?

Google books
The Google books data API has no specific rights data returned in its API. Instead the data API uses IP address to determine the location of a user and restricts the results returned accordingly. A gbs:viewability additional field returns the level of access allowed to an item, from details on to full text access.

Hathi Trust
The Hathu Trust bib_data API returns both a rights code and textual information on rights relating to a digital object according to US jurisdiction.

Internet archive
The Internet Archive's Open library pages provide API access to data regarding texts in the archive. The REST API does not currently return rights or access information.

5 – Circulation and request information for pre 1920 material in Cambridge

Circulation transactions from 01/01/2008 to 01/11/2010 by publication date (all material types except serials)				
Period	**University Library**	**English Faculty Library**	**Philosophy Library**	**History Faculty Library**
Total (all)	**480807**	**166598**	**31490**	**150950**
Total (1850-1920)	***6198***	***3102***	***122***	***1225***
%	***1.29***	***1.86***	***0.39***	***0.81***
breakdown:				
1911-1919	2912	784	17	**354**
1901-1910	2666	1077	24	**246**
1891-1900	401	602	20	**233**
1881-1890	229	219	41	**173**
1871-1880	40	189	8	**64**
1861-1870	28	152	5	**70**
1851-1860	17	71	1	**77**

- The above information is best used to gauge general demand. Due to differences in loan length and material available for circulation, it is not suitable for comparative purposes

- SQL used to extract the above data:

```
SELECT BIB_TEXT.TITLE, MFHD_ITEM.ITEM_ENUM, MFHD_ITEM.CHRON,
LOCATION.LOCATION_NAME, [start] AS StartDate, [end] AS EndDate,
BIB_TEXT.BEGIN_PUB_DATE, BIB_TEXT.BEGIN_PUB_DATE,
CIRC_TRANS_ARCHIVE.CIRC_TRANSACTION_ID, LOCATION_LIMIT.LIMIT_NAME,
BIB_TEXT.BIB_FORMAT FROM ((LOCATION INNER JOIN ((CIRC_TRANS_ARCHIVE INNER
JOIN ((BIB_TEXT INNER JOIN BIB_MFHD ON BIB_TEXT.BIB_ID = BIB_MFHD.BIB_ID) INNER
JOIN MFHD_ITEM ON BIB_MFHD.MFHD_ID = MFHD_ITEM.MFHD_ID) ON
CIRC_TRANS_ARCHIVE.ITEM_ID = MFHD_ITEM.ITEM_ID) INNER JOIN MFHD_MASTER ON
BIB_MFHD.MFHD_ID = MFHD_MASTER.MFHD_ID) ON LOCATION.LOCATION_ID =
MFHD_MASTER.LOCATION_ID) INNER JOIN LOCATION_LIMIT_LOCS ON
LOCATION.LOCATION_ID = LOCATION_LIMIT_LOCS.LOCATION_ID) INNER JOIN
LOCATION_LIMIT ON LOCATION_LIMIT_LOCS.LOCATION_LIMIT_ID =
LOCATION_LIMIT.LOCATION_LIMIT_ID WHERE (((CIRC_TRANS_ARCHIVE.CHARGE_DATE)
Between [start] And [end]) AND ((LOCATION_LIMIT.LIMIT_CODE) Like "UL*")) GROUP BY
BIB_TEXT.TITLE, MFHD_ITEM.ITEM_ENUM, MFHD_ITEM.CHRON,
LOCATION.LOCATION_NAME, [start], [end], BIB_TEXT.BEGIN_PUB_DATE,
BIB_TEXT.BEGIN_PUB_DATE, CIRC_TRANS_ARCHIVE.CIRC_TRANSACTION_ID,
LOCATION_LIMIT.LIMIT_NAME, BIB_TEXT.BIB_FORMAT,
MFHD_MASTER.NORMALIZED_CALL_NO, MFHD_MASTER.DISPLAY_CALL_NO HAVING
(((BIB_TEXT.BEGIN_PUB_DATE)>"1850" And (BIB_TEXT.BEGIN_PUB_DATE)<"1920") AND
((LOCATION_LIMIT.LIMIT_NAME) Like 'UL*') AND ((BIB_TEXT.BIB_FORMAT) Not Like "as"));
```

Rare books fetching in the University Library (pre 1850 material)							
	2004	**2005**	**2006**	**2007**	**2008**	**2009**	**2010**
Aug	4068	4320	4121	4197	4133	5132	4724
Sep	2888	3595	2386	3330	4033	3153	2896
Oct	3003	3471	4853	3138	4081	2851	3914
Nov	4317	3607	3783	4006	3931	4549	3724
Dec	2868	2686	3006	3530	3849	2799	2597
Jan	3080	3031	3262	3526	2823	3160	2784
Feb	3690	3691	3902	4093	4119	3647	4397
Mar	3205	2967	4036	5166	4944	3779	3806
Apr	3297	4300	4416	3105	3474	3065	2976
May	4071	2841	3439	3037	4918	4190	3783
Jun	2933	3184	3401	4705	3703	3335	2805
Jul	4353	4169	4922	3429	3739	3066	4670
Total	**41773**	**41862**	**45527**	**45262**	**47747**	**42726**	**45086**

Appendix #6 – Survey results

Two surveys were composed to assess potential need and pricing requirements for a digitisation-on-demand service. One was aimed specifically at Cambridge Librarians. It contained within it a request to forward the smaller second questionnaire to academic staff. 61 academic staff and 16 librarians responded. Results from both are presented below.

Digitisation-on-demand – Academic responses

Question 1. In place of, or in addition to, requesting a book from UL stacks, would you be interested in a full text digital copy (i.e. as a PDF or text file) – assuming this is available under UK copyright legislation?		
Answer Options	Response Percent	Response Count
Yes	91.8%	56
No	8.2%	5
answered question		61
skipped question		0

Question 2. How much would you be willing to pay for a full-text copy of a work to be scanned on demand?		
Answer Options	Response Percent	Response Count
£ 10- 15	66.7%	36
£ 15-25	25.9%	14
£ 25 -35	5.6%	3
£ 35 +	1.9%	1
answered question		54
skipped question		7

Question 3. How long would you be willing to wait for such a copy?		
Answer Options	Response Percent	Response Count
24 hours	10.2%	6
2-3 days	20.3%	12
3-5 days	25.4%	15
1 week or more	44.1%	26
answered question		59
skipped question		2

Question 4. If a full copy is not available – would you be interested in a partial copy (i.e. one chapter) as a PDF, image or even plain text?		
Answer Options	Response Percent	Response Count
Yes	62.1%	36
No	37.9%	22
answered question		58
skipped question		3

Print-on-demand – Academic responses

Question 5. In addition to, or in place of a digital copy, would you be interested in a bound print-on-demand version of a digitised request?		
Answer Options	Response Percent	Response Count
Yes	65.5%	38
No	34.5%	20
	answered question	58
	skipped question	3

Question 6. If so, how much would you be willing to pay for a printed copy of a digitised work?		
Answer Options	Response Percent	Response Count
£ 10- 15	42.9%	18
£ 15-25	33.3%	14
£ 25 -35	16.7%	7
£ 35 +	7.1%	3
	answered question	42
	skipped question	19

Additional comments from academics

Quality …

"I have bought a number of scanned / pdf / print-on-demand reprint books and the quality is often very poor. Some of the books I have bought through Amazon have been so poorly scanned as to be useless, but it then becomes difficult to say who exactly is at fault, especially if the original book scanned is in poor condition, difficult to know what your rights are as a customer, and what the policy on returns might be. There are likely to be some difficulties with a percentage of scans and this ought to be considered an important problem for developing a reliable and useful service,"

Cost …

"The question asked whether I was interested in access to a digital copy 'in addition to, or in place of' a hard copy at the UL. I answered yes - but think it important to add that I think it is important for the soft copy to be additional to, and ***not*** in place of, hard copies. Nearly no-one likes reading lots of text on a screen, so for the university to hold only soft-copies of material simply off-loads the cost of printing onto individual academics, and we all end up with reams of A4 printed matter we don't need."

"I should add: I don't think the university should be charging academics for access to research materials. Seriously, the two questions about funds struck me as ridiculous."

Value as a replacement service …

"I can't do this because we aren't given the option of either 'in place of' or 'in addition to'. I would be interested in various possible options for the latter but absolutely not in anything relating to the former: if I want something from the UL, I expect to be able to get it, not to have to wait to purchase a digital copy."

"If I wanted a scan of a chapter or even of a whole book, under the current system I could easily go into the UL and do it myself. Your minimum price (£10) seems too high. For £10 you could even buy a large proportion of the UL stock on Abebooks."

Digitisation-on-demand - Librarian responses

Question 1. In the past year, approximately how many times have your users requested a digital copy of an entire work which was otherwise unobtainable?		
Answer Options	Response Percent	Response Count
0 times	62.5%	10
1-5 times	25.0%	4
5-10 times	6.3%	1
10 -15 times	0.0%	0
More than 15 times	6.3%	1
	answered question	16
	skipped question	0

Question 2. In the past year, how many full or partial digital versions (PDF) of a work have you created yourself or within your library to fulfill a reader request?		
Answer Options	Response Percent	Response Count
0 times	62.5%	10
1-10	31.3%	5
10-30	0.0%	0
More than 30 times	6.3%	1
	answered question	16
	skipped question	0

Question 3. As Librarians, would you be interested in a digitise on demand service from the University Library for full or partial copies of a work?		
Answer Options	Response Percent	Response Count
Yes	93.8%	15
No	6.3%	1
If yes, please state why (replace lost / damaged stock etc)		14
	answered question	16
	skipped question	0

Maybe- it would depend on factors such as cost and whether it would be available through ILL.
To provide more convenient access to the collection for users who prefer not to use the UL.
If the price was competitive it would provide an alternative to buying second hand out of print books from online suppliers. The quality of the works supplied would be more reliable than from second hand book suppliers.
It would be more preferable to ordering from the British Library, especially as there is no Cambridge inter-library transfer system in place
A digital copy would be great to provide extra copies of hard-to get works in heavy demand where our library has only one or no copy of the book but the title is available elsewhere in Cambridge. Would this only be available to individuals (i.e., the reader making the request) or could we make a digital copy available at a library level?
LOst or damaged items, certainly, but also to provide access to work out of print and unavailable from second-hand sources
Sheer demand for certain books (e.g. required reading for essays) when we just can't afford more copies.
I probably would be more interested if I knew more, but I am confused between digitise on demand and print-on-demand (see below). At the moment, I think the latter would be more useful for our particular service. Would digitise on demand be for individual use, or could it be made more widely available eg. through the e-books project?
-Replace lost/damaged books

-Order out of print/unavailable books

-Perhaps (if the cost was reasonable) to free library staff time from doing our own digitising of

books/chapters
Possibly. To replace lost/stolen stock
We buy a lot of OP items, some of which have to come from abroad and may take several weeks or months to come.
Occasionally to scan a chapter for student access via CamTools if item was on a reading list but out of print.
Replace stock, source items otherwise out of print and unavailable. Preserve rare books.
replace lost/damaged stock

Question 4. How much would you be willing to pay for a complete copy of a work to be scanned on demand?		
Answer Options	Response Percent	Response Count
£ 5-10	21.4%	3
£ 10- 15	28.6%	4
£ 15-25	28.6%	4
£ 25 +	21.4%	3
	answered question	14
	skipped question	2

Question 5. How long would you be willing to wait for such a copy?		
Answer Options	Response Percent	Response Count
24 hours	0.0%	0
2-3 days	21.4%	3
3-5 days	35.7%	5
1 week or more	42.9%	6
	answered question	14
	skipped question	2

Print-on-demand – Librarian responses

In addition, or in place of a digital copy, would you be interested in a print-on-demand version of a work?		
Answer Options	Response Percent	Response Count
Yes	81.3%	13
No	18.8%	3
	answered question	16
	skipped question	0

If so, how much would you be willing to pay for a full bound copy of a work to be printed?		
Answer Options	Response Percent	Response Count
£ 5-10	7.7%	1
£ 10- 15	15.4%	2
£ 15-25	38.5%	5
£ 25 +	38.5%	5
	answered question	13
	skipped question	3

Have you ever used PDFs or digitised texts from the following major public domain digitial library collections?

Answer Options	Response Percent	Response Count
Hathi Trust	0.0%	0
Google books	31.3%	5
Internet Archive	0.0%	0
none	68.8%	11
	answered question	16
	skipped question	0

If so, how have you found the quality of the digital copies available

Answer Options	Response Percent	Response Count
Acceptable	83.3%	5
Below requirements – if so, please state why in the comments:	16.7%	1
Comments:		3
	answered question	6
	skipped question	10

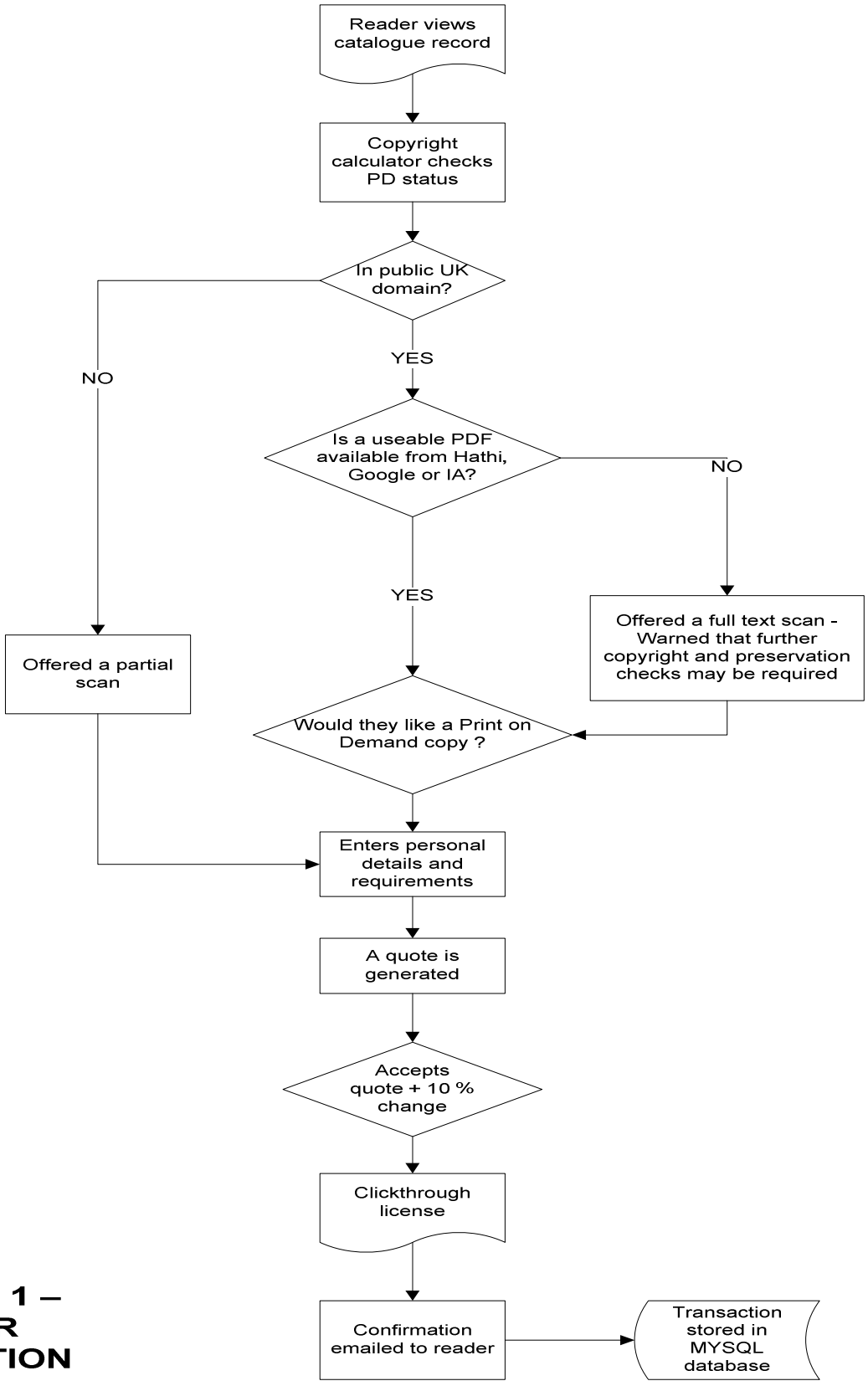
In my experience resolution is a big issue. Quality which is acceptable for ordinary text quickly makes a mess of mathematics or music typesetting. However, as the resolution increases, the download and processing times increase.

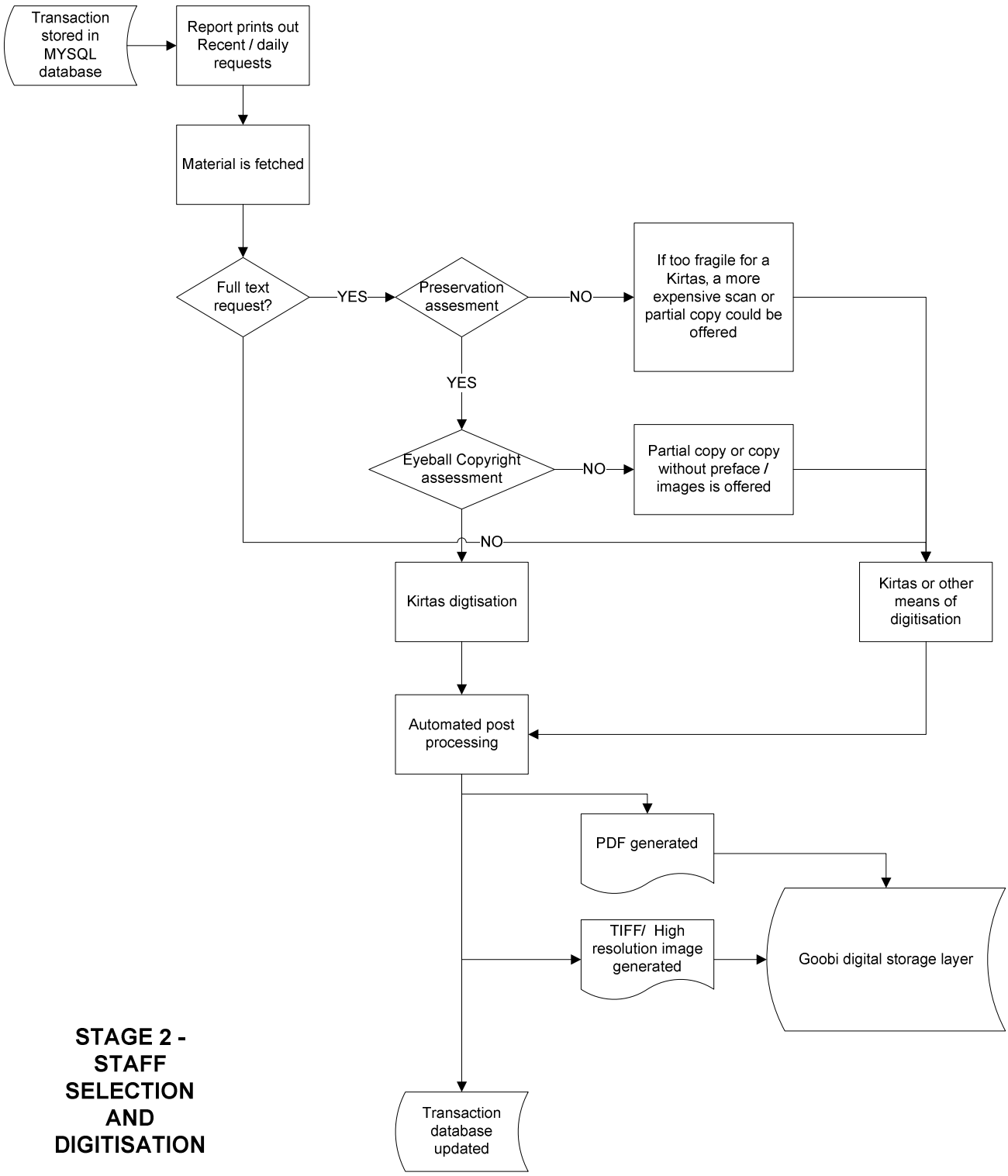
I have not yet used these services since I wasn't aware they were available but will now investigate them.

We have bought commercial print-on-demand in the past and found the quality of the copies available there well below requirements (unformatted pages left in, and over a dozen pages with the scanned fingers of the person who had scanned the book) so would prefer to go through a Cambridge scheme for such a service.

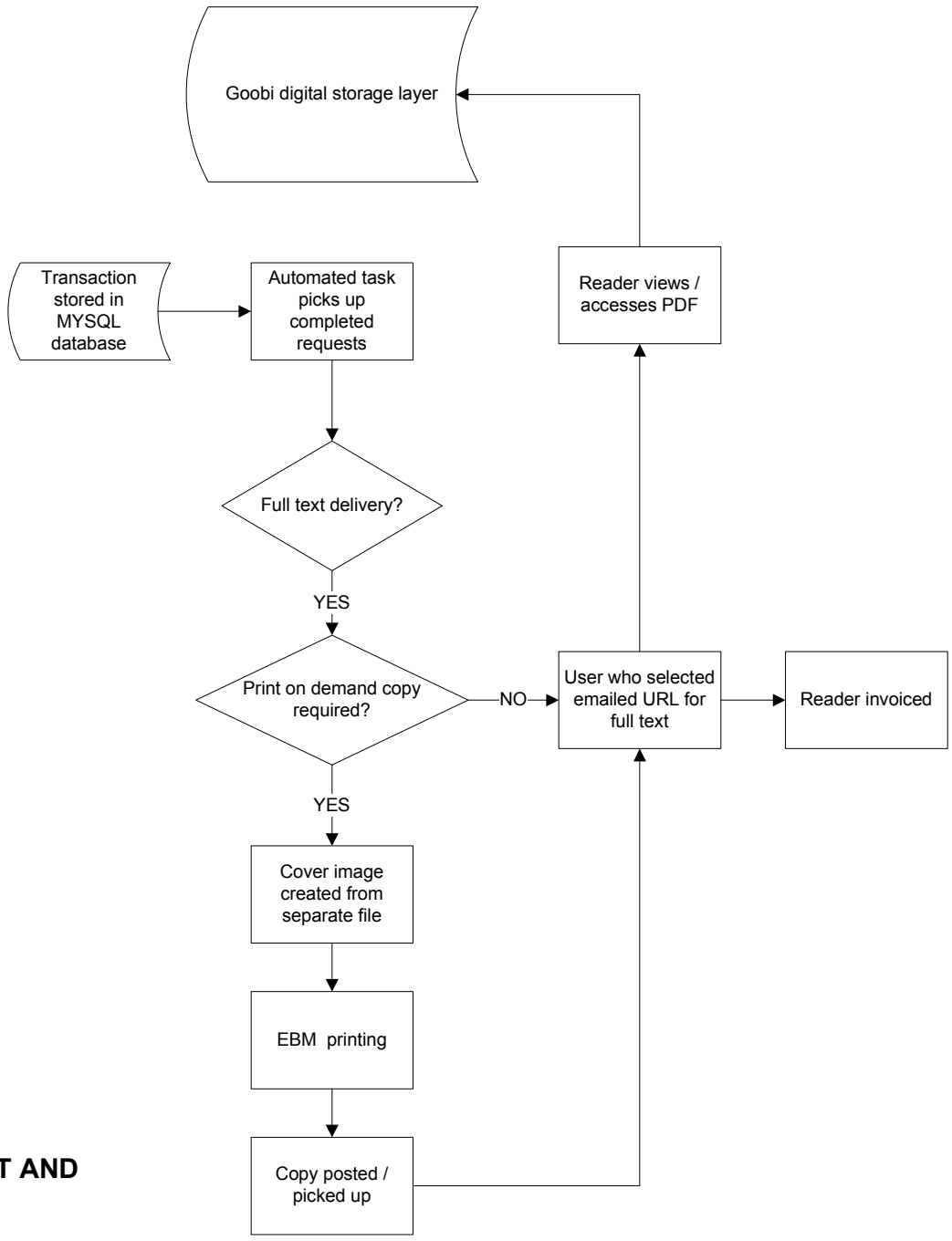
Wouldn't read the whole book, but quite useful to get an idea of contents or to look at a particular extract.

Appendix #7 – Theoretical workflows for CUL digitisation-on-demand service





STAGE 2 - STAFF SELECTION AND DIGITISATION



STAGE 3 – PRINT AND DELIVERY

Appendix #8 - Comparative costs for print and digitisation-on-demand

No comparative service exists within the UK academic library sector. The national Archive operates a photocopy and digitisation on demand service for archival material. Library services in the US, Canada and Australia are examined with information taken from publically available websites.

University of Utah – U.S.A.
The University of Utah Marriott library publishes prices on the library website. They aim to recover all costs from charges. Book prices vary and are listed with individual titles priced on a per page basis usually around $0.05 (US) a page.

For self-published items they charge an initial set up / alternation fee $25 per title with Per page charges (proof copy and all final copies) $0.05 (US) per page for printing.

They provide also offer a 5% discount for faculty, staff and students.

Cost to print a 400 page work:

- $20 or £12.5 GBP
- $45 or £28.2 GBP for a self published work

McGill Libraries – Canada
McGill Library is operating on a cost-recovery basis and provides both Digitisation and Print-on-demand . Current prices* are: (Canadian dollars). They charge a flat rate per request.

- Digitisation on Demand using Kirtas bookscanner (downloadable PDF) $10.00 or £6.16 GBP
- Print-on-Demand using Espresso book machine (print reproduction) $29.00 or £17.8 GBP

National Library Of Australia
The National Library of Australia offers a 'Copies Direct' service for full or partial requests dependent on the copyright status of a work. It charges a flat rate of $A13.20 per 50 consecutive pages for a Photocopy. It can also provide a high quality Photograph/scan at $A35-$A45 per image.
Cost to print a 400 page work:

- $A52.8 or £32.4 GBP

National Archives - UK
The UK National Archives run a Documents Online service. As they are not funded by the UK for the provision of downloads, they charge a nominal fee averaging around £3.50 for most wills and documents.

Cambridge University Library - UK
Cambridge University Library has several current methods for digitisation available. They use imaging mechanisms designed for rare books and standard digital scanner / photocopiers to provide a range of options.
Cost to print a 400 page work:

- Photocopy / scan (post 1900 material): £265.00
- Overhead color scan (pre 1900 material 300dpi): £1,298.50
- Studio image (pre 1900 material - 600 dpi): £4,107.50

Bibliography

2,559,807 itemsWelcome to Ebook and Texts Archive, n.d. http://www.archive.org/details/texts.

Anderson, Chris. “Wired 12.10: The Long Tail,” n.d. http://www.wired.com/wired/archive/12.10/tail.html.

“Archives Database, How does it work,” n.d. https://stadsarchief.amsterdam.nl/english/archives_database/how_does_it_work/index.en.html#1RHf.

“Bain & Company: "Publishing in the digital era" < Bain briefs < Publications,” n.d. http://www.bain.com/bainweb/publications/brief_detail.asp?id=28123&menu_url=briefs.asp.

BARBER, JOHN. “Meet publishers' enemy No. 1; Sci-fi novelist Cory Doctorow is shaking up the traditional book-selling model.” *The Globe and Mail*, November 14, 2009, sec. Weekend Review.

Bentley et. al. “Driving UK Research – Is copyright a help or a hindrance?.” Text, n.d. http://www.bl.uk/news/2010/pressrelease20100722.html.

“Biodiversity Heritage Library - Licensing and Copyright,” n.d. http://biodivlib.wikispaces.com/Licensing+and+Copyright.

“Booksellers at risk as digital growth accelerates, survey says | theBookseller.com,” n.d. http://www.thebookseller.com/news/136289-booksellers-at-risk-as-digital-growth-accelerates-survey-says.html.

“Copies Direct - Help | National Library of Australia,” n.d. http://www.nla.gov.au/copiesdirect/help/index.html.

“Copyright Infringement | Damages claims | UK intellectual property lawyers law firm,” n.d. http://www.gillhams.com/articles/373.cfm.

Dean, Jonathan. “Books live forever in POD future..” *Bookseller*, no. 5177 (May 6, 2005): 22-23.

Dempsey, Lorcan. “Libraries and the Long Tail.” *D-Lib Magazine* 12, no. 4 (4, 2006). http://www.dlib.org/dlib/april06/dempsey/04dempsey.html.

“DIY Book Scanning | A forum dedicated to book scanning, open source, DIY digitization.,” n.d. http://www.diybookscanner.org/.

“Ebook Library Blog » “Beguiled by Bananas” – A Statistical Analysis of Patron-selected vs.Upfront Acquistion presented at Charleston Conference 2009,” n.d. http://blog.eblib.com/?p=2373.

“Ebooks Winners & Losers | Monday Note,” n.d. http://www.mondaynote.com/2010/11/14/ebooks-winners-losers/.

“Five laws of library science - Wikipedia, the free encyclopedia,” n.d. http://en.wikipedia.org/wiki/Five_laws_of_library_science.

“Google Books adds Creative Commons license options - Creative Commons,” n.d. http://creativecommons.org/weblog/entry/16823.

Grogg, J.E., and B. Ashmore. “Google book search libraries and their digital copies.” *Journal of Library Administration* 47, no. 1 (2008): 125-140.

“Harvard Libraries Launch Scan and Deliver Service - HCL News - Harvard College Library,” n.d. http://hcl.harvard.edu/news/articles/2009/scan_and_deliver.cfm.

“HathiTrust Rights Database | www.hathitrust.org,” n.d. http://www.hathitrust.org/rights_database.

“Henry Porter | As I start to write my latest book, I fear for the future of publishing | Comment is free | The Observer,” February 7, 2010. http://www.guardian.co.uk/commentisfree/2010/feb/07/henry-porter-publishing-ebooks.

“In Google Book Settlement, Business Trumps Ideals - PCWorld Business Center,” October 30, 2008.

http://www.pcworld.com/businesscenter/article/153085/in_google_book_settlement_business_trumps_ideals.html.
"JISC MOSAIC," n.d. http://www.sero.co.uk/jisc-mosaic.html.
"JISC national e-books observatory project » JISC national e-books observatory project: Key findings and recommendations," n.d. http://www.jiscebooksproject.org/reports/finalreport.
"JISC OpenBibliography: CUL data release | Open Biblio (graphic) Projects," n.d. http://openbiblio.net/2010/10/05/jisc-openbibliography-cul-data-release/.
Lefevere, Francois-Marie, and Marin Saric. "United States Patent: 7508978 - Detection of grooves in scanned images," March 24, 2009. http://patft.uspto.gov/netacgi/nph-Parser?Sect1=PTO1&Sect2=HITOFF&d=PALL&p=1&u=%2Fnetahtml%2FPTO%2Fsrchnum.htm&r=1&f=G&l=50&s1=7508978.PN.&OS=PN/7508978&RS=PN/7508978.
"Long Tail - Wikipedia, the free encyclopedia," n.d. http://en.wikipedia.org/wiki/Long_Tail.
"miscellaneous factZ – The online home of Rufus Pollock » Blog Archive » Size of the Public Domain II," n.d. http://rufuspollock.org/2009/07/16/size-of-the-public-domain-ii/.
"miscellaneous factZ – The online home of Rufus Pollock » Blog Archive » Size of the Public Domain III," n.d. http://rufuspollock.org/2009/11/26/size-of-the-public-domain-iii/.
"miscellaneous factZ – The online home of Rufus Pollock » Blog Archive » The Size of the Public Domain," n.d. http://rufuspollock.org/2009/06/12/the-size-of-the-public-domain/.
"Moore's law - Wikipedia, the free encyclopedia," n.d. http://en.wikipedia.org/wiki/Moore%27s_law.
"Morgan Library makeover moves out miles of books | coloradoan.com | The Coloradoan," n.d. http://www.coloradoan.com/apps/pbcs.dll/article?AID=20109230376.
"No Sweat of the Brow Copyright - Gutenberg," n.d. http://www.gutenberg.org/wiki/Gutenberg:No_Sweat_of_the_Brow_Copyright.
"Open Content Alliance (OCA) » Contributors," n.d. http://www.opencontentalliance.org/contributors/.
"Open Knowledge Foundation Blog » Blog Archive » Public Domain Calculators at Europeana," n.d. http://blog.okfn.org/2010/05/12/public-domain-calculators-at-europeana/.
Oppenheim, Charles. "Legal issues for information professionals VI: copyright issues in digitisation and the hybrid library.." *Information Services & Use* 20, no. 4 (2000): 203.
Padfield, Timothy. *Copyright for archivists and users of archives / Tim Padfield.* 2nd ed. London: Facet, 2004.
Pedley, Paul. *Essential law for information professionals / Paul Pedley.* 2nd ed. London: Facet, 2006.
Pollock, R., and P. Stepan. "The Size of the EU Public Domain" (2010).
"PublicDomainCalculators/About - Open Knowledge Foundation Wiki," n.d. http://wiki.okfn.org/PublicDomainCalculators/About.
Reddy, P., J. Fan, J. Rowson, S. Rosenberg, and A. Bolwell. "A web service for long tail book publishing." In *International Conference on Information and Knowledge Management, Proceedings*, 45-48, 2008. http://www.scopus.com/inward/record.url?eid=2-s2.0-70349237449&partnerID=40&md5=b6ae5a01ac9c28e1b0b8ab3f1a5b0b7a.
Reddy, Raj, and Gloriana St Clair. "The Million Book Digital Library Project." *I*, February 12, 2001. http://www.rr.cs.cmu.edu/mbdl.htm.
"Rights Management | www.hathitrust.org," n.d. http://www.hathitrust.org/rights_management.

Schloman, Barbara F. “Creative Commons: An Opportunity to Extend the Public Domain.” *Online Journal of Issues in Nursing* (October 13, 2003). http://www.nursingworld.org/ojin/infocol/info_12.htm.
Schmidt, J., and L. O'Neill. “The 'DOD' and 'POD' project in context at McGill: Part of digitizing collections to preserve content, provide access and enrich research.” *Serials* 22, no. 3 (2009): 224-229.
“Stephen King's "Full Dark, No Stars" - WSJ.com,” n.d. http://online.wsj.com/article/SB10001424052702304173704575578241730802982.html.
“The Internet Archive Keeps Book-Scanning Free,” n.d. http://www.wired.com/entertainment/theweb/multimedia/2008/03/gallery_internet_archive.
“The Library of Congress Revives Public Domain Works via CreateSpace Print on-Demand and Amazon Europe Print on-Demand - MarketWatch,” n.d. http://www.marketwatch.com/story/the-library-of-congress-revives-public-domain-works-via-createspace-print-on-demand-and-amazon-europe-print-on-demand-2010-10-05?reflink=MW_news_stmp.
“Tower Project - Introduction,” n.d. http://www.lib.cam.ac.uk/deptserv/towerproject/.
“Update on Google Books Statement | IFLA,” n.d. http://www.ifla.org/en/news/update-on-google-books-statement.
“Welcome to the Shared Digital Future | www.hathitrust.org,” n.d. http://www.hathitrust.org/about.

Video:

- **Copyright calculators** - http://vimeo.com/15678944
- **Espresso Book Machine** - http://www.youtube.com/watch?v=OIq0VqF0MnA
- **Kirtas 2400 scanner** - http://www.youtube.com/watch?v=l2cP14mEQKI