A digital future for yesterday's news

Marten Düring, Estelle Bunout, Steven Claeyssens, Sally Chambers, Pim Hujnen, Jaap Verheul, Clemens Neudecker

Corpus creation and digitized newspapers - perspectives from research and libraries

Digitized newspapers represent one of the richest resources available for (digital) humanities research. Even though full-text search technologies are improving access to these large scale digital collections considerably, their size leads to researchers getting lost in this wealth of information. The creation of sub-corpora, based on specific research questions, seems to reflect the current practices of humanities researchers. Yet, how can these sub-corpora best be created? Is this something that researchers can do - or learn to do - themselves? Or is further technical assistance required? Could API's be developed to facilitate this? What about the need for researchers to build corpora using resources from different libraries and collections? Is sub-corpus building a one-off activity or something that is done iteratively throughout the process?

From a library perspective, key drivers for digitization are preservation, representativeness and improved access and usage of their collections. Many libraries also need to serve the general public, in addition to the research community. Providing a range of user-friendly access possibilities to the digitized newspaper collections, such as APIs, search interfaces, bulk downloads, Linked Data and International Image Interoperability Framework (IIIF) is key. Furthermore, enriching the source collections with additional layers of metadata, such as OCR, OLR and Named Entities, is a standard service that libraries could provide. What is the role of the library in the creation of sub-corpora? If libraries prepare sub-corpora they think could be of interest to the researcher, such as images, obituaries or advertisements, are they actually what the researcher wants or needs? Or, should the creation of sub-corpora be done on a case-by-case basis, together with the researcher? Yet, is a case-by-case basis a sustainable solution that could become a standard service in a library? Or, is it a worthwhile investment bringing libraries closer to their research partners? Furthermore, how do libraries manage the evolution of the collection? Could additional annotations created by researchers, be included as an additional layer in the collection ecosystem?

In this short paper we argue that researchers and libraries are experiencing similar challenges and with their complementary skills, closer collaboration will lead to a more fruitful experience for

all. We will explore how we can find a balance between the need of tailor-made solutions for humanities researchers, with the need of libraries to provide generic services, tailored to serve as many users as possible. Is it possible to translate humanist research questions into computational queries? Could we rethink the creation of sub-corpora - as an intellectual task - which is an integral part of the research process? Is there an opportunity to pool the resources that researchers and libraries have to improve access to digitized newspapers for us all? For example, could researchers and libraries preparing joint project proposals, be a practical solution? Finally, with all the effort that is going into sub-corpus creation could we consider publishing these 'humanities datasets' in an appropriate research data repository? Not only would this raise the visibility of both research and the digitized collection, but it could also increase the potential for other researchers to reuse these carefully crafted data sets.

Transparency as a prerequisite for Digital Source Criticism for Digitized Newspapers

This papers argues that the concept of transparency is a means of empowerment for researchers to adjust their research methodologies to the digital world and to make the most of inherently incomplete or biased collections as well as the necessarily imperfect methods for their enrichment. Here we understand transparency as any effort to provide 1) information which helps assess the provenance and quality of a born digital or digitized source. This includes digitization, enrichment, completeness and inherent biases, 2) information which helps reveal novel opportunities when working with digital or digitized material, 3) information which documents research decisions and methods.

Historians are trained to get the most out of the necessarily limited source record which is available to them: they carefully evaluate texts, weigh ambivalence and contradiction and take into account blind spots. Digitized source collections and emerging toolsets for their analysis, offer new opportunities and similar challenges which need to be mastered. Digitized newspapers in particular are a good example for their richness and complexity. Their analysis requires transparency with regard to legal concerns, the processes of digitization and datafication by means of manual, semi-automatic or automatic processing but also with regard to their presentation through user interfaces. We argue that transparency can only be achieved through close cooperation between content-providing

institutions such as libraries and researchers: Both need to decide together which information is required, researchers are called to formulate clear requirements and to expand their existing skill-sets; content providers are called to provide transparency-relevant information in an accessible manner.

We continue by highlighting selected aspects of transparency which are of particular importance for the domain of historical newspapers but may also apply to other sources.

Digitization policies on the side of libraries are driven by pragmatic considerations as much as demand by the public including researchers. This however does not imply that only the most popular or prominent titles are being selected for digitization and raises a number of crucial questions for researchers: Which opportunities do large digitized collections offer for novel research in their field? How can we create communication channels between researchers and content providers to jointly shape future and ongoing digitization projects? But also: what are the consequences of a full reliance on digitized collections? Which perhaps crucial sources would be missing and how can such gaps be closed?

With regard to an existing corpus, the simple question “what is (not) there?” poses a considerable challenge with regard to large (newspaper) collections: digitized collections entail several layers of information, on the digitization process, content, layout, text recognition, any form of manual or automated enrichment but also in terms of provenance of the collection, the institutional context of its (re)compilation. An understanding of “what is there” is crucial for an assessment of the relevance of any scholarly finding.

Transparency is a two-way street: to be meaningful, it requires the provision of accessible information as well as sufficient technical understanding so as to make methodological choices. We argue that the required transparency is best achieved through exchanges between the libraries and the researchers as early as possible in their respective workflows.