Digitization of copyright protected newspapers in France

Bruno Sagna - Bibliothèque Nationale de France

The Bibliothèque Nationale de France (BnF) as a national library is in charge of the legal deposit of newspapers in print or digital format. Thus the Legal Deposit Law of the 1st of August 2006 is requiring the BnF to "crawl" the available to the public websites of news organizations. About 100 news websites are harvested daily. Further newspapers' sites are harvested by various intervals. As the result of an agreement with publishers the files of 22 daily newspapers are crawled in depth day by day.

Since 2005 the BnF has undertaken a large scale digitization program covering 30 newspapers published between 1799 and 1944 out of which the running newspapers *La Croix, Le Figaro and L'Humanité*. 1944 is a mile stone in the history of French newspapers. Since copyright issues must be attended. Moreover the huge task of digitizing 80 Mio pages just for the published newspapers of the 3rd French republic has to be taken into account. This is comparable with the 5.5 Mio pages in the *Gallica*.

Due to the lack of any legal agreement with collecting societies the BnF is taking newspapers for collective works, which are public domain after a period of 70 years after publishing the first issue as defined in Art. L. 113-2 paragraph 3 of the French intellectual property law. Concerning the year of 2015 the newspapers, which are online available online outside the BnF, are not exceeding the date of December 31st 1944. But the worldwide online availability is going on year by year like a moving wall.

Nevertheless the law entitles the BnF to digitize al the document, which are archived and stored by the legal deposit for preservation purposes. You may have access to digital copies of copyright protected materials only in the premises of the BnF Research Library unless an agreement exists with publishers as it is the case for *"Le Monde diplomatique"* for the period 1954 - 1978. The following issues of this newspaper are licensed commercially.

Regarding digital newspapers a decree issued on the 16th of September 16 2014 by the French Ministry of Culture extends the consultation of Internet Archives beyond the BnF to 26 libraries across the whole country and is enabling broader access to digital newspapers - especially born digital newspapers. The countrywide responsibility for the dissemination of these materials was assigned to the BnF.

Legal issues are extremely relevant due to the high usage of newspapers as primary information resources by researchers. In July 2015 the European Parliament voted in favor of Julia Reda's own-initiative report on updating and harmonizing the EU's copyright framework. The commissioner of digital economy and society in Europe, Günther Oettinger, is expected to bring forth a legislative proposal by the end of the year.

In this context the French position can be considered as cautious with a national consensus concerning intellectual property, a position reiterated by French MPs in the European Parliament. There is no support of the proposal to modify copyright rules from a term of life-plus-70 years to life-plus-50 years and collective licensing models as known in Nordic countries are seen as quite foreign.

France does not have any legislation yet regarding TDM, one of the core topics as demonstrated by *The Hague Declaration*. In 2014 the CSPLA (Conseil supérieur de la propriété littéraire et artistique) mandated by the Ministry of Culture to write a report about TDM practices and intellectual property infringement submitted the conclusion and pointed out, that any legislative change could only take place at the European level. However the SNE (Syndicat national de l'Edition) has opposed any exceptions of TDM and proposed license agreements.

Regarding the current accessibility of content the claim should be: content first, but readers first as well! The quality of the user's experience is linked to technical choices of formats. The BnF has chosen the METS/ALTO format in order to provide two main requirements: First to deal with text operated by automated procedures (OCR) including the accuracy rates or variants of words suggested by the OCR engine. Second to keep the link to the source image by encoding the layout information (coordinates) on the block, on the line, and even on the word level.

The participation to the European project *Europeana Newspapers* made improvements of techniques of OLR (Optical Layout Recognition) possible, which is concerned with the detection of articles and the recognition of their content types. Named Entity Recognition is also to be considered within the broader discussion of TDM. From that background the interface of the digital library is the key for improvements like the standardization of calendars, because the choice between list view and calendar view are enabling searches within the scope of a given year – and not to forget the comfortable facilities of zooming and downloading.