2 Doing digital research at KBLab: A practical introduction to using the National Library of Sweden's data lab

Chris Haffenden
National Library of Sweden

Justyna Sikora
National Library of Sweden

The emergence of digital heritage data and the rapid development of new AI tools for computational analysis are transforming GLAM institutions, particularly in the design of digital research infrastructure. Researchers in the digital humanities and social sciences increasingly expect to access collections at unprecedented scales. This chapter addresses such expectations by providing a hands-on guide to KBLab, the data lab at the National Library of Sweden (KB). It outlines the lab's resources, including access to KB's digitized collections and AI models like KB-BERT, and showcases innovative development projects like *Bildsök*, which makes visual archives more accessible. The chapter also details the steps to initiate research collaborations and discusses best practices for utilizing KBLab's tools effectively. By bridging technical insights with practical applications, it serves as a comprehensive starting point for conducting large-scale digital research at KB and beyond.

1 *Introduction: Library labs as digital research infrastructure*

The emergence of digital cultural heritage has significantly expanded the scope and possibilities of humanities research. With the initiation of mass digitization programs and new collecting practices for born digital material in the GLAM sector—i.e. **g**alleries, **l**ibraries, **a**rchives and **m**useums—the range and volume of heritage data potentially available to researchers has increased exponentially over the past few decades (Bingham and Byrne 2021). In tandem with this digital expansion, recent developments in Machine Learning and Artificial Intelligence (hereafter ML and AI) have enhanced

HUM INFRA Chris Haffenden & Justyna Sikora. 2025. Doing digital research at KBLab: A practical introduction to using the National Library of Sweden's data lab. In Gerlof Bouma, Dana Dannélls, Dimitrios Kokkinakis & Elena Volodina (eds.), *Huminfra handbook: Empowering digital and experimental humanities* (NEALT Proceedings Series 59), 17–56. University of Tartu Library. DOI: 10.58009/aere-perennius0171

the tools available for computational analysis of these materials. Researchers in the digital humanities and computational social sciences now expect access to GLAM collections at unprecedented scales. Rather than focusing upon the individual object, i.e. a book or newspaper, these large-scale approaches routinely focus upon thousands—or even millions—of documents to uncover new patterns or trends (e.g. Underwood 2019, Hurtado Bodell et al. 2024).

Such expectations have presented GLAM institutions with significant infrastructural challenges, not least national and research libraries with rapidly expanding electronic legal deposit collections. How can these organizations ensure that their "collections as data" are accessible to researchers (Padilla 2025)? A widespread response has been the establishment of GLAM Labs, which function as interdisciplinary spaces—both physical and digital—where data scientists and collections specialists can collaborate to address the complexities of digital heritage (Mahey et al. 2019). In particular, the Library Lab has become an established feature of the information and research landscape, with examples ranging from the US Library of Congress' LC Labs to the BNElab at the National Library of Spain. As Ryan Cordell has highlighted, these labs have provided opportunities for libraries to experiment with ML to enrich collections, "making them more useable for scholars, students, and the general public" (Cordell 2020: 1). Library Labs are thus utilizing AI to enhance the library as a digital research infrastructure (Börjeson et al. 2024).

This chapter focuses upon the research possibilities enabled by KBLab, the data lab at the National Library of Sweden (*Kungliga biblioteket*, hereafter KB). Established in 2019, KBLab initially aimed to:

1. offer access to the library's digital collections in structured form for large-scale research; and
2. experiment with AI tools to innovate the library's working practices.

The emergence of transformer-based language models (Devlin et al. 2019) soon expanded the lab's mission, enabling it to contribute to an improved national AI infrastructure for the Swedish language (Haffenden, Fano, et al. 2023: 44–45). By leveraging KB's high-quality, language-specific collections, KBLab has trained novel AI models such as KB-BERT. These models enable the analysis of Swedish text, sound, and image data, significantly enhancing the library's role as a research infrastructure. This ensures a productive symbiosis between the lab's functions: digital heritage enables the development of better Swedish AI tools, which in turn enhance the accessibility and searchability of KB's collections. KBLab thus constitutes an important entry point for researchers seeking to analyze and generate new insights about Swedish heritage data.

The sort of infrastructural work carried out at the lab creates a distinctive position for KBLab in the broader ecology of Swedish support for digital research encapsulated by Huminfra. Whereas the various nodes for digital humanities based at universities tend to work very closely with research projects, often including research engineers for ongoing hands-on support, our emphasis on producing new Open Access AI models for Swedish material necessitates a more arm's length approach towards collaborating with such projects. Insofar as we concentrate on providing access to the collections and using these as the basis for collections-based AI development (Börjeson et al. 2024: 572–576), KBLab most closely resembles the other GLAM lab within Huminfra, the AI lab at the National Archives of Sweden.[1] While the National Archives' lab has prioritized models for handwritten text recognition (HTR) to improve access to archival material, KBLab's focus on text, sound and images reflects the multimodal nature of KB's collections.

This chapter offers a practical introduction to accessing and using KBLab's various resources. We provide an orientation for digital researchers interested in making use of these in their projects, whether it be using the lab as a physical site to access the library's collections or deploying our AI models as research tools to analyze data. While these models for computational research are best utilized by scholars with some prior knowledge of programming and data analysis, our overview also aims to inspire those with more qualitative backgrounds to explore new research opportunities. The first part of the chapter explains what data is available at KBLab and how it can be accessed in the lab environment. The second part describes the various AI models we have released, offering specific use cases and examples of how these might be used to conduct digital research, as well as what existing projects at the lab have engaged with. The final part explains the application process for initiating a project based at the lab. Our aim in producing such a hands-on guide is to give sufficient context and knowledge to understand the possibilities of KBLab. By bridging technical insights with practical applications, the chapter serves as a comprehensive starting point for conducting large-scale digital research at KB and beyond.

2 *What data is available at KBLab? Identifying and preparing research data*

KB's collections range from books and manuscripts to television programmes and radio broadcasts. However, not all of the items in the collections are digitized, and not all that is digitized is equally easily accessible for researchers.

1 See https://huggingface.co/Riksarkivet.

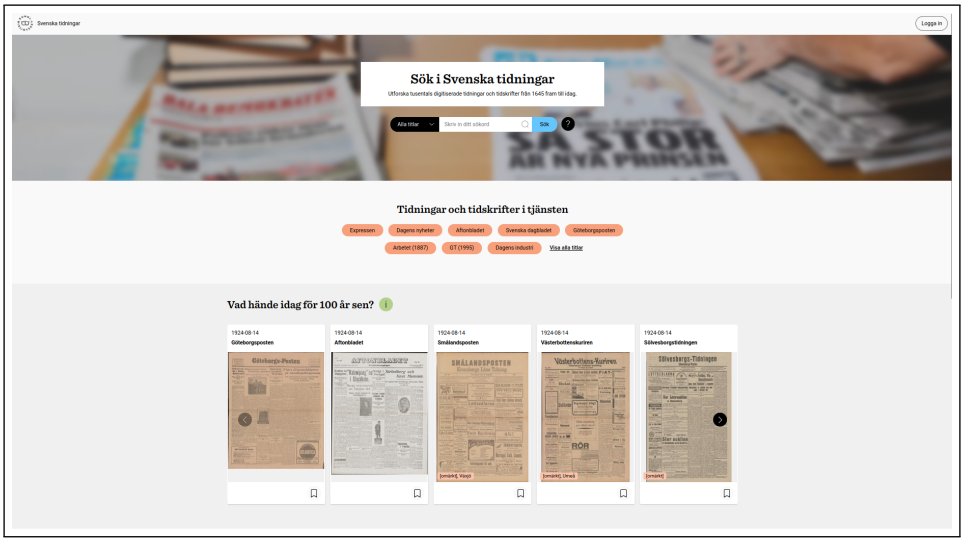


Figure 1: Searching for digitized newspapers with the library's search service

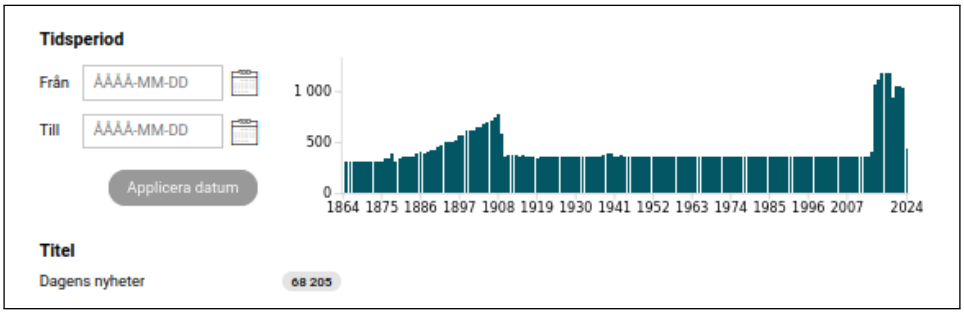


Figure 2: The availability of *Dagens Nyheter*, 1864-2024

This means that it is important to know how to search for the desired materials and to understand how ready the data is for research. Some items might be readily accessible and easy to use, while others may require additional steps or specialised tools for access and analysis. Recognizing these differences is essential, as it enables more efficient navigation through the resources and the selection of appropriate methods for working with the data.

2.1 *How to find the data?*

Let us take the newspaper archive as an example. This collection includes more than 4,376 titles from 1600 to the present, yet not all of the issues of

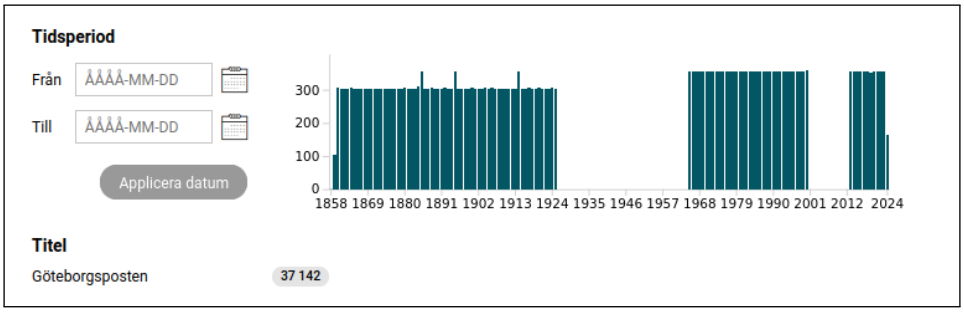


Figure 3: The less complete digital availability of *Göteborgsposten*, 1858-2024

these newspapers have been digitized. What is currently available in digital format is the entire collection of the newspapers up to 1906, along with everything from 2014 onwards, but the coverage for the intervening years is incomplete. The library's newspaper database is continually updated with newer material, though it takes approximately four months to digitize incoming material and make the most recent newspapers available.

A useful tool for checking the availability of the newspaper data is the library's online service: *Tidningar*.[2] This serves as a comprehensive database for all digitized newspapers in KB's collections (see Figure 1). Due to copyright restrictions, it is not possible to see the actual content of newspapers that are less than 100 years old outside of the library itself (though several Swedish university libraries have a designated computer to access this restricted material). However, even in the limited mode available to all users online, this is a helpful tool for checking the coverage of the digitized materials, as well as conducting preliminary searches to determine if the newspapers contain the desired content (i.e. to see if a particular word or term appears over time).

For example, a quick search shows that the newspaper *Dagens Nyheter* is fully available from 1864 to 2024 (see Figure 2); whereas *Göteborgsposten* can be accessed from 1859 to 2024, but the years 1925-1967 and 2001-2012 have yet to be digitized and are thus not digitally available (see Figure 3).

For exploring materials in KB's collection beyond the newspaper archive, the Swedish union catalogue, *Libris*, is an invaluable resource.[3] This contains freely available digital material as well as information about nearly 13 million items, including books, e-books and images from around 600 Swedish libraries (see Figure 4).

2 See https://tidningar.kb.se/.
3 See https://libris.kb.se/.

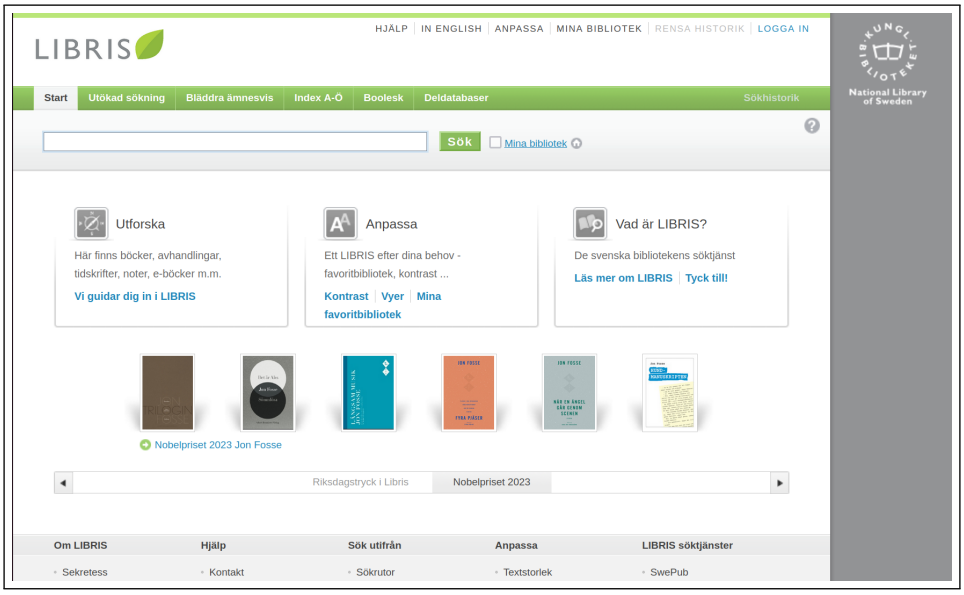


Figure 4: Libris as entrance point for search and discovery

2.2 *Current state of the data and challenges*

Knowing what data is available for research is a crucial first step. But it is equally important to consider the readiness of the data for research, which includes being aware of the format of the material, potential optical character recognition (OCR) errors associated with the digitization process and the extent to which the data has been cleaned, organized and prepared for analysis (Sikora & Haffenden 2024).

Some challenges concerning the quality of the newspaper data and other digitally available sources stem from the particular terms of Swedish legal deposit. Under current regulations, physical copies are prioritized, which means that digital versions of books or other materials do not need to be deposited if a printed version has already been submitted. This is unfortunate, as valuable metadata is lost during the digitization process. Moreover, the quality of digitized materials tends to be lower than their digitally-born equivalents. Direct access to digital versions of materials, when available, would help to mitigate these quality issues.

We can again take newspapers as an example to illustrate this point. Once the newspaper issues are submitted to KB as physical copies, they are forwarded to the digitization centre for scanning. Only after this step, and following the application of OCR models, do the materials become available as digital copies within KB. In practice, this means that much of the

information which is typically taken for granted when reading a physical newspaper is lost during digitization. This includes details such as the reading order of the text, the number of articles on a page, and more specific metadata such as distinguishing between text boxes in the body of an article, identifying which text boxes belong to a specific article, locating the title of an article or captions for images on a page. This loss of structure can complicate tasks like content analysis, archival research or simply navigating the digital format in a way that mirrors the original print experience (Sikora & Haffenden 2024: 62–63).

The final product of the digitization process is raw outputs from the OCR engine, in the form of scanned images and ALTO files. ALTO stands for Analyzed Layout and Text Object and is a metadata format used for encoding information about both layout and text of digitized documents, e.g. newspapers and books, by providing details such as the spatial coordinates of text elements. To enable easier navigation and processing of the materials, KBLab has developed a set of additional files derived from the content of the ALTO files. These are designed to organize and summarize key information from the OCR-outputs, and are useful for researchers who want to work with textual data on a larger scale, since they simplify the process of accessing and analyzing the content. However, it is important to note that these texts are *not* pre-processed and include OCR-errors that may require manual correction (see Figure 5). Consequently, users need at least a basic understanding of programming to be able to retrieve, process and analyze the data.

2.3 *What other data is available beyond newspapers?*

Moving beyond the newspaper data, KB's digital collections also include a number of digitized books, the Swedish Government Official Reports (SOUs) and parliamentary data, among other things. We provide more detail about how to access such data in Section 6 below.

Since material less than 100 years old is still protected by copyright and cannot be made publicly available, KBLab has undertaken various collaborations to make derivatives of this data accessible. One such initiative was a project with Språkbanken Text to create the Kubord and Kubord2 datasets.[4] These include word frequency analyses of the biggest Swedish newspapers, syntactic annotations of the texts at the word level and frequencies of collocations for individual words. The datasets can be downloaded from Språkbanken's webpage or explored graphically via the online tool *Korp*.[5]

4 See spraakbanken.gu.se/resurser/kubord and spraakbanken.gu.se/resurser/kubord2.
5 See https://spraakbanken.gu.se/korp/?mode=kubord.

Figure 5: OCR success and failure on a newspaper page. The OCR software correctly recognized the text in the yellow box: "DAGENS NYHETER". However, the text in the red box—"ÅNGMASKINER och"—was misread as: "tciåmasy\ver oc\v".

Another collaborative effort to enhance the availability of KB's collections involved creating vector models based on the same newspaper data as the Kubord2 datasets. These models can be used to track semantic changes and relationships between words (Forsberg et al. 2023, Bouma et al. 2024). As with the datasets mentioned above, the vector models are freely available online (Språkbanken Text 2024).[6]

6 See https://spraakbanken.gu.se/resurser/kubord-fasttext.

Figure 6: Accessing digitized manuscripts via Manuscripta

2.3.1 *Audio data*

Besides offering access to text data, KBLab has also created an audio dataset entitled RixVox (Rekathati 2023b,a). The underlying material for this dataset consists of Swedish parliamentary debates made available via the Swedish Parliament's open data initiative.[7] The debates were aligned with the corresponding transcripts from written protocols. Additionally, the debates are annotated with various metadata such as gender, age and name. The final product is a dataset consisting of 5,500 hours of Swedish audio material, which is freely available on the data science community platform Hugging Face.[8]

2.3.2 *Manuscripta*

A further source of data is the range of material published in KB's service *Manuscripta*.[9] This is a database of medieval and early modern digitized manuscripts from Swedish libraries, currently including around 800 manuscripts. The images are available in high quality according to IIIF standards and are freely available to download and make use of in research (see Figure 6).

7 See: https://www.riksdagen.se/sv/dokument-och-lagar/riksdagens-oppna-data/.
8 For this audio dataset, see: https://huggingface.co/datasets/KBLab/rixvox.
9 See: https://www.manuscripta.se/.

3 *How to access data via KBLab?*

Members of research projects that have been granted access to KB's digital collections are welcome to visit KBLab in Stockholm to prepare for working with the library's resources. (For more details about how to apply, see Section 6, below). During the onboarding process, we provide researchers with an overview of the infrastructure and logins to our systems. Since it is not possible to move copyrighted material beyond the lab's internal network, we have significant computational capacity in-house to ensure that researchers have sufficient compute resources to carry out their work. This includes a server environment and two NVIDIA DGX A100 servers for more computationally heavy analysis (Börjeson et al. 2024: 570).

The first stage of our onboarding involves signing a formal user agreement, which outlines the terms and conditions for using the Lab and our code of conduct.[10] We then set up an account on a local workstation with the Linux-based Ubuntu system. This provides researchers with the flexibility to create and manage their own software environments according to their own specific needs. On the workstations, our Application Program Interfaces (APIs) can be accessed to explore data, computation can be performed, and both data and results stored. We also have a local GitLab to allow researchers to manage and archive their code.

After the introduction, researchers can start using the library's resources and data, for example via our dedicated graphical user interfaces (GUIs): *Datalab* (datalab.kb.se) and *Betalab* (betalab.kb.se). Both interfaces offer similar functionalities, with one key difference: the data available in the beta version is, for the most part, not under copyright protection, which allows researchers to use it outside of the library and in advance of any visit to KBLab. This provides an opportunity to learn how to navigate the webpage, search for resources and learn how to download the data. Doing this beforehand can save time familiarizing yourself with the interface at the lab. We also offer an overview of the API's and GUIs as part of the onboarding process, along with providing detailed written instructions about how to use these.

3.1 *Datalab*

After logging in to one of the GUIs—*Datalab* or *Betalab*—searches of KB's digital collection can be conducted using the menu to the left or by filtering the packages in the search bar (see Figure 7). A package can constitute, for

10 For details of this code (in Swedish), see: https://www.kb.se/samverkan-och-utveckling/kb-labb/samarbeta-med-kb-labb.html.

Figure 7: Accessing data via the Lab's GUI

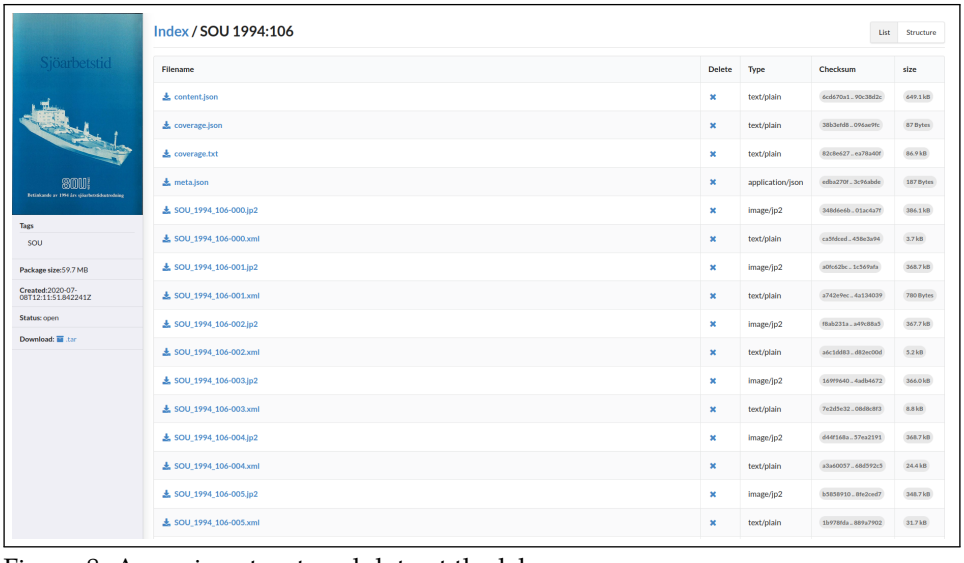


Figure 8: Accessing structured data at the lab

example, a complete newspaper issue, a full book or a SOU. Packages can be filtered based on various tags, such as title, creation date or resource type.

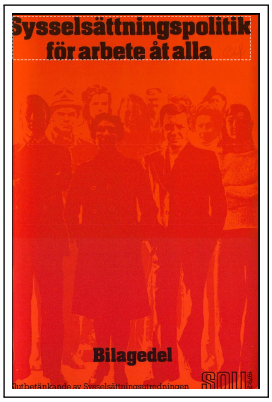
There are two ways to navigate the search: through the "package" or "page" format. Selecting the "page" option displays an individual page from newspapers or books that meet the search criteria. Results are shown as JPG files, and clicking on one of these opens a scanned image of the page. Alternatively, choosing the "package" option combines pages into a single package, such as an entire newspaper issue or a complete SOU document. In this mode only the first page of the package is shown during the search. When using the "package" format, two additional modes are available: "list" and "structure".

The "structure" tab shows all scanned images within a package, providing a visual overview of the entire contents. On the other hand, the "list" tab displays all files that are part of this package, including, but not limited to, the digitized pages in JPG format (see Figure 8). This view allows access to the following files:

`content.json` During digitization, the content of a page is divided into individual boxes containing text or images. The content.json file provides information about all the boxes present on each page within a package. The file format is JSON, where one entry in a dictionary corresponds to one OCR-box. In the example below, metadata about two boxes is shown (see Figure 9b). The "content" key contains information about the text recognised by the OCR engine, while the "box" key provides the coordinates for the box's position on the page. Additionally, there are two other keys: "type", which indicates whether the content was recognized as a picture or text, and "id", a unique identifier for each bounding box. The "id" is a Uniform Resource Identifier (URI), consisting of the following components:

- *https://betalab.kb.se*—the link to the API;
- *sou-1979-27*—the identifier for the entire package;
- *#1-1*—the first number corresponds to the part of the package, and the second number specifies the page; and
- *cblock_0-block_0*—the identifier for the specific box on the page.

Stable URIs are an essential means for researchers using the lab to be able to find their way back to the same point in the collections, and ultimately to demonstrate that their results are reproducible (by providing others with a means of verifying the data).



a.


```
[
  {
    "@id": "https://betalab.kb.se/sou-1979-27#1-1-cblock_0-block_0",
    "@type": "Text",
    "box": [
      "0",
      "11",
      "1745",
      "319"
    ],
    "content": "Syssels\u00e4ttningspolitik\nf\u00f6r arbete \u00e5t alla\n"
  },
  {
    "@id": "https://betalab.kb.se/sou-1979-27#1-1-cblock_1-block_1",
    "@type": "Text",
    "box": [
      "1720",
      "2619",
      "35",
      "100"
    ],
    "content": ":6Z61\n"
```


b.

Figure 9: The cover image of an SOU (a) and its corresponding "content" information in the lab environment (b)


```
title:      "SOU 1979:27"
year:       "1979"
created:    "1979"
```


Figure 10: Lab metadata in JSON format

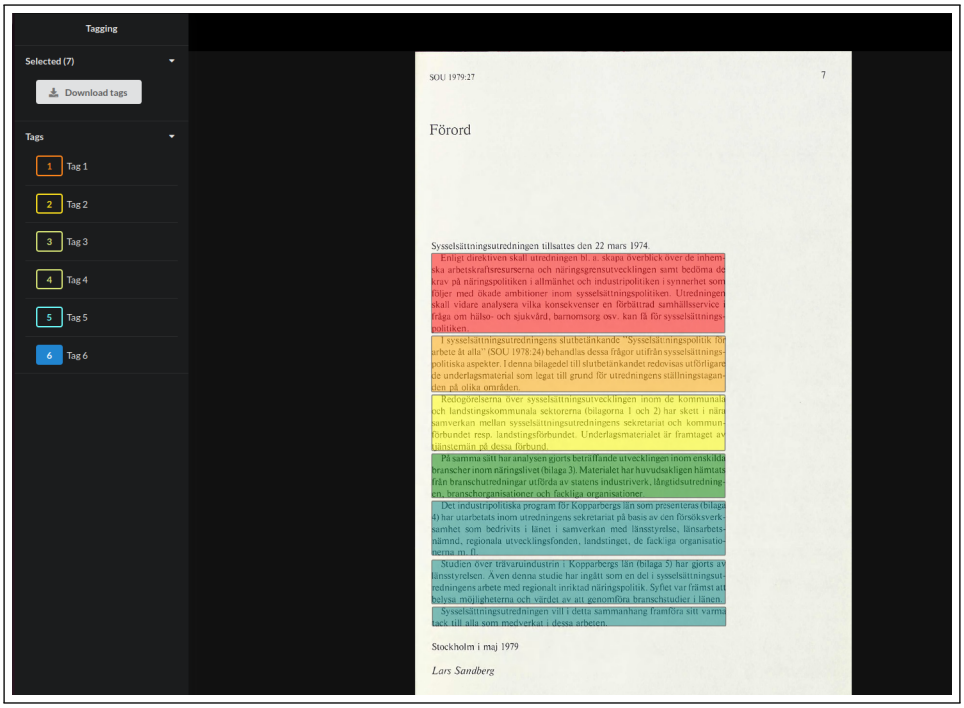


Figure 11: Annotating data via the lab's GUI.

`meta.json` This file consists of metadata about the package, such as the title, publication year and details about the year when the item was created (see Figure 10). The metadata also serves as tags that can be used to filter and search for specific pages and packages.

XML files These are the raw output files created during the OCR process and serve as the basis for creating content.json files.

`structure.json` . This file provides detailed information about the structure of a newspaper. It outlines which text blocks or elements are part of specific pages, offering a clear view of the layout and arrangement of the newspaper's content.

JP2 files These are digitized pages in JPEG 2000 format, representing the scanned pages of the material.

In addition to exploring the data graphically, the GUI also provides a simple annotation tool (see Figure 11). It can be opened by using the "Shift + T" shortcut while in the "structure" view of a package. The tool allows users to annotate six different classes and export the results into a CSV file. The annotation file includes the ID, type, tag, box and content fields. Due to copyright restrictions, annotations of the collections in *Datalab* can only be stored locally on lab computers.

3.2 *KBLab's Python library*

Datalab is an important part of KB's infrastructure, providing a graphical interface for exploring digital data. However, for researchers who want to conduct large-scale analysis, manually calling KBLab's API to download and process large volumes of data programmatically might be more suitable than examining individual objects. Another alternative is to use the python "kblab-client" package, which can be installed with pip.[11]

To familiarize yourself with the API calls, the package can be tested outside of the library on the beta version of the webpage, *Betalab*. Instructions and examples for using the APIs programmatically are available on our GitHub page, where you can access Jupyter notebooks demonstrating how to filter, download and process data at the lab.[12]

11 See: https://pypi.org/project/kblab-client/.
12 See: https://github.com/kb-labb/kblabb-examples/tree/master/api.

4 *Digital tools and AI models*

Curating digital collections and producing datasets for researchers on-site while enhancing access to non-copyrighted materials are central tasks at KBLab. Another key objective is utilizing these collections to develop digital research tools, including collections-based AI models trained on the library's data. Due to practical constraints, such as the limited number of workstations available, only a select number of projects can be hosted at KBLab at any given time. However, by producing and openly releasing AI models, we provide access to the benefits of these resources for a broader audience beyond the library itself. Using Swedish-language data to train high-quality AI models and making them freely available represents an important step toward democratizing AI for smaller languages (Haffenden, Fano, et al. 2023: 44–45). The following section provides an overview of these models and the research they enable.

4.1 *Language models for text analysis*

Our primary focus has been on training and fine-tuning encoder models, such as BERT (Devlin et al. 2019), which produce contextual representations rather than generating text. Unlike encoder-decoder architectures like GPT, which are designed for text generation, encoder models are better suited for tasks such as classification and retrieval. Their smaller size and fewer parameters make them more computationally efficient and cost-effective, which is particularly beneficial for digital researchers working with limited computational resources.

Although the original BERT model was released in 2018, it remains highly relevant today. Subsequent advancements, such as ModernBERT, have introduced improvements in speed, longer context windows and more efficient training methods (Warner et al. 2024). While generative models are optimal for open-ended text generation, our focus at the lab remains on developing smaller, fine-tunable encoder models that require fewer computational resources but still perform effectively in a range of applications.

Different models are better suited to specific tasks, depending on the research needs. The performance of KBLab-trained models, alongside other Swedish language models, can be evaluated through the Superlim leaderboard, a benchmark designed to compare Swedish models across multiple tasks (Berdicevskis et al. 2023). Since KBLab's establishment in 2019, we have trained multiple models in various modalities. The first of these was KB-BERT, a transformer-based language model trained on approximately 15–20 GB of Swedish text, including digitized newspaper archives dating

back to 1945, social media texts, Wikipedia articles and more (Malmsten et al. 2020). The variety of these textual resources helped capture different styles and uses of the language, which led to the creation of a diverse corpus for training a robust Swedish model. KB-BERT's compactness and versatility for applications within Natural Language Processing (NLP) means that this model is still widely used.[13] It can be utilized for a range of NLP tasks such as part-of-speech tagging (POS) and named entity recognition (NER).

Training a high-performing language model requires significant computational resources and large datasets. However, fine-tuning a pre-trained model is considerably less demanding and can be done on a consumer GPU or via an online cloud platform such as Google Colab. Due to copyright regulations, KB's data cannot be uploaded to these platforms, so we provide local GPU access at the lab for researchers who need to fine-tune models for tasks such as classification. For those lacking the resources or expertise to fine-tune models themselves, we also offer several pre-trained, off-the-shelf models ready for immediate use.

Among these is a BERT model for Swedish NER, designed to identify entities such as locations, organizations, and personal names. For example, given the sentence:

> *Göran Persson och Carl Bildt träffades på ett café vid Stureplan den 5e juli 1995.*
> 'Göran Persson and Carl Bildt met at a café by Stureplan on 5 July 1995.'

The NER model can identify the following entities:

- Person: Göra Persson, Carl Bildt
- Location: Stureplan
- Date: 5 July 1995

This capability is particularly useful for anonymization, where personally identifiable information (PII) must be removed or altered to comply with data privacy regulations such as GDPR. Identifying sensitive entities in datasets enables researchers to work with anonymized data while maintaining compliance with GDPR.

4.1.1 *Sentiment Analysis*

Another fine-tuned model available at KBLab is a sentiment analysis model for Swedish, capable of classifying text as positive, negative or neutral.[14] This

13 To access this model, see: https://huggingface.co/KBLab/bert-base-swedish-cased. For user statistics that offer a sense of how widely the model is used, see: https://github.com/kb-labb/huggingface_stats.

14 To access this model, see: huggingface.co/KBLab/robust-swedish-sentiment-multiclass.

model was fine-tuned using a diverse mix of sources, including Trustpilot reviews, news headlines and other texts (Hägglöf 2023). Unlike rule-based sentiment systems, which rely on predefined word lists, transformer-based models consider the semantic context in which words appear, mitigating errors that arise when sentiment is determined solely by individual word associations. Using a BERT model therefore enables a more linguistically sophisticated approach to identifying and analyzing the sentiment of a given text.

Beyond NER and sentiment analysis, fine-tuned models can solve a wide range of classification tasks. For social scientists, classifiers can be valuable for opinion mining: extracting and categorizing public sentiment from social media, surveys and other sources. More generally, digital humanists can use classification models to organize large text corpora into categories to enable further analysis. In literary studies, author attribution models can help determine the authorship of anonymous or disputed texts through stylistic analysis. These diverse applications illustrate the broad research potential and the versatility of fine-tuned NLP models.

4.1.2 *Sentence-BERT and topic modeling*

The original BERT model generates contextual embeddings for individual tokens within a sentence, making it ideal for token-level tasks such as NER, part-of-speech tagging and question answering. However, for some research applications, encoding larger chunks of text at once is more useful. This can be achieved using Sentence-BERT (SBERT), a modification of BERT that generates sentence-level embeddings well-suited for semantic similarity, text clustering and retrieval tasks.

We have trained two Swedish SBERT models at KBLab, both freely available on Hugging Face (Rekathati 2021b). The first KB-SBERT model processes texts up to 256 tokens, while the updated version extends the context window to 384 tokens to enable analysis of longer text chunks at the sentence level.[15] Sentence-BERT excels at detecting paraphrases and retrieving relevant documents in response to queries. While the standard BERT model is optimal for fine-grained word-level analysis, SBERT provides better overall sentence-level understanding.

4.1.3 *Transformer-based topic modeling with BERTopic*

One notable application of Sentence-BERT is BERTopic, a transformer-based approach to topic modeling (Grootendorst 2022). BERTopic clusters text by first generating sentence embeddings, reducing their dimensionality, and

15 To access these models, see: huggingface.co/KBLab/sentence-bert-swedish-cased.

then applying clustering algorithms to group similar topics. This process enables researchers to uncover coherent and meaningful topics in large text corpora. Unlike classical topic modeling methods such as Latent Dirichlet Allocation (LDA), which require extensive text preprocessing—including stop-word removal, stemming and lemmatization—BERTopic can operate effectively with minimal pre-processing, working directly on raw text (Fano & Haffenden 2022). This is possible because SBERT retains the contextual meaning of words within the text rather than treating them as isolated tokens. BERTopic thus provides a more contextually-aware analysis of topics.

BERTopic can be applied in various domains of digital research. It can be used for analyzing historical documents, for instance in helping identify thematic trends and semantic change in newspapers, speeches and letters. It can also be used in literary research as a means of detecting recurring themes across different works. Within the social sciences, BERTopic can be deployed to analyze social media posts and news articles to allow researchers track shifts in public discourse and sentiment over time.

To facilitate research that uses BERTopic, we provide various training resources, including workshops and Google Colab notebooks with detailed instructions and adjustable scripts.[16] These require no specialist software or hardware to use, and can be tested by users with little prior programming experience. The notebooks include detailed instructions and clarifications alongside the code blocks, making them accessible even to those new to coding. The script is also highly adjustable: allowing researchers to swap out datasets, adjust topic parameters, and customize model outputs, making it a flexible and user-friendly introduction to topic modeling.

4.1.4 *Use case: topic modeling at the library*

We have explored applying BERTopic at the library as an alternative method for classifying textual collections (Malmsten et al. 2025). Currently, subject classification is carried out manually by specialist librarians using the Swedish subject headings system, *Svenska ämnesord* (SAO). This controlled vocabulary is designed to provide consistency and clarity in cataloging. However, manual indexing poses various difficulties, as indexers may interpret classification guidelines differently, and the system itself has grown organically over time rather than being optimized for emerging research needs. These classification challenges for librarians can affect how easy it is for users to navigate through the collection.

16 To access this online workshop, see: https://colab.research.google.com/drive/10kB3wfoHSfZE48vEKmznIw-ff36uR8gs?usp=sharing.

To address these challenges, we have experimented with automated methods for subject analysis. Instead of assigning predefined categories, we use transformer-based topic modeling, allowing the classification and ordering to emerge from the contents of the works themselves. Since BERTopic clusters documents based on their textual properties, it can be used to generate data-driven topics rather than relying on pre-established subject headings. Adopting this method could reduce inconsistencies in cataloging and improve discoverability. We discuss this project in greater depth in an Open Access IFLA volume on the use of AI in libraries (Malmsten et al. 2025).

4.2 *Speech recognition models for audio data*

At KBLab, we develop AI models and digital tools not only from textual collections but also by leveraging KB's extensive audio-visual archives. The collections include thousands of hours of legally deposited resources, such as TV and radio broadcasts, which provide an ideal material for training Automatic Speech Recognition (ASR) models.

State-of-the-art ASR models such as Wav2Vec 2.0 (Meta) and Whisper (OpenAI) are particularly valuable for transcribing spoken language into text (Baevski et al. 2020, Radford et al. 2023). These models play a crucial role in making audio-visual collections more accessible, enabling automated transcription for research, cataloging, and search applications. In addition to providing accurate transcriptions, ASR models can be used to develop audio-driven search tools, allowing users to efficiently retrieve specific content from vast audio and video datasets, such as podcasts, radio programs or historical broadcasts.

For speech models to perform reliably across different accents and linguistic variations, they must be trained on a diverse range of audio data. Similar to large language models (LLMs), ASR models trained primarily on English-language data often underperform when applied to smaller languages such as Swedish. Even multilingual models, which are trained on several languages, are typically exposed to limited amounts of Swedish data, making them less effective at capturing the nuances of regional dialects and spoken variations. Recognizing this limitation, we have focused on training speech recognition models specifically for Swedish using local radio broadcasts from KB's archives, ensuring stronger performance across a range of dialects. By prioritizing regional diversity, we aim to improve linguistic inclusivity and enhance the overall democratic accessibility of our ASR models (Haffenden, Fano, et al. 2023: 44-45).

Our Wav2vec2 model called VoxRex is trained on more than 10,000 hours of P4 radio material from the last 20 years (Malmsten et al. 2022). P4 is

Table 1: Comparative WER on Swedish dialects for Meta's XLSR, the monolingual VoxPopuli model, and KBLab's VoxRex models (Malmsten et al. 2022: 2). VoxRex-A, -B, and -C refer to versions trained on different amounts of speech data and varying numbers of training updates. WER refers to the percentage of errors registered by each model, as explained on page 36.

	Model				
Region	XLSR	VoxPopuli	VoxRex-A	VoxRex-B	VoxRex-C
Dalarna	5.62	3.17	2.74	2.74	1.67
Göteborg w. env.	5.64	3.26	2.86	2.82	1.78
Mellansverige	5.62	3.30	2.85	2.83	1.79
Norrland	6.27	3.68	3.12	3.20	1.95
Stockholm w. env.	4.57	2.64	2.26	2.30	1.43
Västergötland	5.53	3.26	2.78	2.81	1.76
Västra sydsverige	7.62	4.25	3.77	3.84	2.10
Västsverige	5.40	3.16	2.65	2.71	1.60
Östergötland	5.65	3.23	2.81	2.77	1.68
Östra sydsverige	6.68	3.80	3.29	3.23	1.94

a Swedish public local radio channel, which consists of around 25 radio stations and therefore provides diverse representation in terms of dialects. The broadcasts contain a large variation of different types of content ranging from music, news and sport reports to people calling in and on-site reporting.

One of the common metrics used to evaluate the performance of ASR models is Word Error Rate (WER), which measures the accuracy of transcriptions by calculating the percentage of words that were incorrectly predicted by the ASR system compared to a reference transcription. As we can see in the table below, the models perform differently on various regional variations of Swedish with Southern and Northern dialects posing more difficulties than, e.g., the Stockholm accent (see Table 1). This proves the need to train the models on more balanced and varied datasets in order to mitigate this effect. In comparison to Meta's speech model XLSR, our Wav2vec 2.0 large VoxRex Swedish model (C) provides an improvement of up to 5.52% on the South-Western dialect ("Västra sydsverige").

KBLab is conducting a project to train updated versions of Wav2Vec 2.0 with larger amounts of data, as well as datasets enriched with additional dialects. Besides releasing Wav2Vec 2.0, the project has also fine-tuned a Swedish Whisper model, KB Whisper, which offers state-of-the-art performance for Swedish speech to text. The work was conducted on the EURO HPC supercomputer Leonardo in Italy.

4.2.1 *Choosing a model for ASR*

Which model should you choose for your research? The choice between Wav2Vec 2.0 and Whisper mostly depends on the specific requirements of the ASR task, given differences in how they handle formatting, accuracy and potential errors like hallucinations.

Wav2Vec 2.0 is designed to output transcriptions that closely mirror the spoken materials, without introducing words or phrases that were not explicitly stated. This makes it ideal for tasks requiring word-for-word accuracy, such as legal transcriptions, technical documentation or academic research. However, Wav2Vec 2.0 typically produces transcripts in lowercase and without punctuation, which may require additional processing to improve readability and format the text appropriately. Additionally, Wav2Vec 2.0 is generally less prone to hallucinations—errors where the model "imagines" content that wasn't actually spoken—making it more reliable for tasks that require verbatim transcription that are faithful to the original speech.

On the other hand, Whisper is renowned for its ability to paraphrase, making it well-suited for tasks such as real-time captioning. Whisper automatically adds punctuation and capitalization, which significantly improves the readability of the transcripts. However, Whisper might sometimes introduce paraphrasing or include words that weren't explicitly spoken, which could be problematic in scenarios where absolute precision is crucial. For researchers working with audio data who might be interested in using either of these models for transcription, we have produced a freely-available tutorial notebook on Google Colab with code and instructions.[17]

4.2.2 *Model-based applications at the library*

Beyond providing a means to democratize AI, collection-based Swedish models can also be used to make the library's collections more accessible to researchers and anyone interested in exploring KB's materials. One initiative aimed at demonstrating this potential involved developing a demo platform called *Mediesök*, which showcases how speech models can make audio materials searchable in new ways (see Figure 12).

Audio materials legally deposited at KB often lack detailed metadata, which is important for understanding the content of these materials. Typically, the only information available is the time and name of the radio broadcast. To allow users to search through the materials without needing to listen to each audio clip individually, an alternative approach involves

17 See: https://colab.research.google.com/drive/1RCP53jqClJz0zDX_VBT_K04BevUKk842?usp=sharing#scrollTo=ib2Owc2w3gHa.

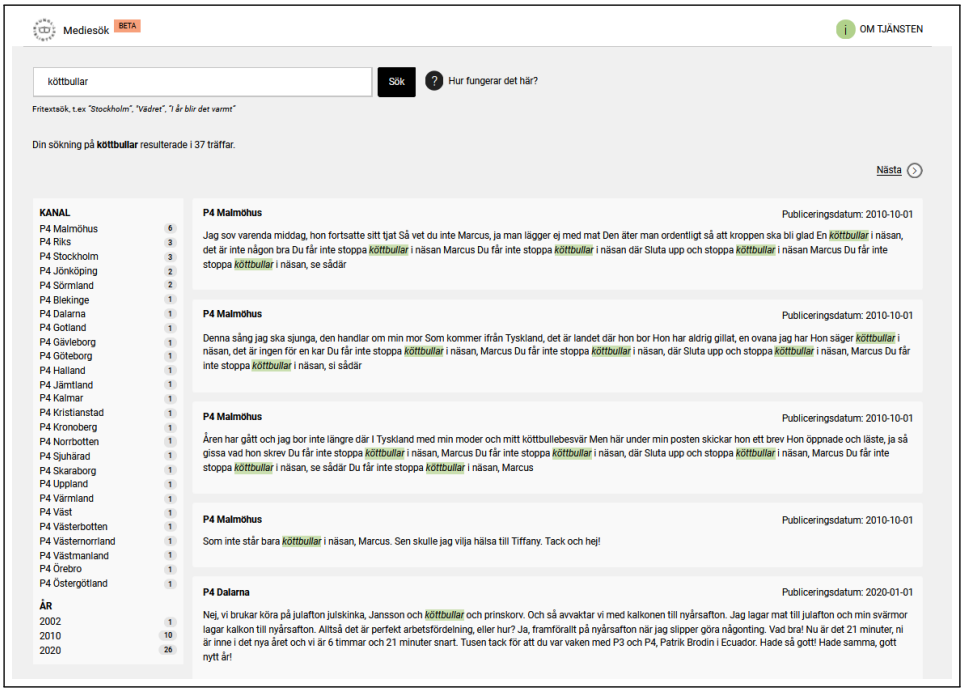


Figure 12: Locating *köttbullar* 'meatballs' in the radio archives. Making audio-visual data amenable to contents-based search via the *Mediesök* demo.

generating textual materials from the audio data via transcription.

To demonstrate this idea, we transcribed 10 hours of P4 radio broadcasts using WhisperX, which is a version of the Whisper model that includes word-level time-stamping, making it possible to easily navigate through the transcriptions and corresponding audio clips. This feature enhances the usability of audio archives, allowing users to effectively search for specific content within the audio materials. *Mediesök* enables researchers to carry out free text searches upon audio material that was previously challenging, if not impossible, to search through, short of listening through each file in real time.

4.3 *Visual models for image analysis*

Besides audio and textual materials, we also work with image collections at KBLab. Similar to our efforts with the audio materials, AI tools can significantly enhance the searchability of visual heritage. One way that we have shown how AI can be integrated into the library's working practices is

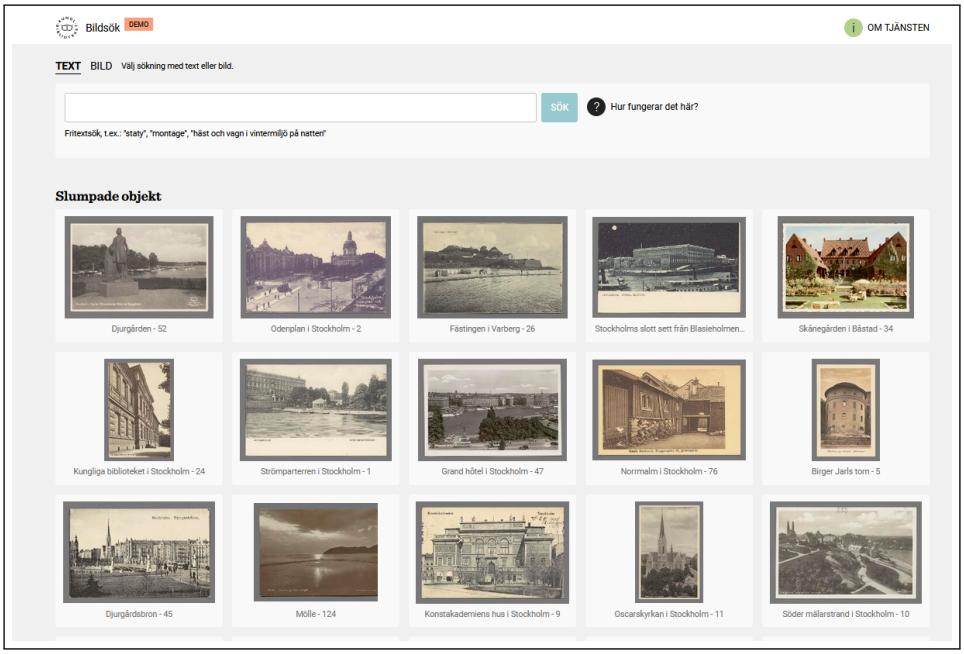


Figure 13: Image search with the help of AI

through the creation of a demo tool called *Bildsök*—i.e. image search.[18] This allows users to explore image collections in new ways: either by uploading an image and searching for similar ones, or by entering free text to find images that match the desired search terms (Haffenden, Rekathati, et al. 2023). It is particularly useful for material such as ephemera that often lacks metadata at the level of the individual object. Deploying an AI model to analyze the visual content of such material sidesteps the absence of metadata to make it searchable according to users' interests.

The model enabling this demo is Swe-CLIP 2M, a model that links images to their textual descriptions. Originally, this model was trained on English data by OpenAI, but its capabilities have since been adapted to other languages, where we helped to produce a Swedish version (Carlsson et al. 2022). The model progressively learns to associate text with images more effectively, enabling it to match the best images to specific text queries. The *Bildsök* demo currently features a collection of more than 17,000 postcards to showcase this functionality (see Figure 13). As an example of how this works, we can observe that searching for a church does indeed return a number of images depicting a sacral building (see Figure 14).

18 See: https://lab.kb.se/bildsok/.

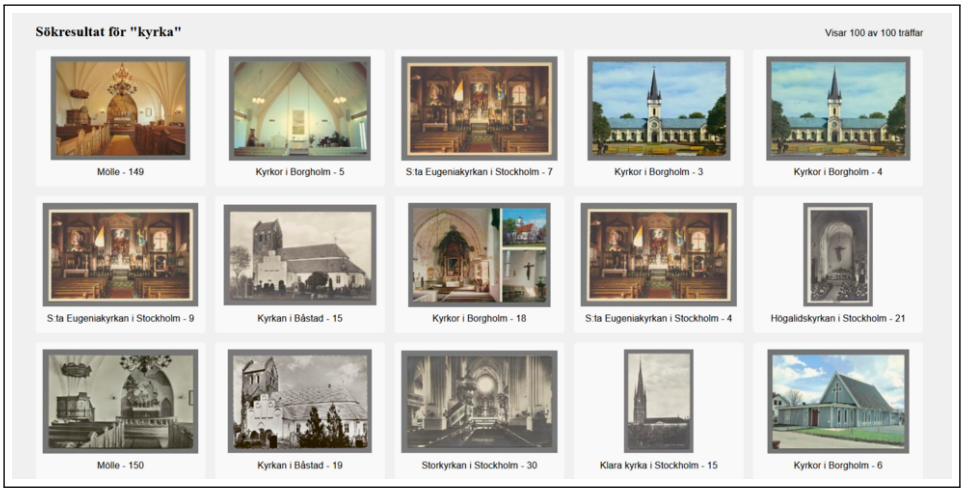


Figure 14: Unlocking image collections with AI-based search

The flexibility of the image search engine lies in its underlying method. When a text query or an image is entered into the search box, it is first converted into a numerical representation or vector. This vector is then compared with the vectors of the items stored in the database. The comparison is made based on cosine similarity—a measure of how closely two vectors align in multidimensional space. The search results are ranked according to their similarity scores, with the most similar items appearing at the top of the list. We have described this demo in more detail in a post on our blog (Haffenden, Rekathati, et al. 2023).

4.4 *Development work and future models*

We are currently engaged in a range of development projects aimed at improving the digital infrastructure for Swedish language processing, leveraging access to the EU's EuroHPC supercomputing infrastructure.[19] A key initiative within this framework is the ongoing effort to train a Swedish Whisper model mentioned above, which is designed to support more diverse and representative speech-to-text applications. By fine-tuning Whisper on a broader range of high-quality training data, we aim to enhance its ability to transcribe spoken Swedish with greater precision and cultural nuance. The new series of ASR models resulting from this work was made freely available via our Hugging Face page in spring 2025, providing researchers and practitioners

19 See: https://www.eurocc-access.eu/why-does-the-royal-library-of-sweden-need-hpc-video/.

with improved tools for audio data processing (Vesterbacka et al. 2025).[20]

Another major focus of our development work is the creation of an updated and improved generation of BERT models for enhanced text analysis. A key constraint of encoder models is their context window, which determines how much textual information a model can process at once before encoding it into numerical representations. At present, KB-BERT has a maximum token length of 512, limiting the amount of text it can process per input. Our goal is to train new models with extended context windows, allowing them to process significantly larger text segments in a single pass. This will be particularly valuable for tasks such as classification, information retrieval and NER, as it enables models to capture and analyze broader linguistic patterns without requiring extensive text chunking or segmentation.

5 *Use cases: What sort of research is possible using KBLab's AI models?*

AI enables large-scale research on the library's collections across diverse academic areas, from social sciences, linguistics and history to literary studies, ecology and beyond. The integration of AI-based technologies into library collection management is transforming how researchers interact with and utilize these resources. By combining humanities perspectives with data science methods, researchers are unlocking new analytical possibilities. One striking example is the "Living with Machines" project, a large-scale collaboration between universities, libraries and research institutes in the UK.[21] This multidisciplinary initiative brought together historians, geographers, data scientists and computational linguists to analyze nineteenth-century newspapers, studying the human impact of the Industrial Revolution. By using newly digitized materials alongside the British Newspaper Archive, the project led to the development of novel tools and assessments of OCR quality, among other outcomes.[22]

A significant aspect of this project was the creation of collections-based historical language models, which were later applied in various studies, such as research on "atypical animacy" (Coll Ardanuy et al. 2020). By utilizing BERT's masking function as a register of linguistic change, researchers examined how the semantic associations of words evolved over time. This approach allowed them to trace shifts in meaning and animacy in an innovative manner (see Figure 15).

20 See: huggingface.co/collections/KBLab/kb-whisper-67af9eafb24da903b63cc4aa.
21 See: https://livingwithmachines.ac.uk/.
22 For further results, see: https://livingwithmachines.ac.uk/achievements/.

Original sentence: And why should one say that the machine does not live?
Masked sentence: And why should one say that the [MASK] does not live?
Predictions with scores: *man* (5.0788), *person* (4.4484), *other* (4.1866), *child* (4.1600), *king* (4.1510), *patient* (4.1249), *one* (4.1141), *stranger* (4.1067), ...

Figure 15: Using a historical language model to register the surprise of semantic innovation (Coll Ardanuy et al. 2020: 4536)

A further instance of the creative deployment of AI within machine learning-based humanities was the project's use of Word2Vec vector models trained on historic newspaper data to track the semantic change of terms related to mechanization, such as "gear", "fellow" and "wheel" (Pedrazzini and McGillivray 2022).

Similar methodologies have been employed at KBLab, where collections-based models have been used in multiple projects. These examples showcase the potential for using AI as the basis for innovative research on GLAM collections, which would otherwise be impossible to conduct at scale. These AI models can facilitate the identification of patterns, extraction of insights and visualization of results, thereby broadening the scope for interdisciplinary collaboration. In the following section we outline instances of such work from projects carried out at the lab.

5.1 *Televising information*

One of the first projects at KBLab, "Televising Information: Audiovisual Communication of Swedish Government Agencies," led by Emil Stjernholm (Lund University), focused on KB's audio-visual collections. The study examined *Anslagstavlan*, a bulletin program produced by Sveriges Television (SVT) since 1972, which government agencies have used to communicate with the public (Stjernholm 2023). By collaborating with the lab, the project applied ASR models to transcribe spoken content from the video recordings, making large-scale analysis of this archival material feasible.

Audio-visual materials in KB's collections are stored as continuous recordings of radio and television broadcasts, which posed a particular challenge. The first step of the project involved manually locating and extracting the relevant *Anslagstavlan* segments from the recordings. Once isolated, these clips were processed using KBLab's Wav2Vec 2.0 large VoxRex model (see Section 4.2), which transcribed the spoken content into text. Manual corrections were then made to enhance accuracy. This approach enabled the researcher to study how Swedish government agencies have communicated with the

public over time in a way that would have been prohibitively labor-intensive using traditional methods.

The result was a 160,000-token corpus derived from 600 episodes of *Anslagstavlan*. Using AntConc, a text analysis tool, the researcher conducted concordance and keyword-in-context analyses, identifying changes in communication strategies and thematic trends over time. The study could thereby offer a new perspective on Sweden's public information practices and their evolution.

This project aligns methodologically with the previously mentioned *Mediesök* demo, as both employ machine learning to transform complex datasets into accessible research materials. Without automated techniques like ASR and text analysis, studying extensive video content like *Anslagstavlan* would have required exhaustive manual effort. By leveraging AI, the project could embark upon large-scale analysis of historically significant audio-visual data.

5.2 *Mining for meaning*

Since 2019, KBLab has hosted "Mining for Meaning: The Dynamics of Public Discourse on Migration," a six-year initiative led by researchers from Linköping University. This project analyzes large text corpora—including parliamentary speeches, web data and newspapers—to examine shifts in the discourse on migration and integration.[23]

Through computational analysis, researchers track the development of sentiment on these topics over time within three domains: the public sphere, the media and politics. By treating newspapers as historical records of collective discourse, as "social sensors", the project applies machine learning techniques to analyze meaning-making processes in society.

One part of the project involved investigating how sentiment towards migration has shifted over time in the national news, using this as a proxy for tracking societal changes in attitudes (Hurtado Bodell et al. 2024). They use a corpus derived from the four largest Swedish newspapers—*Aftonbladet*, *Dagens Nyheter*, *Svenska Dagbladet* and *Expressen*—from the period 1945–2019, which they filtered to create a sub-corpus with "immigration-rich" articles, where each text contains a sufficient percentage of tokens connected to the topic of immigration. Researchers developed a semi-supervised extension to classical topic modeling, called seeded topic modeling, which directs the model's focus using predefined "seed" words, such as "refugees" and "asylum seekers".

By applying machine learning to large-scale newspaper archives, the project enables sociological analysis of historical discourse, providing in-

23 See: https://liu.se/en/research/computational-text-analysis.

sights into how Swedish society has framed migration issues over the decades. It demonstrates how innovative combining sociology and data science approaches to large digital collections can offer new ways of approaching how society makes sense of social changes and trends over time.

5.3 *Welfare state analytics*

Another example of a project conducting large-scale text analysis of digitized collections is "Welfare State Analytics: Text Mining and Modeling Swedish Politics, Media & Culture, 1945–1989" (WeStAc), a collaboration involving researchers from Umeå University, Uppsala University and Aalto University in Finland.[24] The project focused on digitizing and analyzing Swedish texts from 1945 to 1989 to create datasets from Swedish governmental reports, newspapers and literary journals, and thus explore societal and discursive shifts in the domains of politics, media and culture.

Techniques like word embedding and probabilistic methods have been employed to track the evolution of concepts such as emancipation and individualization between 1945 and 1989. One case study analyzed how the notion of "political" evolved over time in post-war Sweden (Norén et al. 2023). This was done by creating and analyzing bigrams and LDA topic models based on a corpus of 27 million tokens derived from articles from the newspapers *Dagens Nyheter* and *Aftonbladet*. The analysis could demonstrate the expansion of the term "political" into broader social and cultural discussions, while noting how core topics like international conflicts, party politics, elections and economic policy remained dominant. This suggests how large-scale approaches that examine the library's collections as data can fruitfully explore change and continuity over time, given a project team with the right skill set (see Section 6 below).

5.4 *Multimodal topic modeling*

Topic modeling has proven to be a valuable tool in digital humanities research, and it could be fruitfully deployed in the GLAM sector too. We have previously described how Sentence-BERT models can be used in BERTopic pipelines to organize and analyze large volumes of text data. Beyond textual analysis, BERTopic also supports multimodal topic modeling, which incorporates both visual and textual signals to improve classification and retrieval processes. This method is based on the CLIP model, used to create the *Bildsök* demo discussed earlier, and further demonstrates how machine learning can enhance search capabilities in visual collections.

24 See: https://www.westac.se/en/.

	Topic	Count	Representation	Visual_Aspect
0	-1	6834	[person, horse, carriage, drawn, church, riding, going, front, street, the]	
1	0	650	[walking, down, street, couple, city, people, building, large, tower, it]	
2	1	570	[church, tower, clock, top, with, front, in, on, of,]	

Figure 16: Topic modeling and thematic analysis of image collections

By integrating multimodal AI models, cultural heritage institutions can significantly enhance the searchability and classification of their collections. Multimodal topic modeling enables automated tagging and categorization of archival materials by considering both images and associated text descriptions. For example, a model like CLIP, which simultaneously processes visual and textual inputs, can detect similarities between a painting's image and its descriptions, fostering richer semantic connections within a database. This approach allows for more nuanced and intuitive explorations of cultural collections (see Figure 16 for AI-based thematic clustering examples).

Combining machine learning techniques further expands the potential applications of topic modeling. For instance, ASR models can transcribe audio and audio-visual materials into text, enabling topic modeling on previously inaccessible material. This facilitates large-scale analysis of radio and TV archives, which contain valuable but underutilized historical insights. By converting spoken content into structured, searchable text, institutions can significantly enhance the accessibility and usability of extensive audio-visual archives.

5.5 *OCR and post-processing improvements*

A persistent challenge in working with digitized data is the quality of OCR outputs. As mentioned earlier, KBLab's data undergoes minimal processing beyond correcting the most common OCR errors. Without systematic post-correction or accounting for potential misspellings, OCR inaccuracies can distort research findings. In the previously described projects, two studies relied on digitized newspaper data to create datasets. To curate these datasets, text blocks containing the term "political" or tokens related to immigration were extracted and used as a basis for corpus construction. However, when working with sources not originally produced in digital form, additional processing steps are required to handle misspelled or fragmented words, making them more difficult to identify and analyze.

This is a common problem with OCR, and researchers have recently developed various models to mitigate these issues. One potential solution involves training a new OCR model specifically for Swedish, which could yield more accurate results. However, while this approach is promising, it is also resource-intensive and costly when applied to large datasets. A more feasible alternative is to improve the quality of existing OCR outputs through post-processing techniques. Methods such as spell-checking, context-aware corrections and machine learning models trained to recognize common OCR errors can significantly enhance the accuracy of datasets, leading to more reliable research outcomes.

One study exploring these possibilities involved fine-tuning a transformer model to perform post-OCR correction on Swedish newspapers from the nineteenth and twentieth centuries (Löfgren & Dannélls 2024). The researchers selected the ByT5 model, which operates at the character level and was pre-trained on the multilingual mc4 dataset. Fine-tuning was conducted using a manually annotated subset of digitized newspapers from KB's collections. The best-performing model achieved a 36% reduction in character error rate, demonstrating the potential for automated OCR correction to improve textual accuracy. This model is intended to be integrated into Språkbanken Text's automatic processing pipeline and is freely available on Hugging Face.[25]

5.6 *Layout analysis*

We have tested layout analysis in a project that experimented with using a Document Image Transformer (DIT) model for filtering the editorial materials from the rest of the contents (Sikora & Haffenden 2024: 64–65). Likewise,

25 To access the model, see: https://huggingface.co/viklofg/swedish-ocr-correction.

similar efforts have been undertaken in another project for dividing the advertisements from the rest of the text blocks by training multiple multimodal models on various mixtures of text, images and metadata (Rekathati 2021a).

Beyond improving OCR quality, the data readiness of the digitized newspapers could also be enhanced by reconstructing the structure of the original publications. Studies using KB's newspaper collections typically operate at the text block level due to the absence of metadata linking blocks to specific articles. This lack of structural information means that researchers cannot easily determine which text blocks belong to which articles, how many articles are contained within a newspaper issue or the correct reading order of the texts. However, recent advances in machine learning provide new opportunities to address these challenges.

One approach has been to apply machine learning techniques to reconstruct the layout of digitized newspapers. In a recent project at the lab, a Document Image Transformer (DIT) model was tested for filtering editorial content from other textual elements (Sikora & Haffenden 2024). Similarly, another project sought to separate advertisements from editorial texts by training multiple multimodal models on a combination of textual, visual, and metadata inputs (Rekathati 2021a). These initiatives demonstrate how computational methods can be used to organize and segment complex historical documents more effectively.

A key prerequisite for these methods is the annotation of training data, which allows machine learning models to learn how to differentiate between various types of text blocks. A useful strategy involves leveraging existing datasets to improve corpus quality, extending to the development of classifiers for filtering specific content within large corpora. For example, researchers can annotate a sub-corpus of gold-standard text blocks related to a specific topic, which can then be used to fine-tune a machine learning model capable of automatically identifying relevant text blocks in broader volumes of data.

By employing supervised learning techniques, researchers can develop classifiers that accurately identify relevant content, even within vast and noisy corpora. This approach not only improves the accuracy of topic-specific datasets but also significantly reduces the time and effort required for manual data curation. Ultimately, such processes can improve the overall quality and usability of digitized resources, making them more accessible and valuable for research.

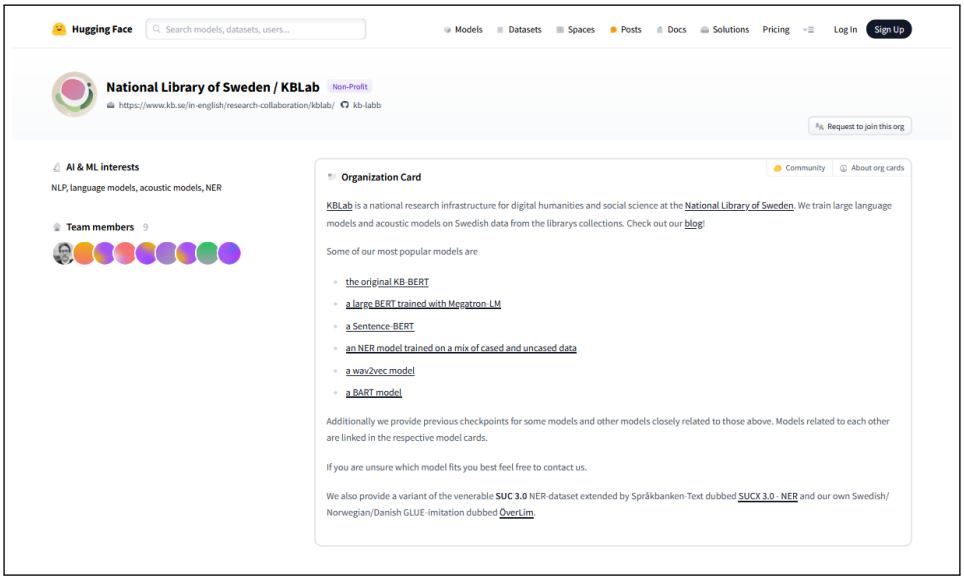


Figure 17: KBLab's models available Open Access via Hugging Face

6 *Initiating research via KBLab: How and where to get started?*

Having outlined the available data, models and research possibilities at KBLab, we now turn to practical steps for initiating a research project. This section provides guidance on different entry points, depending on your research interests and the level of engagement you seek with the lab's resources.

6.1 *Accessing AI models*

Our AI models are freely available via our Hugging Face page (see Figure 17).[26] If you plan to use these models in your research and have questions about their training data or suitability for your specific use case, we encourage you to reach out via email to kblabb@kb.se. Our data scientists can provide orientation and guidance, although we may not always have definitive answers to every inquiry. Documentation and pre-print articles related to the models are linked on our Hugging Face page where available. We also welcome feedback from researchers who have used our models—sharing your experiences helps improve and refine our resources.

26 See: https://huggingface.co/KBLab.

6.2 *Betalab for open data access*

For researchers working with copyright-free material, such as nineteenth-century Swedish newspapers, Swedish Government Official Reports (SOU's) and parliamentary records from the twentieth century, remote access is available via *Betalab* (see Section 3.1). This interface allows you to search, annotate and download data in a structured format. Additionally, *Betalab* serves as a preparatory environment for researchers planning an on-site visit to KBLab, allowing for script development and preliminary analysis before engaging with restricted materials at the lab. Contact us via `kblabb@kb.se` for further dialogue about access.

6.3 *Conduction research on-site*

For researchers requiring access to copyrighted materials, such as twentieth-century Swedish newspapers, KBLab provides a secure on-site environment in Stockholm. Due to copyright restrictions, this cannot be accessed remotely or exported beyond the lab's internal network. Researchers planning to work with this data should consider the following:

Firstly, since copyright restrictions mean access is only available on-site at KBLab's locale in Stockholm, all research needs to be conducted at the lab. It is not possible to move data beyond the lab's internal network, making it comparable to an archaeological site, i.e. amenable to research but essentially immobile.

Secondly, since working with processing structured data presumes a degree of technical competence, large-scale research at the lab requires programming proficiency in Python or R. While our visual search interface is sufficient for some tasks, most research projects working at scale must include team members with data science expertise. If your project lacks these skills, consider collaborating with digital humanities research infrastructures (e.g., GRIDH at Gothenburg University, CDH at Uppsala University) or university-based computer science departments. Our data scientists at the lab do not generally work hands-on in external projects, partly due to resource constraints but also because we are sceptical towards the outsourcing of technical skills that form an integral part of doing digital research (Börjeson et al. 2024: 571).

Thirdly, we provide high-performance computing resources for compiling and processing large datasets, which raises the question of funding. We judge the particular needs of a project on a case-by-case basis depending on its particular demands, but each project generally pays some form of annual overhead fee to use the lab. These costs tends to be included within

project budgets financed by funding institutions like the Swedish Research Council (VR) or Riksbankens Jubileumsfond (RJ). Fees may be waived in cases where the project contributes significant infrastructural value to the library, though such a judgement is a result of KB's process of assessing incoming collaboration requests.

6.4 *Applying for research collaboration*

To initiate a research collaboration at KBLab, you should visit the library's website for further instructions.[27] This process is designed as a dialogue rather than a formal funding application: it is *not* comparable to the type of application that you might submit to VR or RJ and it will suffice to send us a brief outline of your project plans (ca. 1 page). By outlining your research goals and specifying which parts of KB's digital collections you intend to work with, we can assess the feasibility of your proposed project and explore the possibilities for collaboration. Once this dialogue is complete, you will receive a decision about the terms of a collaboration. We recommend beginning this process well in advance of external funding deadlines to ensure smooth planning.

6.5 *Blog for ideas and examples*

For insights into how our AI tools are being used, as well as updates on new models, we invite you to explore the KBLab blog (see Figure 18).[28] This includes case studies on topic modeling with transformers (Hägglöf & Sikora 2023), classical topic modeling approaches (Fano 2021) and the creation of a Swedish speech corpus from parliamentary debates (Rekathati 2023a). We also welcome guest contributions from researchers who have utilized KBLab's resources, such as a recent Master's project featured on the blog (Nachesa 2023). If you are interested in sharing your experiences here, please reach out to us!

7 *Conclusion*

In this chapter, we have provided a hands-on orientation to using KBLab at the National Library of Sweden, highlighting both its role as a research infrastructure and the innovative ways AI is being used to enhance digital

27 See: https://www.kb.se/om-oss/forskningsprojekt-i-samverkan.html.
28 See: https://kb-labb.github.io/.

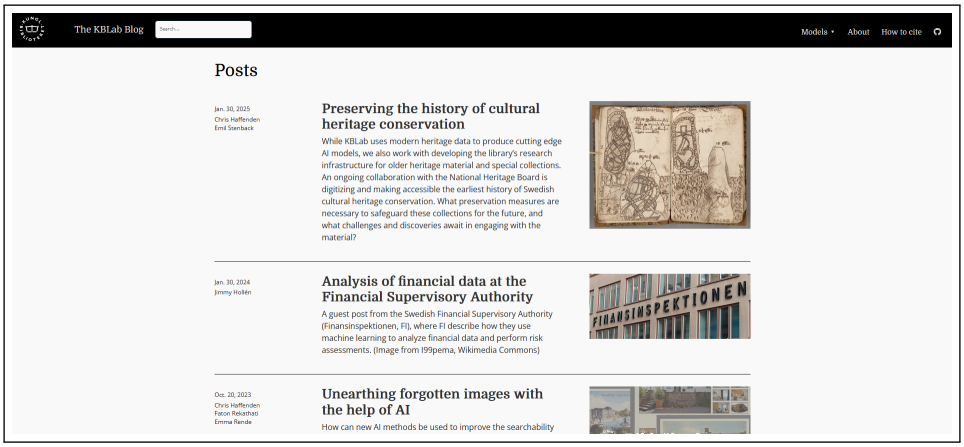


Figure 18: Sources of inspiration from the KBLab blog

heritage accessibility. The rapid development of AI tools has not only improved the searchability of heritage collections but has also created novel opportunities for digital researchers to engage with these collections at scale.

By showcasing how various research projects at KBLab have utilized AI-based methodologies—whether for text analysis, OCR correction or multimodal exploration—we hope to have illustrated the lab's potential for fostering new forms of scholarly inquiry. Beyond demonstrating what is currently possible, this chapter also invites researchers to think creatively about how AI and digital tools might further expand the scope of historical and cultural analysis in the future.

As AI continues to develop, so too will the possibilities for digital research. We encourage scholars from diverse disciplines to explore KBLab's resources and reflect upon how emerging AI tools might shape their own research. If this chapter has sparked new ideas or raised questions, we welcome further dialogue—please feel free to reach out and engage with us.

References

Baevski, Alexei, Henry Zhou, Abdelrahman Mohamed & Michael Auli. 2020. *Wav2vec 2.0: A framework for self-supervised learning of speech representations*. https://arxiv.org/abs/2006.11477v3 (8 April, 2025).

Berdicevskis, Aleksandrs, Gerlof Bouma, Robin Kurtz, Felix Morger, Joey Öhman, Yvonne Adesam, Lars Borin, Dana Dannélls, Markus Forsberg, Tim Isbister, Anna Lindahl, Martin Malmsten, Faton Rekathati, Magnus Sahlgren, Elena Volodina, Love Börjeson, Simon Hengchen & Nina Tah-

masebi. 2023. Superlim: A Swedish language understanding evaluation benchmark. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 8137–8153. Singapore. DOI: 10.18653/v1/2023.emnlp-main.506.

Börjeson, Love, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati, Hillevi Hägglöf & Justyna Sikora. 2024. Transfiguring the library as digital research infrastructure: Making KBLab at the National Library of Sweden. *College & Research Libraries* 85(4). 564–582. DOI: 10.5860/crl.85.4.564.

Bouma, Gerlof, Markus Forsberg, Justyna Sikora & Emma Sköldberg. 2024. Konsten att bedriva svensk ordforskning utan att kränka upphovsrätten. *Proceedings of the Huminfra Conference* (*HIC 2024*). 161–167. DOI: 10.3384/ecp205022.

Carlsson, Fredrik, Philipp Eisen, Faton Rekathati & Magnus Sahlgren. 2022. Cross-lingual and multilingual CLIP. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 6848–6854. Marseille, France. https://aclanthology.org/2022.lrec-1.739/.

Coll Ardanuy, Mariona, Federico Nanni, Kaspar Beelen, Kasra Hosseini, Ruth Ahnert, Jon Lawrence, Katherine McDonough, Giorgia Tolfo, Daniel CS Wilson & Barbara McGillivray. 2020. Living machines: A study of atypical animacy. In *Proceedings of the 28th international conference on computational linguistics*, 4534–4545. Barcelona, Spain (Online). DOI: 10.18653/v1/2020.coling-main.400.

Cordell, Ryan. 2020. *Machine learning + libraries: A report on the state of the field*. Report commissioned by LC Labs. Library of Congress. https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf.

Devlin, Jacob, Ming-Wei Chang, Kenton Lee & Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1* (*long and short papers*), 4171–4186. Minneapolis, Minnesota. DOI: 10.18653/v1/N19-1423.

Fano, Elena. 2021. *Topic models for Statens Offentliga Utredningar*. The KBLab Blog. https://kb-labb.github.io/posts/2021-05-04-topic-models-for-sous/.

Fano, Elena & Chris Haffenden. 2022. *BERTopic for Swedish: Topic modeling made easier via KB-BERT*. The KBLab Blog. https://kb-labb.github.io/posts/2022-06-14-bertopic/.

Forsberg, Markus, Justyna Sikora & Emma Sköldberg. 2023. *Words unboxed: discovering new words with Kubord*. The KBLab Blog. https://kb-labb.github.io/posts/2023-08-29-kubord/ (8 April, 2025).

Grootendorst, Maarten. 2022. *BERTopic: Neural topic modeling with a class-based TF-IDF procedure*. https://arxiv.org/abs/2203.05794v1 (8 April, 2025).

Haffenden, Chris, Elena Fano, Martin Malmsten & Love Börjeson. 2023. Making and using AI in the library: Creating a BERT model at the National Library of Sweden. *College & Research Libraries* 84(1). DOI: 10.5860/crl.84.1.30.

Haffenden, Chris, Faton Rekathati & Emma Rende. 2023. *Unearthing forgotten images with the help of AI*. The KBLab Blog. https://kb-labb.github.io/posts/2023-10-20-unearthing-forgotten-images-with-the-help-of-ai/.

Hägglöf, Hillevi. 2023. *A robust, multi-Label sentiment classifier for Swedish*. The KBLab Blog. https://kb-labb.github.io/posts/2023-06-16-a-robust-multi-label-sentiment-classifier-for-swedish/.

Hägglöf, Hillevi & Justyna Sikora. 2023. *Scientific discourse with BERTopic*. The KBLab Blog. https://kb-labb.github.io/posts/2023-03-17-scientific-discourse-with-bertopic/.

Hurtado Bodell, Miriam, Måns Magnusson & Marc Keuschnigg. 2024. Seeded topic models in digital archives: Analyzing interpretations of immigration in swedish newspapers, 1945–2019. *Sociological Methods & Research*. 00491241241268453. DOI: 10.1177/00491241241268453.

Löfgren, Viktoria & Dana Dannélls. 2024. Post-OCR correction of digitized Swedish newspapers with ByT5. In *Proceedings of the 8th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature (LaTeCH-CLfL 2024)*, 237–242. St. Julians, Malta. https://aclanthology.org/2024.latechclfl-1.23/.

Mahey, Mahendra, Aisha Al-Abdulla, Sarah Ames, Paula Bray, Gustavo Candela, Sally Chambers, Derven Caleb, Milena Dobreva-McPherson, Katrine Gasser, Stefan Karner, Kristy Kokegei, Ditte Laursen, Abigail Potter, Armin Straube, Sophie-Carolin Wagner & Lotte Wilms. 2019. *Open a GLAM lab*. Digital Cultural Heritage Innovation Labs, Book Sprint, Doha, Qatar, 23-27 September 2019. DOI: 10.21428/16ac48ec.f54af6ae.

Malmsten, Martin, Love Börjeson & Chris Haffenden. 2020. *Playing with words at the National Library of Sweden: Making a Swedish BERT*. https://arxiv.org/abs/2007.01658v1 (8 April, 2025).

Malmsten, Martin, Chris Haffenden & Love Börjeson. 2022. *Hearing voices at the National Library – a speech corpus and acoustic model for the Swedish language*. https://arxiv.org/abs/2205.03026v2 (8 April, 2025).

Malmsten, Martin, Viktoria Lundborg, Elena Fano, Chris Haffenden, Fredrik Klingwall, Robin Kurtz, Niklas Lindström, Faton Rekathati & Love Börjeson. 2025. Without heading? Automatic creation of a linked subject

system. In *New horizons in artificial intelligence in libraries*. Edmund Balnaves, Leda Bultrini, Andrew Cox & Raymond Uzwyshyn (eds.). Berlin, Boston: De Gruyter Saur. 179–198. DOI: 10.1515/9783111336435-014.

Nachesa, Maya. 2023. *For how long is a person recognisable by their voice?* The KBLab Blog. https://kb-labb.github.io/posts/2023-07-04-for-how-long-is-a-person-recognisable-by-their-voice/.

Norén, Fredrik, Johan Jarlbrink, Alexandra Borg, Erik Edoff & Måns Magnusson. 2023. The transformation of 'the political' in post-war Sweden. In *Digitised newspapers: A new Eldorado for historians? Reflections on tools, methods and epistemology*. Estelle Bunout, Maud Ehrmann & Frédéric Clavert (eds.). Berlin: De Gruyter Oldenbourg. 411–436. DOI: 10.1515/9783110729214-019.

Padilla, Thomas. 2025. □Humanities data in the library: Integrity, form, access□. *D-Lib Magazine* 2016(23). https://www.dlib.org/dlib/march16/padilla/03padilla.html (8 April, 2025).

Radford, Alec, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey & Ilya Sutskever. 2023. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th international conference on machine learning* (ICML'23). Honolulu, Hawaii, USA.

Rekathati, Faton. 2021a. *A multimodal approach to advertisement classification in digitized newspapers*. The KBLab Blog. https://kb-labb.github.io/posts/2021-03-28-ad-classification/.

Rekathati, Faton. 2021b. *Introducing a Swedish sentence transformer*. The KBLab Blog. https://kb-labb.github.io/posts/2021-08-23-a-swedish-sentence-transformer/.

Rekathati, Faton. 2023a. *Finding speeches in the Riksdag's debates*. The KBLab Blog. https://kb-labb.github.io/posts/2023-02-15-finding-speeches-in-the-riksdags-debates/.

Rekathati, Faton. 2023b. *RixVox: A Swedish speech corpus with 5500 hours of speech from parliamentary debates*. The KBLab Blog. https://kb-labb.github.io/posts/2023-03-09-rixvox-a-swedish-speech-corpus/.

Sikora, Justyna & Chris Haffenden. 2024. AI, data curation and the data readiness of heritage collections: Exploring the Swedish newspaper archive at KBLab. *Proceedings of the Huminfra Conference* (*HiC 2024*). 60–67. DOI: 10.3384/ecp205009.

Språkbanken Text. 2024. *Kubord-fasttext* [*Data set*]. DOI: https://doi.org/10.23695/sp99-9h02.

Underwood, Ted. 2019. *Distant horizons: Digital evidence and literary change*. Chicago: University of Chicago Press.

Vesterbacka, Leonora, Faton Rekathati, Robin Kurtz, Justyna Sikora & Agnes Toftgård. 2025. *Welcome KB-Whisper, a new fine-tuned Swedish Whisper*

model! The KBLab Blog. https://kb-labb.github.io/posts/2025-03-07-welcome-KB-Whisper/.

Warner, Benjamin, Antoine Chaffin, Benjamin Clavié, Orion Weller, Oskar Hallström, Said Taghadouini, Alexis Gallagher, Raja Biswas, Faisal Ladhak, Tom Aarsen, Nathan Cooper, Griffin Adams, Jeremy Howard & Iacopo Poli. 2024. *Smarter, better, faster, longer: A modern bidirectional encoder for fast, memory efficient, and long context finetuning and inference*. https://arxiv.org/abs/2412.13663v2 (8 April, 2025).

List of abbreviations

AI	Artificial Intelligence
ALTO	Analyzed Layout and Text Object (metadata schema)
ASR	Automatic Speech Recognition (the basis for speech-to-text applications)
BERT	Bidirectional Encoder Representations from Transformers (a language model)
CLIP	Contrastive Learning-Image Pretraining (a multimodal model)
CPU	Central Processing Unit (a standard processing chip for a computer)
GLAM	Galleries, Libraries, Archives and Museums
GPT	Generative Pre-trained Transformer (the type of language model underpinning ChatGPT)
GPU	Graphics Processing Unit (a processing chip used to train AI models)
GUI	Graphical User Interface
IFLA	International Federation of Library Associations and institutions
KB	*Kungliga biblioteket* 'the National Library of Sweden'
LLM	Large Language Model
ML	Machine Learning
NER	Named Entity Recognition
NLP	Natural Language Processing
OCR	Optical Character Recognition (process of converting image of text to machine-readable text)
PII	Personally Identifiable Information
RJ	*Riksbankens Jubileumsfond*
SBERT	Sentence-BERT language model

SOU	*Statens offentliga utredningar* 'Swedish Government Official Reports'
VR	*Vetenskapsrådet* 'Swedish Research Council'
WER	Word Error Rate (metric for evaluating speech-to-text performance)

Corresponding authors

Chris Haffenden
KBLab
National Library of Sweden
chris.haffenden@kb.se

Justyna Sikora
KBLab
National Library of Sweden
justyna.sikora@kb.se