Ebooks: rather electronic or book? Extending legal deposit to ebooks at the Bibliothèque nationale de France

Sophie Derrot, Clément Oury

Submitted on: 5/30/2014

Ebooks: rather electronic or book? Extending legal deposit to ebooks at the Bibliothèque nationale de France

Sophie Derrot
Legal deposit department, Bibliothèque nationale de France, Paris, France.
sophie.derrot@bnf.fr

Clément Oury
Legal deposit department, Bibliothèque nationale de France, Paris, France.
clement.oury@bnf.fr

Abstract:

Even though the ebooks market has developed more slowly in France than in other countries, especially in northern America and northern Europe, the number of titles proposed is now increasing. The ebook production and distribution chain has been clarified, while major standard ebooks formats (EPUB, PDF) and metadata schemes (ONIX) have emerged. A growing number of French citizens are currently using ebooks to access knowledge, culture and entertainment. Being able to collect this new kind of resource in order to preserve it in the long term has thus become a major challenge for the national library of France.

In order to tackle this issue and to set up a complete ebook deposit workflow, the Bibliothèque nationale de France (National Library of France or BnF) benefits from its experience in managing digital documents (for its digital library, Gallica) and from its tradition of legal deposit. It also seeks to capitalise on the long-standing relationships it has built with the publishers and the ebooks distributor. However, the Library needs to deal with the dual nature of ebooks, which are on one hand digital documents, collected in the name of the internet legal deposit, and on the other hand a new form of a kind of document that has been acquired by the library for centuries.

This paper aims to present the questions that the BnF needed to answer when faced with the need of collecting ebooks, and the solution it is currently adopting. Technical implementation has just begun and the outcomes – a complete entry, cataloguing, access and preservation workflow – are expected at the end of 2014 or the beginning of 2015.

Keywords: Ebooks Preservation Digital Legal Deposit.

A favourable context

The late emergence of an ebook market in France

At the very beginning of 2000, French media announced the birth a totally new kind of publication: the electronic book or ebook. Devices called “Softbook”, “Cybook” or even “Rocket e-Book” were supposed to change the publishing industry and promote a brand new way of reading: the screen was expected to replace paper. French companies were in many cases leading this so-called revolution. However, of the numerous devices that were launched at the beginning of the 21th century, very few survived. Too complex, too expensive, they came in fact too early and were not able to create a market beyond a very narrow circle of early adopters. The economic failure of the first generation of ebook devices had a strong impact on the future development of ebook market in France: the good old trustworthy book was therefore, for at least a decade, considered by most French readers and many publishers far more reliable than its electronic cousin.

Progressively, technological innovation (such as the use of e-ink) and the increasing popularity of new kinds of devices, such as tablets, led to an evolution of this situation of distrust. The French market changed. New purely online publishers appeared. The best-established publishers started delivering their books directly in both forms. They launched the digitization of their older collections, often fostered by financial support from the State. Standardization became a driver of technological innovation: the jungle of formats was progressively cleared in favour of a few reference formats: PDF for computers and tablets, EPUB and its proprietary equivalents (such as KF8 and iBooks) for ebook readers. On the back-office side, standardization of exchange of metadata is under way, thanks to the wide adoption of the ONIX format[1]. Finally, standardization is also taking place from the point of view of the ebook production and selling channel. Schematically, three main stakeholders have emerged.

Current ebook production and distribution chain

At the beginning of the chain, the publisher is in charge of producing the book: intellectual production (relationships with authors) and “material” production of the book in its paper or electronic form (this part may also be outsourced to a contractor). At the other end of the chain, the digital bookseller is in charge of promoting and selling the ebook to readers. In the middle, the role of the e-distributor is to make a bridge between these two stakeholders: it receives the ebook from the publisher, checks its format, verifies and enhances the quality of its metadata, and sends it to the bookseller. The e-distributor is also frequently in charge of managing the financial flows between the different links of the chain. This summary does not encompass all the complexities and subtleties of the actual production system, and this distribution of roles is always subject to change. However, the respective responsibilities of each actor are today better defined than in previous years.

Compared to its international equivalents, especially in northern America and northern Europe, the ebook market in France is not yet mature. The imbalance between the availability of books in printed and in electronic form is lessening, but there is still a strong difference in terms of demand. However, several figures are showing a positive trend: in March 2014, a poll indicated that 15 % of French citizens were reading ebooks, compared to 5 % two years before[2]. Thus, a growing number of French citizens is using ebooks to access knowledge,

[1] ONIX for Books, an XML standard designed to support computer-to-computer communication of bibliographic information. See http://www.editeur.org/83/Overview/

[2] Regular polls about ebooks usage and practices are funded by three of the main professional stakeholders in the books chain: the SOFIA, the SNE, and the SGDL. See

culture and entertainment. Being able to collect this new kind of resource in order to preserve it in the long run has thus become a major issue for the national library of France.

Management of digital material at the BnF: from digitization to digital legal deposit

Management of digital documents is not a new task at the BnF. The library has been involved in the field of digitization since 1992 through its digital library, Gallica. Gallica has now reached nearly 2 million documents: books and periodicals, newspapers, engravings, manuscripts, objects, sound recordings, audiovisual material.

The BnF is also in charge of receiving digital documents through legal deposit. Legal deposit is a long-standing mission of the Library. It was first established for printed books in 1537, by the Renaissance king François Ier, a lover of art and literature. Over the centuries, the legislation has evolved to cover different publication types and forms, thus adjusting to all major technological and social changes. This is particularly true during the 20th and 21st centuries, when the development of many media innovations created new forms of publication, which have gradually been included in the scope of legal deposit legislation. The BnF received its first digital resources when audio and video material, subject to legal deposit respectively since 1938 and 1975, become distributed in digital form (on CDs and then on DVDs). In 1992, an extension of the law covered the case of software and databases. Finally, on 1st August 2006, the most recent evolution included all publications of the French Internet in the scope of the legal deposit – this law is now part of the "Heritage code" [1].

Internet legal deposit covers all kinds of resources distributed online: from news websites to public parts of social networks, from institutional portals to blogs. Indeed, legal deposit refuses to select the documents to be preserved according to their intellectual or artistic value. However, this new extension brings two fundamental changes to the principles of legal deposit. First, the term "deposit" itself is not really relevant anymore. The BnF (along with the National Audiovisual Institute, or INA, for websites related to TV and radio) is in charge of collecting documents by automated means. Harvesting software, known as a "robot", is used to retrieve content online. Second, comprehensiveness, which is the traditional objective of legal deposit, is not an achievable goal anymore, as the library is faced with the huge and ever-changing space of the internet. Instead of comprehensiveness, the BnF follows an objective of representativeness: it seeks to constitute a collection that mirrors what is available online at a given date. A combination of broad crawls of the .fr domain and focused crawls of websites selected by BnF librarians or partner organizations has been put in place in order to achieve this goal[3].

As most ebooks are promoted and sold online, they are supposed to enter BnF holdings through the legal deposit of the internet[4]. So, from a legal point of view, ebooks have clearly the same status as any other publications hosted online. However, from an ontological point of view, they are digital equivalents of a kind of documents acquired by the library for centuries. This dual nature of ebooks – electronic and books – explains why the BnF decided to dedicate them a specific entry and treatment channel, different from the one of web

http://www.sne.fr/img/pdf/SDL/2012/Barometre_SofiaSneSgdl_Les%20usagesdulivrenumerique_mars2012.pdf and http://www.sne.fr/img/pdf/Evenements/Assises/Assises-21mars2014/Barometre-SNE-Sofia-SGDL-des-usages-du-livre-numerique-21-03-2014.pdf.

[3] See Sylvie Bonnel and Clément Oury, "La sélection de sites web pour une bibliothèque encyclopédique", to be published in the proceedings of the IFLA 2014 Conference in Lyon.

[4] Ebooks that are delivered to the public on physical media (e.g. CD, USB key…) already enter BnF collections through audiovisual legal deposit. But this currently represents a very small part of the ebooks market.

archives. Thus, the library built upon its triple experience of receiving printed books, of producing digitized books, and of collecting online publications in order to design a complete ebooks deposit workflow.

First experiments in the collection of digital publications

The first experiments towards the collection of digital publications at the BnF started around 2000, even before the law was changed. These investigations were performed within the digital library department, which was also in charge of managing Gallica. The two ways of getting digital content were investigated by BnF librarians: automated harvesting of internet content and direct deposit by newspapers or ebooks publishers. However, the first approach (online harvesting) was quickly considered a priority and most effort was put into setting up a complete web archiving workflow. Several reasons explain this choice. First, websites were seen as the most at-risk document; the web was a very young publishing space and it appeared critical to preserve its earliest forms.

They were also practical reasons. On one hand, thanks to international cooperation, tools for web archiving were already available; on the other hand, few ebooks were available, the ebook distribution chain itself was not mature and deposit solutions were not considered stable enough. There were finally legal constraints. In France, laws voted by the Parliament are enforced by a decree. The decree on digital legal deposit was only published in December 2011. So, between 2006 and 2011, there was a legal possibility to harvest freely available content (such as the huge majority of websites), but no means to ask publishers to deposit content distributed under payment, such as ebooks. This explains why the BnF did not start designing its ebook deposit workflow before 2012.

The BnF approach to the legal deposit of ebooks

Cooperation as a key

Cooperation with publishers was immediately identified as a critical element in order to build an efficient and durable system. The BnF benefited from its long-standing relationship with the Syndicat National de l'Édition (SNE), the main union of French publishers. In September 2012, during a meeting between representatives of SNE, BnF and the French ministry of Culture, it was decided to set up two joint working groups:

- a legal working group, whose mandate is to identify if the current decree on internet legal deposit takes sufficiently into account the specificities of ebooks;
- a technical or functional working group, which is in charge of designing the principles of a deposit system.

The functional working group started in March 2013. Its first decisions were related to the way of retrieving content: a system of direct deposit was preferred to web harvesting. The BnF could have chosen to crawl ebook hosting platforms, as it currently does for news websites [2]. However, in several cases, ebooks distributed online are not directly hosted on the website of the online bookstore. The bookstore is only the place where financial transaction occurs; the document itself, or a link to download it, is then sent to the purchaser by another means (e.g. by email). In that case, web harvesting would not have been efficient. Moreover, harvesting ebook hosting platforms would have made it impossible to benefit from the relationships established between the library's teams in charge of the deposit of printed documents and the publishers. Here therefore, the principles in use for printed books legal deposit prevailed over the logic of internet harvesting: in "traditional" legal deposit,

booksellers are not in charge of sending documents to the library; neither should they be for digital legal deposit.

The functional joint working group thus preferred leaving the publisher in charge of the deposit. However, for practical reasons, it was proposed that a publisher could give a mandate to its e-distributor to actually perform the deposit. For the BnF, working with distributors appeared very quickly as the best solution as:

- there are few distributors compared to hundreds of publishers;
- the BnF benefits from a first set of quality controls performed by distributors, both for the ebooks and their metadata;
- distributors receive ebooks without DRM; they are therefore able to send them to the library without DRM (see below).

Analyzing existing (internal and external) transmission channels

The discussions of the joint working group were complemented by several enquiries performed by a dedicated team at BnF, grouping together experts from the legal deposit and IT departments.

This group interviewed the main French distributors, as they were supposed to become the first ones to perform actual deposits. It studied the data and metadata formats they were using, how they were transmitting content, and what they could send to the Library.
At the same time, the group also performed an analysis of the BnF internal entry chains for digital documents (documents from BnF holdings digitized by private contractors; audio and video documents, master versions of books sent by publishers to BnF for visually impaired people[5], etc.). The goal of this enquiry was to identify what tools, what entry tracks already developed for similar kind of documents could be re-used for an ebooks deposit workflow.
These studies led to the design of a first draft of a deposit workflow in the summer of 2013. More detailed analyses followed during the second semester of the year; their conclusions will be presented below. Concrete developments started in the first months of 2014.

Implementation: a work in progress

Defining the scope

As already stated, the landscape of commercialized ebook formats, which was quite dense a few years ago, has tended to become simplified, to the advantage of standardized formats (EPUB) or closed-source formats of the market leaders. Even if it does not solve every issue, this simplification is welcome from a legal deposit point of view, as the 2006 law theoretically extends the perimeter of the objects concerned to all produced and commercialized formats.

The efforts are concentrated on files which can be easily defined as “books” as they share the characteristics of printed books. Therefore, ebooks in formats such as TXT or DOC were excluded from the beginning, as they are more often production formats than publication formats; they are not found in commercial markets and distributors do not work with them. At the other end of the ebook chain of production, physical reading devices and their software

[5] According to the law, whenever an association of visually impaired people requires it, the BnF is entitled to ask from the publishers the master digital version of a printed book (ideally a XML version) and to transfer it to the association. The association is then entitled to transform the book in an accessible digital version.

are not concerned either by this deposit track. As discussed below, DRM and closed-source formats too have to be excluded from the process. Once these exclusions have been made, what is left? During the initial discussions, publishers and distributors made clear that the most frequently produced and sold formats were EPUB 2 and PDF; MOBI and EPUB 3 were currently in the minority. PDF and EPUB (versions 2 and 3, fixed-layout and reflowable) are therefore the main target.

First step: receiving ebooks with metadata

Distributors should depose their files on a dedicated FTP platform in the form of a ZIP file, containing both data files (EPUB or PDF) and metadata files (one or more ONIX files). These files come with their MD5 checksum, for the BnF to check their integrity throughout the process. Once these ZIP files are deposed, a first set of checks will be performed by the library, such as virus checks, verification that the files declared are really in the delivered package, etc. They are then opened and their content is more carefully processed. Before admission into the library's collections, a waiting period is defined to allow each distributor to send any bibliographical and/or technical updates to the file and/or its metadata. Once the delay for potential updates has passed and if the package that has been delivered passes the checks, the ebooks receive a legal deposit number and the distributor is informed that the document has been deposited.

One application supervises the whole process, from the reception of the documents to access and expert cataloguing. This ensures information exchanges between the different applications involved and give the process its consistency.

Metadata flow: automation and conversion are the keys

The guiding principle when putting in place the legal deposit of ebooks was automation, including the re-exploitation of metadata created by publishers and distributors for their own needs. The most used format for carrying this metadata in the book trade are ONIX files, which are generally created by distributors from information provided by publishers and are then sent to online booksellers along with ebook files. The advantages of ONIX files attached to ebooks by the distributors are their richness and precision: this information has to be exact because of its commercial purpose. On the other hand, quite a few of the data are useless from a librarian point of view (for example, prices for every country where the ebook is commercialized) and some important bibliographic metadata is often lacking (the ISBN of the printed version is not always provided). To make existing ONIX files as useful as possible for the library, an ONIX model including the BnF specifications was defined by librarians and proposed to publishers and distributors.

As soon as a book has passed all the checks in the delivery zone, its metadata is processed by the entry application. Among other things, this application converts information within the ONIX file from XML to Intermarc, the specific format for bibliographic metadata used by the BnF. Information given by the distributors should be published within the BnF catalogue and be distributed under an Open License.

This bibliographic record is quite rich, as information from the ONIX file is fairly detailed (title, author, publisher, but also abstract, author's biographical note, keywords, etc.). In a second phase of the project, the automatic creation of links will allow a relation to be established between for example the records for two formats of the same work (printed and EPUB, EPUB and PDF) thanks to the ISBN which publishers will hopefully provide within the ONIX file. Using the same principle, ISNI identifiers could be used to link ebook bibliographical records to authority records.

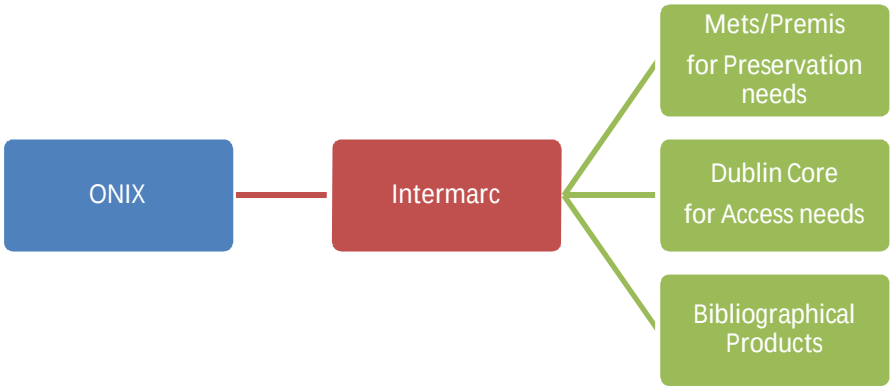


Figure 1. Same information, multiple uses.

All these steps are designed to be automated. For some deposited ebooks, bibliographical records will partially be reworked by cataloguers and will gain more precision. Records from the BnF catalogue are already largely reused by other libraries: ebooks metadata is rather eagerly awaited, as there is currently no specific database for such bibliographical metadata as for printed books. Even publishers are interested in getting this enriched metadata back, to improve their own bibliographical databases.

Data flow: preservation at the heart of the system

Contrary to the metadata files, the ebooks files will not be converted in any way and the original file will be the one that is preserved. Legal deposit is indeed attached to the format of the document and this format has to be kept in its original form.

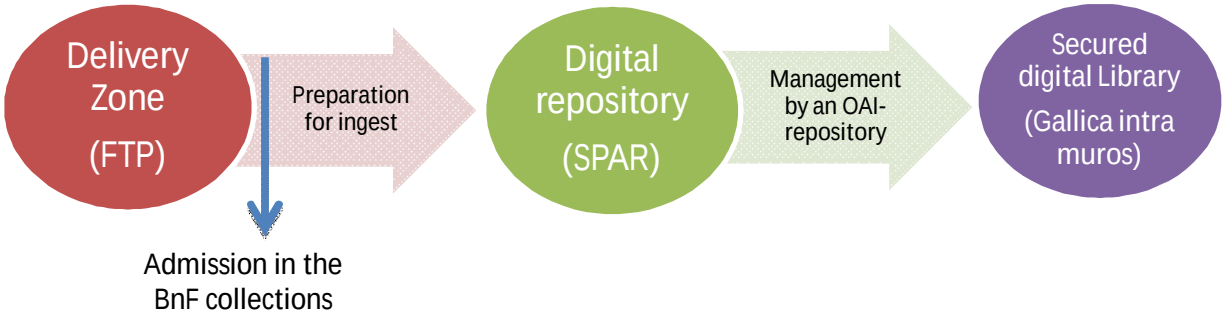
As already shown, ebook files pass a first set of tests at the moment they are deposited on the FTP platform to check their integrity and format. A validation tool developed by the International Digital Publication Forum (IDPF), EPUBCheck, will be used for EPUB files, but the choice has still to be made for PDF. The purpose of these tests is to check if the files are in fact able to be managed by the library: DRM and closed-source formats will not be accepted by the system, both for preservation and access reasons.

The main reason that ebooks with DRM will be excluded is to permit manipulation of the files during the deposit process. Frequent copies – often blocked by DRM – are de facto necessary for deposit, access and preservation processes[6]: setting up a legal deposit system of this kind of protected material didn't seem feasible. Publishers and distributors agreed to this pragmatic position. Closed source formats will not be deposited for the same reasons: preservation systems will probably be unable to deal with them, especially in the long term,

[6] About issues of preserving digital resources with DRMs, see [3].

and access solutions would be difficult to implement. In spite of the ambiguity of this position – in contradiction with the traditional objective of comprehensiveness of legal deposit –, it seemed more reasonable to proceed in this way in order to ensure long-term access and preservation – which are also part of the objectives of legal deposit.

Figure 2. The data flow



Once the ebook files enter the BnF collections, they will be prepared for ingest by the library's digital repository, SPAR (Scalable Preservation and Access Repository)[7]. A dedicated preingest module will be developed for the "Negotiated legal deposit track". The legal deposit is considered to be "negotiated" as the form of the delivery is negotiated between the distributor and the library, so that for example not all formats will be accepted. This module will get the package ready for ingest, by joining together the ebook file (EPUB or PDF), the original ONIX file, possibly a picture file of the book cover, and finally the METS manifest[8]. The role of the SPAR system in the legal deposit workflow is critical. On one hand, the preingest module will provide fundamental information to the General Catalogue, such as the ARK number of the ebook file, i.e. its persistent identifier. On the other hand, SPAR will deliver the deposited ebooks to the access platform and application.

Within SPAR itself, two channels will be set up, according to the ebook format and its components:

- Ebooks whose format is considered "managed" after all quality checks and characterization will be ingested in the first channel.
- Files that are valid from a format specification point of view but which present some preservation threats will be ingested in the second channel. For example, EPUB files containing Flash or Javascript elements are not considered "managed" as BnF does not have sufficient confidence in its capability to preserve them in the long term.

[7] About SPAR, see [4].

[8] METS (Metadata Encoding and Transmission Standard) is an XML format. METS files are wrappers that contain all metadata elements necessary to describe a digital document and preserve it in the long run.

Therefore, the documents won't be available for access before they are ready to be preserved long-term. It is only when the digital repository has treated them that an availability flag will be sent to the catalogue and to the access application.

Access to legal deposit materials is legally restricted to research areas of the library and only authorised persons have the possibility to consult them. The restriction applies to both printed and digital material, including web archives and in the future deposited ebooks. A specific application of the digital library, "Gallica intra muros", provides access to restricted digital material (i.e. under copyright) and is not available outside the library reading rooms. As security is a central concern of both the library and its partners, this aspect of Gallica intra muros has been reinforced. The chosen solution is a "virtual browser" which gives access to material without any possibility for the user to download or modify them. Both PDF and EPUB will be available in the same application, among the other copyrighted digital collections of the library.

The evolution of formats and reading tools will be a crucial point in the coming years. A strong dialogue with partners will be necessary to follow their practices. In addition, the participation of the BnF in international bodies such as IDPF could be a good solution to keep up to date with evolutions in access, as the library already does for preservation matters.

The role of humans: still an open question

It is not currently possible to know what level of human work will be involved in this workflow: will a human quality control be performed on some ebooks before they enter BnF collections? What percentage of books will be catalogued by librarians, if any? What level of quality will be acceptable for catalogue records?

In any case, management of digital along with printed books will represent a dramatic change for the professionals of the legal deposit department. In order to assist them in this evolution, the department launched a series of training sessions called "digital workshops". All members of the legal deposit department are requested to attend these, which deal with questions such as "what is an ebook?" "How does one produce/buy/read/preserve such documents?". This first round of training is intended to give everyone a common basic level of knowledge and awareness regarding e-publications. A second step would be to train them to actually receive, verify and catalogue ebooks – but this will come later, when the critical questions related to the level of human implication in the ebooks workflow have been decided.

Conclusion: lessons learnt and next steps

Developments have already begun and the whole chain should be ready at the end of 2014 or the beginning of 2015. First experiments with volunteer distributors will be carried out at the end of the year. The involvement of publishers and distributors is the first strength of this project. This brings the possibility for the library has to work with a new kind of professionals. Some publishers or distributors belong to the traditional chain of book production and already have the habit of working with the BnF. However other partners are quite new and specific to the ebook market; they are often pure players and sometimes have new practices and new points of view. The reflection on the legal deposit of ebooks has to take into consideration this diversity in the ebook chain of production and commercialisation.

The second strength of the workflow is its capacity to build on many other projects already running within the library. Almost every application needed already exists in one

form or another. The teams concerned work together and the reflection on ebooks benefits from a large panel of expertise – from the ebook market and formats to cataloguing, metadata and preservation . This will greatly facilitate the setting up of the legal deposit of ebooks.

Acknowledgments

The authors thank Peter Stirling for his perfect English and his Scottish accent.

References

[1] Illien G., Sanz P., Sepetjan S., Stirling P. 2012. The state of e-legal deposit in France: looking back at five years of putting new legislation into practice and envisioning the future. In *IFLA journal*, 2012, vol. 38, n° 1. [http://www.ifla.org/files/hq/publications/ifla-journal/ifla-journal-38-1_2012.pdf]

[2] Oury C. 2011. When press is not printed*:* the challenge of collecting digital newspapers at the Bibliothèque nationale de France. *In Proceedings of the IFLA Preconference, newspaper section* (Mikkeli, Finland, August 2012). [http://www.ifla2012mikkeli.com/getfile.php?file=154 or http://halshs.archives-ouvertes.fr/docs/00/76/90/84/PDF/LegalDepositNewspapersBnF_Oury_IFLA2012.pdf]

[3] APARSEN. 2013. *Report on DRM Preservation.* [http://www.alliancepermanentaccess.org/wp-content/uploads/downloads/2014/02/APARSEN-REP-D31_1-01-1_4.pdf].

[4] Derrot S., Fauduet L., Oury C., and Peyrard S. 2012. Preservation is Knowledge: A community-driven preservation approach. In *Proceedings of the 9th International Conference on Preservation of Digital Objects* (Toronto, Canada, October 2012). [https://ipres.ischool.utoronto.ca/sites/ipres.ischool.utoronto.ca/files/iPres%202012%20Conference%20Proceedings%20Final.pdf]