Digital curators at work: analyzing emerging professional identities at the Bibliothèque nationale de France (BnF)

Marianne Clatin, Louise Fauduet, Clément Oury, Jean-Philippe Tramoni

▶ **To cite this version:**

Marianne Clatin, Louise Fauduet, Clément Oury, Jean-Philippe Tramoni. Digital curators at work: analyzing emerging professional identities at the Bibliothèque nationale de France (BnF). IFLA World Library and Information Congress, Aug 2014, Lyon, France. hal-01098526

HAL Id: hal-01098526
https://bnf.hal.science/hal-01098526v1

Submitted on 26 Dec 2014

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Distributed under a Creative Commons Attribution 4.0 International License

Submitted on: 12/26/2014

Digital curators at work: analyzing emerging professional identities at the Bibliothèque nationale de France (BnF)

Marianne Clatin
Bibliographical and Digital Information Department, Bibliothèque nationale de France, Paris, France
Louise Fauduet
Audiovisual Department, Bibliothèque nationale de France, Paris, France
louise.fauduet@bnf.fr

Clément Oury
Legal Deposit Department, Bibliothèque nationale de France, Paris, France
clement.oury@bnf.fr

Jean-Philippe Tramoni
Information Systems Department, Bibliothèque nationale de France, Paris, France
jean-philippe.tramoni@bnf.fr

Copyright © 2014 by Marianne Clatin, Louise Fauduet, Clément Oury and Jean-Philippe Tramoni. This work is made available under the terms of the Creative Commons Attribution 3.0 Unported License: http://creativecommons.org/licenses/by/3.0/

Abstract

Digital curation is the global concept that includes all aspects of work on digitized and born digital cultural objects: from document entry to data description or indexing, storage, dissemination, and long-term preservation. It is an expanding activity, whose rules and best practices are not globally defined yet. The BnF has chosen to rely on observations on the ground to understand how its staff, faced with the urgent need to collect and preserve a fast-growing digital heritage, is developing new tasks and skills. A dedicated group called ORHION, a French acronym for Observatory of Organizations and Human Resources under Digital Influence, has launched a series of studies on professional identities that are particularly affected by the increasing shift to digital activities. It has so far completed its analysis of two subjects: metadata curation and digital entries. In order to assess the BnF's practices, ORHION also relies on other institutions' experiences. It has for instance organized a workshop on web archive curation during this year's general assembly of the International Internet Preservation Consortium, in May 2014.
This paper focuses on the findings of these three use case analyses: the switch from cataloguing to metadata management for description and access; the processes and agents involved in digital entries; and the organization, skills and training of web curating teams. Based on these three experiences, this paper seeks to identify some key trends in digital curation: the notion of digital collection itself is not obvious and needs to be rethought; professional identities are challenged by the apparition of new tasks and the emergence of new actors; stakeholders struggle to understand their own roles and responsibilities in more complex treatment channels. Training staff and sharing the workload between the first pioneers and the rest of the teams are necessary to involve all professionals in digital curation.

Keywords: Digital curation, skills and training, shifting professional identities, human resources, organizations.

1. The BnF approach and the ORHION observatory

Digital curation is an expanding activity within libraries and other heritage institutions. This concept includes all aspects of work on digitized and born digital objects: from document entry to data description or indexing, storage, dissemination, and long-term preservation. In a sense, it is an adaptation to digital documents of the traditional treatment chain used for analogue media. At the National Library of France, digital curation appeared as soon as the library started managing digital documents: at the end of the 90s, when its digital library, Gallica, was launched. The scope of digital material widened quickly. It widened "horizontally": the library took responsibility over new types of digital objects: web archives, born-digital audiovisual content, ebooks… and "vertically": from creation or reception of digital content to access and long term preservation – BnF's digital repository, SPAR, was launched in 2008 [1].

However, the need to consider the impact of these new activities on the library professionals, on the "digital curators", emerged later. When the BnF launched its "mass-digitization" programs, in 2007-2008, hundreds of librarians and other BnF agents were involved in tasks related to the management of digital documents. It became necessary to identify the issues these professionals were facing, the training that was required, and the organizational challenges that were encountered. Throughout 2008 and 2009, the idea of an informal group of librarians interested in sharing their experience of the challenges digital collections posed to their daily activities was put into practice. This group was called ORHION, a French acronym for Observatory of Organizations and Human Resources under Digital Influence; it has progressively developed into something which is officially endorsed for the long term by BnF's top management in 2010 [2 and 3].

The group operates outside existing BnF structures, since its role is to observe and articulate the changes in librarians' practice and skills, not to elaborate the library's strategies. Yet it plays an important role in raising awareness through its four types of action: interviews, working groups, workshops, and information for managers.

In its first years, ORHION dealt with issues such as the definition of the digital collection(s), their management in digital repositories, and their promotion towards readers and internet users. In 2013, an emphasis was put on the human side of the question: ORHION launched a series of studies on professional identities that are particularly affected by the increasing shift to digital activities. The BnF management has proposed four areas of investigation for the period 2013-2015: metadata curation, digital entries, interaction between librarians and IT engineers, and new forms of project management. ORHION has so far completed its analysis of the first two subjects.

In order to assess the BnF's practices, ORHION also relies on other institutions' experiences. It has for instance organized a workshop on web archive curation during this year's general assembly of the International Internet Preservation Consortium, in May 2014.

ORHION observations were not intended to provide a definition of digital curation. However, through these case studies, they carried out a practical analysis of digital curators' activities. This paper will thus focus on the findings of these three recent analyses: the switch from cataloguing to metadata management for description and access; the processes and agents involved in digital entries; and the organization, skills and training of web curating teams.

2. The switch from cataloguing to metadata management for description and access

Studying metadata management for description meant investigating not only the work of cataloguers, but the different positions and skills that make it possible for a document to be found either in the catalogs or in the digital library, Gallica. An emphasis was put on three specific issues: granularity of description; metadata created outside the library; and the interaction of multiple jobs and skills required for creating the metadata linked to the digitization process.

Granularity issues

Granularity in the catalog has been an issue for some time, particularly regarding serials' description. The digital library is concerned as well, since a large amount of documents available on Gallica are newspapers. Without any tables of content or indexes, highlighting these collections means identifying the different issues of a newspaper. Even though this huge task was never done before in the catalog, the competences required for it already exist in the library: it can indeed be compared to the job done by library assistants when previously preparing the microfilming process.

Other types of digitized collections are facing granularity issues, such as pictures by famous photographers, previously collected as slides and now in a digital format. In this case, their description is created by the same cataloguers who already described analogue photographs. While cataloguing a batch of photographs proved to be sufficient in the catalog when a set of 71 slides taken in a theater festival was communicated as a whole to the readers, it is not sufficient in a digital library where the readers expect to find one specific image on a topic – presumably found within different sets of slides. The decision to describe these collections at a picture level meant that the description criteria were now influenced by the access conditions in the digital library, and not only by the catalog. Yet, everyone agreed that the enhancements performed for the digital library should also benefit the original catalog itself, which is still considered the core repository of descriptive metadata on which new services may be built in the future.

Descriptive metadata created outside the library

As a national bibliographic agency, the BnF is responsible for the description of documents within the scope of the national bibliography (produced in France or by French publishers). But what about other kinds of documents?

As other libraries derive records from BnF catalogs, the BnF cataloguers derive records of foreign documents from other repositories, such as WorldCat. Time spared thanks to this can be used for other tasks, such as selection for digitization or web legal deposit. However as far as cataloguing itself is concerned, this process also calls into question the identity of cataloguers. Choosing the best record available among the duplicates in WorldCat may be considered less interesting than creating a record from scratch. On the contrary, some cataloguers consider that saving time on description allows them to concentrate on tasks requiring another kind of analysis, such as choosing subject headings, creating links to authorities, or even creating new authority records. The same issue will occur soon when the BnF will be importing publishers' ONIX descriptions of French ebooks received through legal deposit.

Another challenge to both skills and organization came from ebooks that were individually purchased. Purchasing here consists of selecting a reference in EBSCO international knowledge base. When the document is not already referenced (which is often the case for non-English written documents), the librarians create the short description themselves, according to the criteria already defined in the EBSCO database. Yet, some consider this

description to be insufficient for a library purpose (no subject headings, no link to authority records…). The question then occurs: are cataloguing skills required in such cases? Couldn't the library assistants in charge of regularly checking access to the documents outside the catalog be also responsible for creating the few metadata fields required in these separate tools? And when tens of thousands of ebooks are concerned, shouldn't these professionals be located in the different thematic departments of the library, and not in a single "electronic acquisitions" department?

The interaction of specific skills for description in the digitization process

When Gallica started, some considered that description would soon prove useless, since readers would directly access documents without using the catalog. However, a digital library cannot work without a minimum of metadata (descriptive, technical, and even preservation metadata). Digitizing collections is also quite difficult if they are not catalogued at all. A recent case gave the BnF the opportunity to improve its ability to be creative and adapt its processes.

When the Medals and Coins department decided to speed up its digitization program of more than a hundred thousand non-catalogued Greek ancient coins, it appeared that creating record after record of full description in the catalog wouldn't be practically feasible. A new metadata workflow was tested for this occasion, that is now considered a model for upcoming similar digitization programs: it consisted in defining a "simple" description model that would not prove sufficient for scholarly research purposes (*e.g.* without a full iconographic description), but would be sufficient for the purpose of distinguishing documents in the digital library. The cataloguers then described them in a CSV document rather than in the MARC catalog, which was not designed to allow repetition of thousands of identical pieces of information from one record to another. However, using another tool than the catalog meant that a lot of other actors in addition to the cataloguers had to be part of the process in order to have the description finally included in the digital library: curators specialized in a specific type of documents; bibliographical coordinators; IT staff involved in mass data treatment in the catalog; professionals dealing with the photographing process itself; OAI repositories' functional managers; and developers of the digital library. What was pointed out by all these contributors was the importance and professional enrichment of cooperation between different skills and people who did not usually have the opportunity to work together at the library.

3. The processes and agents involved in digital entries

After metadata specialists, the next individuals considered were professionals dealing with the handling of digital material. A workgroup was convened to define more precisely the domain to be studied. It was decided to conduct this inquiry on five different channels:

- the digitization process for various material on physical media;
- the harvesting of web sites (and especially online press content) in the framework of legal deposit;
- the legal deposit of digital audiovisual material;
- the legal deposit of ebooks;
- subscription to digital periodicals.

The channel for the entry of physical objects was kept as a reference for comparison purposes. The investigation had to focus on the processes and agents involved in these five channels, from initial entry to long-term storage, using the same methodology as in previous ORHION studies. Cataloguing and communication were excluded for the scope of this investigation.

Non-linear workflows

The first findings were a confirmation of some basic discrepancies between physical and digital processes.

The physical entries channel is characterized by the linearity of the entry process. The tasks are common, so this process is globally similar for all physical media, from books, periodicals or maps to CDs. Within this process, the verification tasks (for instance checking the condition of a physical object) are often implicit. Lastly, the handling of physical objects is a type of task that is assigned to specific actors.

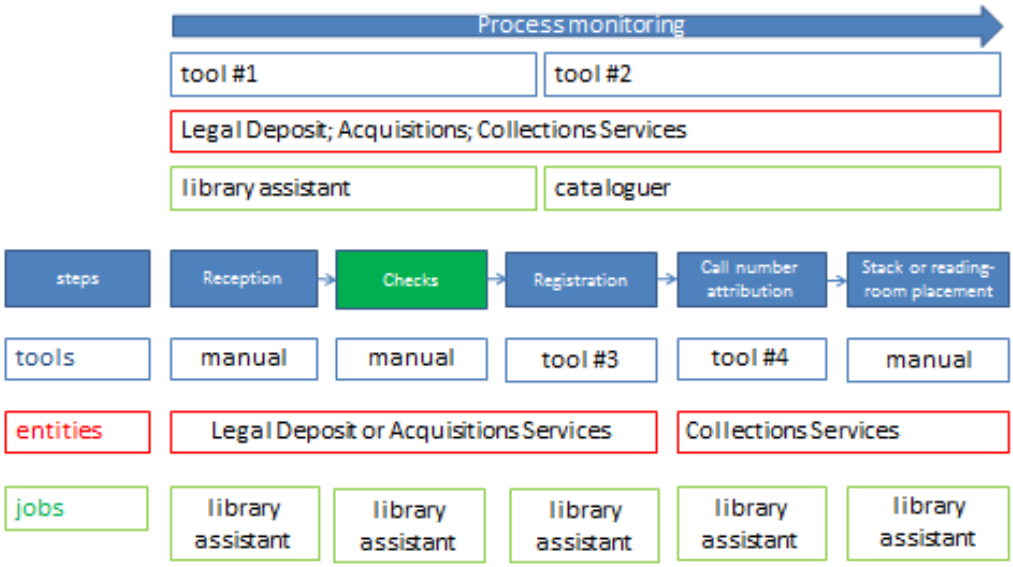

Figure 1: Entry process for physical documents

On the other hand, the digital entry channels are much less sequential processes. While the quality assurance of material received was implicit in the physical entry channel (when dealing with documents, professionals check if they are damaged or not), it becomes mandatory to make it explicit with digital entries, as the steps are automated. Audit or control tasks show up defects that require rework and therefore create loops in the entry process.

In addition, a specific channel is necessary for each type of digital material: no single process model has yet emerged. Finally, this implies different tasks, with different goals, different methods and different actors.

The entry channel for digitized material (in the diagram below) is typical of these observations. There are many checks and verification steps, implying various steps backward and contrasting with the linearity of the physical entry channel. In addition, this channel has to manage two parallel streams. Hence a more explicit management of document flow is required to deal with this complexity.

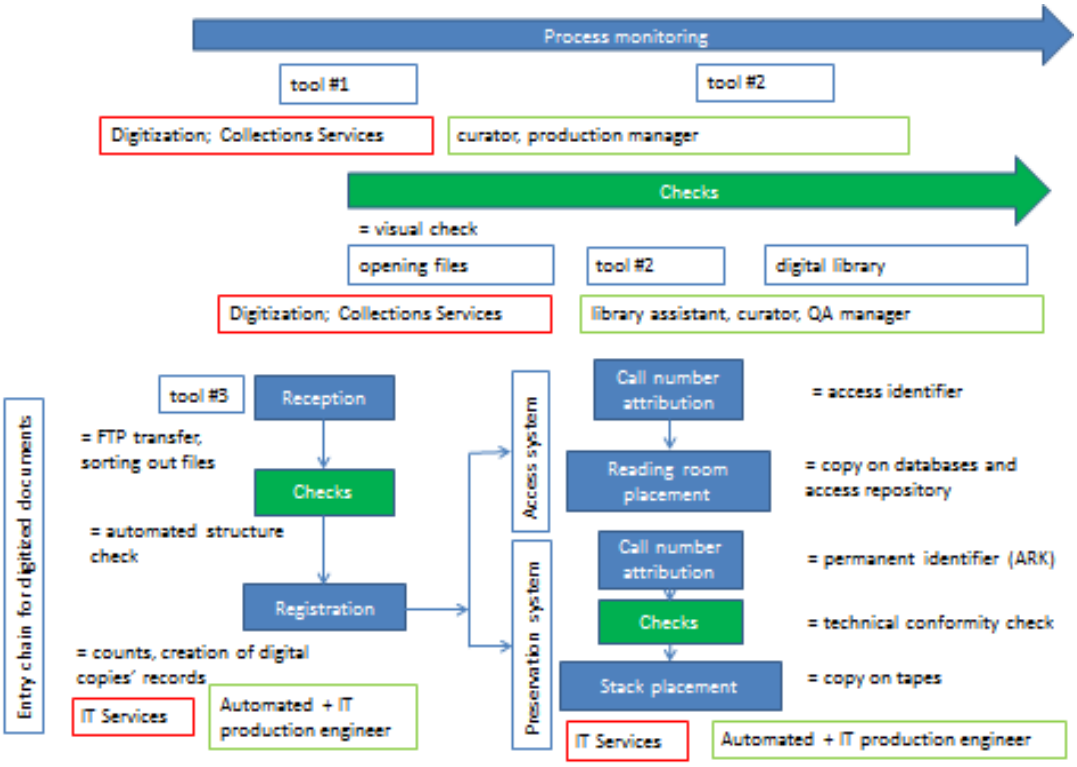

Figure 2: Entry process for digitized documents (QA = quality assurance)

An evolving distribution of roles and responsibilities

The study showed an evolution in the sharing of roles and responsibilities. Within the physical entry channel, library assistants are the key actors of the process – this is not the case for digital documents. For instance, for the online press entry channel, the responsibilities are distributed between engineers managing the flow of IT operations and librarians working both before and after them in the process, under the supervision of the Digital Legal Deposit team. However the process is not fully automated, and it was necessary to define precisely what had to be checked, how, by whom and where in the process.

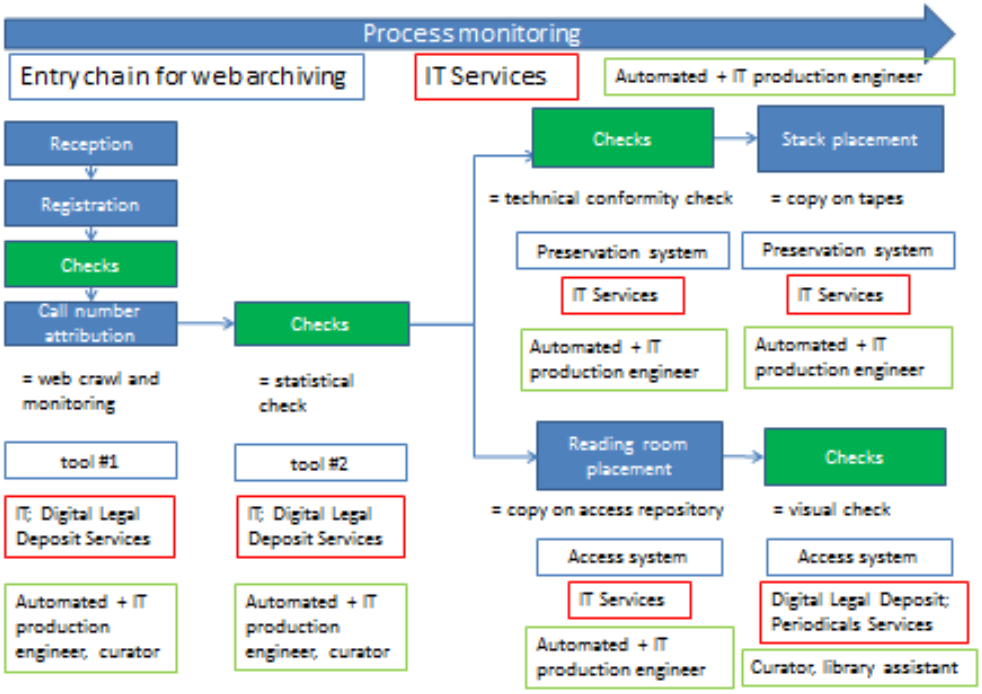

Figure 3: Entry process for online newspapers

For this entry process as for the others studied, the role of IT engineers has thus changed. Their role is not only to support the librarians by providing them tools, but to directly manage the collection – this situation can be compared to that of digital preservation, where the direct administration of the repository is performed by engineers (implementing requirements and specifications expressed by librarians).
Even though the role of library assistant has changed, it is no less important, as they need to check the quality of the collections at each step of the process. However, sampling is often necessary due to the huge amount of documents. To define the library assistant job itself and its responsibilities, it becomes necessary to establish a typology of these tasks: visual checks, verification of compliance or completeness, audit…

Problems of professional identity definitions

Interviews with professionals working on the five selected channels have highlighted some concerns which were later debated in a seminar. In addition to the two issues that were already identified (emergence of new actors and the transformation of the job of library assistants), two others were specifically raised: the sense of losing a connection with the document and the lack of perception of the entry process as a whole.
The sense of losing a connection with the document is especially strong for actors controlling the flow of digital documents through aggregated data – some of them are even essentially dealing with financial and administrative statistics. These actors may perceive their work as being too abstract. On the other side, the quality assurance through visual checks of documents, despite its tedious side, allows a sense of ownership of the collection.
To compensate for this loss of meaning, managers have to figure out how to distribute the processing of physical media and digital objects between teams. Dividing the workload and allocating it to separate groups of actors may lead to the perception that some tasks or channels are more highly regarded than others: a risk then exists of creating a gap between teams within the organization. Furthermore, it is perceived that reusing the skills and expertise acquired from handling physical entries (for instance knowledge of the periodicals

publishing system) may be beneficial and more efficient for monitoring digital entries (for instance harvesting of online press in the context of legal deposit).

Finally, due to these revisions in the entry process, it is becoming difficult for the staff to develop a perception of this process as a whole. To reduce this lack of vision, the building of specific training programs has been a solution implemented not only for people directly involved in digital activities, but also managers and other departments.

4. The organization, skills and training of web curating teams

As the general assembly of the International Internet Preservation Consortium (IIPC) was being held in Paris this year[1], the members of ORHION thought it would be an excellent opportunity to compare and discuss practices with other institutions. Given the observatory's focus on changing professional identities, it submitted a workshop proposal on web archives curators. They were defined for this occasion as the people involved in developing and using web archives who are neither technicians dealing with operations and progress in hardware and software, nor researchers building knowledge out of them; people who are the necessary bridge between these two, and are often a bit of both.

The web curators' collections

First and foremost was the debate on the nature of web collections and the nature of the web curator role. Building a collection is a time-honored curator's vocation. When it comes to web archives, has going about it changed?

When institutions have a legal deposit mandate to collect the web, it is often tied to a national or regional area. However, collecting documents according to their area of publication makes less sense in the internet environment, defined by infinite and international hyperlinks. Readers may have a hard time recomposing global coverage across institutions since many of them have to, or choose to, restrict access to their reading rooms.

Others have underscored the need to take into account the technical specificities of a web crawl from the beginning and to integrate it as a criterion in the curators' crawl requests. Frequency and depth of crawl can be left to the curators to decide, and limitations of the crawling software should be understood and factored in at the time of selection (according to local implementations, some types of video content, content protected by password, and so on, cannot be collected).

Whether sites at risk of disappearing should be preserved above others is a subject of debate. It seemed a more important factor in the beginnings of most programs, but has somewhat been played down. The most important goal has been building a meaningful collection, coherent within itself and supplementing existing collections' strengths. Columbia University Libraries[2] for instance have a Human Rights Web Archive, the Avery Library Historic Preservation and Urban Planning collection, the Burke Library New York City Religions collection, and so on.

Most institutions have tried to build on their traditional areas of excellence and have curators select sites to collect thematically. Even so, ensuring the correspondence between curators' domains of expertise and their web selections can be tricky, as classification of websites is not always obvious. At one time, the Library of Congress tried out a "miscellaneous" web

[1] Program and presentations from the IIPC GA are available at http://netpreserve.org/general-assembly/2014/overview.

[2] https://library.columbia.edu/bts/web_resources_collection.html

collection, hoping to broaden selections and curators' involvement, hoping that over time reasoned collections would naturally emerge, but this was not the case. Indeed, the biggest problem with collection building seems to be for the curators to add this task to their responsibilities, due to lack of time, resources or interest. Even when they are motivated and undeterred by technology, they still tend to work in short bursts where they initially pick out websites on a theme, then their involvement declines over time.
As technical restrictions mean selections as to what can be collected have to be made even in a legal deposit context, the question of readers' best interests and input becomes prominent. Yet readers are still few as these types of collections are slowly getting discovered by researchers and the general public, a great part of the content is still available online and access and analysis tools are being refined. Curators have to guess what will be of use to the future generations, not a process usually found in libraries that receive legal deposit. Some institutions wonder whether direct readers input might be the way to go; Archive-it, the service provider arm of the non-profit internet library The Internet Archive, has developed such a program with K-12 students[3].

The web curators' partners

These collection building choices have been formalized to a greater or lesser degree and communicated to the public[4]. They influence the way that institutions are organized to accomplish web collecting, preserving and access. Core web archiving teams are often part of dedicated digital services, especially when the programs are still in development. They can also be part of the legal deposit services (National Library of France, Royal Library of Denmark…), the collection services (Royal Library of the Netherlands, the British Library…) or acquisitions services (Library and Archives Canada). They are LIS and/or IT specialists, and collaborate with IT specialists, other curators as part of a selectors' network, and/or researchers.

Web curators have multiple strategies to make their collections known to researchers: it often falls to them to help translate the social sciences, history, or literature questions the researchers are trying to answer into queries that can be run in the web archive. They are the intermediaries between readers and IT specialists who can develop analysis tools, as the web collections' uses evolve from browsing documents and content site by site to data-mining.

The web curators' training

The curators working on web archives are rarely in direct contact with the general public in the reading rooms, and rely on colleagues for advocacy and troubleshooting. Yet, there are no specific courses on designing and managing web archives. Training is then tiered: the core team, requiring a higher degree of technical knowledge, needs to be in contact with colleagues from the seventy or so institutions engaged in web archiving, through international consortia and events such as IIPC[5] and iPRES. In turn, they train the curators associated with web collection building and coordinate their activities over time. The Library of Congress has for instance made web archives part of the core training on curators' duties. To enable all personnel to raise awareness of web archives and answer readers' queries, general information sessions should also be organized.

[3] K-12 Web Archiving, https://archive-it.org/k12/.

[4] Some web collection development policies examples are available on the IIPC website: http://netpreserve.org/collection-development-policies.

[5] The BnF, for example, held an IIPC-sponsored workshop on integrating a web archiving program in one's organization in November 2012. Report available at http://netpreserve.org/sites/default/files/resources/Putting%20it%20all%20together.pdf.

5. Lessons learnt about professional identities

Each of these three studies illustrates a different aspect – at a different stage – of digital curation: the second study deals with entry, the first study with bibliographic description, the third study with the initial and final steps of the chain: selection on one hand and promotion on the other hand. Also, different kinds of professionals were under scrutiny: library assistants, cataloguers and curators. And yet, is it possible to take a broader view and go beyond the specificities of these use cases, not striving to define precisely what digital curation is, but at least identifying some of its key trends?

A first conclusion of each of these studies is that the notion of a "digital collection" is not obvious. Digital documentation makes it necessary to reassess the notion of collection granularity: each subset of a document (a slide in a batch of pictures, a web page in an archived website…) is a document in itself, which may need a specific description. Professionals even tend to lose the very notion of document: when someone's activity is essentially related to figures and reports, is this still managing library collections? As librarians seeking to promote web archives acknowledged, readers may also have a hard time understanding the kind of documentation they are dealing with when faced with digital collections. Finally, the sheer diversity of digital documents (digitized and born-digital; acquired by the library temporarily or for the long-term) makes it difficult to figure out the unity of the institutional digital collection.

There are subsequently questions on the roles, responsibilities, and even professional identities of people in charge of digital curation. It is stated that librarians' "traditional" knowledge is still invaluable when it comes to digital documents: for instance cataloguing remains a critical part of librarians' activity (at least in a national library such as the BnF). The knowledge of news publishing required for library assistants to manage serials is also precious for their digital equivalents. However, typical distribution of roles between professional profiles is challenged: the assignment of descriptive metadata isn't the duty of cataloguers anymore; engineers replace library assistants at the heart of the document entry process. Curators working on web archives are still wondering what level of technical knowledge they need to acquire so as to adequately fulfill their mission. As this last example shows, organizational charts are often called into question by the emergence of new activities and evolving share of responsibilities.

As a matter of fact, treatment chains tend to be longer, more complex, and to involve more stakeholders. As they are more recent, these chains are less well-known and less intelligible for the various actors. Therefore, there is a risk that actors no longer recognize their place in the workflow, and that they don't understand why they are asked to perform one task or another. Training programs thus appear to be critical in order to tackle this issue. Digital curation of course requires technical training on digital documents specificities. It may however be considered that digital curation requires less a technical knowledge than a global understanding of production channels.
Finally, several concerns were raised about the way to articulate work on analogue and on digital documents. Whatever the domain may be (metadata description, digital entries, web archives), the activity was at first experimental, with highly motivated pioneers. Involving whole teams in the activity is then as difficult as it was to initially launch it: however, sharing the workload between the first pioneers and the rest of the staff appears as a key to avoid creating a gap between the library's missions towards physical and digital collections.

Acknowledgments
The authors wish to thank ORHION members and those who kindly reviewed the papers.

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

[2] Bermès E., Fauduet L. The Human Face of Digital Preservation: Organizational and Staff Challenges, and Initiatives at the Bibliothèque nationale de France. In *International Journal of Digital Curation*, 2011, vol. 6, n. 1 [http://www.ijdc.net/index.php/ijdc/article/view/175/244].

[3] Clatin M., Fauduet L. Watching the library change, making the library change? An observatory of digital influence on organizations and skills at the Bibliothèque nationale de France. *In Proceedings of the 78th IFLA Conference* (Helsinki, Finland, August 2012). [http://conference.ifla.org/past-wlic/2012/150-clatin-en.pdf].