# Text, data and link-mining in digital libraries: looking for the heritage gold

Emmanuelle Bermès

IFLA WLIC 2017 WROCŁAW

**Submitted on:** 21.11.2017

**Digital Humanities**
**https://dh-libraries.sciencesconf.org**

# Text, data and link-mining in digital libraries: looking for the heritage gold

**Emmanuelle Bermès**
Direction des services et des réseaux, Bibliothèque nationale de France, Paris, France.
E-mail address: Emmanuelle.bermes@bnf.fr



## Abstract:

*With the availability of new massive digital collections, innovative ways of exploring library data are emerging. Researchers are starting to investigate the use of powerful analysis tools that go beyond what the human eye can see, beyond what the human mind can process. Text and data mining techniques offer new opportunities for new types of research. Since a few years now, the BnF has seen its digital collections driving the interests of the early-adopters of new data management tools. These digital studies may be at the core of our users' practice in the future; they may become instrumental in defining what a national library is. That's why in 2016, the BnF started within its 4-year internal research programme a new project called CORPUS, aimed at designing a future service for providing access to digital corpora for researchers.*

**Keywords:** Text and data mining, digital humanities, digital scholarship, digital libraries.

The founding decree of the BnF, dated 1994, says that one of our core missions is to "provide access to collections for a wide audience". Such a target cannot be reached nowadays without providing access to digital material: the BnF, along with its many partners in France, has launched a major digitization initiative, leading to creation of Gallica[1], our digital library which holds 4 million items publicly available online and 4.5 million in its intra-muros version, displayed on site. Born-digital material has also been taken into account as a new type of legal deposit, with the start of web archiving first experimental, then at scale when the French Law on Intellectual Property Rights made it a requirement in 2006: our web archives collection now reaches almost 800 Terabytes. Even the analog collections lead to a significant increase in the amount of data available: our general catalogue holds 20 millions records, and counting.

With the availability of these new massive digital collections, new ways of exploring the data are starting to emerge. Researchers from a variety of academic fields, mainly in the Humanities but not only, are starting to investigate the use of powerful analysis tools that go beyond what the human eye can see, beyond what the human mind can process. Text and data mining techniques offer new

[1] http://gallica.bnf.fr

opportunities for new types of research: not only are they questioning the content of documents, but also how they fit in a wider documentary landscape, based on contextual or extracted data.

Since a few years now, the BnF has seen these new enquiries emerge among the academic community. Because our digital collections are so massive, they are driving the interests of the early-adopters of new data management tools. As librarians, we are also very curious of these new techniques and of the results they are providing. These digital studies may be at the core of our users' practice in the future; they may become instrumental in defining what is a library, and particularly a national library in charge of its country's heritage, in the forthcoming years.

In 2016, the BnF started within its 4-year internal research programme a new project called Corpus[2], aimed at designing a future service for providing access to digital corpora for researchers. The idea of this project was born when the librarians from the Technical Services unit at the BnF started struggling with repeated enquiries from researchers to participate in call for projects, whether at European or National level, as a data provider. The challenges in terms of legal framework, technical infrastructure, human resources, know-how, organisation and others were many. Moreover, each project was peculiar, thus demanding that BnF created a dedicated workflow for each of them. This way of working was not sustainable, as the amount of research being carried out on digital collections was increasing. Therefore, we decided to grant ourselves four years to experiment with actual research projects, explore the needs of our patrons and partners and then design a new service, dedicated to the study of digital corpora by scholars.

Today, as we have almost reached the middle of the Corpus project, we already have a good sense of what we would like to build. The BnF has confirmed its willingness to set up such a service by carving it as an objective in its 5-year performance contract signed with the Ministry of Culture. How to build it remains a major challenge.

## 1 THE CHALLENGE: DIGITAL COLLECTIONS AND WHAT THEY MAY (OR MAY NOT) TELL US IF WE MINE THEM

The scope of the Corpus project is threefold:

- Digitized material, produced from analog items such as books, serials, prints and photographs, maps, video and audio recordings on a variety of media and finally objects such as coins and medals, globes, masks and puppets etc. As mentioned above, the amount of those digitized items now reaches 4.5 millions at the BnF, if we include in-copyright material. Most of the textual content has been ocerized. Born digital material of the same nature, like ebooks and digital photographs, entering the collections by legal deposit, acquisitions or donations are also included in this dataset. This collection is available freely on the web for it greatest part (public domain material) an on site for copyrighted material entered via legal deposit.
- Web archives, harvested from the live web since 1996 (if we include retrospective collections, what we call the "web incunabula", that were retrospectively acquired from the Internet Archive). These archives are the result of 3 types of harvesting: a yearly broad crawl of the French web, deeper targeted crawls based on curator's selections along the year, and agreements with 30 newspapers publishers who deliver daily a digital version of their issues. All this content amounts to almost 800 Terabytes of data, billions of URLs. This collection is available only on site for researchers, due to the French legal deposit law.
- Metadata, created by the BnF or aggregated form other sources (including the publishers themselves) to describe the collections. All the metadata that the BnF disseminate is available under an open licence since January 2014. They are distributed through a variety of channels, such as Z39.50, SRU and SPARQL protocols, OAI-PMH repositories, data dumps and the

[2] http://c.bnf.fr/fom

linked data website data.bnf.fr. One could question the term "collection" for this dataset but because these metadata cover the legal deposit in as much a comprehensive way as possible, they actually have value in themselves to explore a nation's heritage. This value is rendered yearly in the *Observatoire du dépôt légal*[3], a digital report that describes every year the landscape of publishing in France, based on the legal deposit.

There is a fourth digital corpus that has been deemed out of scope of our project: the e-resources (e-journals, e-books) that are acquired from e-publishers and available only through third-party platforms (Elzevier, Springer etc.). This type of material is not specific to the BnF: mainly, it falls outside the scope of the legal deposit because very few of those publishers are French. In addition, in France, the ISTEX initiative aims at creating a national archive with TDM capabilities for this kind of material. Finally, the BnF is only granted with access, not with the actual files and data, for this content. For all these reasons, we decided that providing access and analysis services for e-resources was out of our scope.

Looking at the 3 collections that are actually in scope, the first thing that stands out is that they are built and managed in very different ways, using different tools and by different teams. Trying to have a global approach on these 3 datasets and to mutualize processes and services is already a challenge in itself. Even if we don't consider the issue of TDM, these 3 collections are accessed and their content or data retrieved in ways that are totally distinct, making them silos for the end-user. So the first question of the Corpus project is whether having a data-oriented and service-oriented approach of these collections makes sense and how to bridge between them.

What they have in common is that they are massive and they are organic. By organic, I mean that because of the way these collections have been built, it is not only each item that has meaning, but also the collection as a whole. Because the legal deposit aims at covering the national domain in as much a comprehensive way as possible, the resulting collection of data and metadata can answer questions such as: what is the trend of children and youth publishing in France across the past ten years? How many times was a specific wooden engraving reused in 16th century books published in Lyon? When did the word "migrants" started to replace "immigrés" in the French news media? What are the links between institutional websites and amateur websites related to the First World War?
All these questions could have been answered without actually mining the data, but it would have required years of research and a tedious analysis. Moreover, they require to have already a sense of the answer before actually starting to investigate the question. Digital tools make it much easier and leverage a new field of exploration for the collections. Then they start revealing the potential for types of studies that were not within our reach with former methods and tools.
For example, the Commonplaces project (Roe et al., 2016) uses an algorithm to match similar passages in a corpus of 200.000 literary and scientific texts printed in Britain in the 18th century. The objective of this research is to identify "commonplaces", i.e. most cited pieces of texts, without prior knowledge of what popular extracts might be. The product of this analysis is, in itself, a new corpus, a database that could be used to ask further questions and conduct more in-depth analysis. It is unlikely that such research could have been conducted with such a wide scope, on an interdisciplinary corpus, without data analysis tools. Unexpected patterns may be revealed by such an approach, that may not have been foreseen by researchers.

On the other hand, one could also object that there is a lot of uncertainty in the level of relevance of such research. The commonplaces may actually reveal very "common": is it surprising that the more cited extracts are taken from the Bible of from Shakespeare's plays? Moreover, a lot of the findings extracted using such tools may be the result, first of the nature of the collection and second of the interpretation of the researcher who created the algorithm.
Can digital collections or datasets accurately reflect the society that built them? How much do they tell us about our perception of the information society and its many forms ? How much is hidden in

[3] http://c.bnf.fr/fJK

bias implied by the tools we use to create, collect and access them ? How much will be lost, or has been lost already in the digital memory gap ? We won't know for sure until researchers actually decide to study those collections as organic datasets, and not as groups of isolated items.
There is still a fair amount of scepticism, among researchers but also collections holders, regarding our ability to extract relevant information from this type of research. A great part of the investigations is still dedicated to exploring the collection structure itself and evaluating the methods and tools that can be used to analyse it, a task that can only be achieved through a strong collaboration between librarians and researchers. By any means, techniques like text and data mining should not be considered as an end in themselves, but rather as a mean among others to reach a goal that has to be defined prior to setting up tools.

## 2 THE OPPORTUNITY: NEW METHODS FOR EXPLORING DIGITAL CORPORA

Regardless of these epistemological considerations, the existence of both data and tools creates the opportunity for a new type of research that we have seen arising already in the past few years at the BnF. It is interesting to note that these first experiments have started as separate projects, without any global view on the library setting up text and data mining as one of its priorities. The reasons why the library engaged in these projects in the beginning were very diverse: sometimes it was with the hope of improving collection management tools and systems for our own benefit, sometimes it was to increase the visibility of a specific collection… But the general situation is that it was always following up an encounter between a team of researchers and a team of library staff, both sharing common interests.
I will cite 3 examples of these encounters that represent 3 different situations:
a. The dataset is identified as a relevant source for research, and the use of data technology is seen as a mean to improve, deepen or broaden the possibilities of analysis of this source;
b. The research is focussed on building a tool, and the dataset is used to train or test the relevance and efficiency of the tool;
c. The research is defined with a specific objective or question, and both the dataset and the tools are built on the way to try and answer the question.

### 2.1 The corpus as a source

The first example is the Europeana Newspapers european project (Moreux, 2016). The main topic of this research is to explore one of the most valuable sources of the 19th and 20th centuries available from digital libraries: digitized collections of newspapers. This material is a relevant source for historians but also for genealogists, sociologists, researchers in literature, arts or in history of science… All the bits and pieces of old newspapers, from ads and announcements to illustrations, from "faits divers" to "feuilletons" are precious insights from the past.
So, the Europeana Newspapers project targeted this specific material in order to make it possible to extract even more information from it than used to be possible when it was not yet in digital form. This research included improving data extraction techniques such as OCR (optical character recognition) and OLR (optical layout recognition). It also explores the use of quantitative metadata extracted from the corpus for different purposes, ranging from statistics that give librarians a better knowledge of the collection to data visualizations that can be interpreted by researchers to better understand the story of these publications.
The project led to the creation of a prototype that demonstrates search capabilities offered by the enriched corpus. The deliverable of the project, as is often the case with this type of project, is not a final research conclusion: it's rather a new service, based on preprocessing of the corpus using dedicated tools, that will empower future researchers in their work.

## 2.2 The corpus as a sandbox

The second example I will present illustrates the interest of researchers from scientific fields for digital library collections that may challenge the tools they are building, because compared to other datasets that can be found elsewhere, they have specificities: they come with context and high-quality metadata, they are heavily structured and interlinked, they cover a broader range of types, periods and spaces.
In 2014, the ETIS laboratory, an IT research team from the university of Cergy Pontoise, proposed to investigate assisting image indexation by automatically extracting labels or keywords from the images. This image mining project was conducted on a subset of 4000 annotated images from the BnF's Picture collection[4] (Picard et al., 2015). The researchers from ETIS were interested by the fact that, compared to generic images collections used in computer vision benchmarks, a cultural heritage collection raises more challenges, due to the specific knowledge needed to extract semantics from the material. The diversity of the material is also a challenge: while it may be quite easy to search for a specific coin or seal pattern by similarity, recognizing the similarity between the idea of a "horse", represented on a medieval coin or a manuscript illumination or a 19th century photograph is much more challenging.
From the point of view of the Corpus project, it is interesting to note that stakeholders in this project had different, if not opposite goals: the researchers wanted to test and train their tools and were as interested by success as by failure, while the librarians hoped that this research could lead to the design of a production tool to help them with cataloguing masses of images. But the deliverable of the project is a prototype of the tool and an in-depth analysis of the success and failure of the algorithm, not a live cataloguing service… Managing the expectations of the library when it engages in a research project is a topic to be considered in itself !

## 2.3 The corpus as an interface

The third project I will take as an example was conducted within the Labex "Les passés dans le présent" by two libraries, the BnF and the BDIC (Bibliothèque de documentation internationale contemporaine[5]) from 2013 to 2016. The academic part of the project was carried out by researchers from Télécom Paristech, an engineering college. Called "Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre", this programme intended to dive into the practices of web users in order to analyse how they take hold of the massive digitized heritage collections that libraries are disseminating online. The 1st World War was used as the example of topics web users would be interested in, so this project is later referred to as « WWI ».
The use of web archives as a source for this study was not necessarily intended in the beginning, but appeared as a natural development when the researchers realised that the BnF, thanks to the web legal deposit, was in capacity of creating, collecting, preserving and giving access to a relevant corpus of websites. The second phase of the project was dedicated to this object (Baudouin, Pehlivan, 2017), with the creating of a series of data visualizations describing the links between institutional websites and amateur websites related to the 1st World War, and an in-depth analysis of the "Pages 14-18" forum[6].
The organization adopted by the team in charge of the project is really interesting : the 1st World War web archive collection was created on purpose for the project, but using the regular workflow for legal deposit of websites. Some tools were built by the BnF to extract the corpus and its metadata, others were built by the researchers to analyse it. Finally, one of the strong incentives for using web archives was to benefit from a secure environment from the legal point of view, but the findings of the project were that the quality of the dataset was better because of the way it was built, following library standards. Here the corpus was not a subject of interest in the first place, like in the Europeana Newspapers project, but it was an interface between the researchers and the library, the added value of the library lying beyond the data and the tools, in its processes, skills and know-how.

[4] http://images.bnf.fr
[5] http://www.bdic.fr/
[6] http://pages14-18.mesdiscussions.net/

By being strongly involved in the project from a technical point of view, the BnF generated new skills in managing and giving access to its web archive collection, thus making it more immediately rewarding, from the library's point of view, than the ETIS project mentioned above. However, this was achieved not as a deliverable of the project but as a commitment of the institution to deliver a service to researchers: it's really the premises of the Corpus project.

The 3 projects I took here as examples were all prior to the idea of the Corpus project and contributed to its creation because they raised challenges that were new to the library. From basic technical questions (how to set up a server for the researcher to store his/her data?) to most philosophical ones (should the library engage in projects that don't deliver ready-to-use tools?), from epistemological issues (what can we expect from the quantitative analysis of a corpus that has been transformed to digital by the library?) to organisational ones (what resources can the library provide to researchers to help them define their methods?) these projects raise all the issues that the library will be confronted with when creating a new service for researchers on digital corpora.

## 3 THE LANDSCAPE: THE WIDE ARRAY OF NEEDS AND POSSIBILITIES

Based on these first experiments, the BnF decided that meeting its users' needs when it came to access its digital collections for research purposes should become one of its priorities for the future.
The decision to launch the Corpus project was prompted by the fact that it appeared difficult, based on our experience, to imagine that we would be in capacity to continue serving researchers if this kind of demand was to increase in the future. The BnF being a partner in each project, building ad hoc tools and providing ad hoc manpower is only possible for a limited number of projects every year. The approach "first come, first served" would have to be the rule, which is not satisfactory.
Also, this decision came at a moment where the audience of the library reading rooms on site was dramatically decreasing, while the online audience had been regularly and massively increasing for a decade. The national library was working on redefining its public service policy and targeting the new usages of researchers was part of it: the idea was to encourage the development of digital corpora studies and to build the service that would make it sustainable for the library to fuel these studies with adequate data.
So, one of the challenges for the corpus project was to grant ourselves with an organisation that would allow us to serve research projects equally, sustainably and efficiently, without necessarily becoming a full partner of each and every one of them. We need to consider partnership only when there is a strong benefit for the library, as was the case with the WWI project.

In order to achieve this, we decided to experiment, within the Corpus project, with different research projects and researchers teams, to identify their needs and try to comply with them in an iterative way, and then assess the process in order to design the future service. We gave ourselves the challenge to try and address the issue in the broadest way possible: the project is inclusive of all digital collections as stated above, of all areas and topics of research, of all disciplines and methods.
We were helped by the fact that the demand continued to stream in from a variety of partners. In 2015, the OBVIL labex (Observatoire de la Vie Littéraire), already a strong partner of the BnF, asked us to provide a corpus of more than 130.000 OCRed texts from Gallica (ALTO files) in order to conduct big data analysis, notably with the ARTFL laboratory from the University of Chicago, who wanted to expand the Commonplaces project mentioned above. In 2016, we worked with the Web90 team, from the CNRS, who studies the history of the web before 2000 ; as a follow-up to the WWI project, we built a new interface for them to access a selection of web archives corpora, with full text search capacities and extraction of quantitative metadata. 2016 was also the year where we conducted a study of data mining on logs from Gallica, together with Télécom Paristech, in order to better understand how users navigate the website of the digital library (Nouvellet, 2017). In 2017, a project with LIPN, a linguistics lab from the Paris 13 University, on finding neologisms in web archives is funded by the ministry of Culture. The Corpus project partners with a team from the CELSA

(Information Science team based at the Sorbonne University) on an exchange of skills and know-how. Finally in 2018, the BnF will be investigating a benchmark on images mining techniques with a variety of partners.
I could even cite more projects, thus demonstrating that the need for digital corpora is not decreasing among our partners. It is not possible to detail them all here, so I will only share a synthesis of the first findings made by the Corpus projects by gathering feedback on all these experiments.

First, the user him/herself in these projects is a very different entity from what s/he used to be in the library's traditional reading rooms. We often have to deal with a team, where different skills are gathered: junior and senior researchers, engineers, experts in digital content or methodology… Their relationship with the library is also different: they are looking for interactions with librarians who are experts in the field (either in collections or in technology) and if they come to work on site, they need specific places (meeting rooms, access to servers or dedicated terminals…)
Because it is so peculiar for us to interact with these specific users, we have tended to treat them as partners rather than library users: we have set up specific contracts (memorandum of understanding), we have granted them with access to the library's offices and staff meeting rooms, we have provided them with dedicated equipment (terminals, servers, working spaces…). They were hardly considered users and not counted as such, which means that part of the library's rightful mission to give access to its collections is not covered in our audience KPI. This audience may not be significant… yet, but it will be in the near future.

Second, there is a wide array of different types of users and situations, especially with respect to technology.
On one end of the spectrum, researchers in disciplines like history know a lot about the collections but are lacking basic technological skills. They are learning data management and curation on the job in an ad hoc fashion and turn to the library not only for support, but also for ideas on how to deal with the digital collections (Koltay, 2017). They also expect the library to provide them with ready-to-use software or pre-analysed corpora: for instance the Archive Web Labs platform, developed within the Corpus project for the needs of the Web90 team, was mainly used as an interface to query to two selected corpora, rather than as a way to retrieve and analyse raw metadata that was also provided through the platform.
On the other end of the spectrum, IT researchers like the ones from Télécom Paristech, LIPN or ETIS mentioned above are fully equipped to handle the data and want to experiment with their own tools and their own environments. They expect the library to deliver the data as raw as possible after they have selected the subset of interest to them, and they need to be autonomous in terms of infrastructure. For the project to study the logs of Gallica, we partnered with Teralab[7], a cloud infrastructure developed by Institut Mines Télécom, in order to benefit from a flexible enough environment without the traditional limitations of the work stations managed by the library.
Between the two ends, the shades of different situations is infinite. If the BnF is to create a service for researchers in order to empower them to study digital corpora as they wish, we'll have to accommodate all these different situations, from the creation of new corpora to their delivery, from the design of new software to the setting up of flexible work environments, from basic information on the collections to expert counselling on formats and metadata.
The task may seem daunting, but we are convinced it can be addressed step by step. It will be an iterative and experimental process; much of a research in itself.

## 4 BUILDING THE FUTURE: NEW SERVICES FOR DIGITAL LIBRARIES

The Corpus project tries to be comprehensive in terms of the nature of issues that have to be dealt with when setting up such a service in a national library. We have identified the following topics :

[7] https://www.teralab-datascience.fr/

- **Legal aspects**: the first domain where researchers need advice from the library is to understand the legal constraints that they need to take into account. They don't necessarily have legal training, so the library needs to clarify the conditions for use and reuse of the corpora. The library may have to take actions to secure from a legal point of view the actions taken by researchers (anonymisation of personal data, contracts, secure environments). Some actions fall outside the scope of the current legal framework and may even require that we consider an evolution of the law[8].
- **Organisational aspects:** the library needs to set up an organisation that is ready to take into account new demands. The roles must be clarified and sometimes created out of the blue. The process need to be clearly communicated both internally and to external users so that they can activate it. These organisational aspects are not limited to technical and scientific library staff: we also need administrative processes to write and sign contracts, to charge for services, to grant specific permissions, etc.
- **Human resources**: the library needs to define how it will mobilize its staff, especially if the demand is growing. Sometimes the automation of processes will be efficient in order to save time, but in other cases the demands are too specific and require dedicated work. Some interventions can only be done by experts who already have an important workload. In relationship with the processes in place, we have to determine who can be solicited, when, how and to what extent.
- **Skills and know-how**: we have to investigate the nature of skills that have to be developed in order to deliver the service. Some are natural for the library (knowledge of the collections themselves), some are more specific (knowledge of processing tools). Then there is the question of how to educate the researchers if they need it: is it necessary to involve experts in training programmes? Is it the role of the library? How can educational structures help with this task?
- **Infrastructure**: digital collections are not virtual; they are not delivered through thin air, they need machines, servers, storage space, networks… The infrastructure created by the library is stable and secure, but lacks the flexibility requested by those researchers who are equipped already. To what extent can or should the library develop the capacity to host data, services and software in personalized environments for its users? Can such an infrastructure be sustainable for a library? Is it worthwhile considering to share it with other institutions?

We are far from having answered all these questions but we have already defined a way forward for some of them.

## 4.1 Ready-to-use APIs and datasets

The data dissemination strategy already in place at the BnF involves APIs such as IIIF, available for Gallica since 2016, other protocols for the metadata (Z39.50, OAI-PMH, SRU…) and data dumps such as the ones available on the data.bnf.fr website. Some of the requests made by researchers can be fulfilled already by accessing these data.
In 2016, we started documenting all the available datasets and APIs on Github when we organised the 1st BnF Hackathon[9]. In 2017, this initiative will become even more systematic, with a dedicated website for this purpose.
While this initiative is not specific to researchers, it has always been demonstrated that making the data available in an easy-to-use and well-documented manner creates opportunities for academics to use it and ask new questions, as is the case with data.bnf.fr (Glorieux, 2016 ; Langlais, 2017). So why not extend this approach to corpora that have been produced and pre-processed for a dedicated project? In many cases, like the Europeana Newspapers project, the OBVIL initiative or even the

[8] An exception for text and data mining was voted in the french "law for a digital republic" in 2016, but it was never actually implemented. The BnF continues to investigate TDM using other legal tools such as the "code du patrimoine" that defines the conditions of access to legal deposit material, and of course on public domain material.

[9] https://github.com/hackathonBnF/hackathon2016/wiki

corpora that have been extracted for the Web90 team, the extraction work realised by the library and the pre-processing realised either by the researchers or the library is reusable in other contexts. Some of these projects like the Commonplaces study even set the reuse by others as one of their objectives. Dissemination of the results of this work with a clear reuse policy can be a starting point for a digital scholarship initiative, as it is the case for the British Library for instance[10] (McGregor et al., 2016).

### 4.2 Towards an on-site laboratory

If the British Libary's initiative take the form of multiple events and projects aggregated online, the idea of a "lab", dedicated to digital scholarship and open to innovative uses and reuses of digital corpora available in libraries, is definitely trending. The Library of Congress released a report encouraging such an initiative in late 2016 (Gallinger, Chudnov, 2016). The Center for Digital Research in the University of Leiden is another example of the kind of services that libraries can build to support researchers in the development of adequate skills for text and data mining studies (Oudenhoven, 2017).
When it comes to designing the future digital scholarship service at the BnF, we have to take into account that part of the digital collections are accessible only inside the premises and that some of our researchers are in demand of work spaces and expert support: it's the reason why we see our future "Lab" as an actual space, within the research library. However, that space may be quite different from a regular reading room: working in groups should be possible, as well as teaching and other forms of gatherings (co-design workshops…)
We are looking for partnerships to make this space more attractive, notably with laboratories and universities who already have TDM studies going on. Their resources will also be helpful to design and build the future IT infrastructure of the Lab.
By the end of the year, we are also conducting a study, interviewing researchers and experts both inside and outside the library, in order to better understand their expectations.

## 5 CONCLUSION

As more and more data and digital collections are available in libraries, what role can these libraries play in the digital scholarship landscape is becoming a key issue. National libraries such as the BnF are particularly concerned because of the nature and specificity of the collections they hold. After several years of experiments, the BnF works within the Corpus project to improve its knowledge of researcher's needs and design the processes and infrastructures needed to meet their demands. This initiative will lead to significant use cases demonstrating the potential of TDM for the study of digital corpora, to a strong policy of data dissemination and, in the near future, to the creation of a Lab dedicated to digital scholarship.

## References

(Baudouin, Pehlivan, 2017) Valérie Beaudouin, Zeynep Pehlivan. *Cartographie de la Grande Guerre sur le Web : Rapport final de la phase 2 du projet "Le devenir en ligne du patrimoine numérisé : l'exemple de la Grande Guerre"*. Bibliothèque nationale de France; Bibliothèque de documentation internationale contemporaine; Télécom ParisTech, 2017. https://hal.archives-ouvertes.fr/hal-01425600

(Chambers, 2016) Sally Chambers, *It's not about the catalogue, it's about the data - Catalogue 2.0: The future of the library catalogue*. University of Gent, Feb. 2017. https://biblio.ugent.be/publication/8511250/file/8511251.pdf

[10] https://www.bl.uk/subjects/digital-scholarship

(Gallinger, Chudnov, 2016) Michelle Gallinger and Daniel Chudnov, *Library of Congress Lab : Library of Congress Digital Scholars Lab Pilot Project*. Library of Congress, Dec. 2016. http://digitalpreservation.gov/meetings/dcs16/DChudnov-MGallinger_LCLabReport.pdf?loclr=blogsig

(Glorieux, 2016) "Data.bnf.fr, les documents" in *J'attends des résultats*, juin 2016. https://resultats.hypotheses.org/795

(Johnson, 2016) Rob Johnson et al., *Text and data mining in higher education and public research*. Report commissioned by the Association des Directeurs & personnels de direction des Bibliothèques Universitaires et de la Documentation (ABDU), December 2016. http://adbu.fr/competplug/uploads/2016/12/TDM-in-Public-Research-Revised-15-Dec-16.pdf

(Koltay, 2017) Tibor Koltay, "Data literacy for researchers and data librarians". *Journal of Librarianship and Information Science*, 2017, Vol. 49(1) 3–14. http://journals.sagepub.com/doi/abs/10.1177/0961000615616450

(Langlais, 2017) Pierre-Carl Langlais, "Les bibliothèques numériques sont-elles représentatives ? " in *Sciences Communes*, April 2017. https://scoms.hypotheses.org/799

(McGregor et al., 2016) McGregor, N., Ridge, M., Wisdom, S., Alencar-Brayner, A. "The Digital Scholarship Training Programme at British Library: Concluding Report & Future Developments". In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, 2016. http://dh2016.adho.org/abstracts/178

(Moreux, 2016) "Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment : Facilitating Access for various Profiles of Users" in *IFLA News Media Section, Lexington, August 2016, At Lexington, USA*, Aug 2016, Lexington, United States. https://hal-bnf.archives-ouvertes.fr/hal-01389455

(Nouvellet, 2017) Adrien Nouvellet et al., "Modélisation des comportements à partir de l'analyse des logs de Gallica", *Journée d'étude « Quels usages aujourd'hui des bibliothèques numériques ? Enseignement et perspectives à partir de Gallica »*, BnF, Paris, 3 mai 2017. http://www.bnf.fr/fr/professionnels/anx_journees_pros_videos/a.video_170503_05_table_ronde_4.html.

(Oudenhoven, 2017) Martine Oudenhoven, "On the role of a university library in the TDM landscape" in *FutureTDM*, June 2017 http://www.futuretdm.eu/blog/role-university-library-tdm-landscape/

(Picard et al., 2015) David Picard, Philippe-Henri Gosselin, Marie-Claude Gaspard. "Challenges in Content-Based Image Indexing of Cultural Heritage Collections." In *IEEE Signal Processing Magazine*, Institute of Electrical and Electronics Engineers, 2015, 32 (4), pp.95 - 102

(Roe et al., 2016) Roe, G, Gladstone, C, Morrissey, R et al 2016, "Digging into ECCO: Identifying Commonplaces and other Forms of Text Reuse at Scale". In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków. http://dh2016.adho.org/abstracts/343