IFLA WLIC 2016 COLUMBUS

Submitted on: 08.08.2016

Born digital news collections: New formats, new approaches

Elisa Villanueva
Collections department, Koninklijke Bibliotheek, The Hague, Netherlands.
mevillan@uc.cl

Jasper Faase
Collections department, Koninklijke Bibliotheek, The Hague, Netherlands.
Jasper.faase@kb.nl

Abstract:

Nowadays most newspapers (or almost all of them) have a digital version or are definitely turning to digital. Furthermore, there are many news websites which provide news content. Additionally, with the rise of Web 2.0 many websites are based on user generated content, and others are almost built on the user's posts and comments as their main sources.

Most of the libraries face challenges collecting, preserving 'born digital' news and newspapers which puts this part of our cultural heritage at risk. .The aim of this paper will be to present a desktop research into the approach other libraries or cultural heritage institutions took (or are taking) during the creation of their born digital news collection, and then conclude with some considerations regarding which could be the best practices and the main aspects to take into account.

Keywords: News, newspapers, born digital, libraries, collections.

1. Newspapers: the memory of democracy

National libraries have always been responsible for collecting, preserving and giving access to their national heritage. These are the main centers of knowledge and information of a country and their main task is to preserve its nation's memory and cultural manifestations. Important part of their collections are newspapers, which constitute an essential testimony of a country's history. A famous journalist named Alan Barth times ago mentioned 'News is only the first rough draft of History'[1]. These are considered as an essential testimony of the "memory of democracy". Then, preserving news is not only thought for the exclusive use for

[1] A. Barth, *The New Republic*, Volume 108. (Washington D.C.: Republic Publishing Company, 1943), p. 677.

researchers, but also for law, citizen, community, etc. Nowadays, with the technological advances much more traditional newspapers (or almost all of them) have a digital version, a web site or are definitely turning to digital, and many newspapers are born digital. Furthermore, there are many news websites which provide news content. Additionally, with the rise of Web 2.0 many websites are based on user generated content. Most of the current press have turned digital and almost most of the national libraries are not capable of collecting, preserving and giving access to these properly:

> In the electronic era, the effective archiving of online news is simply not happening on a meaningful scale. The nascent legal deposit and web harvesting programs of various national libraries are either not yet scaled to archive significant amounts of electronic news content, or are not designed to capture online news content in formats that current research practices require.[2]

1.1. New features, new participants

As it has been described above, the process of collecting and preserving born digital news is quite a challenge because there are many new and different features to take into account (if compare it with print newspapers): formats, harvesting technologies, frequency capture, copyrights, completeness, accuracy, technological obsolescence, upload system, access, reproduction restrictions, metadata, storage, etc. As an effect, more and diverse actors and parties will be involved or more aware in the challenge of collecting and preserving born digital news. Editorial systems, newspapers publishers and news websites, news producers, companies that work as news publishers mediators or providing services regarding data management and distribution (e.g. ProQuest, Lexis Nexis), institutions related to libraries (e.g. CRL, IFLA, Educopia) journalism (e.g. Reynolds Journalism Institute) or digital preservation institutions. Negotiations, agreements and partnerships between some of these parties have been developing through the last years, but still there is not a leading standard established of how would be the best way to approach this challenge.

It is important to consider that because it is digitally originated it is impossible to apply the same procedures as for print newspaper. Selection, acquisition, preservation and access are radically different for digital content, and this would have a direct effect in the whole process of developing a born digital news collection.

1.2. Methodology

For this reason, the aim of this paper will be to present a desktop research into the approach other libraries or cultural heritage institutions took (or are taking) during the creation of their born digital newspaper collection, and then conclude with some considerations regarding which could be the best practices and the main aspects to take into account. This was done through interviews to different professionals that work in libraries and related institutions throughout the world. The selection of this institutions was based in the criteria if they had a certain approach with the plan of collecting born digital news content. The following libraries where interviewed: National Library of Denmark, National Library of South Africa, National Library of Swiss, National library of Luxemburg, National Library of Australia, National Library of Croatia, National Library of France, British Library, National Library of Sweden, National Library of France, Kentucky University Library, Texas University Library and the

[2] Center for Research Libraries, 'Focus on Global Resources', <http://www.crl.edu/focus/article/9559> (4 April 2016).

Library of Congress. Furthermore, in order to get a broader background about the position of researchers towards the current and future practices regarding digital news collections and preservation, other people from institutions related to libraries (CRL, IFLA), journalism (Reynolds Journalism Institute) and digital preservation were consulted. The collected data will be distributed per institution, according to the main stages involve in the usual process libraries has to go through when developing its collections: Selection, acquisition, preservation and access.

1.3. New formats, new approaches

Before we focus on each institution's model, it will be important to provide a brief background regarding the process of collecting born digital news content. For this purpose we will start with some definitions and naming the actors involved in this process.

Born digital content could be quite simple but at the same time vague concept to define. Every material that is considered to be born digital is because it has been originated as a digital product since its beginnings. There are several confusions about which type of material is considered born digital, digital, if these are the same, etc. Born digital content is distinct from digital content, which is created through the digitization of analog content. Examples of born digital content include word processing documents, spreadsheets, and original images produced with digital cameras.[3] Having this definition it would be possible to narrow our scope to only news content that where originated as digital products, leaving behind those which have their basis as analog content and then being digitized. Even though, this could seemed to be repetitive, it is important to set this difference because both type of content have a different lifecycle, which will have different effects in the further treatment for being collected, accessed and preserved. Born digital news displays many formats variations, because these configure not only traditional plain text, but also images, videos, user generated content in social media, and at last, but not less important, embedded metadata. In other words, 'News production is no longer the periodic, linear process with a single, fixed output, response, and update. Therefore devising effective strategies for preserving news in the electronic environment requires an understanding of the "lifecycle" of news content.'[4]

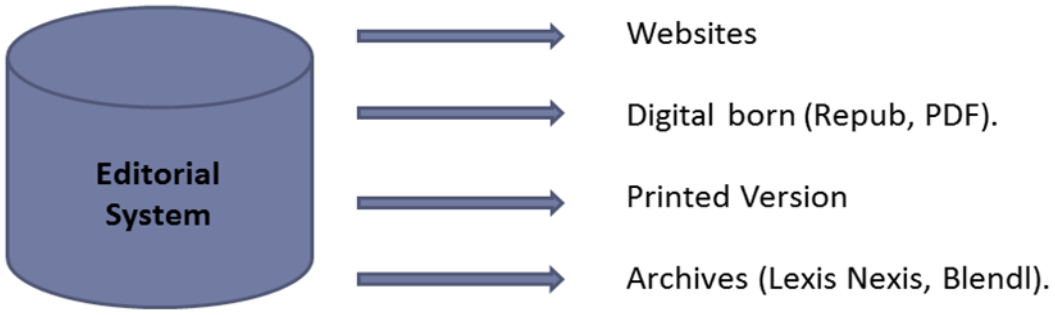

Figure 1: Born Digital news content production

[3] Federal Agency Digitization Guidelines Initiative, 'Born Digital' <http://www.digitizationguidelines.gov/term.php?term=borndigital> (5 April 2016).

[4] Center for Research Libraries, 'Preserving news in the digital environment: Mapping the newspaper industry in transition', April 27, 2011, p.4.
<https://www.crl.edu/sites/default/files/reports/LCreport_final.pdf > (7 March 2016).

Other aspect of born digital news which could be a future challenge for libraries is the constant changing. Digital news which are displayed in web sites have the facility to be frequently updated, which brings more problems for these to be accurately collected. Moreover, ephemerality is another important aspect of born digital news content that would have direct influence in the developing of a collection. Most of the content online disappears after a certain time, so if it was not captured since its' early publication it will be lost. There is a big gap of lost valuable part of our social history, and this is going to continue growing. This exists because of the big amount of news content that is published daily since long time ago until nowadays, and that has not been properly preserved.

> But today's great library is being destroyed even as it is being built. Until you lose something big on the Internet, something truly valuable, this paradox can be difficult to understand. Transformative technologies in any era are met with initial skepticism, and that attitude often fuels indifference about initial preservation efforts. Historians and digital preservationists agree on this fact: The early web, today's web, will be mostly lost to time.[5]

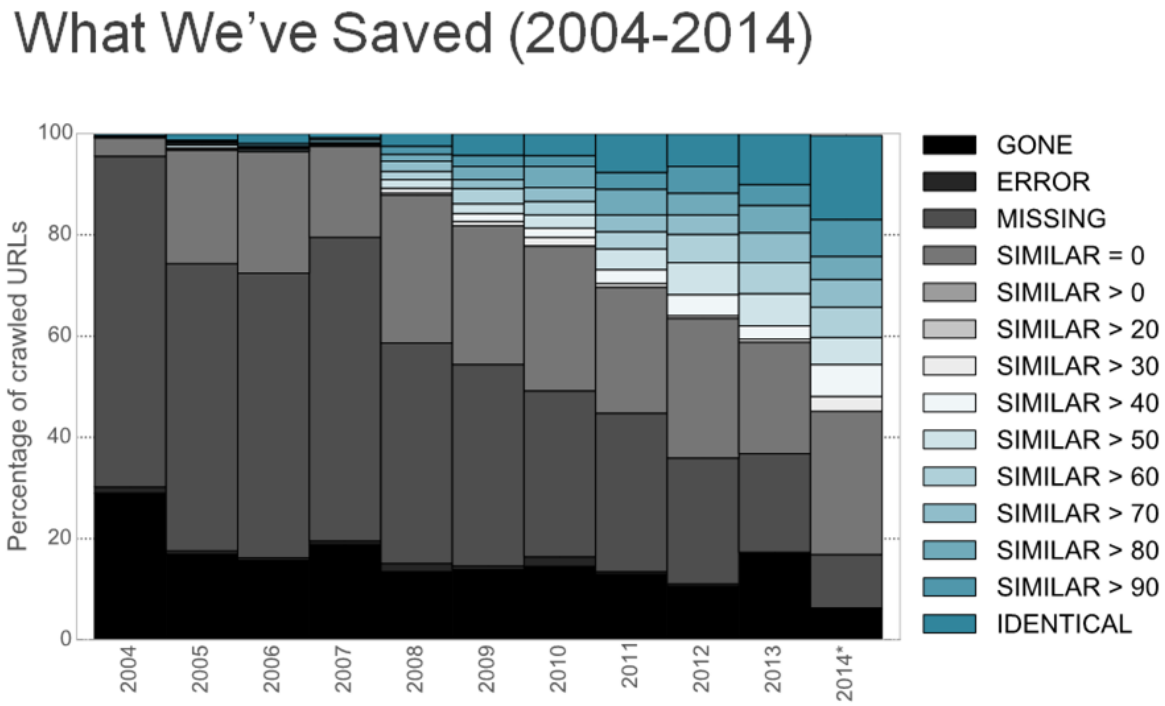

Figure 2: British Library percentages of crawled URLS. Source: Andy Jackson, British Library Web Archiving Technical Lead.[6]

Even though this statement is mainly about web archiving, born digital newspaper is going through a similar way. There is a gap and this, as the same as it has been experienced with web archiving will continue increasing if we do not find a proper method to collect and preserve this material.

[5] The Atlantic, 'Raiders of the lost Web', <http://www.theatlantic.com/technology/archive/2015/10/raiders-of-the-lost-web/409210/> (5 April 2016).

[6] International Internet Preservation Consortium, 'General Assembly ', <http://netpreserve.org/general-assembly/ga2015-schedule> (27 April 2016).

2. Results summary

As it will be described in the appendix of this paper, some institutions are dealing with born digital news content, while others are still collecting only the analog versions. Even though the Bibliotèque nationale de France, The National library of Australia, The National Library of Sweden, and The National Library of Croatia had made great improvements, there is not a leading institution capable to do this properly in order to set standards. One of the frequent excuses towards not collecting born digital news were resources and time, but most of them assured they had a plan regarding born digital newspaper in mind.
Summing up, currently there is not an ideal perspective when developing a born digital newspaper collection. Those institutions that had a certain approach with born digital news content had made many taught decisions in order to capture at least a brief scope and not with the accuracy and completeness digital news would require. Most of them are aware about the loses certain decisions will brought to their collection and national heritage, but still made them in order to do at least something about it. One of the core decisions is related with the selection criteria and covering scope, which in general terms, has to be with the immense scope of news production Web 2.0 allows, the legal situation of each country, copyrights, the relation with publishers and technology. Moreover, the frequency of capture has a direct impact in the accuracy and completeness of the collection. As it is well known, born digital news are not static plain texts, they are continuously evolving so it is not easy and possible to cover all the scope. Veracity and completeness regarding news content will not be full field, and this could have a serious repercussion for the reconstruction of our history in the future, whether it is for practical purposes (legal issues, politics, etc.) or for research. That is why preservation is also a big challenge for libraries and cultural and heritage institutions. This stage is closely related with diversity of formats, harvesting and preservation technologies and at last, but not less important; technical obsolescence. Because of the difficulties and developed technology this labor requires, some institutions are being counseled by others more technical, as an example The Internet Archive and its popular Wayback Machine. Also related with technical obsolescence is access. Probably today, or close to the date, it will be possible to have access to most of the born digital news content already archived, but this probably would not be possible in the future because technologies evolve in an uncontrollable speed. Even today, not because technical advances, but because of copyrights many of these are not possible to accessed or sometimes with some restrictions (only onsite, limited time, etc.). Furthermore, some material had difficulties to be visualize as it was in its origins. Because of diversity of formats, some of them could not be harvested properly and some fields could be lost in the process, as an example playing back a video related to a news article.

2.1. Best practices, solutions and models

2.1.1. CRL

After understanding the general scene in which today born digital news collections are being developed, it would be relevant to give a brief description about what other institutions or researchers related to this topic think might be relevant to do or to take into account. James Simon, Vice president of CRL (Center for Research Libraries), told us that his institution since long time ago have been interested in the preservation of print and electronic materials. During the mid-2000 they collaborated in advisory capacities about licenses with publishers and they could obtain agreements under paywalls, with big newspapers for academic subscription. The aim is to support libraries by promoting solutions and new models in order to achieve their main tasks. They worked in partnership with the Library of Congress about

the different technical platforms involved in the newspaper online publications, the different preservation approaches for different formats (produce online and PDF versions), etc.:

> The decline of the newspaper industry combined with the ascent of digital media for news reporting and distribution means that a significant portion of the journalistic record is now at risk. Recently, the Library of Congress (LC) National Digital Information Infrastructure and Preservation Program (NDIIPP) held a workshop to explore possible strategies for collecting and preserving digital news on a national basis. For purposes of discussion, LC defined digital news to include, at minimum, "digital newspaper Web sites, television and radio broadcasts distributed via the Internet, blogs, pod casts, digital photographs, and videos that document current events and cultural trends." The workshop, held September 2 and 3, 2009, brought together about 30 invited specialists in the field: broadcasters, producers, distributors, and archivists, as well as researchers who depend upon digital news. Attendees heard presentations on existing LC programs that preserve television, radio, and newspapers. Presentations also featured a variety of archiving programs at individual universities, state and local institutions, and media organizations.[7]

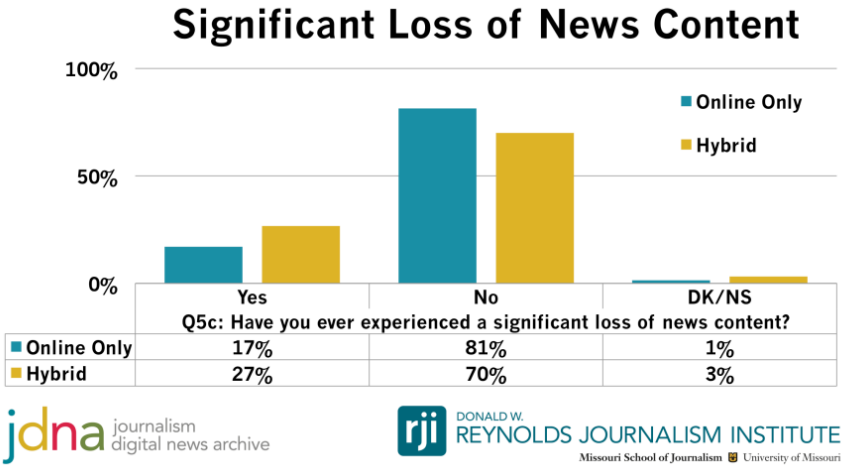

Figure 3: Survey to students and staff of the Missouri Journalism Intitute. Source: IFLA News media Conference (Hamburg 2016) in Edward McCain's presentation *The New Missouri Method*.[8]

These experiences allowed him to conclude that it is of great importance building closer relations with the publishers. Because libraries are not the primary market for them, these should try to offer something attractive by way of some kind of retribution (for example backup). Furthermore, he considered that the restricted access to legal deposit and the limited infrastructure usually libraries count with are the most important barriers when thinking about this type of collection development. He mentioned that National Libraries play an important role in the preservation and access to newspaper, while research libraries should support them with research that appoints to support this task.

[7] Center for Research Libraries, 'Focus on Global Resources', <http://www.crl.edu/focus/article/6318> (4 April 2016).

[8] D. Carner & E. McCain, 'The new Missouri Method', < http://blogs.sub.uni-hamburg.de/ifla-newsmedia/wp-content/uploads/2016/04/McCain-%E2%80%93-Carner-Teaching-Personal-Digital-Archiving-to-Journalism-Students-Slides.pdf> (23 May, 2016).

As it has been described in the beginning of this paper, understanding born digital news content lifecycle is essential in order to manage its further treatments for being collected, preserved and to provide access to this type of content. Libraries by themselves, as it could be perceived by the collected data, cannot fulfil this task in part because of the complexity of this lifecycle. That is why, one of the actions that could appoint to accomplish this situation, as Simon said, is to establish a closer relation with the publishers:

> Further, because the page image files, as output by the publisher, have only minimal metadata attached, by archiving them a library does not reap the benefits of the extensive annotation and coding of the content files that takes place within the editorial and digital asset management systems of the publishers, and which provides useful information about authorship, rights, provenance, and subject matter of the content. Perhaps the publishers, for instance, could export a uniform XML package at the issue or article level, perhaps captured on output from the pagination or editorial system. This is the moment in the lifecycle of the news item when the annotation is richest and the data most highly structured.[9]

Building a closer and more communicative relation with publishers could bring many benefits for the collection and preservation of born digital news content. One of these benefits could be regarding the process of acquisition. If news content is directly transferred from the publisher to the library, instead than being harvested from a web site, these could be archived at the most rich level of metadata possible, that then it would be lost when being uploaded to the web site. It would be just a matter of communication and setting strategy. The articles and their metadata would not require extra work from the publishers, because they own the material just as it is needed. All these different possibilities of formats in which one publication is being shaped let us make us the question of what should be consider the ideal publication format or level of capture when collecting born digital newspapers. Nowadays, most of the institutions consulted are dealing with web archiving as one of the most frequent collecting methods. This other perspective let us think in how the news production system could be taken advantage of in order to obtain the more contained source of other. That is why, beside the already mentioned and more frequently used formats (web archiving, PDF, RSS feeds, snapshots and XML), databases could also be considered as an ideal candidate for being collected. Additionally, in another report done by the same institution (CRL) tells it may be a better strategy to work with major news organizations and concentrate on capturing articles and other categories of discrete news objects, rather than entire sites.[10] Actually, this have been applied in the selection criteria of the legal deposit act of The National Library of Sweden[11]. This decision appoints to the safety of the metadata but also as a way to deal with the constant updates digital material goes through during one day, particularly news content. The following paragraph summarizes their selection criteria:

> The starting point was that the web pages and similar dynamic material should not be included, but only unchanging electronic documents or more precisely "a defined unit

[9] Center for Research Libraries, 'Preserving news in the digital environment: Mapping the newspaper industry in transition', April 27, 2011, p.4.
<https://www.crl.edu/sites/default/files/reports/LCreport_final.pdf > (7 March 2016).

[10] Center for Research Libraries, 'Preserving news in the digital environment: outline for an agenda for North American libraries', June 20, 2013. <https://www.crl.edu/sites/default/files/d6/attachments/tg/Section%203%20brief%20rev%202013.pdf> (7 March 2016)

[11] National Library of Sweden, 'Legal Deposit of Electronic Materials in Sweden', http://www.kb.se/dokument/Pliktleverans/Eplikt_enskilda_eng140917.pdf (18 April 2016).

> of electronic materials with text, sound or image that has a predetermined content intended to be presented at each use".[12]

However, part of the news suppliers where not satisfied with this plan, they considered it was "not practical and above all economically indefensible". They explained it was hard to pick what is to be sent on a daily basis because it is increasingly common for an article produced for print to be changed when it is published on the Internet.[13] Therefore, part of the curating decisions is being translated to the news producers, and this responsibility could turn to be problematic at some level. That is why it is extremely important to think this relation as a giving and gaining system. In which the effort of one party would be reattributed at some point by the other, beyond the legal aspect.

2.1.2. Dodging the Memory Hole

The Donald W. Reynolds Journalism Institute of the University of Missouri organizes some activities regarding the necessity to archive digital news: Dodging the Memory Hole. During one of these meetings[14] there was an interesting conversation between news executives:

> A roundtable discussion of news executives also revealed opportunities to engage in new types of relationships with the creators of news. Particularly, opening a dialog with the maintainers of content management systems that are used in newsrooms could make the transfer of content out of those systems more predictable and archivable.[15]

Other important point of the discussion was that one of the most difficult challenges is to extract the information in an archive and made it available to everyone. It is important to take into account that nowadays, because of the metadata embedded in content, search algorithm are getting more sophisticated. Then, it is essential to provide metadata because journalists, researchers or people in general will use data mining as a main scope when doing a search through an archive. All the news executives present in the meeting were conscious about the importance of having an assessment or establishing partnerships with institutions as libraries or archives, because these have the expertise to determine where the content go, what goes on it and who should have access. The importance of a strong relationship with their local library or with any institution that could give them assessment for the preservation and access to their archives. Then, flat PDF would no longer be relevant to collect, but XML rich in data would be. That is why, they mentioned it would be important to follow the right process for embedded metadata, for keeping it all of this and getting it back. Most of them agreed that setting standards or an institutional solution would work, because there is not a factual solution for setting an ideal agreement, a sort of partnership that benefits everybody:

> We have clear data that if content is not captured from the web soon after its creation, it is at risk. Which brings me to where I think our main challenge is with collecting born-digital news: library acquisition policies and practices. Libraries collect the majority of their content by buying something–a newspaper subscription, a standing

---

[12] P. Nilsson, *Collecting bits and pieces – the development of methods for handling e-legal deposit of on-line material at The National Library of Sweden*, (IFLA Conference, August 2014, p.6.) http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-nilsson-en.pdf > (11 April 2016).

[13] Ibid., p.7

[14] *Dodging the Memory Hole: Saving Born-Digital News Content*, held at RJI on Nov. 10–11, 2014.. <https://www.youtube.com/watch?v=IN-TXReeayY> (29 March 2016).

[15] Library of Congress, 'Dodging the memory hole: Collaboration to save the news', <https://blogs.loc.gov/digitalpreservation/2014/12/dodging-the-memory-hole-collaborations-to-save-the-news/?loclr=blogsig> (28 April 2016).

> order for a serial publication, a package of titles from a publisher, an access license from an aggregator, etc. The news content that's available for purchase and printed in a newspaper is a small subset of the content that's created and available online. Videos, interactive graphs, comments and other user-generated data are almost exclusively available online. The absence of an acquisition stream for this content puts it at risk of being lost to future library and archives users.[16]

Edward McCain, part librarian, part journalist, and founder of the "Dodging the Memory Hole" talked according to his experience, which is mainly related with the current situation in some regions of the United States. In his opinion, the biggest challenge regarding collecting born digital newspapers has to be with electronic deposit copyright legislation. Content is owned by the publishers, and usually they have other priorities before that working for long term preservation. He believes that the ideal situation when negotiating with publishers would be with the intervention of a third party. He puts the situation in these terms: Publishers have the necessity to make a profit with their content. Memory institutions (libraries, archives, museums) are in charge to keep the content, and are mission driven. The third part should be a nonprofit mediator, who works for the good of the other two parties. A kind of cooperative to put the content together, a centralized platform that could work in partnership with a public library for example. Currently, there are news banks (example: Lexis Nexis) that could fit in a similar model as the mentioned, but the problem with these kind of organizations is that these don't work for small newspapers.

McCain's model with Knight Foundation

Is a model in which small newspapers could work in partnership with public libraries in order to preserve and give access to born digital news content the best way possible. The main objective is to build a support network based on the building of trust that they can work together. Make comfortable the newspaper community in order to make them include preservation as part of their priorities. If they begin considering information for 100 small towns, with this start it would be a valuable corpus of information. The relevance will be put in the connection in between the parties (university libraries, newspaper and public libraries), but also the connection in between the communities. In his opinion, nowadays, when there is not a unique way of consuming digital content, everyone adapts it to their own devices, channels, etc. That is why collecting XML will allow a more rich data for future research (data mining) and also these are built by an algorithm, which could be apply certain speed of capture for special events (e.g.: Brussels attack) or just leave it as it is when nothing relevant is happening. As a sort of closing opinion, McCain considers that when dealing with born digital newspaper collections the main problem is not technical, but more about resources. Who is going to pay?, how much is going to cost? Furthermore, he think maybe is not so effective to have very defined standards, but to continue being flexible in this approach. Even though, currently the Web Archiving model is ok, this is not ready to use it efficiently. In the case of born digital news content, it does not take advantage with the production system. That is why he believes in the importance of being focus in content management.[17]

[16] Library of Congress, 'Dodging the memory hole: Saving Digital news', <https://blogs.loc.gov/digitalpreservation/2015/06/dodge-that-memory-hole-saving-digital-news/ > (28 April 2016).

[17] News Challenge, 'Knights news', <https://www.newschallenge.org/challenge/how-might-libraries-serve-21st-century-information-needs/submissions/helping-local-communities-curate-their-collective-cultural-heritage> (12 April 2016)

2. Conclusions

Library curators have distinguished because they have to make constantly decisions regarding selection. Born digital newspaper will demand much more decision making because the process of collecting digital content requires much more procedures, time and technologies rather than when collecting print. Also, nowadays with the rise of Web 2.0 much more material is being published, so more relevant information could be positioned as a subject of interested for being collected. Selection of titles or websites is not the main barrier when making a selection, but also the frequency of capture and the format (PDF, XML, Snapshot, etc.) of preservation. Even though it is intended to harvest the most complete scope possible, there are not technologies and resources capable to do it all. The digital world moves faster than us, so at some point we have to decide what do we want to loose and what do we want to preserve. And not only about the content but also about format. Digital content is not plain but dynamic. The text is displayed in diverse formats, sometimes appears or disappeared, it has images, videos, sounds and movement. What about accuracy and completeness regarding the consumption experience? Based on the experience of the institutions consulted, preserving the same layout will not be always possible.

As it has been analyzed, it is possible to conclude that meaningful partnerships and negotiations since the beginning will be a meaningful contribution for the collection development. Nowadays it seems to be mandatory for memory institutions to act in negotiations beyond their cultural ethic values or mission goals when negotiating with publishers. They have to have something to offer publishers but they should be aware of it and expose it during negotiations, because this are profit orientated organizations. Commonly a library or an archival institution gets involve with the relevant parties after a document is published. It could be very convenient and also save the institution from future problems if some sort of standards could be established in advance (e.g.: regarding metadata). Other aspect to take into account is the importance of gaining a deep understanding of the news content lifecycle in order to preserve documents rich in metadata. That work is already done and it could be easily lost, but at the same time easily gain if we collect the material during the right part of its cycle. In this same perspective, thinking about collecting databases from the Editorial producers could be one possibility in order to cover a relevant scope through a more practical and direct method.
In order to close, the following conclusions mentioned in the different panels during the event *Dodging the memory Hole II* are very illustrative and relevant suggestions to take into account when dealing with born digital newspaper collections: ‘The necessity to educate news creators about the idea that preservation is also part of the creative process; the importance to be in public policy agenda; and to share the value about preservation’.

Acknowledgments

Frederick Zarndt (IFLA News Media), James Simon (CRL), Edward McCain (Missouri Journalism Institute), National Library of Denmark, Swiss National Library, National Library of South Africa, National Library of Luxembourg, British Library, Library of Congress, National Library of Australia, National Library of Croatia, National Library of Sweden, Biblioteque nationale de France, National Library of Germany, University Library Kentucky and University Library of Texas.

References

Collection development standards per institution

National Library of Denmark
http://www.kb.dk/en/kb/service/pligtaflevering-ISSN/lov.html

Swiss National Library
file://fs-srv-p100/users$/EVi030/Downloads/FAQ_e-Helvetica_Deposit_e.pdf
https://www.nb.admin.ch/nb_professionnel/01693/01696/01707/index.html?lang=en

National Library of South Africa
http://www.nlsa.ac.za/downloads/legaldep.pdf

British library
http://www.webarchive.org.uk/ukwa/info/about#what_uk_archive

Library of Congress
https://www.loc.gov/acq/devpol/neu.pdf

National Library of Australia
http://pandora.nla.gov.au/selectionguidelines.html#newspapers

National Library of Sweden
http://www.kb.se/dokument/Pliktleverans/Eplikt_enskilda_eng140917.pdf

National Library of France
http://www.bnf.fr/en/professionals/digital_legal_deposit/a.digital_legal_deposit_web_archiving.html

National Library of Germany
http://www.dnb.de/EN/Netzpublikationen/netzpublikationen_node.html

Kentucky University Library
https://kdnp.uky.edu/project/

Texas University Library
http://texashistory.unt.edu/explore/collections/TDNP/

Bibliography

Alverson, J., Leetaru, K., McCargar, V., Ondracek, K., Simon, J., and Reilly, B., 'Preserving news in the digital environment: Mapping the newspaper industry in transition', Report from the Center for Research Libraries (CRL), April 27, 2011.

Barth, A., *The New Republic*, Volume 108. (Washington D.C.: Republic Publishing Company, 1943), p. 677.

Carner, D., McCain, E., Zarndt, F., 'An international survey of born digital legal deposit policies and practices' IFLA Conference Paper, 2015.
'Missing links: The digital news preservation discontinuity 'IFLA Conference Paper, August 8 2014.

Nilsson, P., 'Collecting bits and pieces – the development of methods for handling e-legal deposit of on-line news material at The national Library of Sweden', IFLA Conference paper, August 16, 2014.
http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-nilsson-en.pdf

Oury, C., 'All we need is news preservation: harvesting digital newspapers at the Bibliothèque nationale de France', IFLA Conference Paper, August 11, 2014, <http://library.ifla.org/1026/1/170-oury-en.pdf>

Potter, A., 'Dodge that Memory Hole: Saving Digital News', The Signal Digital Preservation, 02 June, 2015,
http://blogs.loc.gov/digitalpreservation/2015/06/dodge- that-memory-hole-saving-digital-news/

Reilly, F., 'A New template for the preservation of Electronic news', IFLA Conference Paper, August 16, 2014,
<http://www.ifla.org/files/assets/newspapers/Geneva_2014/s6-reilly-en.pdf>
'Preserving news in the digital environment: outline for an agenda for North American Libraries', (An outline for a libraries agenda, based on the report, "Mapping the Newspaper Industry in Transition" from the Center for Research Libraries, June, 2013).

Skinner, K., Schultz, M., 'Guidelines for Digital Preservation Readiness', Educopia Institute, March 4, 2014.

Websites

Center for Research Libraries,
'Focus on Global Resources', <http://www.crl.edu/focus/article/9559> (4 April 2016).
'Preserving news in the digital environment: outline for an agenda for North American libraries', June 20, 2013.
<https://www.crl.edu/sites/default/files/d6/attachments/tg/Section%203%20brief%20rev%202013.pdf> (7 March 2016).

'The "State of the Art": A Comparative Analysis of newspaper Digitization to Date', , April 10, 2015.

Dodging the Memory Hole: Saving Born-Digital News Content, held at RJI on Nov. 10–11, 2014..
<https://www.youtube.com/watch?v=IN-TXReeayY> (29 March 2016).

Federal Agency Digitization Guidelines Initiative, 'Born Digital'
<http://www.digitizationguidelines.gov/term.php?term=borndigital> (5 April 2016).

International Internet Preservation Consortium, 'General Assembly ',
<http://netpreserve.org/general-assembly/ga2015-schedule> (27 April 2016)

Library of Congress, 'Dodging the memory hole: Collaboration to save the news',
<https://blogs.loc.gov/digitalpreservation/2014/12/dodging-the-memory-hole-collaborations-to-save-the-news/?loclr=blogsig> (28 April 2016).
'Dodging the memory hole: Saving Digital news',
<https://blogs.loc.gov/digitalpreservation/2015/06/dodge-that-memory-hole-saving-digital-news/ > (28 April 2016).

News Challenge, 'Knights news', <https://www.newschallenge.org/challenge/how-might-libraries-serve-21st-century-information-needs/submissions/helping-local-communities-curate-their-collective-cultural-heritage> (12 April 2016)

The Atlantic, 'Raiders of the lost Web',
<http://www.theatlantic.com/technology/archive/2015/10/raiders-of-the-lost-web/409210/> (5 April 2016).

The Internet Archive, 'Heritrix', <http://crawler.archive.org/index.html> (27 April 2016).
'Way back Machine', <http://archive.org/web/>(27 April 2016).

Appendix

3. Results

National Library of Denmark

The National Library of Denmark only collects paper. This year they will start with a pilot project in which they will be collecting 4 titles PDF s during 3 months. These will be uploaded through an aggregator system by the publishers. For negotiations with publishers they negotiate through a mediator organization called Copydan[18], and usually negotiations are easy to deal. Certainly. They have finished the pilot and are now on the verge of asking vendors to provide quotes for delivering 58 titles on pdf. It is the daily pdf's of the major Danish newspapers that they want. Even though National Library of Denmark currently do not collect digital newspaper, they have a Web Archive. This has a very restricted access because they believe the information is very personal so it is only accessible for scientists and academics with a special permission.

Swiss National Library

The Swiss National Library is going through a similar situation. They only have a few snapshots of Swiss newspaper web sites in their Web Archive Switzerland, but they mentioned that they do not harvest them in the frequency a newspaper would require. They are planning to start a project on e-papers, but because of other priorities they had to postpone this project.

National Library of South Africa

The National Library of South Africa is in the same position as Swiss, they only collect analog newspapers. They said electronic resources are still a challenge for them because of resources.

National Library of Luxembourg

The National Library of Luxembourg only collects paper as well, but they are planning to harvest some web sites of news publishers. Meanwhile, they asked print newspaper publishers to handle them their PDF versions for preservation instead of re scanning the paper versions. Even though they are not currently collecting digital newspaper, they have program two pilot project plans for this year (2016). One is about new agreements with publishers (regarding PDF paper versions for preservation). The other is about web harvesting and for hopefully making an agreement with the publishers to give access to this material, not only behind the desk, as it is common for the legal deposit material.

British Library

Even though the British Library collects preferably print newspaper, under their legal deposit law[19], they also collect digital news content in news website. They consider news websites the ones which organizations define as one, for example community websites, but they do not

[18] Copydan, <http://www.copydan.dk/> (27 April 2016).

[19] Legislation UK Government, 'Legal Deposit Libraries Act 2003', <http://www.legislation.gov.uk/ukpga/2003/28/contents> (18 April 2016).

collect blogs, personal web sites, or any web site which is built on user generated content. They do a daily or weekly capture for news website, and it is very important for BL to archive the news website the same day of publication. They do all the harvesting work, so they do not have to negotiate with publishers in order to accomplish this. Even though, this is a benefit of harvesting web sites, this it is also a difficulty when making decisions about selection or towards the cover scope, because they do not have a recognized ISSN[20].

The British Library use a special curatorial tool in order to cover the news web sites scope intended for harvesting. They are still catching up with the system, so even though they actually do not cover the intended scope, they have covered around 1800 news website. The harvesting frequency is randomly once a day. One of the problems they experienced when using harvesting technologies was that in some websites they could capture videos but they could not play them back after. Media is considered part of the scope of news websites but not the main core as it is content. That is why they do not capture websites which are only TV or radio. Meanwhile, the British Library, gives access to digital news (through their main web archive) on site, only in legal deposit libraries. A small number are free of copyright and available through remote access.

Library of Congress

The Library of Congress collects news websites as well. Their selection criteria is based on material which do not has a traditional newspaper analog version, and also having in mind websites which preferably present new stories and interesting thoughts about general news. They do not have to pay licenses, they just notify in advance to the publishers that they are going to crawl their websites. The Library of Congress frequency of capture is variable, usually is once a week or they use RSS feeds (Huffington Post). They do not have a current plan of news preservation but they are willing to preserve these not only for current access, but for the future. Library of Congress do not give access yet, but they are willing to do that, and this will be on site.

National Library of Australia

The National library of Australia is actively archiving news web sites. They have been collecting daily the Sydney Morning Herald, one of the major Australian newspapers, since June 1st, 2009. Furthermore, with the introduction of electronic legal deposit law, they started archiving three major news sites: The Guardian Australia, The Canberra Times and the news.com.au website. Unlike the Sydney Morning Herald, they do not have a copyright license for these, so they will not be accessible outside the Library. The general selection criteria for digital news content would be: Sites for newspapers that mainly duplicate the information provided in print will not be selected for preservation; sites providing selected features and stories to promote print newspapers or others electronic news services are not considered suitable for preservation by the National Library. Snapshots of certain examples may be taken to illustrate the use made of the Internet; dial-up, commercial services are not selected; newspapers available online only will be assessed against the guidelines and preserved if they meet standards as authority, quality and originality of content[21]. Meanwhile, the National Library of Australia harvests each day to a level of one hop. They use the HTTrack harvester. They capture the front page and the content lined from that front page.

[20] International Standard Serial Number.

[21] PANDORA, 'Selection guidelines for archiving and preservation by the National Library of Australia', <http://pandora.nla.gov.au/selectionguidelines.html#othernatsig> (28 April, 2016)

This procedure is schedule to run overnight using a scheduling mechanism, but during week days they usually replace the overnight harvest with a manually initiated harvest around 10:00 a.m. to capture the main morning news content. The weekend harvests retain the overnight harvest. Each harvest does require a quick quality assurance process fix that get the banners displaying properly. The National library of Australia has a different access system. For example, a couple of years ago the Sydney Morning Herald introduced limited viewing of the online newspaper, a limit of 30 free views per month. The collection's page is sort alphabetically, not chronologically, so unfortunately, the dates are all over the page. They do a new collection page for each month to make the display pages manageable. The other newspapers, with no copyright license are collected, preserved but not accessible yet. The following text is located beside these titles:

This resource has been restricted in Pandora. Access to this item is restricted for 70 years from the date of archiving. During this period it is only accessible on restricted computers within the National Library's Special Materials Reading Room.

National Library of Croatia - University Library in Zagreb

The National and University Library in Zagreb (Croatia) collects both print and born digital news. The Croatian Web Archive (HAW)[22] collects and archives online newspapers, new portals, local news, blogs with their in-house software for selective archiving. Digital content is part of their collection since the early 1990s. Since then, new systems, procedures and policies have been developed and adopted. Furthermore, the education and training of the staff who work with digital resources is permanent. The frequency of harvesting depends on the structure of the site. Usually, they are archived daily, weekly or once a month. They are archiving two major Croatian online news portals daily (Večernji.hr, Jutarnji.hr). They guarantee that that relation with the publishers are good, this is mediated by the ISSN Centre for Croatia. They do not collect content that is under pay walls because of the technical issues involve. They try to harvest and preserve news websites to be as similar as possible to the original. Other experienced institution doing digital news preservation is The National Library of Croatia. This has three different approaches for harvesting procedures: Selective harvesting/archiving (in house software developed by their partner University Computing Center-Srce); National domain harvesting (Heritrix[23], Wayback machine[24]); and Thematic harvesting (Heritrix, Wayback Machine). They told us that they try to collect as much items as possible, but they do not collect any resource more than once a day. They perform mostly manually quality control, but they have several modules for system monitoring. For the selective archive they use Checking large archived copies is a tool that signals staff which archived copies exceeds size of 500 MB. Also they apply daily report serves for possible duplicates notifies of similarly collected samples. For example, when the last two copies are similar in more than 80% it is likely that resources are online but not updated. Finally, the automatic monthly report is a tool for checking the availability of the resource at its live URL. For the Domain harvesting, because the content is massive, quality assurance is more complicated and in a lower quality. After each domain harvesting, the staff manually checks a sample of 100 archived websites. Regarding preservation for the future, they collect HTML web sites as they are, they respect the content and layout as complete as possible. They experienced different kind of problems regarding the type of harvesting. For example, during the selective harvesting the problems usually are caused by flash, JavaScript and Drupal,

[22] Croatian Web Archive, <http://haw.nsk.hr/en> (28 April 2016).
[23] The Internet Archive, 'Heritrix', <http://crawler.archive.org/index.html> (27 April 2016).
[24] Ibid., 'Way back Machine', <http://archive.org/web/>(27 April 2016).

password protected resources, videos on YouTube, or similar streaming audio or video files. On the other hand, for the domain harvesting, the problems are caused by poorly configures websites. On the contrary, The National Library of Croatia Archive is publicly available, and it includes remote access. Some publishers demand that the content could only be accessed within the premises of the National and University Library of Zagreb.

Regarding presentation of the digital content, they have different approaches for different types of harvesting. For Selective harvesting, full-text indexing has been implemented, search may be performed by any word in the title, URL, keywords, etc. Advance search and browsing through subject categories is also possible. For harvesting the national domain, the harvested copies are available through the Wayback Machine with interface in Croatian language. These could be accessed by entering the correct URL and choosing the year and date. For thematic Harvesting, these copies could be searched through a particular thematic collection. All these copies are searchable through the general catalog as well.

National Library of Sweden

The National Library of Sweden, most of the digital news content they archive is under their Legal Deposit Act (2012) , so it is attached by the legislation criteria, which in general terms do not consider entire websites but separated articles. Documents that are subject to Swedish legal deposit are: publicly accessible in Sweden on electronic networks: available exclusively in web, not analog form; documents that have a defined electronic format and are comprised of any combination of text, sound, and image; documents with an abiding, not a transient form altered each and every time they are accessed; and that should be regarded as Swedish. They also collect other newspapers out the legislation. Their relation with the publishers is basically regulated by the legislation, so they do not have to deal directly with them in order to archive news. They believe that the biggest challenge regarding this task would be: Developing methods for the archiving system and interpretation of legislation (abstract) and to decide which type of documents is subject to be archived in the legal deposit, because the boundaries are not so clear. They think that the best way to manage the relation with the publishers would be through an aggregator that sets everything. Publishers have interested to gain profits because of their publications, so it is hard to make them understood and work for preservation. The National Library of Sweden have experienced that one of the most efficient methods to capture born digital news content is through RSS feeds updates. They also do this through FTP (also some material that could not be properly uploaded through RSS feeds is done through this canal, and they have a webpage design for small publishers to upload their material. Even though, The National Library of Sweden has been doing web harvesting since 1990's, and all is catalogued (National Union Catalogue–Libris) they are not able to give access to this material because of the national legislation.

National Library of France

The National Library of France has a digital legal deposit which does not replace the printed legal deposit. The BnF collects both the printed edition and the digital edition of a same newspaper. Currently for example it collects the daily newspaper *Le Monde* in its printed publishing and its website www.lemonde.fr. They also the collect of about 20 titles of paid newspapers, mainly daily local press. For example, the BnF collects the *title Ouest France* in its printed publishing, the website www.ouest-france.com and the paid PDF publishing. Sometimes librarians buy both printed and digital version if it is necessary: it depends on the consultation requirements. The acquisitions can double the digital legal deposit for the same

reason. Their selection criteria is part of the documentary policy for web archiving in general. Therefore, the department in charge of the press includes the born digital news. The publications involved in the digital collection of daily newspapers are political and general information press, whether local or national, specialized, or whether pure players (eg rue89.com, mediapart.fr). It concerns 3 types of press: information portal, national press and local press. The selection vary depending on the type of press: regarding information portals, only the portals with more audience (based on the ranking OJD) and more content are selected in this collection. A dozen portals are concerned. All titles of the printed national press with web site are selected. Plus other national titles but the list is not intended to be exhaustive (pure players, web sites, according to OJD ranking). For the local press, the titles referenced on the specific site PQR66[25] are selected. Added to this sample, the titles present in the paid press collection, as well as titles overseas whose content is consistent. To complete the daily press collection, the department in charge of the press selects also: websites of political and general news requiring a frequency less important than daily, sites of press agencies which contents are rarely available for free, sites related to journalism as profession: professional associations, information centers, observatories and blogs of journalists. These different types of websites are distinguished in our selection tool by special themes and keywords, and different parameters of crawl. The other collection departments select also news websites about their disciplines. And the legal deposit service can crawl some news websites during the broad crawls on the French domains.

For the BnF a digital newspaper is very composite. It's a collection of texts inside a website linked with images, videos, blogs, etc. It can be also a platform of information which give a view of articles from several media. Actually the most part of newspapers are linked to others platforms.

The crawling system, for the daily press collection BnF uses special parameters for the crawler. The crawler copies the first page (URL) and one page linked to this first page. It permits to collect the most articles and others elements of contextualization as some commentaries, videos, blogs, etc. For the other titles, the parameter frequency and depth can be different. A newspaper website in BnF web collection depends on the parameters of crawl.
Because they collect both digital and print versions of a newspaper, there is no deal with completeness between printed and digital version. Librarians try to identify newspapers that change support and to integrate them in the digital collection. And it is possible to collect in parallel the two versions of a same title when the articles aren't both in the version.

Regarding the relation with publishers, the BnF needn't permission to collect websites but the crawler is identified when a website is collected. Thereby the publishers can contact the BnF if the BnF crawler disturb the operation of their websites. That is why the relation with the publishing community is limited. For the paid press, it is necessary to contact the publishers to have access to their titles. The publishers are subject to legal deposit law even for the paid press. They have to give the elements to connect the newspapers or the articles. The printed legal deposit and the acquisitions are in relation with the persons in charge of the subscriptions. For the legal deposit, the contacts with the publishers are good: they agree with the principle of the legal deposit in general, including the digital legal deposit. But it is not easy for them and for the library to find a convenient solution to collect their titles. The BnF

[25] E-marketing, PQR66, <http://www.e-marketing.fr/Definitions-Glossaire/PQR-66-242803.html> (28 April 2016)

contact them for each title and tries to find a technical solution to collect their paid website. Often the BnF deals with the technical service of the newspapers or the marketing service for the authentication. And sometimes the publishers subcontract the website publishing : the BnF deals in that case with the company. This technical collect is very unstable : the collect can stop if there is a little change in the URLs of the websites, in the way to connect the websites, in the subscriptions. That is why only twenty paid titles are collected. And sometimes BnF postpones the collection of some titles because the technical solutions are very complicated. Concerning accuracy, for the legal deposit, the way the BnF collects the digital newspapers and gives access to the collection permits to restitute the same layout of the living web and the links inside the sites. It is impossible to collect all the elements and to reproduce in the archives all the reading experience but it gives a good picture of how the websites could be used. For example, the web archives show that it is possible to add some commentaries or share information on social networks even if it isn't possible to do it concretely in the archives or see all the commentaries. Furthermore, to help the future researchers, librarians try to document their selection criteria to explain what their collections contain.

For acquisitions, the digital newspapers take part from databases. The access is completely different from reading experience. The research is enriched by a full text indexation for example. The researcher can access information article by article.

For long-term preservation BnF don't have any agreement to preserve the content over time except the legal deposit law. They can collect websites without the agreement of the publishers. The crawler is clearly identified and webmasters can access a page of information about the digital legal deposit[26]. The publishers are informed that BnF can collect several time their online newspapers. On the other hand, for the acquisitions, there are agreements with the publishers or with the aggregators of newspapers. These agreements contain conditions about the offer, the retention period, the way to access the collection, but there isn't a standard agreement: it depends of the digital offer. The legal department checks all the agreements.

Because of the legal deposit legislation, born digital content is collected by the crawler Heritrix, in different ways : focused crawls, events crawls (eg elections) and broad crawls. The paid press is collected by the robot too even by FTP. The publisher deposits the files by FTP and the BnF crawls the server FTP with Heritrix. For the digital legal deposit, BnF uses a daily schedule for the daily news (about 100 titles) and other frequencies (weekly, monthly, biannual and annual) according to the information refresh on the sites. If a title isn't free, the selector chooses an annual frequency. If the title is free, but with few content a biannual frequency can be enough. During special events, like elections, it is possible to adapt others frequencies (eg several times per day).They said it is impossible to capture all the digital information. Even though that impossibility, BnF tries to find a balance between the quality of the content collected, the technical obligations and the period of time during which the information is available. For example, the first tests used the parameter one page plus two clicks but the crawl didn't finish in time: the collect must be finished in 23 hours. Since then, the BnF has given a limit of time and of budget (in URLs) to the harvester. Even though, the level of capture (the depth) is important and the moment the crawl starts too. The robot captures less or more information according the publishing model of the titles. We need to

[26] Biblioteque nationale de France, <http://www.bnf.fr/en/tools/a.dl_web_capture_robot_eng.html> (11 April 2016).

realize the collect of the daily press in the same harvest definitions (one for the daily press and one for the paid press) to use fewer engines and simplify the process but it will be better to adapt the collect to each title. The digital legal deposit produces ARCs until 2014 and since 2014 WARCs as its international web partners. The WARC format is a standardized format to ISO (ISO 28500: 2009) since 2009. It's impossible to capture all the versions. BnF chose the richest version and sometimes there's no choice: the robot copies one version online. The daily newspapers are harvested in the same harvest definitions which start always at the same hour. The harvest definition on the paid press starts at 2pm and the harvest definition about news press starts at 10am. The legal deposit department performs a quality control. But it's very different from the quality control of the printed press. For the printed version, the control is made daily issue by issue. For the digital version, quality assurance procedures are performed daily on a representative sample of the collection. There are two kinds of quality controls: statistical quality control: the reports and metrics of each crawl are analyzed in order to identify if something went wrong: e.g. if too few or too many URLs have been collected for a website; and visual quality control: the archived website is visually checked; against its on line equivalent when possible.

The BnF preserve all the URLs on the news websites. They reflect the way the information is published by the reporters but also how it's used by the readers, how they react with the blogs, the commentaries, etc. The harvesting technologies don't work with flash websites. Moreover it's difficult to subscribe some websites. The other most important problem come from https, javascript and flash format. According to the version of the software environment, https behave differently: sometimes Heritrix can harvest and sometimes it can't. Heritrix interprets the javascripts and creates wrong URLs which disturb the logs of the publishers. Heritrix can't harvest flash format at all. Access in the National Library of France is restricted for the legal deposit by the law to the interior of the "research" levels of the BnF. There's no exception between the printed legal deposit collection and the digital collection.

For the acquisitions (database), the readers can access in all the BnF's reading books. It depends on the agreement with the publisher. Recently the BnF gives access of the web archives in regional libraries, partners of the BnF, with the same conditions. The readers can access the BnF web archives from their regional library. The researcher expectations are included in the criteria selection (developing crawls to preserve the traces of important nodes and networks; archiving the most popular sites and also those that break new ground…) and in the practical way of accessing the collections. The access tool takes into account the hypertextual dimension of the internet. The Wayback Machine, used in the main consultation interface of the BnF web archives, allows researchers to navigate within the archives as they would have done on the live web. This spatial navigation is enhanced by a temporal exploration. Starting from a given site, it is possible to go back in time and analyzed its successive transformations. This kind of indexing presupposes that, to discover a site, its address is already known. To mitigate this problem, a full text indexing would allow users to search for pages and files depending on their textual content using keywords. Due to the huge volume of the collections (21 billion files, 470 TB of data), and the difficulty of handling multiple temporal layers, this full-text indexing has so far only been performed in a very limited fashion, which represents a real obstacle to the use of the collections. A research project on a small part of the web archives has done during 2015 but it's not a public service. In parallel, in order to overcome the difficulty for the uninitiated of distinguishing what is proposed in the web archives from that directly accessible in the web, the BnF has put in place a presentation in the form of "Guided Tours", conceived as flagship products on subjects which can be easily understood and which are representative of national,

political and cultural memory, where the phenomenon of disappearance is clearly apparent. There's one special guided tour on the digital press collection concerning the daily news collection and the paid press collection. For the legal deposit, to facilitate the access of the digital press collection and to encourage the researchers to consult the web archives, the websites described in the press guided tour are described in the General Catalogue. It represents about a hundred titles. There's a link between the exemplary in the General Catalogue and the web archives access tool which works provided that the reader is onsite at the library. The reader can find a record for each version : one record for the printed publishing, one record for the website, one record for the PDF format (when it's crawled). But it isn't possible for the librarians to describe all the harvested titles in the General catalogue. All the staff of the collection department in charge of the press was formed to this specific collection in the web archives. The news collection is also presented to the library staff in charge of the bibliographic information in the reading room. Afterwards they can promote the digital news collections to the readers during their research. The publication of the guided tour was accompanied by an information campaign on the BnF public website[27] and the internal website dedicated to professionals in the form of short articles. The digital legal deposit promotes also the news collection to the librarian partners in France and in the consortium of the web archiving by email or during seminars. The collection policy is affected because web is an important component of the documents. Some documents migrate their support and some are born digital. The librarians try to capture these changes, all these "news" expression. But they needn't to choose between paper or digital as it was explained above.

National Library of Germany

The German National Library has been collecting e-paper editions of daily newspaper since 2010 under a legal basis. The legal deposit on 2006 has been extended to the collection of media works or "inmaterial form", online publications. "All commercial and non-commercial publishers in Germany are obliged to submit two (mandatory) copies of their works to the DNB. In this case of online publications, only one copy need to be provided."

In conjunction with a service provider, they developed an automated process in order to handle the large amounts of data. This is now put into routine and used to collect the editions of 930 daily newspapers including 18 Sunday newspapers form the publisher servers, to convert them into PDF/A format which is suitable for long-term preservation, and to give access in the catalogue and archive. Users can have access to these in the Library reading rooms.

One of the main reasons why the collection policy changed (from microfilming to e-paper acquisition) where practical and economic reason. When an e-paper version corresponds full to the printed edition it would be not necessary to microfilm. In addition, an automated workflow to collect e-papers is less expensive than dealing with microfilms. They only collect e-papers provided as a download PDF. According to them the advantages of using this format are the following: it is suitable for preserving text, image and layout, and it could be converted in PDF/A 1 b format for long term preservation; it combines information and layout (unlike ePub or other text-focused formats); it is possible a full text search in one or over all e-paper issues and further automatic processing is possible; it ensures up-to-date

[27] Biblioteque nationale de France, <http://www.bnf.fr/fr/collections_et_services/livre_presse_medias/a.archives_internet.html> (April 11 2016).

availability; the document quality is far better than the quality of microfilms. Nevertheless, this format present a few disadvantages: preservation of e-papers is more complex than of microfilm because the big amount and diversity of pictures, fonts, text-parts and the format itself is quite complex; the establishment of a particular workflow for acquisition, collecting, cataloging, archiving and preservation of e-papers is necessary. The task of collecting e-papers is complex because configuration and distribution differs a lot among publishers. This situation will demand a lot of work to the newspaper publishers as well as to the Library.

The access is onsite in the libraries of Leipzig, Frankfurt and main. The searches can now be make in individual editions of e-papers or access them directly from the catalogue. These cannot be offered to external users, nor to be saved or copy for copyright reasons The waiting period before an issue can be accessed is about one week for digital versions. They consider that the automatic collection of electronic editions, including all associated metadata, has considerably improved the bibliographic information about daily newspapers.

Until now they considered that the results of this project are convincing. Even though this automated workflow creates some work but much less than previous workflows would demand. This has enabled the DNB to expand newspaper collection within an acceptable budget. Based on their experience, they consider that for publishers and the legal deposit libraries it would be attractive to simplify the delivery procedures as much as possible. Their aim is that newspaper publishers are obliged to submit only one mandatory e-paper copy to the DNB and the DNB will ensure access to these in their legal deposit libraries. They believe a closer co-operation would be necessary.

Kentucky University Library

They collect both (paper and digital), but their primary focus at this time is on born-digital newspaper content. This allows them to skip any digitization necessary to make a paper copy keyword searchable online. Their selection criteria starts from the base that all the content must be from Kentucky, first and foremost. They cannot collect them all because many of the large dailies, and some of the small rural newspapers owned by outside (the state) conglomerates, will not allow them to harvest. These choose third=party vendors who monetize collections behind a paywall. Then, they are left to harvest and preserve those titles that care about the preservation of their collective history.

Their relationship with their state press association is quite good and with individual publishers vary. This goes back to being understaffed. If one doesn't have the time to devote to relationship building, it's hard to have one. Regarding to the main challenges when dealing with their publishing community they share an experience with the clipping service from whom they originally set up a harvesting script for Kentucky newspapers decided, quite hastily, last year to begin charging for "electricity and broadband per title." Their rate was unacceptable, so they dropped them from their harvesting schema. It has been a struggle to maintain up-to-date harvests from participating newspapers because they are sorely understaffed. Without a dedicated manager to oversee relationship build, technical support to the publishers, and harvesting workflows, it's virtually impossible to maintain. As a result, they are, and will continue to, suffer significant gaps in the historical record.

In consideration with accuracy (preserving the same layout, the same reading experience) they strive for this in their born-digital PDF harvests the same as they demand it of their print-t0-digital newspaper presentations. They have not yet crossed the bridge of HTML

based news content, so they can't speak to that beyond it would be ideal to present it as it was originally presented on the web to the users.

They have an standard agreement with publishers. It is a Deed of Gift from those publishers who are participating in our digital newspaper program.

They provide access to newspapers – born-digital contemporary and historic – through kdnp.uky.edu. Right now, these are materials they already had as a result of their participation in NDNP, other historic newspaper digitization projects, and the born-digital content harvested from the aforementioned clipping service. Any print-ready PDFs they gather at this time is done so manually. They have manage to cover the scope they intended to in their collection. Certainly, HTML content will be the most complicated to handle based on the linked layers and third party components such as video or advertising.

TIFFs are their preservation format. JP2s are generated from their NDNP content, and PDFs are used quite often for access and printing purposes.

Their materials are put into the KDNP (Internet Archive), but they also add them to their local repository.

When they devised a daily crawler for the clipping service's servers, they experienced no problems. The crawler ran at night when there was no one at work (in either location) and network use was at its lowest.

The materials at present are on the Internet Archive. The user interface is built on Blacklight hosted locally here at UK Libraries. Materials located in Chronicling America through the Library of Congress are all in the public domain. For copyright materials, they host them locally via KDNP. Users are free to quote from the newspapers, but published image use require permission from them and from the publisher. They do not allow hosting by third parties of copyrighted materials.

Texas University Library

They collect and preserve PDFs from the print master. The capturing frequency is once every hour. Before they were collecting microfilms, so PDFs appear to be the most efficient transition in order to proceed with their born digital newspaper collection.

Their relation with publishers is mediated with Texas press Association, they work together in a sort of partnership. They have an agreement for long term preservation. For access they have an agreement of 2 years of embargo, this works good for them because they spend a similar period of time for the whole collection procedure.

According to their experience, they consider that the main challenges in their task is the interaction with publishers. Because of the different interests of both parties (library and publishers) it is difficult to make them understand the relevance of preservation. In some occasions, they even were reticent with the idea of letting the library to preserve their contents. That is why they believe that a communication is the key issue when dealing with publishers.