Abstract:

Scriptorium platform was launched by the State and University Library Lausanne (BCU Lausanne) in December 2012 in order to give an open access to older editions of major newspapers of Swiss Canton de Vaud (http://scriptorium.bcu-lausanne.ch). Within four year, a corpus of more than 3 million pages has been digitized and processed. It is dedicated to newspapers published from 1732. The first collection scanned and launched on the platform concerned the daily paper intitled 24 Heures (still published by Tamedia SA) and its ancestor Feuille d’Avis de Lausanne, from its first foundation in 1762 to 2001. It includes now 11 different collections grouping about 100 different titles, some of them representing a huge number of issues, other only a few ones, the whole corpus coming to nearly 200'000 issues (single documents). The more recent collections, like Vigousse (satirical paper) or the University of Lausanne periodicals (Uniscope and Allez savoir!) concern year 2015. Scriptorium is gradually enriched by adding new collections of papers.

Scriptorium gives free access to the full text (5 billion words!) and provides very fine research functionalities. Scanned pages, parts of pages or complete issue can be downloaded or shared using specific links generated “on the fly”. Personal accounts, with a log, permit to set bookmarks, to make specific annotations on documents and to customize the platform with own settings. Scriptorium works on MediaInfo system (http://www.mediainfo.com/), developed by Geneza, a company based in Slovenia. Since March 2016, a new fully html5 and responsive version of the platform has been launched.

The paper will focus on the main aspects addressed by the digitization project, in the years 2011-2016: specific context, partners, problems of copyright, difficulties and troubleshooting faced, costs, feedback, consultation statistics, etc.

Keywords: newspaper, digitization, Switzerland, archives, web, viewer, opa, BCU, Lausanne
Goals of the Scriptorium project, general context

Scriptorium platform was launched by the State and University Library Lausanne (BCU Lausanne) in 2012 in order to give an open access to older editions of major newspapers of Canton de Vaud, one of the major Swiss states, counting 750’000 inhabitants.

Scriptorium project is part of a wider ambition of the BCU Lausanne in matter of digital services: since 2005, digital library is part of the strategy of the institution, including digitization of out-of-copyright books and newspapers.

Created in the middle of the 16th century, almost 5 centuries ago, the BCU Lausanne has a double mission: it plays both roles of the University Library and State library, the Canton de Vaud being one of the 25 states forming the Swiss Confederation, each canton being in charge of culture (no general Culture Ministry in Switzerland!). That’s why our institution is in charge of the legal deposit of all books and papers published and printed in Vaud canton. And that’s why our goal in matter of digitization is devoted in priority to documents related to Vaud canton.

For what concerns digitization of ancient documents belonging to our cultural heritage, which refers more specifically to my job as rare books curator, two major projects were led by the BCU Lausanne in the past ten years.

The first project involved the digitization, in collaboration with Google, of more than 100’000 volumes published before 1870, including almost all books published and printed in Vaud canton from the 17th century held in our library. It was discussed in 2006 and realized during the years 2007 to 2010. This important corpus of printed books is available for free consultation through the Google Books platform. Specific links to the scanned version have been added to the bibliographical records in our online catalogue.

The second project was devoted more specifically to the newspapers published in Vaud from the first ones, printed in the 18th century, until nowadays. Such a project involved creating a platform allowing us to give a friendly access on the web to the newspapers we wanted to digitize.

This platform, launched in December 2012, was christened “Scriptorium” in reference to the monks from the Middle Ages who patiently copied and transmitted to us the texts of the Antiquity.

We think that the massive digitization of printed books, review and newspapers belonging to our cultural heritage represents a challenge for an institution like ours. It will give to contents concerning our region something like a “second life”. It will allow, too, new perspectives in the data mining of huge quantity of written contents and stimulate historical and sociological research and inquiries.
Why opt for a dedicated platform for publishing digitized papers? Given the dimensions of the pages, a lot bigger than those of standard books, no easy to display on a screen, given that a newspaper’s page contains different individual articles, given that a single title frequently includes thousands of issues, we decided to opt for a specialized tool, not a simple viewer.

Our aims for such a tool were the following:

On user’s side:
- Simple and clear display of the collections available on the platform
- Easy and attractive display of the content of an issue
- Easy and attractive display of the content of a page or part of a page (zooming)
- Easy text search in whole or part of the contents (indexing)
- Possibility to generate a link for a specific collection, issue, page or part of a page
- Possibility to share contents via social networks
- Free download for documents or part of a document
- Open Access for everyone

On the administration side:
- Scalability of the system (corpus of contents is supposed to grow quickly!)
- Scalability in upgrading: html5 (no flash!), responsive design, etc.
- Metadata import as automatized as possible
- Minimized licence and maintenance costs: Scriptorium projects don’t benefit from special funding, so the costs involved are to be taken on by the operating budget of the library.

The first collection scanned and launched on the platform was the daily newspaper entitled 24 Heures (still published by Tamedia SA) and its ancestor Feuille d’Avis de Lausanne, from its foundation in 1762 to 2001. The corpus now includes 11 different collections grouping about 100 different titles, some of them representing a huge number of issues, other only a few ones, the whole rising to nearly 200’000 issues (single documents). The most recent collections, like Vigousse (satirical paper) or the University of Lausanne periodicals (Uniscope and Allez savoir!) concern year 2015. Scriptorium is gradually enriched by adding new collections of newspapers.

Within four years, a corpus of more than 3 million pages has been digitized and processed. According to our estimation, this corpus represents about 50% of the global mass to digitize. It concerns newspapers published from 1732 until 2015.

Scriptorium gives free access to the full text (5 billion words!) and provides very fine research functionalities. Scanned pages, parts of pages or complete issue can be downloaded for free or shared using specific links generated “on the fly”. Personal accounts, with a log, permit to set bookmarks, to make specific annotations on documents and to customize the platform with own settings. Scriptorium runs under MediaInfo system, developed by Geneza, a young company based in Slovenia.media.
Some problems to solve when digitizing a new collection…

One of the main problems to resolve when starting a newspaper digitization project concerns the copyright. It is very important to identify the right holders in order to clarify this point when starting a project related to a paper still published. In the context of Scriptorium, we had to lead such discussion for a handful of titles. Signing a agreement with the owners of the title was essential for us, because, according to guidelines elaborated by the Swiss National Library in collaboration with Presse suisse (Schweizer Presse, the organization representing the Swiss newspapers publishers), the publisher will assume the collective responsibility for all copyright holders involved in a specific paper (redactors, photographers, etc.). Without such an agreement, it would be totally impossible for an institution like ours to digitize any newspaper!

Given that we belong to the administration and that our aims are not commercial ones, given that, too, we put a great attention not to despoil anybody publishing on Scriptorium issues of papers still commercialized, we got from each right holder consulted a very good feedback, some of them even accepting to contribute financially to the costs of digitization. These discussions are finalized in an agreement signed by both partners.

The problem of newspapers no more published can be more complicated. We make, for each specific title, a serious inquiry in order to find who the last owner of the paper was. If it is impossible to identify the legal owner, we discuss each case in order to take a decision. At this time, we didn’t have to renounce to any project.

Another problem when starting a newspaper digitization project is to identify the possible gaps in the holdings available at the library, what is very frequent, alas. As far as we can, we try to locate the missing issues in other Swiss libraries or archives centres. Such mixed provenance of the issues must be very tidily integrated to the metadata related to the collection loaded, in order to give the users a correct information about the location of the original.

By the way, the metadata concerning a newspaper is not simple to collect! The title can change frequently, as do the format or periodicity, different publishers and/or printers can follow one another. In the majority of cases, the traditional metadata available in the library catalogue are totally insufficient!

And, last but not least, the recurrent problem for us is to get enough money to realize new projects devoted to specific corpus (papers from a specific region or town, papers devoted to a specific field, etc.). We try to be proactive with regional administrations, interesting a city, for instance, to contribute financially when publishing on Scriptorium the local papers. We must stress, in this concern, the support provided by the Swiss national library, whose practice is to contribute for 20% of the cost.

Moreover the global cost projection of such scanning operations has to be estimated as precisely as possible, which is not so simple, no having any information about the exact number of pages involved. The best practises in this domain recommend adopting the 1 m. = 20’000 pages ratio.
Scanning and pre-publishing operations

The scanning and processing operations are not carried out in house but outsourced to a Swiss company specialized in mass scanning operations, ASSY S.A., located only a few miles away, allowing the BCU Lausanne to take in charge the transport of the volumes from the library buildings to our provider.

Our provider produces greyscale or colour scans (depending on whether the original is in black and white or contains pages printed in colour) in JPEG2000 format (multi-level, multi-tiles with lossless encoding) of each page. This format allows that same file contains optimized byte-stream for extremely quick extraction of required resolution and part of the image that needs to be displayed for the end user in Scriptorium. With such approach, we avoid the need to display a whole image to visitors, but instead display only the part where he is zooming. Such format and definition has a cost: the 3,3 millions of pages and the indexes related to these pages (5 billion words!) occupy 44 TB of space disk. Moreover, our provider is in charge of the naming and structuration of the files, a very important step in order for the further loading operations in Scriptorium. Improving image quality (page split, deskew, crop) final quality check and OCR processing is also part of his job. These activities represent 2/3 of the cost per page involved, the last third covering the loading of the collection on Scriptorium.

Loading a collection on Scriptorium

The loading operations, including the attribution of correct metadata to each issue published, according to the datas provided by the BCU Lausanne, is at the present state taken in charge by Geneza, the provider of MediaINFO system, on which Scriptorium is based. In a next future, these operations will have to be carried out by the library.

The BCU Lausanne has the possibility to add, for each collection or title, a specific information page (history, particularities, precision about collections scanned, etc.). The logos of the partners involved in each project can be displayed on the dedicated page.

Why choose MediaINFO system?

When the BCU Lausanne decided to give a free online access to the main digitized newspapers published in our region, we knew we couldn’t propose a standard image viewer.

In this context, we had 3 options :
- build a custom-made platform by ourselves
- opt for the most spread system for the visualization of digitized newspapers, Olive Active Papers
- opt for a new system developed by a team of engineers located from 2009 in Slovenia, named MediaINFO, soon used by some National Libraries in Europe (Norway, Slovaquia) and implanted in Switzerland (e-publishing of Presse polytechniques et universitaires romandes, online publication of Domaine public and L’Emilie, two Swiss papers still alive).

Given that creating a custom-made system would be very complicated with the few resources of our IT, we decided to opt for an existing system. We asked tenders to both companies.
involved, Olive and Geneza (MediINFO). Not having many financial resources and having to digitize and publish very important quantities of documents (millions of pages), we opted for MediaINFO, which offer granted the best quality/price ratio. We had though to give up one major feature proposed by Olive: that is the manual structuration of each page in articles, advertisements, obituaries, etc. Such a structuration, very time consuming, have a high cost, and we opted for a system which is avoiding such manual, but still offering a lot of interesting features for the user and up-to-date perspectives regarding the last computer science and web developments. From this year, a fully html5 and responsive version of the platform has been launched.

On the technical side, MediaINFO runs on Linux servers and use Lucene (Apache SOLR) for building and mining the words indexes, allowing stemming, and proximity factor settings, etc.

Next updates expected (next months):
- Possibility for administrators (i.e. librarians) to generate periodically “story pages” with contents extracted from to documents loaded in Scriptorium.
- Chart giving the user an idea of chronological occurrences of searched words within the corpus.

**On the financial side**

On the financial side, the project Scriptorium is not a cheap one in the scale of a library like the BCU Lausanne.

For the past 3 years, a budget from about 100’000 euros was devoted to digitize and load on Scriptorium new collections of papers. This sum doesn’t include the annual administration and maintenance fees of the platform (about 25’000 euros).

MediaINFO users have to pay a licence in proportion with the number of pages published. Our first licence concerned 2,5 million pages and cost about 100’000 euros; the present one will run until 3,5 million pages. With 3,3 million now, we have to anticipate for 2017 a new upgrade of our licence.

As we can see, the Scriptorium project represents no doubts for the BCU Lausanne a big financial investment. We think though it is worth going on with it until a complete panel of the papers published in Vaud canton from 18th to 20th century will be available.

**Feedback, consultation statistics**

The feedback received from users is globally positive. Negative comments were mostly due to:
- the use of flash in the previous version of the platform
- bad understanding of the user interface
- slowdown of the system due to a lack of computer resources during the first 6 months.

The new version of the platform, html5 compatible, is considered by the great majority of the users as a great improvement.
In 2015, following statistics have been collected:
- number of users that have reached Scriptorium main page: 87'077 (238 per day)
- number of generic searched from browsing window: 1'155'938 (3’167 per day)
- number of searches inside specific object, searching for highlights: 700’335 (1’919 per day)
- number of objects (issues) that were opened: 766’290 (2’099 per day)
- number of requests for download of whole object: 15’930 (44 per day)
- number of requests for download of selected pages: 26’095 (71 per day)
- number of requests for download of selected zones: 2’995 (8 per day).
Scriptorium features

Starting from the general homepage
The user has the possibility to make a search within all the collection available or to select one or more collections or sub-collections.

For example:
- **Nouvelliste Vaudois**
- **Le Grutli** (sub-collection of **Presse socialiste**) collection

Within this selection, one’s can select a period …

… getting a set of issues corresponding to the filters defined:
Within the selection, one’s can add a full-text argument to search for:

The matching issues obtained can be displayed in 3 different forms…

… and sorted in 3 different ways with ascending/descending option
Let’s have a look to one of the matching results...

UNION OUVRIERE DE LAUSANNE
Assistance des Débuts
Kremlin, 15 janvier, 5 et 6, à 19 h. 30
Salle du Palais, de la Place
Pour une durée intéressante.

Le Congrès socialiste international de l’ISPO.

LA FÊTE DU 1° MAI
A la maison de la justice de l’ouverture
Le 1er mai est un jour de fête
C’est la fête Syndicale, déclaration
et l’affirmation de la classe ouvrière.
Sous le signe de la lutte pour la liberté
et l’égalité des droits des travailleurs.

Le Congrès international de l’ISPO.

La mairie de la ville de Lausanne.

Le 1er mai, jour du
Congrès socialiste international de l’ISPO.

Le 1er mai est un jour de fête
et de célébration de la classe ouvrière.
La lutte des travailleurs pour la liberté
et l’égalité des droits.

Le Congrès socialiste international de l’ISPO.

Le 1er mai, jour du
Congrès socialiste international de l’ISPO.

Le 1er mai est un jour de fête
et de célébration de la classe ouvrière.
La lutte des travailleurs pour la liberté
et l’égalité des droits.
Displaying next or previous matching document is possible by using the command “Doc, precedent / Doc. Suivant”

Displaying next matching result within the same document is possible by using the “arrow” command (1/5 to 5/5).

Operation possible on a displayed document (issue):
5 options are available (symbols in left column)

- Allows user to display the metadata relative to the document
- Allows user the make a full text search in the document
- This command is used to download entire document, chosen pages or specific part of a page
- This command offers the possibility to choose a display for the pages
PDF obtained:

**Jour du redde rationem de la part des capitalistes allait bientôt venir.**

À la notification du décret de 1904, les ouvriers des diverses usines de Gênes, de Sampierdarena, de Sestri Ponente et d'autres pays environnants, se mirent en grève.

Les ouvriers du port de Gênes, ainsi que les conducteurs des tramways de la ville suivirent cet exemple.

Plus de vingt mille ouvriers étaient en grève.

Assistés par leur député, le camarade Pietro Chiara, ainsi que de Bisolati et Agnini, et quelques députés républicains, les ouvriers se sont tenus tranquillement les bras croisés, repoussant les invitations des patrons à la reprise du travail sous peine de renvoi général.

Une commission élu par le conseil des ouvriers s'est rendue à Rome auprès du président du Conseil des Ministres, Saracco, lequel, après quelques hésitations, s'est décidé à accorder aux ouvriers la composition de la Chambre de Travail et la restitution de tous les objets saisis par la police. La grève a duré que trois jours et la victoire a été complète.

Les ouvriers ne demandaient rien de plus que le respect du droit de coalition sanctionné par la loi.

Ils ont triomphé.

Nous nous félicitons avec nos camarades d'Italie des résultats superbes de leur propagande.

Angletiere. — Fascisme avec Indépendance. — Il y avait d'abord quatre millions d'Indiens assistés et menant de faim, les journaux annoncent qu'il y en a plus que 3,950,000.

Il n'en reste que ce nombre par la raison bien simple que 700,000 sont morts de faim.

**La Grutli**

Le Melocht militariste continue ses exploits. Et tandis qu'il succède jusqu'à la moitié, la force productive du peuple suisse, on ne trouve pas deux ou trois millions pour le développement de l'instruction publique ou pour une amélioration de la législation protectrice du travail.

C'est une honte.

Il en est toujours ainsi, tant que le peuple, en son ignorance et en sa stupidité, continuera à envoyer aux Chambres des colonels, des majors, toute la clique dorée sur épaule et sur caquette.

Un peuple a les législateurs qu'il mérite.

**Italie.** — Une conversation. — M. Merlini, l'avocat défenseur de Brescic, qui a toujours milité dans les rangs anarchistes, vient de faire son adhé-
Generating a link to specific contents:
Command (on the top of the screen) at right offers users the possibility to generate a link to a specific content and to share:

Link to copy or send via social networks:
http://scriptorium.bcu-lausanne.ch/s/0jNovPMLXD

Last point: advanced search makes possible researches in metadata fields related to a document (not an article!). The user can define a specific proximity factor value and activate or not the stemming in order to harvest more or less results.
Personal notes
Enjoying a personal account (in a next future every person registered in our library system will get such an account), the user has the possibility to make annotations related to document

GENÈVE
Quand des enseignants osent faire la grève
Mercredi après-midi 4 mai, dans douze de sept collèges du Cycle d’orientation de Genève, c’était donc la grève. Décidée par les assen- de chaque établissement, suivie par plus de moitié des enseignants, elle avait été déclenchée par la Direction générale du Cycle d’orientation (CO) et le chef du Département de l’instruction publique. D’une manière générale, tout en reconnaissant le bien-fondé des revendications de la FAMCO (Fédération des associations de maître du Cycle d’orientation), ils considéraient que la grève n’était pas le moyen approprié d’influer sur l’opinion publique. L’événement reste en soi considérable. Il nous邀 à sa source pour tenter de compr