# AI, Data Curation and the Data Readiness of Heritage Collections: Exploring the Swedish Newspaper Archive at KBLab

Justyna Sikora[1,*,†], Chris Haffenden[2,†]

[1]KBLab, Kungliga biblioteket, Karlavägen 100, 115 26 Stockholm, Sweden
[2]KBLab, as above

**Abstract**

The increasing availability of digital material and tools for large-scale computational analysis has produced a growing interest in big data approaches in the humanities and social sciences. However, the vital role of *data curation* as a precondition for such projects remains underappreciated. This paper details the work of KBLab at the National Library of Sweden in testing AI tools to help curate the digitized newspaper archive and make it more amenable to quantitative, machine learning-based research. It provides a description of the library's newspaper data to offer orientation to researchers interested in the material, before turning to recount the results of our exploration with automated data curation. It concludes by sketching possible next steps for these exploratory efforts, as well as situating this project within a broader recent turn to conceptualize and prioritize the notion of *data readiness*. Its principal argument is in drawing attention to data curation as an essential part of any digital research project, not something prior to or external from the research process.

**Keywords**

Data curation, data readiness, digitized newspaper archives, document AI, digital research infrastructure

## 1. Introduction

Digital research presumes digital data. This is a platitude, but bringing it into focus helps illuminate the critical role of infrastructural questions within the research process. While the increasing availability of large-scale digitized corpora and tools for computational analysis has produced a growing interest in operationalizing big data in the humanities and social sciences, significant blindspots remain about the complexities involved in such projects. A specific challenge is dealing with the varying gap between i) the output of the digitization process and ii) the attainment of machine-readable data of sufficient quality to pursue credible research. In short, there persists a lack of recognition of *data curation* as an enabling condition for digital research [1, 2]. This is a problem since, as Lisa Gitelman has argued, "raw data is an oxymoron" [3]; there is no such thing as ready-made data. Instead, data needs to be prepared and curated according

*Huminfra Conference 2024, Gothenburg, 10–11 January 2024.*
*Corresponding author.
†These authors contributed equally.
justyna.sikora@kb.se (J. Sikora); chris.haffenden@kb.se (C. Haffenden)
0000-0002-5561-5163 (C. Haffenden)

to the particular specifications of a given project, making it a necessary practice that invariably demands resources and shapes results.

Data curation "encompasses gathering material, making it discoverable by describing and organizing it, placing it in a context of related information, supporting its use for diverse intellectual purposes, and ensuring its long-term survival" [4]. In the context of digital research, this touches upon a key question of multidisciplinary work focused on big data: how quantitative approaches best interact with more qualitative humanities scholarship [5]? Delineating what she terms the "dangerous art of text mining", Jo Guldi warns that data scientists risk producing "analyses that are empty, biased, or simply false", if they lack awareness of the particular history and context that has formed humanities data [6]. Though a pertinent warning, it tends towards a one-sided version of myopia: emphasizing the shortsightedness of data science while underplaying the occasional data naivety of the humanities. Here we can recall Neil Lawrence's observation that technical project members are "often treated like magicians who are expected to wave a model across a disparate and carelessly collated set of data and with a cry of 'sortitouticus' a magical conclusion is drawn". In outlining the unrealistic expectations placed upon data scientists, he suggested that "[j]ust as extracting drinkable water from the real ocean requires the expensive process of desalination, extracting usable data from the data-ocean requires a significant amount of processing" [7]. Linking questions of digital labour and project efficacy, this highlights the pivotal role of data curation in any effective collaboration between humanities scholars and data scientists.

This paper uses the National Library of Sweden's (*Kungliga biblioteket*, hereafter KB) digitized newspaper archive to discuss data curation as it pertains to making heritage collections available for large-scale research. More specifically, we show how AI tools can be used to automate elements of the curating process and thus cope with large volumes of data. The first part provides a contextual account of KB's data, explaining the what, how and why of the newspaper archive, including problems resulting from the digitization process. The second part details our exploratory work at KBLab, the library's data lab for digital research [8], in testing AI models to curate the digitized newspaper material. We have two key aims with such an account: providing orientation for researchers interested in using KB's newspaper data; and contributing towards a recent trend foregrounding an active approach to data readiness [7, 9, 10].

## 2. KB's collections as data

As a national library, KB collects and hosts a wide range of data sources: from postcards, radio broadcasts and computer games to more conventional print material like books and ephemera. According to Swedish legal deposit legislation, a copy of every published item is to be stored at the library for the benefit of future users. This makes KB's archives an invaluable resource for researchers, both for the historical depth and breadth of the collections.

Thanks to recent digitization efforts, a growing volume of this material is accessible in digital form. One of the services providing access to part of the digitized collection is tidningar.kb.se. Via this website, users can search through more than 300 years of newspaper material. However, for researchers seeking to conduct larger-scale, programmatic analysis, another on-site service is available at the library's premises via KBLab: the datalab.kb.se platform that makes it pos-

sible to engage with the collections as data, including the extensive newspaper archive (See: github.com/Kungbib/kblab/).

### 2.1. Opportunities and challenges with digital heritage data

Access to the digitized materials via KBLab enables a new type of research focused on quantitative methods *at scale*. While undoubtedly providing new opportunities, this can create challenges for more qualitatively-inclined scholars, since identifying relevant data for a project can be difficult given the extent of the collections. Moreover, finding pertinent data may only partially solve the problem of locating suitable material. Considering the content of the data is obviously the first step when investigating potential research material. However, another crucial part is determining the data readiness of these sources - i.e. is the data ready to be used as a dataset for research? Are there gaps or duplicates? What sort of data wrangling might be required before it is fit for purpose [7, 10]? Being able to use the data is thus as important as finding it; beyond locating sources, researchers also need to consider any problems these might pose in terms of data handling.

It might be thought that coming to the library premises and accessing the digital collections would be the final step in defining a research topic and seeking suitable resources for a project. Yet taking care of these issues does not necessarily mean the most challenging part of the work is complete. As the data is not always available in a user-friendly format for qualitative scholars, more preparation may be needed before any data analysis can be started.

If we consider text data, it is easy to assume that searching it would be as straightforward as skimming a newspaper to find relevant articles about the topic of interest. Instead, a researcher arriving at the lab to work with the digital collections will encounter a potentially daunting view of rather complex data structures, for parts of the collection that might not previously have been explored. While familiar ground for a data scientist or statistician, dealing with structured data might be a new experience for a humanities scholar with a qualitative background. This reinforces the point above, that finding appropriate research data also requires thinking about the data readiness of the available sources.

In sum, the turn towards quantitative methods from the data sciences can provide a means of dealing with the issue of scaling up research, creating new pathways for humanities research. But it also introduces pressing new challenges and skill requirements not previously part of the humanities toolbox, such as assessing material in terms of data readiness.

#### 2.1.1. What does the newspaper data look like at KBLab?

One of the most requested parts of KB's digital collections is the newspaper data. The library's holdings consist of over 1,900 titles, spanning from 1645 to the present. The newspapers available via datalab.kb.se are scanned and processed using optical character recognition (OCR). As a result of this process, the pages are broken down into bounding boxes and the corresponding text. This also means the data is not structured into clear units such as articles. Instead, it is available as consecutive blocks of text along with their corresponding scanned images and the coordinates pointing to the OCR boxes containing the texts. In other words, an effect of the digitization process - beyond the introduction of OCR errors [11] - is the loss of various metadata we take for granted, i.e. which parts of a page form part of the same article, and which parts of the newspaper

comprise the same section, e.g. "sport" or "culture". This format currently requires a certain level of data literacy to work with. Initial efforts have therefore been made towards preparing the materials into a structured dataset.

#### 2.1.2. Newspapers' structure and the complex issue of layout

Without the constraints of finite resources and a huge volume of complex data, the ideal scenario would be to provide researchers with structured datasets, with the possibility to search for individual articles about certain topics or with specific keywords. However, at present the OCR processing and the resulting structure of the newspaper data pose a substantial challenge to this goal.

A significant part of this challenge is about the stripping of metadata that results from digitizing the physical newspaper, especially reassembling the OCR boxes into a coherent order. From the perspective of machine learning this is about layout analysis, which is a complex task and where newspaper data is more challenging than other document types. When we consider material like receipts or contracts, one page typically contains coherent information belonging to a single document. A newspaper page, by contrast, may consist of multiple articles scattered over the page, and articles spread across multiple pages.

For layout analysis, working with historical newspapers is more straightforward than handling modern newspapers. With the layout of historical newspapers, the articles tend to be organized in long columns on a page. When we look at a page of a modern newspaper, identifying the boundaries of an article might appear easy at first glance, but this is not always a trivial task even for a human. Page layouts often subvert the left-to-right and up-and-down reading logic. Moreover, article paragraphs may describe different topics, making it more challenging to group them.

The most significant aspect determining the complexity of layout analysis of modern newspapers, though, is the evolution of layouts over time and their variation from one newspaper to another; it is far from a standardized, static problem and there is no one-size-fits-all solution. Simple heuristics cannot be applied to extract the text and reconstruct the articles. A more robust approach is needed to dynamically handle the diverse page design.

## 3. Can AI help with data curation?

With the growing number of documents produced in every aspect of life, the need for automated processing and information extraction is increasingly pressing. Looking at each and every individual document is not feasible anymore. A growing branch of AI that addresses these challenges, and one focused specifically on developing tools for processing various OCR-ed materials, is Document AI [12]. While the greater part of this research has targeted data such as receipts or contracts, any document in PDF format can be a subject for Document AI, including newspapers.

The current state of the newspaper data may, as mentioned, seem chaotic or overwhelming, particularly to those unfamiliar with structured data. To alleviate this, we can test leveraging recent advancements in Document AI. Given the complexity of newspaper layouts, performing

a full-scale layout analysis on this data is overly ambitious. However, one step towards tidying up the collection to present it in a more approachable way is to separate i) body texts from ii) headlines, captions and all other text not considered part of the main text. Insofar as this allows us to obtain the main content of the newspapers, this constitutes an experiment in automating part of the data curation process.

### 3.1. Image transformer and training of body text model

Inspired by the training objectives of language models such as BERT [13], the transformer architecture has been successfully applied to image processing. Several models such as LayoutLM and BEIT have been pre-trained on numerous images and can be further fine-tuned to solve tasks like document image classification or semantic segmentation.

The aim of integrating image transformers in processing the newspaper data is to distinguish the main body of articles from the surrounding page content. This is a first step towards creating a comprehensive dataset based on the newspaper materials accessible to researchers on-site at KBLab's premises.

### 3.2. How was the model trained?

To provide a good variation of newspaper layouts, issues from four Swedish newspapers – *Svenska Dagbladet*, *Aftonbladet*, *Dagens Nyheter* and *Expressen* between the years 2010-2021 – were sampled as training data for the model. Afterwards, the newspapers were annotated for two classes: i) boxes containing body text and ii) the rest of the contents on a page, including headers, captions, images etc. In total 64,837 boxes were annotated. The data was then divided into training and test sets, which were subsequently used to fine-tune a Document Image Transformer (DiT) model [14].

The DiT model was pre-trained using a masked image modeling task, meaning a number of inputs were randomly replaced with a [MASK] token. To conduct the pre-training, the input images were divided into non-overlapping patches and converted into visual tokens. The tokens were obtained from a custom discrete variational auto-encoder (dVAE). In contrast to other image transformers, the DiT model was pre-trained on a dataset consisting of 42 million document images to enhance the performance on scanned data. This approach makes the model especially well-suited for processing the OCR-ed newspaper data. The objective of the pre-training was then to predict the masked discrete visual tokens with the output representation.

In this work, we have fine-tuned the base version of the DiT model, which consists of 12 layers of transformers block with a hidden size of 768, to solve the image classification task. As in pre-training, images are split into patches and tokenized before the fine-tuning phase. The sequences of tokenized patches are then used as an input to the model, which outputs probabilities for an image to contain body text or non-body text. Only image features are taken into consideration, i.e. no additional information about textual content is provided to the model.

### 3.3. Results

After the fine-tuning for 5 epochs, the model achieved 95,5 % accuracy on the test set, which means it correctly assigned body text and non-body text labels to the test examples in 95,5 % of cases. The exemplary output is shown in the appendix below, where examples categorized as body text are marked in green, while the red boxes show the negative predictions (e.g. non-body text). As the accuracy score suggests, the body text was largely correctly recognized, with the model performing particularly well on articles with typical layout, i.e. those consisting of several consecutive paragraphs of text. The model also handles well cases where OCR-boxes resemble the main text but actually contain additional information, such as details about the authors. This applies to the byline in mathptmx [scaled=.90]helvet courier 2, for instance: taking into account the graphic features, the model correctly assigns the "body-text" label to all boxes but the last one.

An important factor that influences the results and makes the classification task more difficult is the specificity of the bounding boxes created by the OCR process. Certain of the body text paragraphs are split into multiple boxes, some of which may only contain one word or a sentence. However, these cases are rather rare. An example can be observed in Figure 2 where the lead has been divided into two parts – a longer paragraph and one sentence. The main part of the lead was correctly categorized, while the last sentence was mislabeled. This occurred most likely because a wide but short box containing one or a couple of sentences resembles more closely an image caption than body text.

Despite the model sometimes categorizing parts of the pages incorrectly, we suggest this is an important first step towards preparing the data for research. As the examples in the appendix suggest, text blocks can easily be separated from the headlines, images and other parts of the pages not considered as the main text based on the model's prediction. For a researcher wishing to pursue an analysis of the newspaper data at scale, and who might thus want this investigation to be based solely on the main body of the newspaper text, this represents a significant move towards greater data readiness for the project. Document AI thus appears a promising tool for helping with data curation.

## 4. What next?

In a perfect world, researchers working on the newspaper archive at KBLab would access structured datasets that are cleaned and adapted to use as is, with minimal need for processing. While we are far from living in such a world, the latest advancements in AI can help boost data curation and bring us closer to this goal. By testing the latest models within Document AI, our exploratory efforts suggest automated curating is a promising way to improve the data readiness of heritage collections for large-scale research.

Possible future work involves creating a pipeline with various models for processing the data. The architecture could include both a module for recognizing the body texts and one for filtering out adverts. We have already experimented with ad classification: multiple text, image and combined models have been trained to classify adverts, with the best model achieving 97,6% accuracy [15]. Since adverts often contain boxes resembling the main body, excluding boxes marked as body text in adverts would help in isolating the actual articles (even if re-assembling

the OCR boxes into these articles remains to be solved). A further option is enriching the image transformer model with additional information, such as the contents of the boxes and geometric information about the OCR-ed boxes, to turn the classification task into a multimodal problem.

More broadly, this paper has suggested the essential role of data curation within digital research projects. Rather than something prior to or external from a project, we suggest these activities should be treated as integral to the research process – both since they constitute a necessary stage in enabling the research, and since they have a concrete effect on the outcomes, i.e. how data is curated shapes the results, as doing it differently produces different outputs. In highlighting such matters, our case study forms part of a broader recent turn to conceptualize and prioritize the notion of *data readiness*. Thinking more about the readiness of data for large-scale digital research is an excellent starting point for sustainable future collaboration between data science and the humanities.

## References


[1] G. Henry, Data curation for the humanities, in: J. M. Ray (Ed.), Research data management: Practical strategies for information professionals, Purdue University Press, West Lafayette, IN, 2014, pp. 347–374.

[2] A. H. Poole, "a greatly unexplored area": Digital curation and innovation in digital humanities, Journal of the Association for Information Science and Technology 68 (2017) 1772–1781. URL: https://doi.org/10.1002/asi.23743.

[3] L. Gitelman (ed.), Raw data is an oxymoron, MIT press, Cambridge, Mass., 2013.

[4] T. Mu noz, A. H. Renear, Issues in humanities data curation (2011). URL: http://hdl.handle.net/2142/30852.

[5] M. Kemman, Trading Zones of Digital History, De Gruyter, 2021. URL: https://www.degruyter.com/document/doi/10.1515/9783110682106/html.

[6] J. Guldi, The Dangerous Art of Text Mining: A Methodology for Digital History, Cambridge University Press, Cambridge, 2023.

[7] N. D. Lawrence, Data readiness levels, arXiv preprint arXiv:1705.02245 (2017). URL: https://doi.org/10.48550/arXiv.1705.02245.

[8] L. Börjeson, C. Haffenden, M. Malmsten, F. Klingwall, E. Rende, R. Kurtz, F. Rekathati, H. Hägglöf, J. Sikora, Transfiguring the library as digital research infrastructure: Making kblab at the national library of sweden, SocArXiv (2023). URL: https://osf.io/preprints/socarxiv/w48rf.

[9] F. Olsson, M. Sahlgren, We need to talk about data: The importance of data readiness in natural language processing, arXiv preprint arXiv:2110.05464 (2021). URL: https://doi.org/10.48550/arXiv.2110.05464.

[10] M. Hurtado Bodell, M. Magnusson, S. Mützel, From documents to data: A framework for total corpus quality, Socius 8 (2022) 23780231221135523. URL: https://doi.org/10.1177/23780231221135523.

[11] J. Jarlbrink, P. Snickars, Cultural heritage as digital noise: Nineteenth century newspapers in the digital archive, Journal of Documentation 73 (2017) 1228–1243. URL: https://doi.org/10.1108/JD-09-2016-0106.

[12] L. Cui, Y. Xu, T. Lv, F. Wei, Document ai: Benchmarks, models and applications, arXiv preprint arXiv:2111.08609 (2021). URL: https://doi.org/10.48550/arXiv.2111.08609.
[13] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, I. Polosukhin, Attention is all you need, Advances in neural information processing systems 30 (2017). URL: https://papers.nips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.
[14] J. Li, Y. Xu, T. Lv, L. Cui, C. Zhang, F. Wei, Dit: Self-supervised pre-training for document image transformer, in: Proceedings of the 30th ACM International Conference on Multimedia, 2022, pp. 3530–3539. URL: https://doi.org/10.1145/3503161.3547911.
[15] F. Rekathati, The kblab blog: A multimodal approach to advertisement classification in digitized newspapers, 2021. URL: https://kb-labb.github.io/posts/2021-03-28-ad-classification/.

## A. Example predictions from the model


**Figure 1:** Predictions on a page of *Aftonbladet*, where the model has correctly identified body text (green boxes) as opposed to other texts such as headers or adverts (red boxes).


**Figure 2:** Predictions on a page of *Svenska Dagbladet*, where the model struggled to separate all body from non-body text, i.e. in the advert.