Research to Support the British Library's Work on "Emerging Formats"

A Report by CIBER Research

John Akeroyd
Nick Canty
Anthony Watkinson

March 2017

CIBER-research.eu.

Index

1. Background

The British Library (BL) is charged through the Legal Deposit Act with collecting the cultural and intellectual outputs of the UK for posterity and an amendment to that Act of 6 April 2013, the Legal Deposit Libraries (Non-Print Works) Regulations 2013[1] came into force, extending legal deposit to include electronic publications, whether offline or online. The BL therefore has a brief to be aware of all digital publications and any perceived trends or changes. They have recently identified what have been termed "emerging formats", where failure to collect and preserve content could result in a significant loss of the UK's published record in that there is no print equivalent in whole or in part. The following in particular, have been identified:

- "Book Apps"
- Interactive narratives;
- Structured data;

Through a Call in January 2017, the BL sought a consultant or consultants to undertake a review of such formats with a view to achieving a better understanding of what they are and any trends in place, and thus to present the perceived challenges to the Library in terms of capture, access, storage and preservation. The call titled "Research to Support the British Library's Work on "Emerging Formats" was issued and the tender was won by CIBER Research, a research and consultancy organisation working in the publishing, information and library fields. The outcome was to be a report with recommendations where appropriate and a subsequent presentation to the various national deposit libraries. This is that report and the summary and recommendations detail what we have found.

2. Methodology

The research has followed a conventional pattern of qualitative and quantitative investigations (given the short duration of the work this was the only feasible option). From a qualitative perspective, we have interviewed or been in contact with at least 30 people (listed in Appendix I), ranging from industry experts, publishing consultants, publishers, agencies, libraries and national libraries. These together have provided a substantial body of evidence which has been distilled into the report and quoted directly where appropriate.

There was also much desk research, building on the literature search undertaken for the initial bid and the results of this are also included in the report in Section 11. However much of this will be known to the library already, especially the work on preservation and other background research. We have not found many recent articles on the topic in question.

And as to quantitative data, this was especially problematic to identify. Publishers, aggregators, and bookstores alike were all loathe to reveal any detail as to sales or even the amount they have for sale and only some very high level figures could be determined and even those are hard to verify. We had more success with the libraries but that represents only a fractional view of the whole picture. Narrowing data down to the UK was also problematic, especially in that many of the companies

[1] http://www.legislation.gov.uk/uksi/2013/777/contents/made

involved are essentially global and do not especially concern themselves with the state of origin. However, we have tried to detail some figures in Section 5.

3. Trends in digital publishing

Below we have provided a review of current trends in publishing and digital publishing, focussing on the formats identified but not exclusively so. We have organised and presented this in terms of the traditional publishing sectors**.**

3.1 Trade Publishing and Apps

For many years, it seemed that trade publishing had dropped behind other sectors in the extent to which they had taken advantages of the opportunities created by the web. However, both major publishers and the small innovative players, not to mention self-publishers, have made up for lost time - Apps are one consequence.

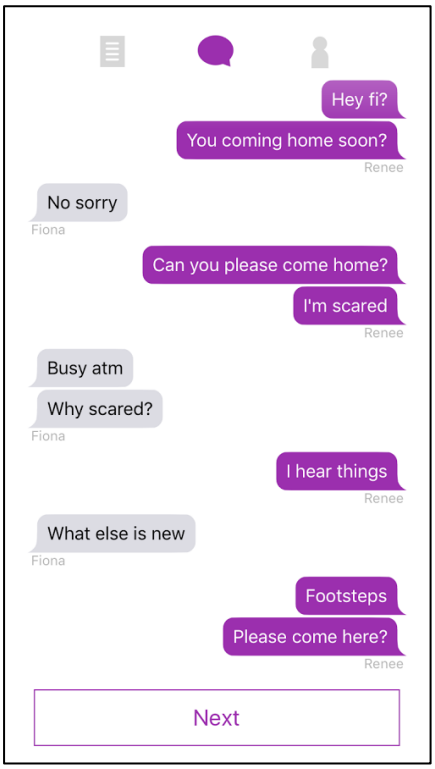
There are contradictory views on the future of Apps: some commentators see them as a transitory technology; for example, 5G will enable the streaming of cloud based content and with the introduction of mobile phones with full 64bit architecture (Google Pixel has this and in future iPhones will), the storage power of these devices will remove the need to have Apps per se, as the phone becomes a processing platform. However, it should also be noted that Apps on mobile phones for Twitter and Facebook already pull content from the cloud so their demise may not be that imminent.

On a more general note it has also been suggested that attention in trade publishing has switched in recent years from digital development to audience development partly because the book, in print and digital form, has survived and not collapsed as some commentators predicted.

For publishers, 90% of their revenue may come from the 'traditional' format books, ebooks, audio books etc but they see Apps, games, websites, video and transmedia as the most difficult 10% of their business to manage. It is this part of the pyramid they are attempting to resource. It is the result of digital convergence whereby the differences between different media are fast disappearing. Leading edge developments include ambient literature, which is based around situated experiences and where stories evolve depending on where you walked. There is an AHRC project looking at this at the University of the West of England under Professor Tom Abba[2].

A further new development is chat stories. These are short tales in the form of chat conversations as experienced on messaging services. These are accessible through Apps in App stores and signing up through websites and work on a freemium model with greater functionality and options for the user depending on payment levels. Stories are told by characters messaging each other. It remains unclear whether this is a game or a book. An example is below.

[2] http://www.dcrc.org.uk/people/tom-abba/



This format is similar to the use of Twitter for story-telling; for example, Such Tweet Sorrow[3], based on Romeo and Juliette, was such a story told by six professional actors over five weeks on Twitter 2010. The actors improvised around a prepared story grid and could interact with each other in their own words and react to followers, fans, real events and comments via Twitter. The play was a cooperation between the Royal Shakespeare Company and the Mudlark Production Company. Other media notably Facebook, YouTube and Xbox Live were used to support Twitter ensuring the story rolled out across the Internet. Tweet Sorrow won a Royal Television Society Award for Digital Innovation. While these may be marginal publishing activities, the challenge will come when a major author decides to create a story on one of these formats and their work becomes worthy of serious review, commentary, entry in prizes and so on.

3.2 Academic Publishing

There has been little change so far in the format of academic e-book; it is the way they are sold - in bundles - that has changed. It is probable that ePub3 may enable the easier embedding of multimedia especially in the humanities but it is a slow process. One of the largest international publishers publishing in most disciplines, with a big investment in the UK has envisaged a strategy under the terms of which book publishing is divided between continuing print books, the print surrogates (that the BL is used to ingesting) and books that contain material that is not replicable in print[4]. It is not known whether this policy has been implemented. All academic book publishers are suffering from declining sales and uncertainty about how to move forward. The context of this thinking is cuts in book publishing staff and in output.

[3] http://wearemudlark.com/projects/such-tweet-sorrow/
[4] http://blog.alpsp.org/2015/02/jon-walmsley-on-changing-face-of.html

Straightforward print humanities monograph publishing continues to be the norm in the dominant commercial publishers and the big university presses, and output continues to grow. There are several smaller university presses which are beginning to emulate their American counterparts and are starting to do something different – see for example the BOOC[5] experiment of UCL press.

In the journal world, over the past five years there has been talk and experimentation usually labelled "the article of the future" which has not necessarily been associated with open access. Indeed, open access publishers have, overall, stuck to a traditional format – partly to appear respectable. Some companies – usually commercial – are beginning to look to new formats.

Regarding interactive narratives, federated closed communities, which are closed working groups administrated on mobile phones and which allow for increased collaboration across individuals and groups, provide an example of how this interactive and collaborative working can extend narratives into new areas beyond story-telling. An example is researchers working on a problem across academic disciplines who write an interactive narrative based around their collaboration. The owner of the network, such as a university, would want to hold the IP. This cannot be done easily today but this collaboration would see an increase in narratives across many different spheres

The products of educational publishing are changing very rapidly and within a very few years the output has changed from a heavy and continuing reliance on print with a whole range of add-ons, which may be digital, to a range of different types of digital offerings some of which are taking full advantage of the affordances of the web. Much of this content no longer has any print equivalent. However these formats are not part of our remit and we have not explicitly explored them in this report but we suggest that the BL investigates this sector thoroughly.

3.3 Structured Databases

The trends in database publishing have been about aggregation - of multiple databases into a single entity and of single global suppliers of multinational services. Lexis Nexis would be an example of the first and Proquest the second. But these database services themselves are under pressure, to add greater value and are adopting innovative ideas such as user generated content and ancillary software services which ensure their data is plugged into the user workflow. The big purveyors we interviewed, seem very confident but is the future looking good despite current reasonable margins? For example, looking at the dynamic end of the spectrum will companies like Euromonitor survive? It is not a simple question: with AI and machine learning the software could work more closely with the data and data analysts would not be necessary. This would have been dismissed a few years ago but now - it's a realistic question. The real value is in the IP rights in the interface between the data and the user, it is no longer just the content.

3.4 Research and Scientific Data

Research data has become of increasing concern over recent years, to an extent the consequence of research funders demanding its availability. We understand that it is without the BL's scope for Legal Deposit but nevertheless is an issue which the BL

[5] https://www.ucl.ac.uk/ucl-press/ucl-press-news/call-for-content-booc

may need to address especially as some is now being published as supplemental data by mainstream publishers. There has also been a great deal of interest in the UK in how research data should be preserved. Some of the data discussed is scientific, and key reports have emphasised "big science" and only mentioned where datasets should be archived. The recent emphasis has been on open data and, to some extent, how such data should be archived and consideration of how supplementary data is related to an article, has been pushed into the background. A new group of the Research Data Alliance[6] has both these questions on their agenda but the information available so far does not seem to be explicit on these points.

Some scientific databases are composed of ingested datasets without any quality control and some act as filters with some element of peer review. The latter, if they are UK based, may come under the remit of Legal Deposit though they probably have not been collected in print.

3.5 Links

More significant is data linked to journals or data embedded in journals. As far as links to "supplementary" data is concerned, the British Library already receives such digital content with the Elsevier e-journals via Portico and it is ingested but not made available. The Portico policy is:

"If we find links in metadata accompanying content sent to us, if those links refer to content also sent to us by the publisher (supplementary files, images, etc.) - then we create a link to the component stored in our archive. However, if the link refers to something not in the content sent to us (for example, a link to a website), we simply preserve the link information. We do NOT attempt to resolve it, or to go harvest what exists at the location to which the link points. For links in non-metadata files (for example, in a PDF file), we do not do anything at all".

Mention should be made of F1000Research where the form of the article has been changed and there are various journals based on the platform of Ubiquity Press, under the heading of meta-journals, where data, software and video are central to the message of the article.

What F1000Reseach[7] are doing highlights the problem of how to handle "links" which, from the earliest days, have been what researchers have found most important in the format of e-journals. Links can be to data and other content available via DOIs or URLs which are held externally but there is a growing move to linking to content which is an aspect of the article and which are essential and intrinsic. The content (which can be characterised as data) is sometimes held on the F1000Research site but may also be sent to Figshare[8]. The policy is not clear. The importance of F1000Research billed by them as more of a platform than a journal, but recognised by researchers as a journal, is that it now has alliances with prominent Open Access funders such as Wellcome.

(*On the specific question of where the links reside, Acreman of F1000 tells us "The British Library get the same packages as PMC so they basically get a copy of all the data/supplementary files within the article package we send them for archiving.*

[6] https://www.rd-alliance.org/ig-data-policy-standardisation-and-implementation-rda-9th-plenary-meeting
[7] https://f1000research.com
[8] https://figshare.com/

We distinguish between source data underlying figures and tables and truly "supplementary" files (i.e. not really needed – our guidelines say: Additional information that is not absolutely required in order to follow the study design and analysis of the results, e.g. questionnaires, extra or supporting images or tables, can be submitted as supplementary material). Supplementary files are dealt with like figures and tables; they don't get a DOI. In a proposal we wrote recently:

"Extra resources that are necessary for the proper presentation of the paper and provided by the authors, as well as embedded materials like figures, images, video, or audio are stored locally on the F1000Research's cloud system alongside the paper itself (and backed up regularly).

For source data, which are essential and mandatory, we normally ask authors to deposit them in an open and structured repository – as outlined in the guidelines[9]. The type of repository depends on the type of datasets (there is a table in the guidelines with different examples). The repository provides a DOI. In addition, on F1000Res only, we can host small datasets ourselves and coin a DOI for them.

Whether hosted in a repository or by F1000Res, the dataset is described in the "Data Availability" section within the article, with details of how to access it and the DOI").

Unfortunately, the status of "supplementary data" is a matter of the policies of individual journals. Are they part of the article or additional to it? The new group mentioned above will be attempting to standardise policies across the publishing community. It could be argued that it is more important to archive and preserve data that may be called "supplementary" but is part of the article.

How to handle text and other data in the humanities and social sciences is, as yet not so much discussed though one assumes that in humanities journals, there must be similar supplementary data and in particular in social sciences simulations (which need further attention). In the past there were experiments where "resources" were digitised and linked to a monograph – fortunately most, if not all of them, were in the USA[10]. There may be a revival

4. Definitions

4.1 Book Apps

We defined an App in our tender as "downloadable programmes which can be 'played' on a suitable device such as a mobile phone or any device compatible with the operating system for which the App. has been written. Most commonly these are iOS - which play on Apple systems, and Android, the underlying Operating System for many mobile phones and tablets". Karch, (2016) writes: "An App is a piece of software. It can run on the internet, on your computer, or on your phone or any other electronic device. The word "App" is a modern usage, but this is really the same thing as a software program".

Thus, an App is something which is downloadable and executable and is only limited by the parameters of the software, the operating system, the device

[9] https://f1000research.com/for-authors/data-guidelines

[10] http://ciber-research.eu/download/Watkinson-Electronic_Solutions_to_the_Problems_of_Mongraph_Publishing.pdf - see section C5

which runs those and the imagination and ability of the App developers. Apps are designed to, and probably do, run best on the specific device for which they are made but through the provision of reading platforms (which may be the App) it is possible to emulate Apps on competing platforms – thus Apps for iOS can be read in Android and vice versa.

Other characteristics of Apps are:

- They are distributed through App stores (related to specific devices), through independent App stores and direct from publishers;
- Apps are developed in varied codes such as C, C++, Java, html5, and proprietary code e.g Objective- C for iOS.
- Apps give more opportunities for interaction and intelligent functions in comparison to say Epub3, though the latter has more universal readers;
- App readers are available for most platforms (iOS, Linux, Windows,)
- Apps tend to be used for children 'books' (heavily illustrated an animated), how to books and travel books which use geo-location data;
- There is significant overlap with games – indeed Google categorises such works as games and not ebooks (and in their ebook store only accept epub3 and PDF files).
- In commercial terms, there are essentially three business models for Apps. i) Premium – paid for, pay for it once and own outright, ii) Freemium such as Candy Crush and supported by micro transactions and iii) Free, forever supported by advertising (eg the Pointless App by Blink Publishing, which contains 100% original content).

When comparing Google and Apple's Apps (Android versus iOS, 2017) "most popular apps are available for both platforms. But for tablets, there are more Apps designed specifically for the iPad while Android tablet Apps are often scaled up versions of Android smartphone apps. Developers at start-ups often focus on one platform (usually iOS) when they first launch their smartphone app because they do not have resources to serve multiple platforms from the get go. For example, the popular Instagram app started with iOS and their Android App came much later".

4.2 Interactive narratives

Interactive narratives could be defined simply as books which enable the reader to interact with the "story" as to goes along so that the narrative can be tailored or customised to suit the wishes of the reader. Like books, there are characters, a plot, incidents and a narrative which gives them literary value. They are usually characterised as being multimedia and providing an immersive experience. Interactive narratives can be in any format and as such, are a particular genre of publication rather than an emerging format. Thus, they may be Apps or enhanced ebooks e.g. Epub3 or indeed websites. The challenges are the same challenges associated with those formats rather than the genre itself.

We have also discounted authoring systems as being out of scope for this work and more generally we do not think that legal deposit covers software per se and authoring systems are just that. They are not required at run time and thus are neutral as to format and readers. Thus, authoring systems are available on multiple platforms for development. There might be problems

with making arrangements with authors who use this software (for self-publishing) if necessary but apart from the likely instability of the software, it is difficult to judge what issues will present themselves for preservation purposes.

4.3 Structured Data

Databases, datasets, structured data are all near synonyms for "one or more large structured sets of persistent data, usually associated with software to update and query the data. A simple database might be a single file containing many records, each of which contains the same set of fields where each field is a certain fixed width" (Database, 2017). The concern in the context of legal deposit is probably more that of function or content than the structure per se. Thus, one National Library advised us that they determine whether a database is collected based on it having 'literary, editorial or intellectual input i.e. that it is not just a set of computer generated data but has some degree of value added through human intervention".

The ISSN standard (ISO 3297) as interpreted in the ISSN Manual (ISSN 2016) defines databases as "Ongoing integrating resources: Ongoing integrating resources are resources that are updated over time and with no predetermined conclusion, for which the updates are integrated into the resources and do not remain discrete" and has inclusion criteria:

- "There is editorial content (i.e., the resource mostly consists of written, textual content, and there is evidence of editorial or journalistic treatment);
- There is identified editorial responsibility (i.e., a statement indicating the name of the publisher / producer, and at least the country of publication). Generally, editorial responsibility will consist of more than one individual;
- There is a consistent title (i.e., a title which remains consistent when the resource is updated) and the title is prominently visible on the resource";

Our research suggests that at the outset the **BL will need to arrive at an agreed definition for databases** which is viable and sensible and can be put into practice.

In the Outsell report of 2006[11] the authors made the distinction between Composite e-publications and Enquiry-driven e-publications. Whilst that still holds true it is the case that more, if not most, new databases are effectively internet or web resources interrogated remotely. The distinctions which might now be made are more the following:

- Those which are effectively static and are not subject to further change or amendment; these are likely to be historic data perhaps derived from digitalisation processes. It is possible that the library already possesses such data in print format. In this case preservation decisions (print or digital preferred) may be a matter of collection policies. Examples could be some of the ProQuest owned Chadwyck-

[11] Refining the map of the universe of electronic publications potentially eligible for legal deposit. Electronic Publishing Services Ltd, an Outsell, Inc. company. 28 November 2006

Healey collections, though they do seem to have occasional updates or "editions": these can be associated with new front-end software and in any case, additional material – new relevant content – may fall into the category of already available in print. There is a growing trend in collections being offered which contain non-text material which may not have been ingested. There are at least two other publishers in the UK who have imprints with similar publications – Bloomsbury and Sage.

- Those which are updated but on a fixed regular basis so that there are defined versions; e.g. magazines which have become website databases will fall into that category;
- Those which are updated constantly so that the only determinant of the state of the database is the time at which it is interrogated; a further complication of the latter is the increasing trend for users to interact and add to that data (user generated content).

The examples provided in the Call are partly at one end of the spectrum and may or may not be within the scope of the Act. There are other examples in other disciplines which we would have reviewed were there more time. A good example is the various databases published by CABI. We are told:

"Our databases contain much more content than the printed journals, and some of our digital subsets have no print equivalent these days at all. The databases also offer more functionality, of course, but you cannot capture that on the printed page (e.g. links to full-text, visualisation of search results). What I think is probably useful is to make a distinction between the raw data and the discovery functionality, so that you can focus on deposit of the content rather than of any associated software. If you get an XML feed from the publisher, it can be registered and stored electronically, and can also be used by anyone in the future, so is quite a future-proof system"

Alas there are many of the databases in classical biology which are often produced by groups of scholars without publishing experience. The documentation relating to these databases – many of them from Natural History Museum – are worth considering further. One wonders how much is easy to retrieve.

5 Size of Publishing Outputs

5.1 Apps

It is not possible to disentangle Apps as books from all other Apps and it is not easy to get definitive figures on Apps. The App stores are inevitably secretive as to their offering.

- As reported in 2017 Android sells Apps from Google Play, which currently has over 2.5 million Apps available[12], most of which will run on tablets. However, some Android devices, such as the Kindle Fire, use separate App stores that have a smaller selection of Apps available. The Apple app store offers 700,000 Apps, 250,000 of which are available for the iPad. Most developers prefer to develop games for iOS before they develop for Android. In the end, most Apps are available on both.

[12] https://web.archive.org/web/20170210051327/https://www.appbrain.com/stats/number-of-android-apps

- As to ebooks in general, the Google Play store now has over 5 million eBook titles, according to Google, up from the four million the company reported via its support site.[13] .

- Blackberry has several thousand ebook Apps in Blackberry OS;

- Windows, according to Microsoft, as of September 28, 2015, has over 669,000 apps available[14] on the Windows Store, which includes Apps for Windows NT, Windows Phone, and Universal apps, which work on both platforms. Games, Entertainment, Books and Reference, and Education are the largest categories by number of Apps and most App developers have just one App.

- Amazon.co.uk delivers about 200,000 Android Apps of which 14,000 are ebooks or ebook readers and 145 are children's apps.

- The educational app store[15] (UK) has over 3000 Apps most of which are ostensibly ebooks but not labelled as such and many of which may be of UK origin.

(None of these figures explicitly distinguishes UK content; a random check suggests that not that many are British).

We are also not able to separate the different amount of "book App" content on say the IStore, available through Android devices and other ways. It is our impression from interaction with publishers that Apple is still the preferred platform. It is much more difficult to build an Apple App; you have to make a case for quality, and you have to follow instructions with great care. It is more expensive but it is more prestigious and there is better marketing. From a preservation point of view each App using this route has common standards which are well documented.

Our informants from within the publishing industry gave us some information about their own experience. A major UK publisher informed us that their overall output of Apps is increasing but only as they acquire new companies that produce Apps. From 2009-2016 they produced 460 Apps but from 2017 they will produce 20/30 apps a year including from their gaming studios and these are profitable.

Children's publishing has long been distinguished as one area of trade publishing where the App format has become established. One well known player in this sector has told us that although their output of books has increased to almost 100 titles a year they are still only releasing one or two a year. Apps are time consuming to build. The margins are probably poor. Some are licensed brand names such as Peppa Pig or Pinocchio. But they are not marketing tools. They do not usually represent another format of a print original; they are new stories and stand on their own – hence they are well worth considering ingesting for serious contributions to the national collection

[13] https://techcrunch.com/2013/03/06/google-play-offers-over-5m-ebooks-and-more-than-18m-songs-one-year-after-its-rebranding/
[14] http://www.techarena.co.ke/2015/09/28/microsoft-we-have-669000-apps-for-phones-tablets-and-pcs-in-the-windows-store/
[15] http://www.educationalappstore.com/

In summary, we tend to the view that the total number of UK Apps which are in effect books or are book like, is **probably no more than thousands** and certainly less than tens of thousands. And it seems true that whilst some Apps have sold well, there are many where sales are very few e.g. less than a hundred. We are also aware that Google routinely cleans out its store so that it is entirely possible that some will be lost at an arbitrary date.

Interactive narratives are a sub-genre of Apps and websites and are likely to number less than a thousand, and for the purposes of this report could be counted along with Apps (or are already captured in the web archive). To give some context to this, Tell Tale USA, a producer of interactive "TV stories" have sold 5-10m+ licenses worldwide and Choice of Games another US company produces 4-5 games a month. Inkle Studio's 80 Days narrative has sold 500k licenses. Approximately 20% of their sales came from the UK making them one of the most important companies in this field.

5.2 Structured Data

In Section 4 we attempted to defined databases and a subset which is clearly relevant, is bibliographic databases and these should be easy to identify. However, they are also conceivably in decline though several factors:

- Data being merged to form greater aggregations given the decreasing costs of storage and ease of search;
- Search engines such as Google Scholar undermine the need for such services;
- Library webscale discovery systems such as Summon, Primo or EDS being used as an alternative by students et al rather than searching the native database, resulting in cancellations of those (this might not be true of all disciplines but is a factor);
- Libraries themselves closing or at least working more collaboratively.

Overall usage of this type of database is hard to determine though research shows that they are no longer the first port of call for many searchers who default to Google or Google scholar.

To establish the size of the database population, we sought numbers from several sources but, as with Apps, this is not an easy task.

Firstly, we approached the ISSN register and they provided details of the number of databases worldwide and subsequently a listing for the UK. The worldwide total was 1475 and the UK total i.e. databases which had been registered in the ISSN Register as originating in the UK and catalogued into the UK national record, was just 34. This is probably as indicative of low cataloguing/ISSN assignment rates as it is an indicator of numbers in existence.

Secondly we approached ProQuest as one of the major global database vendors. They would not tell us the overall total of databases they sell but did indicate that they supply hundreds and that circa 20 were of clear UK origin.

Thirdly we undertook a search of some major UK university library sites and note that the largest had several hundred databases listed, although again UK provenance could not be established; others were open access services which should already be

being collected by the library. Finally, we contacted the UCL Library as one of the major collections in the UK and they reported "we currently have 476 paid for 'database-like' resources (either subscription or hosting fee) however this number does include several ejournals collections and some ebooks". A further breakdown shows from a list of paid for resources "356 are identified as databases".

In summary, our research suggests that given the definitions which we have put forward above, then there are probably **no more than a few hundred databases** existing in the UK which warrant consideration for deposit and some or maybe many are multinational so that UK origin may be hard to determine. And we have also been told that data is rarely deleted, at least from the library type databases, in that it has continuing value. Indeed many databases are amalgamations of historic data.

6 Current Actors

6.1 National Libraries

We agreed with BL, that one strand of research was to understand the policies and activities of other national libraries and to that end contacted major National Libraries including Germany, France, the Library of Congress, Canada, and Australia. In addition, we contacted many smaller countries such as Finland, Sweden, Croatia and Singapore. We also met with the Dutch National Library (KB) collection and preservation team. Everybody we contacted was extremely helpful and open in describing their position, providing public polices where they existed and their strategies on the 'emerging formats' we presented.

The outcomes of the trawl can be covered in two points: firstly, many of the libraries contacted had in place existing digital Legal Deposit regulations which, whether directly or indirectly, would address the formats we raised. In some cases - Finland for example[16]- their collection policy was a detailed and exhaustive list of all the different media covered by Legal Deposit which appears on its website including exclusions. More generally National Libraries had generic digital policies some relatively recent and which were so worded as to potentially include Apps or databases. Thus, the Canadian National Library reported "Canada's legal deposit legislation applies to electronic publications and specifies that they must be accompanied by any software or technical information needed to access them, so the legal framework is in place". Two libraries (Portugal and the Netherlands) had no legal deposit at all preferring or being obliged to live with a voluntary scheme. The generic policies often require a degree of interpretation and it is often left to the Library or Library Officers which content to collect or which to collect at what interval.

Web harvesting was also common with policies which attempt to identify state developed or delivered content and for this to be collected at intervals.
We also noted that the range of what was collected varied considerably with some collecting all media types including films, video and indeed Games whilst other were more focused, as with the BL, on textual material.

[16] https://www.kansalliskirjasto.fi/en/legal-deposit-office

Secondly in terms of the explicit formats we were addressing, the most common response was that libraries were no further ahead in their thinking or practice than the BL itself. Thus, the Bibliotechque Nationale de France (BNF) stated their position and interest in the project thus:

"Apps and databases are all part of our legal deposit mandate, but apps are not designated as such in the law, and things are not clear regarding databases online. There are continuing efforts on the library's end to change legislation to make the legal deposit of born-digital materials more clear and efficient…

- Several entities at the BnF are involved in devising solutions to collect, preserve and disseminated emerging formats. There have been no formalized discussions regarding the ideal distribution of collection management responsibilities at the library level, but we are witnessing a slow realization that there should be.

- We do not wish to build a collection policy based on content types. It does not really factor in (sic) so far anyway: what we do collect at this time depends on the goodwill of the publishers and technological opportunities/obstacles.

-Processes have been designed by the Department of Legal Deposit to collect, preserve and disseminate ebooks, including those with multimedia or interactive content, if they are in the PDF or EPUB formats. The processes have not been tested at scale. So far there is no automated solution in place to identify the books with interactive narratives among the ebooks.

- Collecting apps seems to be the uncontested job of Multimedia Services. Contacts with Apple have been made in 2013-2014 regarding the legal deposit of apps. They have directed us to contact each individual publisher. We have made a few tests with iTunes and Android downloads. We would like to work on a more scalable and efficient solution.

The BnF has a partnership agreement with the BL, and might be able to work together on this topic in that collaboration context".

And the Library of Congress, which had surprisingly given the topic little thought, reported "we are looking to add eBooks and digital audio to that scope in the near future. I think the Library of Congress is somewhat behind the other major national libraries around the world in terms of digital acquisition and what can be acquired through copyright or legal deposit means, and that seems to be borne out by your relationship with the British Library right now, and the exploratory efforts that you're involved in"(Stephen Want; Chief, Copyright Acquisitions Division; United States Copyright Office).

Sweden by contrast with all others, has a policy on these emerging formats and it is, by and large, to reject them. Thus their categories of "Not subject to legal deposit" include "entire databases; coding for these databases" and "Apps" but these are "subject to submission if the content differs from that which is available publicly or in analog form".

In summary, much is line with the BL position: that of having a legal Deposit act in place but which is still being worked on/developed as the library tries to come to terms, with multiple and emerging formats alongside an explosion in

publishing. If there were differences it is that other national libraries have a wider brief than the BL, some including films, multimedia, and Games (though these tend to be more handheld or portable media rather than downloads).

We are inclined to the view that BL has not a great deal to learn from its global partners except perhaps the Scandinavian national libraries (and possibly others such as Japan or Korea - whom we did not approach) who seemed to have given Apps at least more thought (though the challenges faced in these small countries may be proportionally less). However, the research also suggests that there is room for more collaboration over standards, know-how and even preservation.

6.2 Preservation agencies

6.2.1 Portico

Portico is among the largest community-supported digital archives in the world. Working with libraries and publishers, they preserve e-journals, e-books, and other electronic scholarly content to ensure researchers and students will have access to it in the future. Publishers in many sectors particularly scholarly publishers rely on Portico for archiving and preservation and Portico ship files to the British Library for a significant part of their ejournals collection. Portico already has agreements with the publishers concerned to preserve these journals in their dark archive. The BL does not use their services for e-books. Our conversation with Kate Wittenberg (CEO) and Sheila Morrissey was almost entirely about book content.

Sheila Morrissey sent us the following information in answer to some explicit questions about ePub3 files with embedded multimedia and how they handled links:

"We currently get ePub from only one publisher, who also sends us a PDF version of the same content. We simply store the ePub and note it as an additional rendition instance of the artefact. We do NOT go inside and extract embedded multimedia - there is no recursive descent through the contents of the ePub to extract components (image, sounds, movies, etc.,) and characterize them, or preserve them as "stand-alone" units.

Regarding links: If we find links in metadata accompanying content sent to us, if those links refer to content also sent to us by the publisher (supplementary files, images, etc.) - then we create a link to the component stored in our archive. However, if the link refers to something not in the content sent to us (for example, a link to a website), we simply preserve the link information. We do NOT attempt to resolve it, or to go harvest what exists at the location to which the link points. For links in non-metadata files (for example, in a PDF file), we do not do anything at all".

Scholarly editions are one digital area where Portico have been especially engaged. This is not surprising as they are part of Ithaka and a sibling of JSTOR which is primarily a service for the humanities. The publishers, even the larger university presses, tend not to be aware of what preservation of this type of content involves. This generalisation covers UK publishers who are prominent in this area. There are no clear standards. There is some commonality but not total. Part of the problem is the extra functionality which makes these digital editions distinctive. Each publisher tries to do its own

thing. They are optimistic about handling some types of non-text content such as audio.

Portico is also interested in time based media such as exists at the Museum of Modern Art in New York City. We are not clear whether this type of content is of interest to the BL at the current time. An explanation of what is means for conservation can be found at: https://www.guggenheim.org/blogs/checklist/what-is-time-based-media-a-q-and-a-with-guggenheim-conservator-joanna-phillips.You have to get hold of software before collection and it is a whole range of software from browser down.

The other well-known archive is CLOCKSS[17]. This is also much used by scholarly publishers in the UK but it is in addition to another archive such as Portico because its model does not involve preservation for the long term. The emphasis is on distributed holdings. The mission is to build a sustainable, geographically distributed dark archive with which to ensure the long-term survival of Web-based scholarly publications for the benefit of the greater global research community. When a publisher is liquidated, or has some other problem the content can be liberated from the archive.

6.3 Other agencies

ISSN International

Under this heading we should also report communication with the ISSN International Agency in Paris, specifically on the question of databases. They had no doubt that an ISSN should be assigned to a database as, under their definition a database would constitute a continuing resource. This also implies that databases would be catalogued by the relevant National Centre into the national catalogue as a Marc record which in turn is copied into the ISSN master register. ISSN provided us with some detail as to what the register contains and we have reported elsewhere the data they have provided. This suggests that either there are very few databases globally or more likely that the policy is not being widely adopted. Indeed, a parallel search of some major catalogues such as COPAC confirms that.

The National Archives

We discussed the research with the National Archives (TNA) and they were extremely helpful in providing us background information on their collection policies and related issues such as formats, metadata and usage and pointers to their emerging digital strategy.

In our discussions with both BL staff and with TNA we noted the potential at least for overlap. For example, the ONS was cited in the tender call as a publisher of interest, whilst the TNA confirmed that it was within their scope to archive and preserve ONS publications whatever the format. The reasons for this may be clear to both parties and may not be an issue in that providing dual back up has merit. From a user perspective, it could cause confusion. We suggest BL ensure that their strategies and policies are effectively shared with the TNA especially as to potential synergies in digital continuity and collections strategy.

[17] https://www.clockss.org/clockss/Home

We have attached the TNA response as Appendix II.

7 Collection and Preservation

Preservation of **Apps** on the surface would appear as less problematic than other digital publications. However, they fall into the category of continuing obsolescence whereby the technologies on which they depend are constantly being amended and upgraded. So that, whilst the underlying software systems are generally backwards compatible i.e. Apps developed for earlier version of an OS should still play on a more recent version, there is concern that over time and many versions this may not be the case. However, Apps are at least defined entities which can be ingested through transfer, or download or otherwise copied for collection. And although there is a multiplicity of formats (albeit only android and iOS seem to predominate) these might be addressed either through i) emulation software (readers) which work across different platforms so that for example a Windows reader can play Android Apps or ii) the provision of suitable hardware. There is much work already done on emulation, for example the early games collection of Internet Archive[18]. However, whilst there is some good quality emulation software in some instances e.g. the Bluestacks App Player[19] will render Android on Win 7/8/10 and there are others, rendering iOS Apps is more problematic – there are tools such as Cider APK[20] and iEMU APK[21] but these are likely to be unacceptable and not endorsed by Apple i.e. they may even be illegal.

Further input (Nick Coveney, Kings Rd Publishers) suggests "Unfortunately you won't find a decent "one size fits all" emulation solution for Apps… the only Apple endorsed emulator is its own - "Simulator"[22] which sits within their Xcode software and requires an Apple ID for verification in order to open the.Ipa files in a test environment - although I believe this does allow for historic iOS and device displays (to my knowledge this also overrides the need for provisioning profiles usually associated with apps outside an app store for UAT purposes). Although it is worth noting that there are lots of emulator options available for each platform unofficially/unendorsed, I would suggest that the official/legitimate emulators are the best option…."

Another suggestion that has been put to us is that Apps could be combined in an Epub3 wrapper which might enable easier storage alongside all other Epub3 files. But other technical experts have expressed significant doubts over the feasibility of achieving this and even were it possible, the justifications for so doing are also questionable give the possible consequences for rendering the App

Though relatively easy to identify and capture, **databases** present significant challenges in terms of preservation. Most database systems have an extract or export function (for back up if nothing else) which should allow the whole data set, to be transferred to another server or media whether by automated means or through digital media. The problems derive from the following:

[18] https://archive.org/details/softwarelibrary_msdos_games
[19] http://www.bluestacks.com/apps.html
[20] http://www.androidcrush.com/download-cider-apk-for-android/
[21] http://www.androidcrush.com/download-iemu-apk-for-android/
[22] https://developer.apple.com/library/content/documentation/IDEs/Conceptual/iOS_Simulator_Guide/GettingStartedwithiOSSimulator/GettingStartedwithiOSSimulator.html

- The files can be very, very large and require significant storage capability;
- Many databases are not static and are constantly updated so that processes are needed to take snapshots at regular if not frequent intervals which need to be determined;
- Databases may well have bespoke interrogation software which is intrinsic to the database function and will need to be captured and maintained alongside the underlying data to enable the emulation of search at the time of capture, so that there is a need to know:
    - Which search system was in use at the time of capture;
    - On what platform does that run and what is needed to run it;
- Search itself may be heavily personalised or contextualised so that the results of a search will in any event be different on different search instances;
- Databases may contain a significant number of links or indeed may be largely comprised of links which will need to be preserved.

Some datasets are now being made available as "open data" and as such are explicitly designed to permit alternative search engines. In this case, it may be viable just to capture or harvest the underlying data and not be overly concerned as to how search system are being used.

There is also the point that where databases are under constant change, then those changes in themselves need to be recovered to provide an audit trail.

Interactive narratives are essentially produced in App format because they are not meant to be reproduced. The original coding language is not important to developers as the App works with source compiled (machine) code. Human-readable resources (images text, audio etc.) are all used piecemeal and 'built on the fly' for each narrative and the interface between those resources and the ultimate presentation is locked up and unique to each company. This makes preservation challenging.

The main software producers of interactive narratives use Unity, which is an industry-standard tool for creating games. Unity gives access to and executes upon the database structure. Most game writers start in an Excel spreadsheet then export to a form the game/story can read; this data is then manipulated for the reader to use. This makes it hard to archive as there is no underlying 'spinal document'. Given these issues and the varying level of sales suggests that the BL may need to consider its collection policies for this type of content. Indeed It was suggested that using cultural metrics like sales data may help to decide what is preserved. the BL should preserved.

A further challenge for the preservation of interactive narratives is the question of who hosts the content. Random House worked with Failbetter Games to produce the *Black Crown* interactive narrative whose author Rob Sherman approached the British Library about preserving the game. *Black Crown* was hosted on the Failbetter Games platform and Random House believed the content should be hosted by the BL for preservation purposes. The issue was not resolved and the site was taken down and all the digital assets were given to the author.

8 Metadata and file formats

a) Metadata

During our investigations, the issue of metadata was raised frequently. Metadata is a key component in discovery, the supply chain and preservation and is well developed in the library and publishing communities. Thus, a common workflow will be: a book is allocated an ISBN at an early stage; an ONIX record is created by the publisher to support the supply chain; a Marc record is provided to support library catalogues and discovery and Mets data created for digital preservation. There are cross walks between different metadata schema so that for example Marc records might be created from an Onix record. Serials (or continuing resources as they are technically described) are similar, with ISSN as an identifier and standards such as ONIX-p.

To clarify the position, we opened discussions with the ISBN international office, bibliographic data companies Nielsen and BDS, the ISSN Agency and some knowledgeable individuals. Our main conclusion from the research is that in respect of all three formats there is no or very little metadata to be had.Thus, Apps are largely sold through the App stores or via publisher's sites and as such the metadata needed is limited; for example, a short description and thumbnail might be all that is in place. There are no identifiers as such, or more extensive metadata or standards. It also appears that neither BDS nor Nielsen provide App metadata or Marc records or indeed have been asked to do so. ISBN International Office recounted a flurry of interest around 5 years ago but that came to nothing and has not been resurrected. Indeed, as one expert (Macfarlane, EasyPress Technologies) commented "The problem with apps is discoverability and because they're all in a container e.g. C++, consequently, they are difficult to index. You are then reliant on the indexing the publisher provides, which may be nothing at all". Many of the Apps we encountered were categorised as Games by both stores and the publishers; there is metadata associated with Games. There is a catalogue service for example – Spong[23] - and Marc records are frequently created for Games but largely for hand held Games rather than otherwise. So while all Apps do have metadata, these are largely marketing related e.g. around a genre and some marketing wording, short descriptions and an image but very little in way of filing data to help organise the category in the App store. App developers apply search terms and key words but, it has been suggested, no-one actually finds their Games or Apps this way; it is overwhelmingly by word of mouth, from social media, friends and so on.

Databases also suffer from this lack of metadata. As we pointed out in Section 4 the ISSN has relatively few records even though many databases would qualify as continuing resources and nor do static databases command an ISBN. University libraries have significant numbers of databases in their collections but do not generally add them to their OPACs preferring just to provide an A to Z list which would appear to poorly serve the reader. It is probably the case that databases which are aggregations contain publications which already have metadata in their own right.

b) File formats

Ebook formats have been under development for some years with several proprietary formats emerging based on devices and their ecosystems – the classic example is Amazon Kindle which originally supported the AZW e-book format, identical to the Mobipocket (MOBI) format. The main competitor to these proprietary formats was PDF which, given its device neutrality, the availability of many readers and its ability to render pages as in print have ensured its continuing use and success.

[23] http://spong.com/

EPub first appeared as a format as the Open eBook Publication Structure or "OEB", which was originally developed in 1999 with a Version Epub2 in 2007. EPub did not achieve any great success until the most recent version Epub3 which is based on html5 (indeed it preceded). Epub3 has achieved much more success and has become more prevalent as an ebook format and indeed a general document format. As we have noted elsewhere, Google has now standardised on Epub3 and Pdf in its ebook offerings and the library ebook supply company, Overdrive (the largest ebook lending supplier in the world) has also stipulated that they will only now supply Epub3 and Pdf and is attempting to reduce Pdfs. Open access publishers have adopted Epub3 as have many main stream publishers such as the IoP[24]. Only the proprietary ebook suppliers have continued with their own formats (KF8 for Amazon, ibooks for Apple), though there are many Epub3 readers for ithe Kindle Fire tablet. The success of Epub3 is its ability to deal with multimedia, the availability of readers on multiple platforms and, an important point, its inherent accessibility. Epub3 was designed with accessibility in its make-up and had considerable input from the Daisy consortium, to the extent that it is now usurping Daisy as a standard format for accessible publications.

It is now clear that Epub3, has an international future even if British publishers, as distinct from other publishing countries are slow to migrate from ePub2. Other than the proprietary formats it is hard to see there are many competitors (except perhaps the Russian developed fb3[25]. EPub3 can easily embrace embedded content in different media which will make enhanced ebooks a reality ie moving from flat representations of the codex to more dynamic content. The fact that ePub (IDPF) has now been "taken over" by W3C has consequences not yet fully understood but is also indicative. And although book publishing and journal publishing have pursued different journeys, this may now begin to change and there are arguments for ePub3 being used in journal publishing and some key players are adopting it[26].

As to databases, file formats are well rehearsed and are usually one or other of the main document formats e.g. PDFs, html or XML based. Images will often be jpegs or tiffs. Proquest tell us that "The full-text file formats are either PDF or web-based as images or text. The only way to download the full-text in its original form is through PDF file, even for publications where browsing by image (such as Magazine content) is available. Images are HTML files embedded using a Flash-based image viewer. This is how we display many of the publications in PAO and British Periodicals".

And as to preservation, over the last decade there has been a lot of discussion of DTDs which enable preservation since the emergence of the NLM DTD and the more recent JATS and BITS[27]. JATS is becoming the de facto metadata standard for journals and BITS is still waiting for general acceptance for books.

9 Usage

We have been asked to summarise the use of the formats discussed especially by researchers but this would need far more effort than we can provide. Moreover, some of the formats which are in use by researchers and others are without the BL collection policy. So, for example, the BL is not mandated to collect ephemera except

[24] http://iopscience.iop.org/page/ePub3
[25] https://publishingperspectives.com/2017/04/russia-litres-ebook-format-fb3/
[26] https://www.elsevier.com/about/press-releases/corporate/elsevier-embraces-epub3-format,-ensuring-more-enriched-and-interactive-ebook-experience-for-readers
[27] https://dtd.nlm.nih.gov/publishing

incidentally through web harvesting. Games have also been ruled out despite their considerable popularity and the way in which their narratives reflect the Zeitgeist of the time. This has been much discussed because of the importance of ephemera to all types of cultural historians. And both book Apps and interactive narratives as interpreted in this report will be important in this context. However researchers, other than cultural historians and specialists in certain types of literature, are not the users of most examples of these formats. (There is a book App genre which is different; these are Professional Apps as illustrated by the NICE and BMJ Apps).

Works presented by structured databases as exemplified by Lloyds List, and most Lexis Nexis products are very important to both professionals and corporates including, in the case of the latter, publishers themselves. The whole range of scholars also makes use of the online resources which we touched on in our section on trends. We have some hard information about the importance of business intelligence company Euromonitor in terms of sectors, but not of course numbers. Over 50% of their market is to corporates.

In summary usage requires far more investigation than we can provide in such a short report. We suggest that if BL wants to know more it will need to consider further work in this area.

10 Summary Findings and Recommendations

We have largely organised this report based on the questions that were asked of us and have presented the evidence we have gathered as against the formats proposed by BL. In this summary, we look at each format in turn and try and derive some general recommendations.

10.1 Apps

Our research has concluded that there are Apps which are in effect books though the definition may need to be clearer for all concerned. However, we do not believe that there are more than a few thousand, if that, published in the UK by UK companies albeit a small number have become very successful. Those that do exist are often children's books or "how to guides" or topics which somehow use the added functionality that an App provides e.g. interact with a mobile phone. We also heard from several industry experts the view that Apps were declining and may well cease to be of significance in the future being usurped by epub3 formatted ebooks, streaming services or web Apps. Professional Apps may have further to go as might new forms of publishing that are hard to imagine at the moment but bring together all kinds of media together other dimensions such as time and narrative. There are also strong overlaps with Games and indeed many publications are identified as Games by the publishers and not as books.

The main problem with capturing Apps is that they are outside the traditional publishing framework so that there are no obvious sources to check as to their availability. Google Play appears only to list them incidentally and both BDS Ltd and Nielsen do not address Apps in their cataloguing services. And, once identified and captured, Apps raise issue of preservation through the high level of technological obsolescence. However, they are finite artefacts and though they use one or other of the mobile/tablet OS, there are emulators available, though these might be limited.

Interactive narratives are in many ways a sub-genre of Apps. They are equally difficult to pin down and numbers could vary from a few hundred to a few thousand but we think no more than that. They are a growing genre alongside other

personalised offerings but are not a format per se; some appear either to be HTML or HTML5 based in which case they should be harvested by the web harvesting now in place or they are discrete Apps in which case the above applies. We see no merit or need for the Library to collect authoring software; it is not necessary for playback and the legal deposit does not require it.

10.2 Databases

The impression we have from the data available is that there are relatively fewer databases than Apps, perhaps no more than hundreds - though the same caveats about definitions apply. Some and maybe many are multinational and identifying them as UK based will be problematic. But identifying publishers should be that much easier in that there are listings and catalogues which can be systematically addressed. However, the consequent capture of data may be more problematic in that: publishers may be unwilling to collaborate; the file sizes are likely to be large; the UK provenance problematic; and many databases will be ever changing necessitating a programme of snapshots. And once captured, preservation is not trivial given the problem, with some, of the need to preserve the associated search interface which itself will be subject to change. And it may be that some will require international collaboration or collaboration with publishers to resolve.

10.3 Recommendations

In summary, we recommend that the BL:

a) Agree clear unambiguous definitions of these emerging formats with associated collection policies; these will need to be under continuous review

b) Agree with the TNA and other national agencies, collection and access policies, if these are not already known, to avoid duplication or gaps;

c) Undertake a systematic check of appropriate catalogues, directories and stores for all 3 formats to identify publications and then seek legal deposit;

d) In the case of Apps, capture those identified now, even if there is not yet an appropriate preservation strategy in place, in that some are at risk of deletion and loss;

e) Work as far as it can, towards ensuring that adequate metadata standards exist and are implemented for all formats throughout the content lifecycle; that probably implies working with other agencies and groups;

f) Promulgate the policies developed to all those responsible for publications;

g) Ensure that there is, in due course, a technical, preservation infrastructure available capable of dealing with these formats; this might imply international collaboration, with other preservation agencies and publishers themselves;

h) Collaborate with other National Libraries on the outcomes of this report through professional groups and the possible dissemination of this report;

i) Discuss with publishing representative bodies or publishers to discover whether licensing agreements for software and grants of rights from owners of multimedia content, actually acquire the rights needed for the archiving and preservation by another entity.

11 References

Android vs. iOS (2017) http://www.diffen.com/difference/Android_vs_iOS (Accessed: March 20, 2017)

database. Dictionary.com. The Free On-line Dictionary of Computing. Denis Howe. http://www.dictionary.com/browse/database (accessed: March 20, 2017).

ISSN Manual. http://www.issn.org/understanding-the-issn/assignment-rules/issn-manual/

ISO 3297:2007 Information and documentation — International standard serial number (ISSN) http://www.issn.org/

Karch, M (2016) What Are Apps? - Definition and Examples. https://www.lifewire.com/what-are-apps-1616114

Muir, A. (2005), Legal Deposit of Digital Publications 2005. A Doctoral Thesis. University of Loughborough https://dspace.lboro.ac.uk/2134/8469.

12 Bibliography

Caylin Smith There's Ample to Sample: Content Sampling at the British Library. http://openpreservation.org/blog/2017/01/23/theres-ample-to-sample-content-sampling-at-the-british-library/

Deposit of electronic publications with the national library of Australia. Guide to requirements for publishers; June 2016 National Library of Australia.

Gibby, Richard and Caroline Brazier, (2012),"Observations on the development of non-print legal deposit in the UK", Library Review, Vol. 61 Iss 5 pp. 362 – 377.http://dx.doi.org/10.1108/00242531211280487.

Gibby, Richard and Green, Andrew. (2008) Electronic Legal Deposit in the United Kingdom. New Review of Academic Librarianship. 2008 Vol 14 Iss 1-2 p55-70.

Gollins, Tim Parsimonious preservation: preventing pointless processes! (The small simple steps that take digital preservation a long way forward) Online Information 2009 proceedings p 75.

Kirchhoff, Amy and Sheila Morrissey, Sheila (2014) Preserving eBooks. DPC Technology Watch Report 14-01 June 2014; DPC Technology Watch Series. Digital Preservation Coalition 2014 – and Amy Kirchhoff and Sheila M. Morrissey 2014. ISSN: 2048-7916. DOI: http://dx.doi.org/10.7207/twr14-01.

Legal deposits of electronic materials in Sweden. For Individual Suppliers. National Library of Sweden.

Muir, A. (2001),"Legal deposit and preservation of digital publications: a review of research and development activity", Journal of Documentation, Vol. 57 Iss 5 pp. 652 – 682.http://dx.doi.org/10.1108/EUM0000000007097

Nicholas Joint, (2006),"Legal deposit and collection development in a digital world", Library Review, Vol. 55 Iss 8 pp. 468 – 473.http://dx.doi.org/10.1108/00242530610689310

Ojanen, Lauri; Vahtola, Aija (2016) Games in the National Collection. Scandinavian Library Quarterly; Vol: 49 Issue: 4 p34, 2016

Preserving eBooks , DPC Technology Watch Report 14-01 July 2014, authors Amy Kirchhoff and Sheila Morrissey, Series Editors Charles Beagrie Ltd ISSN: 2048-7916. DOI: http://dx.doi.org/10.7207/twr14-01

Sustaining The Value; The British Library Digital Preservation Strategy 2017-2020. British Library 2016.

Wikert, Jo (2012) . EPUB 3 facts and forecasts. Why ebook publishing will look more like software development than print production. Tools of Change. http://toc.oreilly.com/2012/10/epub-3-facts-and-forecasts.html

Wikert, Jo (2012) HTML5, EPUB 3, and ebooks vs. web apps Tools of change http://toc.oreilly.com/2012/09/html5-epub-3-and-ebooks-vs-web-apps.html

Appendix 1 People contacted/interviewed

We are grateful for the input of many people in compiling this report many of whom have given valuable time or committed to extensive reporting. The list below is meant to be comprehensive and we duly apologise if any names have been missed.

Sam	Alloing	Digital Preservation Officer	Koninklijke Bibliotheek
Angus	Beach	Information Architect	NICE
Dale	Beeton	Business Development	Axiell Systems
Gaelle	Bequet	Director	ISSN Agency Paris
Claire	Caulfield	Digital/Print Operations Manager	British Library
Nicola	Cavalli	Publishing Consultant	Italy
Ian	Cooke	Head of Contemporary British Publications	British Library
Sharon	Cooper	Chief Digital Officer	BMJ
Nick	Coveney,	ex-Head of Digital	Kings Road Publishing
Andy	Davis	Publisher Liaison and Non-Print Legal Deposit Implementation	British Library
Diana	Delafini	Head ISBN UK	Nielsen
Jo	Doyle	Marketing Director	JollyBooks.com
Koko	Ekong,	Ebook Technical and Design Manager,	Penguin Random House
David	Espley	CTO (UK)	LexisNexis
Louise	Fauduet	Chef du service Multimédias Département de l'Audiovisuel	Bibliothèque Nationale de France
Trevor	Fenwick	Executive Chairman and Managing Director	Euromonitor
Dan	Franklin		Penguin Random House
Sarah	Gilmore	E resources Team	UCL Library Services
Vesna	Golubovic	Library Advisor	National Library of Croatia
Stella	Griffiths	Executive Director	International ISBN Agency
Jon	Ingold	Director	Inkle Studios
Daniel	Jansson	Technical Business Analyst	Kungl. biblioteket/National Library of Sweden
Jerry	Jenkins	Curator for Emerging Media	British Library

Ruth	Jones	Director Business Development	Ingram
Bill	Karsdorf	VP and Principal Consultant	Apex Co-Advantage
James	Mcfarlane	Director	Easy Press Technologies
Goh Yu	Mei	Associate Librarian	National Library, Singapore
Aaron	Melzak	ProQuest Technical Product Specialist	ProQuest
Iain	Moir	Head of Intellectual property	NICE
Sheila	Morrissey	Senior Researcher	Portico
Maureen	Pennock	Digital Preservation	British Library
Barbara	Sierman	Digital Preservation Officer	Koninklijke Bibliotheek
Joshua	Tallent	Director of Outreach and Education	Firebrandtech
Carol	Tullo	Director of Information Policy	TNA
Johan	Van Der Kniff	Digital Preservation Formats	Koninklijke Bibliotheek
Gert-Jan	van Velzen	Account manager Deposit Collection	Koninklijke Bibliotheek
George	Walkley	Head of Digital	Hachette UK
Stephen	Want	Chief, Copyright Acquisitions Division. United States Copyright Office	Library of Congress
Lesley	Whyte	Director	BDS Ltd
Neil	Wilson	Head of Metadata Services	British Library
Kate	Wittenberg	Managing Director	Portico

Appendix II

We sought answers to number of questions from the TNA and they forwarded this helpful response:

- **The size of publishing in a particular format in the UK (eg numbers of publishers, numbers of publications), and an assessment of publishing trends**

The Public Records Act 1958 provides for the Public Record Office and the public records system, covering administrative and departmental records of Government and the Courts. Now operating under the title of The National Archives, it is responsible for preserving the Government and Court records, including digital records and these are selected from the business records of 23 main Government Departments and around 200 other public bodies (mostly Non-Departmental Public Bodies). The number of individual records runs into the millions. Transfer generally occurs around 25 years after creation (currently, being lowered to 20 years by 2013) so digital transfer is phasing in with various pilots and programmes underway. Note much official information is proactively published online already and is freely available and captured.

- **A review of how Intellectual Property rights, and other legal considerations, apply to works commonly published in those formats**

Material created in Departments and other Crown status bodies is Crown Copyright under Chapter 10 of the Copyright, Designs and Patents Act 1988. Most Crown (c) material, once released into the public domain is licensed under the Open Government Licence and can be freely re-used with minimal conditions. Material acquired by Government (including non-Crown (c) material acquired in the process of archival transfer to The National Archives) can be made available to the public online under the public administration exception, recently expanded by The Copyright (Public Administration) Regulations 2014. http://www.legislation.gov.uk/uksi/2014/1385/regulation/2/made

As the overwhelming majority of the material is unpublished and a proportion may be sensitive, there are Freedom of Information procedures in place for Departments to flag material assessed as exempt from release under FOIA with the relevant exemption (eg personal data, national security).

- **An assessment of how much original content (ie unpublished in other sources) is produced**

It is almost all original, unpublished material: occasionally publications are involved incidentally in the context of other public business but in general selection policy is to avoid duplication of publications owing to the overlap with the LD regime.

- **File formats and structure used in publication, including an assessment of common standards and tools**

The majority are common office formats, data sets though there are more specialist formats used occasionally, eg 3D models. A study of the digital landscape in Government Departments was conducted in 2014 and is published at: http://www.nationalarchives.gov.uk/documents/digital-landscape-in-government-2014-15.pdf.

We have the same issues of the authentic record copy and presentation instance(s) as Legal Deposit libraries. Multi-component records are becoming more prevalent. In addition to identifying and understanding discrete file formats, data streams and code may be essential to acquire, preserve and give access to the record in future.

- **The availability of metadata to describe publications**

Government practice in metadata management varies according to the technical and organisational environment of origin and technical, administrative and resource discovery metadata varies. There are no enforced, nor enforceable metadata standards beyond high level catalogue information. TNA is not able to enforce the deposit of specific metadata, computer programmes, etc. required to deposit and preserve the records as exist under Regulation 17 of The Legal Deposit Libraries (Non-Print Works) Regulations 2013.

- **Current actors in collecting and preservation**

The National Archives guides and supervises the appraisal and selection activity of Government bodies under its Collection Policy. It is the default archive for the public records and has an ongoing working relationship with most of these bodies, particularly 23 central government departments. In the analogue environment, there is a network of local places of deposit appointed to hold and preserve records of local interest (eg lower courts, environmental and NHS bodies). The precise role of these in preserving the digital public records has yet to be finalised.

- **Use of publications, including for research**

The National Archives is traditionally used for historical, genealogical and journalistic research. New users deploying text and data mining and data analytics have expectations to use and re-use the record.

I mentioned that a paper entitled "Digital Cataloguing Practices" at The National Archives will publish on March 31 and so look out for that on the TNA website. A number of the themes are explored. Our Digital Strategy may also help and is just published. Sections 2 and 3 may give you a focus: http://www.nationalarchives.gov.uk/about/our-role/plans-policies-performance-and-projects/our-plans/digital-strategy/