Transfiguring the Library as Digital Research Infrastructure: Making KBLab at the National Library of Sweden

Love Börjeson, Chris Haffenden, Martin Malmsten, Fredrik Klingwall, Emma Rende, Robin Kurtz, Faton Rekathati, Hillevi Hägglöf and Justyna Sikora

This article provides an account of the making of KBLab, the data lab at the National Library of Sweden (KB). The first part discusses the work involved in establishing a lab as both a physical and a digital site for researchers to use digital collections at previously unimaginable scales. The second part explains how the lab has deployed the library's collections as data to produce high quality Swedish AI models, which constitute a significant new form of digital research infrastructure. We situate this work in the context of uneven AI coverage for smaller languages, and consider how the lab's models have contributed to the making of important AI infrastructure for the Swedish language. The conclusion raises the possibilities and challenges involved in continuing this type of library-based AI development.

Introduction

In an era of big data, significant new demands are being placed upon libraries.[1] As the world becomes increasingly amenable to processes of datafication, and more and more previously unquantified aspects of life are rendered into data, the library as a cultural heritage institution has been forced into a period of creative transformation.[2] This is partly a matter of developing collecting practices for the vast amount of material being produced online and exploring sustainable ways to describe and store these web archive collections for future users.[3] But it also involves strategies to meet the needs of users in the present, especially the novel requirements of digital scholarship.[4] Researchers in the humanities and social sciences pursuing

* *Love Börjeson is Director of KBLab at the National Library of Sweden, love.borjeson@kb.se; Chris Haffenden is Research Co-ordinator at KBLab, email: chris.haffenden@kb.se; Martin Malmsten is Head Data Scientist at KBLab and an IT Architect at the National Library of Sweden, email: martin.malmsten@kb.se; Fredrik Klingwall is a Developer at KBLab and the National Library of Sweden, email: fredrik.klingwall@kb.se; Emma Rende is a Product Manager at KBLab and the National Library of Sweden, email: emma.rende@kb.se; Robin Kurtz is a Senior Data Scientist at KBLab, email: robin.kurtz@kb.se; Faton Rekathati is a Data Scientist at KBLab, email: faton.rekathati@kb.se; Hillevi Hägglöf is a former Data Scientist at KBLab, email: hillevi.hagglof@gmail.com; and Justyna Sikora is a Data Scientist at KBLab, email: justyna.sikora@kb.se.*

digital approaches now routinely expect to be able to conduct analysis of library collections at previously unimaginable scales.[5] Such an expectation is particularly evident at research and national libraries with legal deposit material, where it creates distinctive challenges for information systems that have historically favored the analogue object and single item use. How do these libraries go about providing access to their collections as data, when so much of their underpinning socio-technical imaginaries have been centered upon the individual book?[6]

This article explores this question via the organizational form of the *data lab*. Faced with increasing demands for computational access to collections over the past decade, university and national libraries have responded by instituting such labs—with LC Labs at Library of Congress, British Library Labs and Yale Digital Humanities Lab as characteristic examples. Broadly speaking, these amount to the creation of an internal platform where the professional expertise of data scientists can be harnessed towards the informational complexities of digitization and facilitating new forms of digital research. Here we use the example of KBLab at the National Library of Sweden (*Kungliga biblioteket*, hereafter KB) to discuss what is involved in creating such a lab in a library setting. The first part details the infrastructural work required to make KB's digital collections available for large-scale analysis, as well as the practical and technical setup established at KBLab. The second part moves on to explain how the use of collections as the basis for development work with artificial intelligence (AI) has proved foundational in transforming the library into a digital research infrastructure. Though the particular details of KBLab are specific to the Swedish context, we raise broader arguments relevant for a wider international audience of library professionals, digital researchers and policy makers. In sum, the article elaborates on the value of such data labs in the heritage sector, while offering a principal justification for the project of library-centered AI development as a public good.

Literature Overview: AI in the Library

Until very recently there existed a surprising "absence of scholarly research on AI-related technologies in libraries."[7] Yet—and at least in part due to the rapid increase in public discussion of AI prompted by the release of ChatGPT—a greater body of studies exploring the possible roles and applications of AI in the context of academic and research libraries has now started to appear. One strand of such work has approached this subject from a broad perspective, looking at the question of what AI and machine learning could offer libraries in general: this includes Ryan Cordell's recent state-of-the-field report for the Library of Congress and the various publications of the "collections as data" movement propagated by Thomas Padilla and others.[8] An alternative type of study has been those focused more particularly on the AI awareness levels of library staff and the expectations that these professionals have about potential future adoption and application of such techniques.[9] Various studies have also discussed both some of the challenges in enabling a data-driven approach to digital research at scale, and some of the ways in which new forms of AI applications could be integrated in the work processes of academic libraries to make digital collections more amenable for such research.[10]

However, there are two significant dimensions of AI applications in libraries that have received less attention. The first of these is the practical and organizational efforts involved in making it possible for new insights from the field of data science to be explored and experimented with in the information-intensive environment of academic and research libraries in general, and at legal deposit libraries in particular. Here we build upon a recent attempt

to address this lacuna, *Open a GLAM Lab*, the manifesto encouraging the growth of further labs in the heritage sector, by providing a specific case study of such institutional work.[11] The second dimension is the potential for libraries not only to integrate AI techniques developed elsewhere, but also to serve as a site of experimentation for the making and testing of more democratically-inclined AI tools that are transparent and open for scrutiny. In pursuing this aspect, we are reinforcing the suggestion that "data labs at libraries—and especially national libraries—can have a significant role to play in the future of AI.

KBLab as Entrance Point to the Collections for Research

In this opening section, we sketch the practical and organizational conditions that shaped the making of a data lab at the National Library of Sweden. How does KBLab align with and form part of KB's broader mission as a national library? What is it about a national library's collections that is particularly well-suited to the type of work that is possible in such a lab? And what is involved in creating access to this material in a lab environment? Addressing these questions provides the contextual detail necessary to make sense of the subsequent discussion of AI development in the library presented towards the end of the article.

Library Collections as Data

The concern with making collections available is entirely central to KB's purpose as a publicly-funded heritage institution. The library's obligations to the research community in this regard are highlighted in the legal act defining its principal mission, where the opening paragraph describes KB as both a "national library" and a "national research infrastructure."[13] While the concrete tasks that pertain to this—i.e. to collect, describe, preserve, and make accessible material—are outlined in relation to the general good of aiding democratic development, the act specifically connects these activities to the end of safeguarding Swedish research quality. In this sense, KB is bound by law to maintain a close relationship with the shifting and dynamic needs of researchers. In practice, and given today's increasingly digitalized media ecology, this means incorporating the digital into a national research infrastructure and in turn becoming a digital research infrastructure. As we will demonstrate, KBLab comprises an essential component in both of these dimensions.

KB's collections can be characterized in terms of their considerable breadth and scale. While first introduced as a form of official censorship in 1661, with publishers forced to submit a copy of each work to the state for approval prior to public circulation, Sweden's legal deposit act has long since served to make the national library a guarantor of future cultural heritage. The law dictates that a copy of every publication issued in Swedish must be submitted to the library for preservation; since 1979 this has included audio-visual material as well as print, and since 2015 at least a degree of electronic publications.[14] Beyond their historical depth and continuity over time, the collections thus also encompass cultural production from a diverse and shifting media landscape: ranging across newspapers, magazines and books through scientific journals and governmental reports to radio broadcasts, television shows and computer games. To give a sense of the scale involved, KB's physical collections alone now number over 18 million items in the archives.

Although only a small part of these collections has yet been digitized, sufficient volumes of digital material exist to make the "collections as data" perspective highly pertinent.[15] Such a framework entails exciting possibilities but also significant infrastructural challenges for the

GLAM sector. In terms of the former, the creation of high quality, language-specific humanities data opens up new potential for researchers to be able to analyze the contents of digital collections at previously unimaginable scale, and often in previously inconceivable ways. The existence of such data is also especially valuable for the development of AI tools for smaller languages—a point revisted below.

Yet producing and providing access to humanities data is far from a trivial task. The library's collections have a particular history that has shaped their form in the archives, producing data artefacts that need to be managed. To take one key example, we can consider the effects of optical character recognition (OCR) software within the production process of digitized heritage material.[16] Since the particular terms of the Swedish legal deposit law have previously prioritized physical over digital examples, physical newspapers have been submitted to KB and then subsequently digitized. Beyond certain OCR errors with specific Swedish words, an effect of this digitization is the loss of various aspects of metadata that we as humans take for granted. A digital copy is gained from this process, but what is left is a mishmash of text blocks with no clear indication of which blocks belong together, which articles comprise part of the same section, and which texts are editorial content rather than adverts. It is certainly possible to use machine learning to attempt to put Humpty together again and reconstruct the newspaper, but this is a complex and laborious undertaking.[17] Making collections amenable to computational analysis is a qualified task that often demands considerable labor in terms of data cleaning and curation; humanities data is far from ready-made.[18]

It is within this particular context that KBLab came into being. On the one hand, there has been growing demand from scholars within the humanities and social sciences using digital approaches, who wish to be able to access digital collections to conduct large-scale analysis. As the pilot study that laid the ground for the founding of the lab suggested, "researchers, funding agencies and governmental research propositions are also increasingly pushing scholarship in a data intensive direction in order to promote digital scholarship."[19] On the other hand, there is the technical complexity involved in creating an infrastructure capable of providing access to these collections as data, when little has previously existed. Despite any suggestion to the contrary, enabling the production of high-quality datasets fit for research can be a complicated and messy undertaking. We now turn to consider how we sought to address this challenge through the making of a lab at KB.

Designing Technical Infrastructure

When the library formally initiated the project to establish a data lab in 2019, two particular user groups—and purposes—were specified. Internally, KBLab was conceived of as a resource for method development and AI innovation at KB: a means of providing staff and leadership with improved knowledge about the potential for automating various library processes. Externally, the lab was intended to position itself among existing institutions and environments to become an established infrastructure for facilitating and supporting digital research. In the short term, it was to meet the needs of two major projects within the digital humanities and social sciences funded by the Swedish Research Council: "Welfare State Analytics. Text Mining and Modeling Swedish Politics, Media & Culture, 1945–1989" (based at Umeå University)[20] and "Mining for Meaning: The Dynamics of Public Discourse on Migration" (based at Linköping University).[21] That both of these wished to conduct large-scale analysis of mid to later twentieth-century material from KB's collections—principally newspapers, but also

fiction and periodicals—had an important effect on the technical and organizational development of the lab. Since the projects were concerned with the analysis of material still protected by copyright, our initial task was to design a computing infrastructure to provide local access within the library itself.

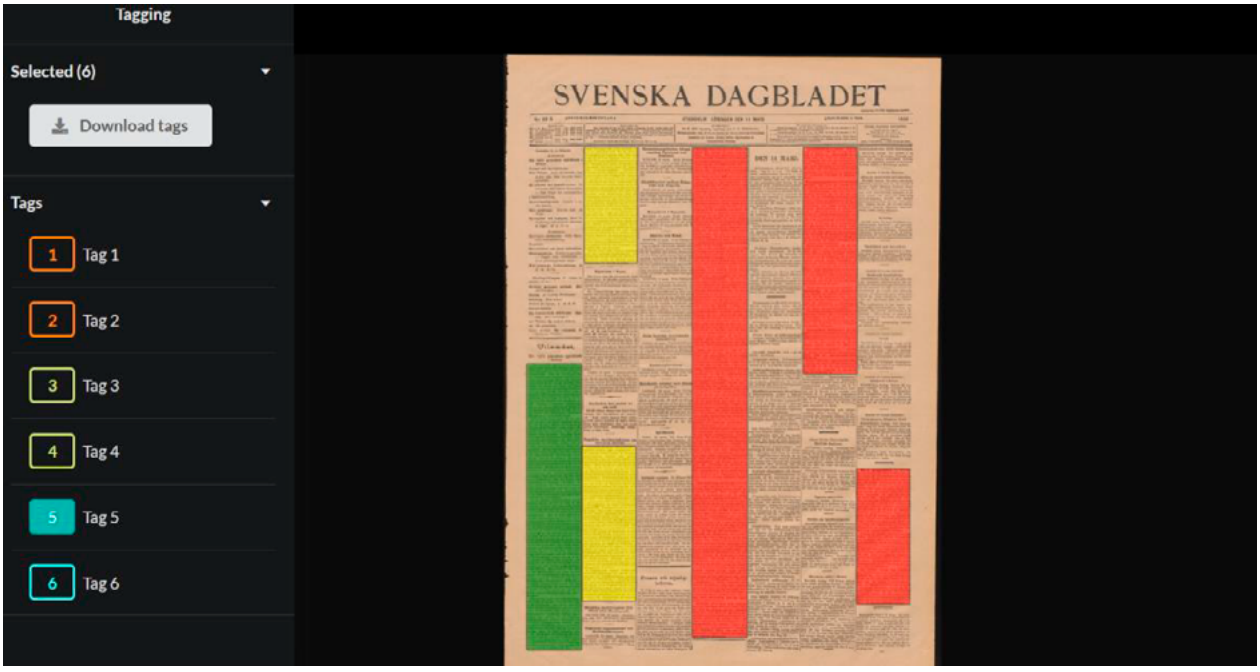
A key starting point in approaching this task was the conviction that a data lab at a library should offer experimental access to the collections. This required a laboratory space that enabled and encouraged researchers to explore. Of course, one way of making digital collections available for further research is to produce predetermined datasets that can then be released to be analyzed and used in diverse ways. However, and beyond the fact that copyright restrictions prevented such an approach in this instance, providing already defined datasets tends towards restricting rather than supporting open-ended critical investigation. Considering the research process as something not necessarily linear—indeed, as often tangential and shaped by serendipitous findings far beyond the initial remit of enquiry[22]— we opted to design an infrastructure where an exploratory working method would be possible, if not inevitable, for the researchers who come to use it. Once a research project is onboarded at the lab, it is granted *unlimited* access to KB's digital collections so researchers can explore and design their own datasets as a result of contact with these collections.

The technical challenge this involved was in how to enable exploration without compromising security. Our solution was to offer indirect access though an Application Program Interface (API). Researchers can conduct searches of KB's digital collections via the lab's API that will give results in the form of a representation of the original data, rather than the unnecessary risk of exposing the library's databases through direct access to the files themselves.[23] These representations take the shape of JSON files, a data form reflecting the longer history of KB's engagement with linked data and one which is particularly apposite for a digital research infrastructure since it is machine-readable.[24] Another key aspect of the linked data model underpinning the lab's design is the presence of Uniform Resource Identifiers (URIs) in the lab environment. By providing stable and persistent URIs for the archival material, researchers are able to find their way back to the same point in the collections and ultimately to demonstrate that their results are reproducible. Through establishing an API and an information model that makes linked data of the digital archive, we could create programmatic access to the library's collections that offers researchers the autonomy to steer their own exploratory processes.

In addition to making the material searchable through the API, we also created a graphical user interface (GUI) for the lab environment (see figure 1). This serves to strengthen the lab's functionality as a research infrastructure in various, overlapping ways. Firstly, it provides a means of validation: in accessing the material in visual form, researchers can verify and navigate among their results. Secondly, it provides a way for scholars within the humanities and social sciences without programming skills to access and interact with the material in the lab environment. This can be particularly important for multi-disciplinary projects that seek to combine the perspectives of data science with more traditional forms of expertise in close reading and analysis within various humanities disciplines. It is also pertinent for mixed-methods analyses that, beyond conducting large-scale computational analysis, wish to incorporate investigation of visual aspects of the material (and therefore also need to be able to see the individual object rather than an aggregation of its textual contents). Thirdly, it allows for the annotation of the material, which can prove a significant element in projects that utilize machine learning by training models based on the collections. There is a specific

FIGURE 1
Visualizing a Lab Environment for Explorative Research, KBLab's GUI

FIGURE 2
Annotating and Extracting Research Data in the Lab Environment



function within the interface that allows users to annotate according to their own chosen labels and then extract the particular text that has been annotated (see figure 2).

We made the lab's GUI available to researchers outside the lab itself through a prototype service called betalab. On the one hand, this forms part of the onboarding process for research projects that have been granted access to use KBLab. Prior to gaining access to the physical premises of the lab, researchers can use betalab to test and accustom themselves to the lab environment—in certain cases, they can even design and prepare scripts to be run on-site once they have access (which can prove a significant time-saving approach for geographically

disparate projects). On the other hand, betalab is also used as an access point for those parts of the digital collections that are available at the lab but not subject to copyright restrictions. This open data includes historical newspaper material up to 1906, the Swedish Government Official Reports (SOU's) and various parliamentary data. If, for example, a research project wished to access newspaper material in machine-readable, structured form to conduct an investigation into nineteenth-century crime reporting, KBLab would provide them with access to this data via betalab. In this sense, it provides an important complement to the primary part of the lab that can be accessed on-site at KBLab's premises.

For research projects that need to use KBLab for large-scale computational analysis of those parts of the collections protected by copyright, we established a computer lab at KB's locale at Karlavägen in Stockholm. This physical manifestation of the lab as research infrastructure is significantly a matter of computing power: since the terms of Swedish copyright legislation mean it is not currently possible for these projects to move the data outside the lab for external processing, we needed to ensure that there were sufficient computational resources in-house to meet the researchers' needs. To this end, we built a local computing infrastructure with three levels: a) powerful workstations at the computer lab; b) a server environment for computation and access to the material via an API; and c) two NVidia DGX A100 servers for more computationally heavy analysis. (We have also since been granted access to the EU's supercomputing infrastructure for our own development work, a point we return to below.)

The guiding principles that shaped the technical work to establish this solution have been pragmatism, flexibility, and a desire to create autonomy for the researchers who use it. The workstations in the computer lab, for instance, use the Linux-based system Ubuntu, as this allows researchers to create and control their own software environments according to their particular needs and preferences. Likewise, to enable researchers to manage their own back-up for code and work-in-progress, we created gitlab, an internal, server-based git function. We elected to start by acquiring consumer rather than enterprise hardware for the lab: in part due to the (relatively) limited resources we had at our disposal, but also because it allowed us to move quickly and adapt according to the shifting needs of researchers as these emerged. The work involved in establishing this setup has depended upon KB's existing staff expertise within IT-architecture and systems design; without the input of an experienced and creative IT-architect, the making of KBLab would not have been possible.

Research Coordination

With the lab established as an entrance point to the collections, another important organizational matter to be dealt with was research coordination. A significant aspect of this involved determining the principles and procedures via which access to the computer lab should be granted. In particular, and given that demand to use KBLab among researchers has consistently been greater than our on-site capacity, how should places at the limited number of workstations be allocated?[25] To address this in a fair and transparent manner that aligns with KB's values and missions as a public authority, we made applying to the lab part of the library's broader process for managing research and development applications.[26] Researchers who are interested in collaborating with KBLab therefore begin by submitting a brief project outline describing what they would like to do in their proposed research. This application is then subjected to an initial screening to confirm that the project actually involves research elements—i.e. that there are questions and hypotheses amenable to further exploration—before any decision is

made about the specific terms of collaboration that might be possible. Important to note in this context is that, apart from confirming the presence of a research question and determining its essential feasibility, we make no judgement upon the substantive content of the research proposal.

The question of sustainable funding is central to the existence of any lab, and this also impacts how new applications to KBLab are handled.[27] While the initial outlay for the lab was financed through a combination of internal funding from the library and external funding from the projects mentioned above, our working assumption is that research projects based at the lab should be self-financing—i.e. that they pay an overhead fee to cover the running costs (technical and administrative) in utilizing the lab, in line with a general Swedish praxis for the use of research infrastructure. Given the configuration of funding for academic research in Sweden, this means researchers have to include a budget post for use of the lab in their applications to research funding organizations such as the Swedish Research Council (VR) and Riksbankens Jubileumsfond (RJ). It also means researchers need to coordinate applying for a place at the lab with the process of submitting a funding application to these organizations.

The advantage of this approach is that it serves as a mechanism for quality control: by granting access to projects that have been awarded funding following a competitive, peer-reviewed process, we can ensure that the research allocated a place at the lab is of the highest caliber. However, a potential disadvantage is that it can favor larger projects proposed by established researchers at the expense of smaller initiatives by less well-established scholars. To counter this, a pragmatic cost-benefit analysis is adopted when considering each potential project, which can allow the overhead fee for use of the lab to be waived in certain cases. For example, if a project involves significant infrastructural gains for the library beyond the particulars of the project itself, then such a solution might be possible. A typical instance where infrastructural positives outweigh any overhead costs is the various Masters projects in machine learning that have been hosted at the lab, which have explored how AI models can help make the library's collections more accessible.[28]

A further dimension that affects how the overhead costs for a potential project at the lab are assessed is the level of data science competence in the project team in relation to the complexity of the proposed research. The underlying issue here is finding productive forms of collaboration between expertise in AI and machine learning, on the one hand, and more traditional qualitative competences in the humanities and social sciences, on the other.[29] Based on our experience, outsourcing the requisite expertise for large-scale data analysis to technical staff outside the project is the *least* effective way of dealing with this question. Such an approach tends to be problematic, partly since it risks making vital technical labor invisible and uncredited, and partly because it lends itself to a situation where researchers in the humanities and social sciences are publishing work where they do not properly understand either the methods used or their subsequent results.

On this basis, we recommend that projects based at KBLab incorporate data science competence within their project team, so that this perspective is represented and accountable at all stages of the research process. In practice, this means we are reluctant to grant lab space to proposals lacking the necessary technical skills, instead referring these to other infrastructural organizations such as the various Swedish centers for digital humanities who can provide greater levels of support. To proposals that have included the necessary expertise, we offer an overhead fee that is adjusted according to the technical complexity and demands of the

specific project: ranging from a standard rate that includes initial support and advice in using the lab, to higher levels when a greater degree of development work is required from the lab's staff to make the project possible. In each case, ascertaining the particular needs and requirements of a proposal presumes an ongoing dialogue with the researcher and deliberation from several of the lab's staff.

Once a project has been offered a place at the lab and received research funding, it is ready for onboarding. This process was designed in accordance with the particular model of explorative research practices mentioned above in the discussion of the lab's technical setup. A thorough introductory phase clarifies the formal terms for using the lab, where researchers sign a personal user agreement stipulating the legal conditions for accessing and using the data available at KBLab, as well as receiving a copy of the code of conduct (see appendix 1). This is followed by a hands-on guide where the researcher(s) will be shown how to access data via the lab's API, how to manage ongoing results and which among the lab's various tools might be of interest. After this introduction, researchers are ready to work autonomously at the lab: beyond consulting with lab staff in the event of problems, they are free to begin interacting with the collections at KBLab according to their particular interests.

Collections-based Models as Digital Infrastructure

Having discussed the making of the lab as a physical site for researchers to access the collections, we now turn to discuss how we have harnessed the collections as the basis for new digital tools that in themselves constitute a significant form of research infrastructure. Whereas the number of researchers who can use the on-site lab is necessarily limited by practical constraints, the creation of such tools that can be distributed beyond the library has enabled the lab to have a far greater reach. In the remainder of this article, we outline our work in producing and releasing collections-based models at KBLab: how have KB's digital collections enabled the emergence of a library-based form of AI development?

Library Collections as Training Data

The past five years have witnessed rapid rates of development within the field of AI and machine learning. For instance, the release of transformer-based language models like BERT has proved the basis for unprecedented performance in many natural language processing tasks.[30] However, the emergence of such AI tools has occurred according to existing global hierarchies of power and resources: they are far from being equally available to all languages and actors. While Google AI developed dedicated BERT models with cutting-edge capabilities for major languages like English and Chinese, other languages had to make do with a less powerful multilingual model. Where big tech companies lacked the commercial interest to train these tools for particular languages, actors within the academy and beyond have tended to take the initiative to produce state-of-the-art monolingual models.[31] For so-called lesser-resourced languages, a significant bottleneck to doing so was the availability of sufficient computational resources and training data. In the instance of Sweden, the first monolingual BERT model was created by the Public Employment Agency using solely data from Swedish Wikipedia which, while better than Google's multilingual model, was still considerably less effective than the English BERT.[32]

Yet the prevailing paradigm for producing state-of-the-art AI models enables national libraries and other heritage institutions to contribute to development in novel ways, especially

in the case of smaller languages. With the emphasis on unsupervised learning in current AI development—i.e. when vast algorithms called artificial neural networks learn through being exposed to huge volumes of unlabeled training data, rather than, as previously, from smaller amounts of (expensively) annotated data[33]—new opportunities have emerged for the custodians of large amounts of high quality, language-specific data.

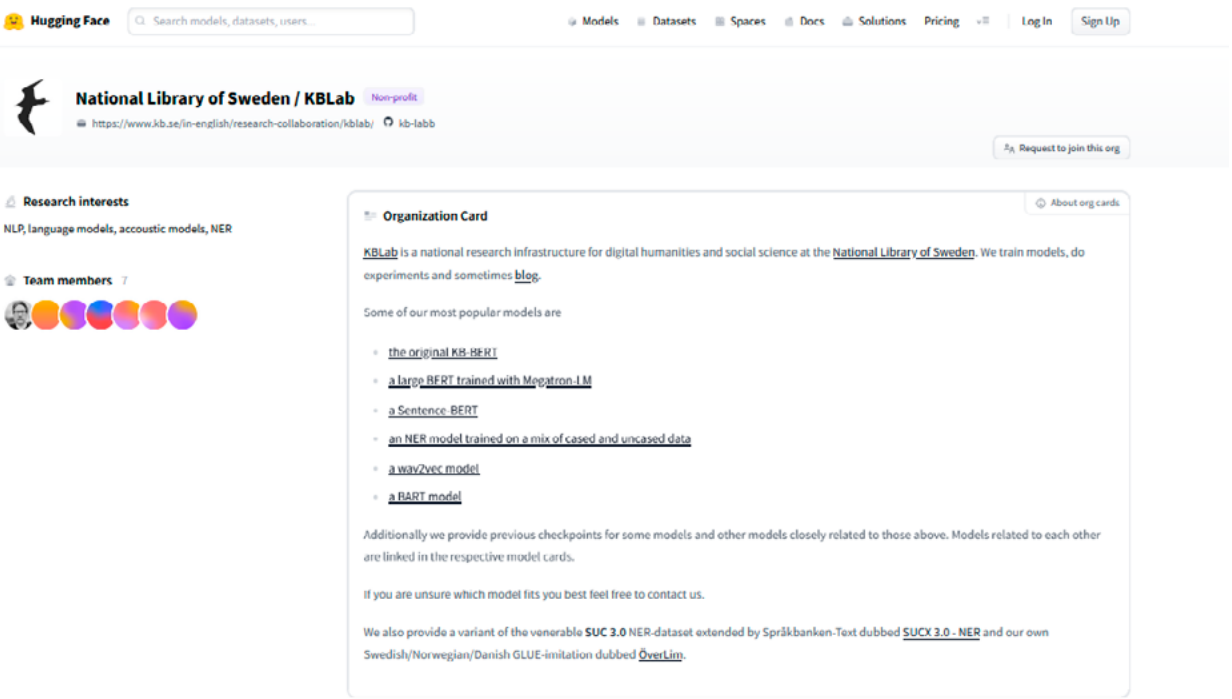
In such a context, the breadth and depth of KB's collections mentioned above becomes a uniquely valuable resource for the making of cutting-edge tools for Swedish AI. Indeed, the fact that legal deposit provides KB with something approaching population data for the language means there is an important *democratic* dimension to harnessing the library's collections as training data. With recourse to a broader and more representative range of data than that available to other actors (who have access chiefly to Swedish data that can be scraped from the web), KB has the potential to produce AI models of greater quality and effect. Given that this data cannot be shared beyond the library due to copyright and GDPR legislation, this creates a powerful rationale for the training of models in-house at KBLab.[34]

Making and Distributing Collections-based AI

Against this backdrop of enhancing the quality of Swedish AI infrastructure towards global state-of-the-art, we have been using the library's digital collections to train new AI tools since the lab was established in 2019. The first phase of this development work focused specifically upon text, with the aim of improving the capabilities available for automated analysis of Swedish text in light of recent innovations with transformer models. Here we turned to the breadth and depth of KB's collections to train a BERT model for Swedish capable of processing "the living language of the national community."[35] To create such data representativity, we produced a large and diverse training corpus that made substantial use of the library's digitized newspaper archives dating back to 1945, as well as more recent online material and social media to capture more colloquial language use. Making this material amenable to machine learning so it could be used as training data also involved painstaking and laborious processes of data curation, which in turn depended upon the specialized competence in data science and programming of the lab's staff. The language model that this enabled, KB-BERT, proved significantly more effective than existing models and has since become the standard model to use for Swedish language processing.[36]

In line with the increasingly multi-modal direction of current AI innovation and the multimedia inclinations of recent humanities scholarship, our development efforts at KBLab have also moved beyond solely text. Here we have been able to take advantage of the diversity of media forms stored in the archive: ranging across a variety of different modes, KB is guardian of unparalleled collections of Swedish text, images, sound, and film, which equates to a considerable range of possibilities for training new models. A pertinent example is the work at the lab in producing improved tools for automated sound recognition (ASR). This involved using the library's enormous, and often largely unexplored, holdings of audio-visual material from the twentieth century. More specifically, we utilized KB's digitized national and local radio programs from the past two decades to produce a corpus of over 1.4 million hours of spoken Swedish, including dialects from all the regions in the country.[37] This was then used as training data for Swedish versions of the wav2vec 2.0 model developed by Facebook (now Meta) AI.[38] As was the case with KB-BERT, the collections-based models that this produced, entitled VoxRex, outperformed existing multilingual and monolingual models for speech-to-

FIGURE 3
Open Access Swedish AI Models Available for Download



text tasks.[39] As we explain below, the existence of cutting-edge tools for Swedish speech-to-text creates a range of synergy effects, both within and beyond heritage institutions.

To ensure the AI tools produced at KBLab can benefit as many as possible, we release our models open access via the data science community platform Hugging Face (see figure 3).[40] To date, we have made available 46 models in this way so that users are able to download and experiment according to their particular interests. In addition to the KB-BERT and VoxRex models mentioned above, these include a canonical Swedish SpaCy, a zero-shot classifier, Sentence-BERT and BERT models fine-tuned for named-entity recognition (NER), as well as Swedish versions of the latest Whisper models for ASR released by OpenAI.[41] Beyond training our own models for Swedish text and sound, we have also collaborated with other actors in developing multimodal tools that connect image and text to enable new forms of image search.[42] As part of the transparent and accountable approach to AI development that we pursue at the lab, we make sure to document the data we have used to train our models through a combination of Hugging Face descriptions, blog posts and research articles.[43] We also share our code via GitHub.[44] In this way, we seek to make it possible for users of the lab's models to understand how these tools have been made, and to consider how, in light of the particular values and emphases contained within the data in KB's collections, the models might need to be adjusted for use in specific applications.[45]

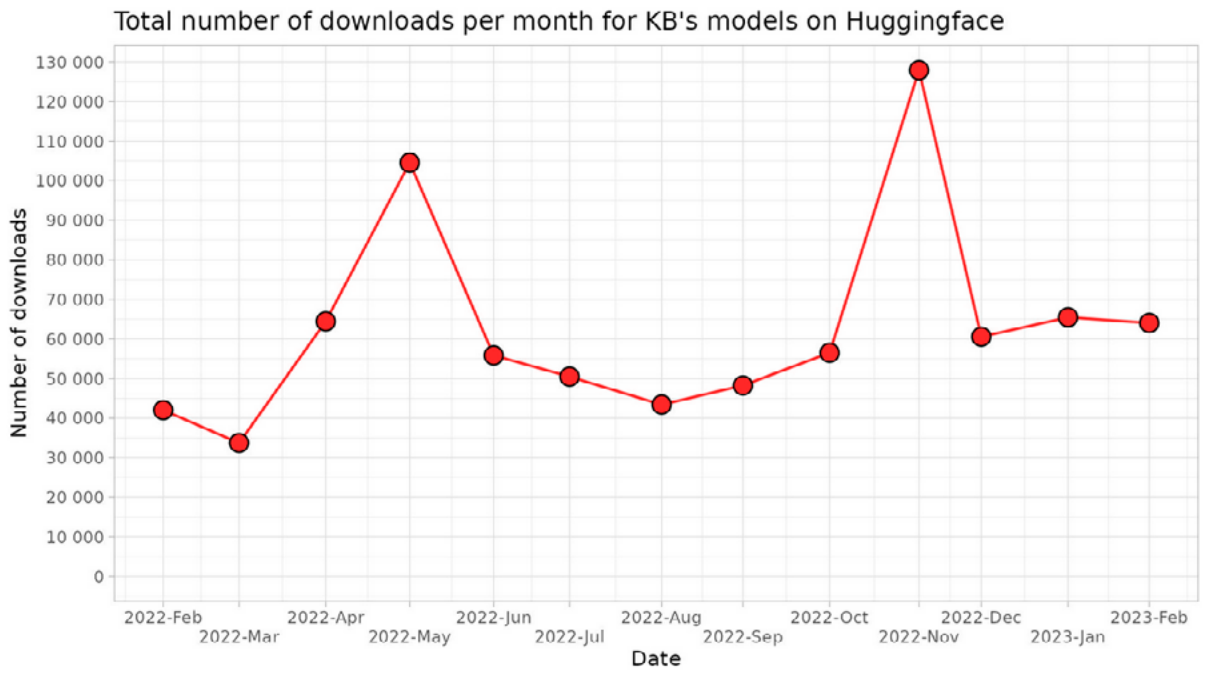
The Value of Collections-based Models in Practice

Turning to tracing the value of the models trained at the lab, these are being put to use in a diverse range of contexts. The initial impetus towards producing these models was as a means for making the library's huge, but often largely uncharted holdings of digital material more accessible; through creating such tools we hoped to help the library better understand and describe

its own collections, while simultaneously improving research access to this material. That this has come to fruition is demonstrated, for instance, in the various Master's projects at KBLab that have shown how KB-BERT could be used for the automated enrichment of metadata in the digital newspaper archive.[46] It is also evident in a pilot project exploring how a topic-modelling approach built upon our Swedish Sentence-BERT, BERTopic, might be used to provide a sort of automated subject headings that offer more fine-grained navigation of the collections.[47] Perhaps most striking, though, is the positive feedback loop created by the lab's development work with sound data mentioned above: first, KB's collections enabled the production of state-of-the-art ASR models; these models can then be used for speech-to-text to make these collections amenable to text searching; and text transcriptions of the radio and television material can be used as new training data for yet another generation of new and better models at the lab. AI development and improved metadata thus work hand-in-hand to improve the accessibility of the material, thereby enhancing the library's function as a research infrastructure.

Beyond the library and research projects based at the lab, KBLab's models have proved valuable for both a surprising array of academic research and for information-intensive organizations outside the academy, in public and private sectors alike. In terms of the former, KB-BERT has now been utilized by medical researchers seeking to develop new lifestyle treatments for diabetes patients; in attempts to automatically identify the presence of implants (i.e. pacemakers or stents) in heart patients prior to MRI scans; and for the classification of legal documents.[48] In terms of the latter, the lab's models have been put to work in automating and streamlining the information handling processes of various public authorities, including local councils, the Swedish Tax Agency (*Skatteverket*), the Swedish Courts (*Domstolsverket*) and most recently, the support function of State administration (*Statens servicecenter*).[49] As a growing number of Swedish organizations and companies start to become aware of the pos-

FIGURE 4
Download Statistics for KBLab's Open Access AI Models for February 2022–February 2023



sibilities presented by AI, they are increasingly turning to the lab's models for easy to access and state-of-the-art performance.[50]

However, the most striking evidence of the scale of the impact of the lab's collections-based models is quantitative in nature, with statistics showing over a million downloads since they were made available on our Hugging Face page (see figure 4).[51] Of course, these figures need to be contextualized: this does not refer to the number of discrete users, but rather the total number of times the models have been used (within, say, a particular application). While there is no way of knowing further details about such usage—beyond what can be traced from citation of the lab's publications, and cases where specific developers have contacted the lab—these statistics can still be taken as a forceful demonstration of the reach of our work in producing new AI tools. As both of the independent evaluations of the lab's first two years have highlighted, the widespread uptake of models trained at the library indicates a pertinent way in which the lab's development work reinforces and furthers KB's democratic commitments.[53] In making and releasing AI models using the library's data, KBLab thus offers a powerful new way of sharing the value of the collections outside the library itself.

Library-based AI Development as a Public Good

The principal merit of this type of AI development is the way it can simultaneously enhance the library's functions as national research infrastructure and guarantor of democratic values.[54] As a form of digital infrastructure, open access collection-based models enable a wide range of AI applications for Swedish, which would have proved difficult, if not impossible, without such infrastructural tools in place.[55] As a means of encouraging and enabling democratic social development, there are distinctive, yet mutually reinforcing aspects, of these models that are worth accentuating.

That we are facilitating an expansion of Swedish AI implementation by making the models freely available is intensified by the particular logic of the models' architecture. More specifically, the relative resource allocation between the pre-training and fine-tuning phases of a Transformer model lends itself to effective dissemination: while pre-training for a general purpose model like KB-BERT is computationally expensive and presumes large amounts of data, subsequent fine-tuning can be carried out with but a fraction of the data and computational resources.[56] This means that a far wider range of social actors outside resource-intensive environments like the university can consider downloading the models, experimenting with them, and applying them to their own particular use cases. Our work at the lab is thereby contributing to a democratizing of both the technology and the library's data.[57] In this sense, KBLab is helping to share the newly-found value of the collections as data—understood as a form of publicly funded and maintained *commons*[58]—to new groups of users than those traditionally reached by the library and beyond.

This form of library-based development can even counter some of the more problematic aspects of an AI future driven purely by private sector actors, particularly the growing deficit of data accountability. As emergent AI technologies become more mature and increasingly governed by commercial concerns, there has been a concurrent move towards treating training data and methods as trade secrets to be protected from competitors. This was exemplified in recent discussions about the lack of transparency surrounding OpenAI's release of GPT-4, with one scientific researcher, Sasha Luccioni, suggesting "it's just completely impossible to do science with a model like this," given the lack of access to details about the data used to

make it.[59] Such opaque practices can also be connected to a wider culture of silence in the tech industry that precludes critical voices about, for example, the overreliance on vast, unaccounted-for web materials in training new models, as the case of Timnit Gebru amply illustrates.[60] By contrast, in adhering to careful practices of documentation, scrutinizing the workings of data representativity, and pursuing more representative models based on the breadth of the library's collections, we are engaged in an accountable form of AI work at the lab that can variously complement and challenge that of private tech companies.[61] Insofar as it is more open and operates according to other imperatives than commercial interest, library-based AI development can constitute one of the much-needed "alternatives to the hugely concentrated power of a few large tech companies and the elite universities closely intertwined with them."[62]

Yet using publicly-funded heritage data as the basis for a more ethical AI development is dependent upon sourcing new forms of resources. Although it was possible to produce cutting-edge tools at the lab when these were of a proportion of a BERT model, the pace of recent AI innovation has led to new models at a scale that makes this far more challenging. To give a sense of the leap in scale: where BERT had hundreds of millions of parameters, GPT-3 has over 175 billion and GPT-4 is suspected to have far more—though, of course, this latter figure remains shrouded in secrecy as yet. While we have the prerequisite training data and specialized expertise in data science to produce larger models, a significant bottleneck has been in locating sufficient computational resources. To solve this, and to be able to further our work in producing state-of-the-art models for Swedish, we sought the help of ENCCS (EuroCC National Competence Centre Sweden) to apply to use the EU's infrastructure for supercomputers, EuroHPC.[63] Gaining access to first HPC Vega (in Slovenia, with 240 GPUs) and now HPC Meluxina (in Luxembourg, with 800 GPUs), has enabled development work of a different scale at the lab.[64] In becoming the first public authority to use these EU-funded development resources, KBLab is furthering the prospect of a Swedish AI that is open, accountable, and democratically inclined.

Finally, contributing to the making of a national AI infrastructure in this way also demands novel collaboration. With the release of KB-BERT establishing the lab as a key player in Swedish language technology, we have since become involved in national and international networks that include a diversity of actors who are engaged in AI questions: from researchers and university departments, to coordinating organizations, public authorities and private companies. Forming new relationships and collaborating with this configuration of groups beyond those that the library has traditionally cooperated with is an important step in trying to work effectively in the rapidly evolving space of AI development. A recent example was the lab's role in a project, together with the National Language Bank of Sweden at Gothenburg University, the Swedish Research Institute (RISE), and AI Sweden, to provide a set of benchmarks for evaluating Swedish language models.[65] By working together to make it easier for users of Swedish AI to determine which models might best fit their purpose, we are helping to make recent innovations more widely accessible. In this way, the lab's research collaboration with external actors is also leading to improved infrastructure.

Conclusion

The establishment of a data lab at the National Library of Sweden has enhanced the library as a digital research infrastructure. As we have explained, various practical and technical considerations shaped the making of KBLab as a physical site where researchers can now access

the collections at unprecedented scales. The library's digital collections have enabled the lab to play an important role in contributing to the development of a national AI infrastructure for the Swedish language. As a closing note, we offer some reflections on the possibilities and challenges facing the lab as a node for library-based AI development.

One of the key justifications for library labs in particular, and GLAM labs more generally, is that they provide new ways of sharing the value of cultural heritage material. Establishing such a lab can lead to snowball effects with various positive, if often unforeseen, consequences. In particular, the consolidation of in-house expertise within data science and machine learning opens up significant possibilities for heritage institutions that are increasingly becoming custodians of large volumes of digital material. Through working in tandem with domain specialists (i.e. librarians, archivists, curators, etc.), such labs can make these collections available at scale to researchers and other users so they can pursue new lines of inquiry. Adopting a collections as data approach also creates significant opportunities to contribute to AI development, especially for lesser-resourced languages that have not been prioritized by major commercial actors. By using high quality, language-specific heritage data to contribute to national infrastructure, and engaging in novel collaboration with external actors, these labs can play a role in democratizing this data, while laying claim to a powerful new form of social relevance in the process. In short, GLAM labs create new and unexpected lives for collections far beyond the heritage sector itself.

Conversely, while it might seem a platitude, it is far easier to start a lab, with all the start-up energy and buzz this entails, than it is to entrench one as a given part of a heritage organization. In part, this is about the thorny question of funding and a systematic tendency to underinvest in digital research infrastructure.[66] But it is also connected more specifically to the difficulty of attracting and retaining highly-qualified staff within publicly-financed AI development, when the demand for this data science expertise in the private sector is intensifying. There are even complexities to be addressed concerning how this expertise should be integrated within the wider organization: should data scientists be centralized within a lab, as is the case with KBLab, or are there arguments for distributing this competence across the organization as a whole? How might fruitful interactions between data scientists and domain experts best be encouraged?[67] In dealing with such questions and seeking to navigate a way forwards through the rapidly shifting terrains of digitalization and AI innovation, there is a compelling need for strategic leadership and direction.

Based on our conviction that the future interactions of AI and the library can be mutually beneficial, we conclude by offering some concrete pointers to any research library considering establishing a data lab or investing in digital transformation connected to AI. The first pointer underlines the importance of *people* and the centrality of an open, interdisciplinary outlook among the project team. Whether recruited internally or externally, the library professionals engaged in this work should be both driven by infrastructural questions and curious about other perspectives, be this data scientists interested in the design and use of research infrastructure or humanities researchers interested in collections as data. The second pointer is about the provision of a realistic time frame to give the project space to experiment before being expected to deliver. This is partly a matter of financial support and ensuring there is sufficient continuity beyond the short-term possibilities of external research grants, but it is also connected to the question of timing for strategic evaluation. (In the case of KBLab, for instance, the external assessment reviewing the lab's establishment took place after two years

before the decision to make the lab a permanent part of the library was taken.) The third pointer is to allow for specialized legal support to assist the project in navigating relevant national and international legislation about data use and sharing. The final pointer is the importance of continued dialogue with various stakeholders, within and beyond the library, concerning the possibilities and risks with ongoing AI development. Working responsibly and transparently to mitigate these risks, library-based AI can be a synergizing venture that significantly enhances the availability, usability, and value of heritage collections for present and future users.

Appendix 1. KBLab's Code of Conduct

KBLab is an open and friendly working environment, where collaboration is encouraged, questions are welcomed, and the commitment to critical, open-ended and independent enquiry is foundational. This openness is essential for new research ideas and projects to flourish.

A prerequisite for such an environment is being kind to one another. To maintain this space for open enquiry, we expect all people connected to KBLab to behave according to the principles of mutual respect and decency. Not honouring these principles can lead to access to KBLab being withdrawn.

If you have any questions about this code, please contact us at kblabb@kb.se.

Notes

1. M.B. Hoy, "Big Data: An Introduction for Librarians," *Medical Reference Services Quarterly* 33:3 (2014): 320-326, https://doi.org/10.1080/02763869.2014.925709.

2. K. Cukier & V. Mayer-Schoenberger, "The Rise of Big Data: How It's Changing the Way We Think About the World," *Foreign Affairs* 92:3 (2013): 28-40, http://www.jstor.org/stable/23526834.

3. N.J. Bingham & H. Byrne, "Archival Strategies for Contemporary Collecting in a World of Big Data: Challenges and Opportunities with Curating the UK Web Archive," *Big Data & Society* (2021), https://doi.org/10.1177/2053951721990409.

4. S. Ames & S. Lewis, "Disrupting the Library: Digital Scholarship and Big Data at the National Library of Scotland," *Big Data & Society* (2020), https://doi.org/10.1177/2053951720970576.

5. For example, T. Underwood, *Distant Horizons: Digital Evidence and Literary Change* (Chicago: The University of Chicago Press, 2019).

6. For the notion of "socio-technical imaginaries," see S. Jasanoff & S. Kim, "Containing the Atom: Socio-technical Imaginaries and Nuclear Power in the United States and South Korea," *Minerva* 47:2 (2009): 119-146, http://www.jstor.org/stable/41821489. For "collections as data," see, for example, T. Padilla, "Humanities Data in the Library: Integrity, Form, Access," *D-Lib Magazine* (2016), https://www.dlib.org/dlib/march16/padilla/03padilla.html [accessed March 8, 2023].

7. A. Wheatley & S. Hervieux, "Artificial Intelligence in Academic Libraries: An Environmental Scan," *Information Services & Use* 39 (2019): 348, https://doi.org/10.3233/ISU-190065.

8. R. Cordell, "Machine Learning + Libraries: A Report on the State of the Field," LC Labs, Library of Congress (14 July 2020), https://labs.loc.gov/static/labs/work/reports/Cordell-LOC-ML-report.pdf [accessed March 8, 2023]; T. Padilla, *Responsible Operations: Data Science, Machine Learning, and AI in Libraries* (Dublin, OH: OCLC Research, 2019), https://doi.org/10.25333/xk7z-9g97 [accessed March 8, 2023]; T. Padilla, L. Allen, H. Frost, S. Potvin, E. Russey Roke & S. Varner, "Final Report - Always Already Computational: Collections as Data," *Zenodo* (2019), https://doi.org/10.5281/zenodo.3152935.

9. A.M. Cox, S. Pinfield & S. Rutter, "The Intelligent Library: Thought Leaders' Views on the Likely Impact of Artificial Intelligence on Academic Libraries," *Library Hi Tech* 37:3 (2019): 418-435, https://doi.org/10.1108/LHT-08-2018-0105; D. Harisanty, N.E.V. Anna, T.E. Putri, A.A. Firdaus & N.A. Noor Azizi, "Leaders, Practitioners

and Scientists' Awareness of Artificial Intelligence in Libraries: A Pilot Study," *Library Hi Tech* pre-print (2022), https://doi.org/10.1108/LHT-10-2021-0356.

10. Ames & Lewis, "Disrupting the Library"; C. Haffenden, E. Fano, M. Malmsten & L. Börjeson, "Making and Using AI in the Library: Creating a BERT Model at the National Library of Sweden," *College & Research Libraries* 84:1 (2023): 30-48, https://doi.org/10.5860/crl.84.1.30.

11. The acronym GLAM stands for Galleries, Libraries, Archives and Museums. M. Mahey, A. Al-Abdulla, S. Ames, P. Bray, G. Candela, S. Chambers, C. Derven, M. Dobreva-McPherson, K. Gasser, S. Karner, K. Kokegei, D. Laursen, A. Potter, A. Straube, S-C. Wagner & L. Wilms with forewords by: T. A. Al-Emadi, J. Broady-Preston, P. Landry & G. Papaioannou, *Open a GLAM Lab. Digital Cultural Heritage Innovation Labs*, (23-27 September, 2019), https://glam-labs.s3.amazonaws.com/media/dd/documents/Open_a_GLAM_Lab-10-screen.9c4c9c7.pdf [accessed March 8, 2023].

12. Haffenden, et al., "Making and Using AI in the Library," 45.

13. The text of this legal act is available (in Swedish) here: "Förordning (2008:1421) med instruktion för Kungl. Biblioteket" (2008), https://www.riksdagen.se/sv/dokument-lagar/dokument/svensk-forfattningssamling/forordning-20081421-med-instruktion-for-kungl_sfs-2008-1421 [accessed March 8, 2023].

14. G. Konstenius, "Plikten under lupp! En studie av pliktlagstiftningens roll, utformning och relevans i förhållande till medielandskapets utveckling" [Eng. Legal Deposit in Focus! A Study of the Role, Design and Relevance of Legal Deposit Legislation in Relation to the Development of the Media Landscape], *Kungl. Biblioteket* (2017),
https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-539 [accessed March 8, 2023].

15. See Padilla, *Responsible Operations* and Padilla et al., "Final Report".

16. M.C. Traub, J. van Ossenbruggen & L. Hardman, "Impact Analysis of OCR Quality on Research Tasks in Digital Archives" in *Research and Advanced Technology for Digital Libraries*, ed. S. Kapidakis, C. Mazurek, C., and M. Werla (Springer, Cham, 2015),
https://doi.org/10.1007/978-3-319-24592-8_19 [accessed March 8, 2023].

17. See F. Rekathati, "A Multimodal Approach to Advertisement Classification in Digitized Newspapers," *The KBLab Blog* (2021), https://kb-labb.github.io/posts/2021-03-28-ad-classification/ [accessed March 8, 2023].

18. Cf. L. Gitelman (ed), *"Raw Data" is an Oxymoron* (Cambridge, Mass.: The MIT Press, 2013), https://doi.org/10.7551/mitpress/9302.001.0001.

19. P. Snickars, "Datalabb på KB: En förstudie" [Eng. Data Lab at KB: A Pre-study], *Kungl. Biblioteket* (2018), 29, https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-339 [accessed March 8, 2023].

20. For more details, see: https://www.westac.se/en/.

21. For more on which, see: https://liu.se/en/research/computational-text-analysis.

22. Cf. J. Law, *After Method: Mess in Social Science Research* (London: Routledge, 2004).

23. A description of the lab's API is available here: M. Malmsten, "KB Data Lab," *GitHub* (2020), https://github.com/Kungbib/kblab [accessed March 8, 2023].

24. For KB's longer engagement with linked data, see, for instance, M. Malmsten, "Exposing Library Data as Linked Data," *IFLA Satellite Preconference Sponsored by the Information Technology Section* (2009),
https://citeseerx.ist.psu.edu/document?repid=rep1&type=pdf&doi=2e0791d88a65cb2517e284c2bfca02b7c6660f30 [accessed March 8, 2023].

25. For further details about demand for KBLab among researchers, see M. Fridlund, "Utvärdering av KB-labb" [Eng. Evaluation of KBLab], Gothenburg University, Centre for Digital Humanities (September 2021), 9, https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-97 [accessed March 8, 2023].

26. For more details about this process, see: https://kb.se/in-english/research-collaboration/criteria-for-collaboration.html.

27. For further discussion of funding a GLAM lab, see Mahey et al., *Open a GLAM Lab*, 29.

28. Examples of these projects include F. Rekathati, "Curating News Sections in a Historical Swedish News Corpus," Independent Master's Thesis, Linköping University, Department of Computer and Information Science (2020), http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-166313 and G. Henning, "News Article Segmentation Using Multimodal Input: Using Mask R-Cnn and Sentence Transformers," Independent Master's Thesis, KTH, School of Electrical Engineering and Computer Science (2022), http://urn.kb.se/resolve?urn=urn:nbn:se:kth:diva-309527.

29. Further consideration of these issues can be found in M. Kemman, *Trading Zones of Digital History* (Berlin: De Gruyter Oldenbourg, 2021), https://doi.org/10.1515/9783110682106 and E. Fano & C. Haffenden, "Digital humaniora eller humanistisk datavetenskap?" [Eng. Digital Humanities or Humanistic Computer Science?] *Samlingsbloggen* (2022), https://www.kb.se/hitta-och-bestall/samlingsbloggen/blogginlagg/2022-04-14-digital-humaniora-eller-humanistisk-datavetenskap.html [accessed March 8, 2023].

30. J. Devlin, M.-W. Chang, K. Lee & K. Toutanova, "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding," *arXiv* (2019), https://arxiv.org/abs/1810.04805v2.

31. For instance, A. Virtanen, J. Kanerva, R. Ilo, J. Luoma, J. Luotolahti, T. Salakoski, F. Ginter & S. Pyysalo, "Multilingual Is Not Enough: BERT for Finnish," *arXiv* (2019), https://arxiv.org/abs/1912.07076; L. Martin, B. Muller, P.J.O. Súarez, Y. Dupont, L. Romary, E.V. de la Clergerie, D. Seddah & B. Sagot, "CamemBERT: a Tasty French Language Model," *arXiv* (2020), https://arxiv.org/abs/1911.03894; and P.E. Kummervold, J. De la Rosa, F. Wetjen & S.A. Brygfjeld, "Operationalizing a National Digital Library: The Case for a Norwegian Transformer Model," *arXiv* (2021), https://arxiv.org/abs/2104.09617.

32. Haffenden, et al., "Making and Using AI in the Library," 35-36.

33. Cf. A. Radford, K. Narasimhan, T. Salimans & I. Sutskever, "Improving Language Understanding by Generative Pre-Training," [pre-print] (2018), https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf [accessed March 8, 2023].

34. Haffenden, et al., "Making and Using AI in the Library," 35.

35. M. Malmsten, L. Börjeson & C. Haffenden, "Playing with Words at the National Library of Sweden: Making a Swedish BERT," *arXiv* (2020), https://arxiv.org/abs/2007.01658.

36. This model is available here: https://huggingface.co/KBLab/bert-base-swedish-cased.

37. M. Malmsten, C. Haffenden & L. Börjeson, "Hearing Voices at the National Library: A Speech Corpus and Acoustic Model for the Swedish Language," *arXiv* (2022), https://arxiv.org/abs/2007.01658.

38. A. Baevski, H. Zhou, A. Mohamed & M. Auli, "Wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations," *arXiv* (2020), https://arxiv.org/abs/2006.11477.

39. Malmsten et al., "Hearing Voices at the National Library." This model is available here: https://huggingface.co/KBLab/wav2vec2-large-voxrex-swedish.

40. This has been the policy for distribution for all of the models released so far, and will continue to remain so for predictive models. The situation with generative models is more complex, since there is a risk for such tools to be used in ways that counter the democratic goals of the library's mission—i.e. in creating misinformation and fake news. On this basis, a more restrictive distribution policy will most likely be pursued in relation to generative models, should they be produced in the future.

41. See organization page here: https://huggingface.co/KBLab.

42. F. Carlsson, P. Eisen, F. Rekathati & M. Sahlgren, "Cross-lingual and Multilingual CLIP," *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (2022), https://aclanthology.org/2022.lrec-1.739/ [accessed March 8, 2023].

43. See, for instance, F. Rekathati, "Swedish Sentence Transformer 2.0," *The KBLab Blog* (2023), https://kb-labb.github.io/posts/2023-01-16-sentence-transformer-20/ [accessed March 8, 2023].

44. See: https://github.com/kb-labb.

45. Fridlund, "Utvärdering av KB-labb," 14.

46. For instance, A. Ekmark, "Text Block Prediction and Article Reconstruction Using BERT," Independent Master's Thesis, Uppsala University, Department of Statistics (2021), http://urn.kb.se/resolve?urn=urn:nbn:se:uu:diva-447248.

47. Cf. E. Fano & C. Haffenden, "BERTopic for Swedish: Topic Modeling Made Easier via KB-BERT," *The KBLab Blog* (2022), https://kb-labb.github.io/posts/2022-06-14-bertopic/ [accessed March 8, 2023].

48. C. Dwibedi, E. Mellergård, A.C. Gyllensten, K. Nilsson, A.S. Axelsson, M. Bäckman, M. Sahlgren, S.H. Friend, S. Persson, S. Franzén, B. Abrahamsson, K. Steen Carlsson & A.H. Rosengren, "Effect of Self-Managed Lifestyle Treatment on Glycemic Control in Patients With Type 2 Diabetes," *NPJ Digital Medicine* 5:60 (2022), https://doi.org/10.1038/s41746-022-00606-9; O. Jerdhaf, M. Santini, M. Lundberg & A. Karlsson, "Implant Terms: Focused Terminology Extraction with Swedish BERT - Preliminary Results," *Eighth Swedish Language Technology Conference* (2020), http://urn.kb.se/resolve?urn=urn:nbn:se:ri:diva-52378; A. Avram, V. Pais & D Tufis, "PyEuroVoc: A Tool for Multilingual Legal Document Classification with EuroVoc Descriptors," *arXiv* (2021), https://arxiv.org/abs/2108.01139.

49. For the work of the Swedish Tax Agency and the Swedish Courts in using lab models, see M. Juhlin, "De samhällsekonomiska effekterna kopplat till Kungliga bibliotekets AI-baserade språkmodeller" [Eng. The Socioeconomic Effects of the National Library of Sweden's AI Language Models], *Policy Impact* report (2022), 36-41, https://urn.kb.se/resolve?urn=urn:nbn:se:kb:publ-692; for KB-BERT being used as the basis of a new search application for precedent in state bureaucracy, see: https://www.statenssc.se/nyheter/nyhetsarkiv/2023-03-14-ai-baserad-soktjanst-ska-underlatta-remisshanteringen-i-staten [accessed March 8, 2023].

50. Fridlund, "Utvärdering av KB-labb," 14.
51. These download statistics are also available here: https://github.com/kb-labb/huggingface_stats [accessed March 8, 2023].
52. Cf. Juhlin, "De samhällsekonomiska effekterna," 20-22.
53. Fridlund, "Utvärdering av KB-labb," 14; Juhlin, "De samhällsekonomiska effekterna".
54. "Förordning (2008:1421) med instruktion för Kungl. Biblioteket".
55. Cf. P.N. Edwards, G.C. Bowker, S.J. Jackson & R. Williams, "Introduction: An Agenda for Infrastructure Studies," *Journal of the Association for Information Systems* 10:5 (2009), https://aisel.aisnet.org/jais/vol10/iss5/6 [accessed March 8, 2023].
56. Haffenden, et al., "Making and Using AI in the Library," 33.
57. Fridlund, "Utvärdering av KB-labb," 13.
58. Cf. D. Harvey, "The Future of the Commons," *Radical History Review* 109 (2009): 101-107, https://doi.org/10.1215/01636545-2010-017.
59. K. Sanderson, "GPT-4 Is Here: What Scientists Think," *Nature* March 16 (2023), https://doi.org/10.1038/d41586-023-00816-5.
60. E.M. Bender, T. Gebru, A. McMillan-Major & S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" *FAccT '21: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency* (2021),
https://doi.org/10.1145/3442188.3445922; T. Gebru, "For Truly Ethical AI, Its Research Must Be Independent From Big Tech" *The Guardian*, December 6 (2021),
https://www.theguardian.com/commentisfree/2021/dec/06/google-silicon-valley-ai-timnit-gebru [accessed March 8, 2023].
61. Fridlund, "Utvärdering av KB-labb," 14.
62. Gebru, "For Truly Ethical AI".
63. For further details about the granting of such access to EuroHPC, see: https://www.eurocc-access.eu/success-stories/success-story-national-library-of-sweden-has-now-access-to-vega/ [accessed March 8, 2023] and https://enccs.se/news/2022/10/national-library-of-sweden-accesses-meluxina/ [accessed March 8, 2023].
64. For instance, R. Kurtz & J. Öhman, "SUCX 3.0," *The KBLab Blog* (2022), https://kb-labb.github.io/posts/2022-02-07-sucx3_ner/ [accessed March 8, 2023].
65. R. Kurtz, "Evaluating Swedish Language Models," *The KBLab Blog* (2022), https://kb-labb.github.io/posts/2022-03-16-evaluating-swedish-language-models/ [accessed March 8, 2023].
See also: https://www.ai.se/en/node/81535/superlim [accessed March 8, 2023].
66. R. Knowles, B.A. Mateen & Y. Yehudi, "We Need to Talk About the Lack of Investment in Digital Research Infrastructure," *Nature Computational Science* 1 (2021): 169–171,
https://doi.org/10.1038/s43588-021-00048-5
67. Haffenden, et al., "Making and Using AI in the Library," 45.