Datalabb på KB

En förstudie

Pelle Snickars

KUNGL. BIBLIOTEKET
National Library of Sweden

Datalabb på KB

For more information:
National Library of Sweden, Executive support
www.kb.se
+46-10-709 33 69
Date:2018-10-19
Dnr: 1.2.1-2017-752
ISSN:

Datalabb på KB

En förstudie

Pelle Snickars

Contents

Förord

”All information med hög efterfrågan finns digitalt tillgänglig. Det mesta går att få tag på snabbt och enkelt, oberoende av tid och plats. Det svenska kulturarvet är sammanlänkat med det globala kulturarvet. Många bibliotek och andra aktörer samverkar internationellt och har tillsammans ökat tillgången till material. Alla samlingar är beskrivna och sammanlänkade.”

Så heter det Kungliga bibliotekets vision för perioden fram till år 2025. Visionen tydliggör vad nationalbiblioteket vill åstadkomma när vi möter de enorma förändringar och utmaningar som biblioteksväsendet och hela forskningssamhället står inför.

Ett av KB:s viktigaste uppdrag är att främja den svenska forskningens kvalitet genom att tillhandahålla en effektiv forskningsinfrastruktur. Forskningen efterfrågar idag allt oftare tillgång till stora datamängder, snarare än enskilda digitala objekt. Det innebär att vi bevittnar ett paradigmskifte i synen vad den digitala utmaningen egentligen består i.

De helt nya metoder som växer fram när forskningen får möjlighet att söka i och analysera stora datamängder håller på att skapa en forskningsdriven datarevolution inom en rad olika vetenskapsområden, inte minst inom humaniora och samhällsvetenskap. Ökad tillgång till stora datavolymer erbjuder också möjlighet att besvara nya frågeställningar. Datadriven forskning och digital humaniora har redan blivit ett par centrala begrepp.

Förutom forskningsvärdet ger det digitala formatet förutsättningar att utveckla nya tjänster, till nytta både för enskilda, samhälle och företag. Här finns ett starkt och positivt samband mellan informationsdigitalisering och samhällelig digitalisering.

Men för att samhället ska kunna ta vara på denna revolutions fulla kraft behövs långsiktigt ökade resurser för skolskalig digitalisering av kulturarvsinstitutionernas samlingar så att kulturarvet systematiskt kan lyftas över i den digitala domänen och omvandlas till forskningsdata.

Kungliga biblioteket och övriga kulturarvsinstitutioner måste också bli bättre på att göra det digitaliserade och digitalt födda materialet tillgängligt i ett format som forskningen efterfrågar. Annars kommer forskningspotentialen till stor del att förbli outnyttjad.

Förväntningarna är redan höga på kulturarvsinstitutionernas kapacitet att tillhandahålla sina samlingar som stora sammanlänkade datamängder. För att möta dessa förväntningar och bidra till en mer effektiv forskningsinfrastruktur har KB beslutat att genomföra en förstudie inför ett framtida inrättande av ett digitalt datalaboratorium på nationalbiblioteket.

Förstudien har genomförts på uppdrag av Kungliga biblioteket av Pelle Snickars och har fått namnet ”Datalabb på KB”. Den kommer att ligga tillgrund för ett inriktningsbeslut senare under hösten 2018.

Intresset är stort för digital humaniora såväl nationellt som internationellt. Förstudien är därför på engelska. Studien har finansierats med stöd av Riksbankens jubileumsfond.

Stockholm den 19 oktober 2018,

Lars Ilshammar

Biträdande riksbibliotekarie

Foreword

All information in great demand is available digitally. Most material is easily and readily accessible, independent of time and place. Swedish cultural heritage is closely interlinked with the global cultural heritage. Many libraries and other stakeholders cooperate internationally, and collectively increase the availability of research material. All collections are catalogued and interlinked."

The above quotation is taken from the vision for the National Library of Sweden for 2025, a vision which is intended to clarify the goals of the National Library as regards the enormous changes and challenges facing the field of library and information science, and, indeed, the whole of academia.

One of the the National Library of Sweden's most important missions is to further the quality of Swedish research by offering an efficient research infrastructure. Contemporary research increasingly requires large data sets rather than single digital artefacts, and this creates a paradigm shift in the view what these digital challenges actually are.

New methodologies are being developed where research gains the opportunity to search and analyse large amounts of data, in turn creating a digital revolution driven by research in a number of different disciplines, not least within the humanities and social sciences. Increased access to big data also offers the possibility to answer new hypotheses. Digital scholarship and digital humanities have already become central concepts in this new digital paradigm.

Besides the value for research per se, the digital format offers possibilities for the development of new services for the benefit of individual citizens and society as a whole, not to mention global business. There is a strong and positive correlation between the digitisation of information and the digitisation of society.

However, in order for society to make use of the full force of this revolution, increased long-term resources are needed for a full-scale digitisation of the collections of cultural heritage institutions, such as archives, libraries and museums (ALM). This will enable the nation's entire cultural heritage to be transferred into the digital domain, transforming it into research data.

The National Library of Sweden and other ALM organisations need to increase their skills in making digitised and digitally originated material accessible in a format useful to researchers and scholars. Otherwise the inherent potential of the material will, to a large extent, remain unrealised.

Expectations on ALM organisations’ capacity to digitise their collections, and make them accessible as linked open data, are already enormous. In order to fulfil these expectations, and contribute to a more efficient research infrastructure, the National Library of Sweden has decided to complete a pilot study pending a future launch of a digital data lab there.

The pilot study, entitled "Datalab at KB", was commissioned by the National Library of Sweden and conducted by Pelle Snickars. The study will form the basis for a policy decision expected to be taken during the latter half of 2018.

The study has been financed with support from the Swedish Foundation for Humanities and Social Sciences (Riksbankens Jubileumsfond).

Stockholm, 19 October 2018

Lars Ilshammar

Deputy National Librarian

Summary

During the last decade digital scholarship has become increasingly prominent within the academy. This type of research is conducted by scholars in totally different academic disciplines and varies in scope—but digital scholarship also shares some features. For example, it is often embedded in *digital methods* that depart from digital evidence in the form of data or *datasets*. Within the ALM-sector such digital scholarship has expanded the focus of digitisation activities towards different forms of explorations. Especially within digital humanities scholarship the systematic intertwining of research questions, digital materials, and tools have stressed the need to reformulate what an apt library and research infrastructure for the humanities (and social sciences) should pertain. The setup of library labs has been one answer.

The primary function of library labs are to deliver digital collections as data (or datasets) to researchers and other interested users, as well as to enable research or tinkering with the same data. Library labs are sometimes devoted to experimentation with provided datasets, but they can also be envisioned as a *core service* that more and more national libraries provide—with the lab (and its services) becoming an integrated part of a developed digital infrastructure. The foundation of a library lab at the National Library of Sweden would hence strengthened and support a commitment to cutting-edge data driven research, and also expand the library research infrastructure substantially.

This report suggest that the National Library should launch a library lab. The main purpose of the library lab—datalab.kb.se—is to support all forms of research on digitised heritage. My advice is that the lab receives five main objectives: (1.) to support digital scholarship through novel applications and methods; (2.) to support digital scholarship by curating, assembling or aggregating datasets; (3.) to support and actively participate in the co-development of research applications with the lab as a dynamic partner; (4.) to support and supplement the National Library's digital development in general by seeing the lab as an internal innovation hub; (5.) to support and enhance digital knowledge of staff working at the National Library.

Furthermore, I suggest that datalab.kb.se should be staffed with a minimum of three persons: a (part time) library lab manager, a data curator (or data librarian) and a developer. Moreover, I suggest that datalab.kb.se should run as a pilot project spanning two plus two years—with a major assessment between the two phases. The principal workload during the setup phase will be to establish a robust web presence at datalab.kb.se (by enhancing functionalities and information at data.kb.se), as well as creating and curating new datasets. Importantly, the lab should not envision itself as a local undertaking, but rather as *national data infrastructure* catering to many different forms of digital scholarship.

Introduction

The Annual Report from the British Library usually offers insights into the many domains and whereabouts of national libraries—not the least in terms of future directions. In the latest report (from 2017/18) it is, for example, stated that the British Library Digital Scholarship team continues to "undertake innovative research with digital collections and open up new datasets for use by researchers." One way to facilitate such digital scholarship is to start a lab, and the British Library set up a library lab environment already in 2013. Ever since the British Library Labs has been inviting researchers, developers and artists "from around the world" to engage in "creative endeavours" using the library's digitally curated collections, content and data. Following the latest Annual Report, the Library Labs team has now "facilitated the use of over 180 terabytes of data including 97 freely available datasets at data.bl.uk."

Digital scholarship, *curated data*, *single datasets*, invited *developers* and *programmers*—these are all present buzzwords and novel categories within the library domain. Previously, computational expertise were necessary and primarily required for internal workflow within IT departments—now such skills and competencies are increasingly turning into a prerequisite for doing actual research in a transforming library infrastructure (increasingly turning digital). This infrastructural and scholarly transformation might appear as swift and sudden. Yet, digitisation activities within the ALM-sector (archives, libraries and museums) has been a harbinger of novel times to come—both in terms of scholarly perspectives and library practices.

National libraries have been digitising their collections for decades—in Sweden digitisation work started already in the late 1990s. For a number of years, collections were digitised primarily for preservational purposes, but after the millenium—due to the rise of the Web and initiatives as Google Books—digital access to library collections steadily became more important. Permission to use library collections was, however, often hindered by copyright legislation, and digital access was hence foremost given to older (textual) collections prior to the 20th century.

Digitisation work performed at the National Library of Sweden has in general been similar to other European countries. The library has digitised a major amount of its audiovisual collections, various selected works from the print collections, and a large amount of newspapers. The latter has been a prioritised category since newspapers are an important research material for many users. Born digital collections have also grown through web archiving activities (Kulturarw3) and audiovisual deposits, and even more so since 2015 when (some) electronic materials became subject to legal deposit. Regarding the digital trajectory that the National Library has undertaken during the last 15 years, preservation was most important at first, then digitisation for access was increasingly advocated. There are, however, also good reasons to question the distinction between digitising for access and digitising for preservation. Some scholars have even argued that the split "is artificial and misleading" since access to collections are usually "a given" and an outcome of *all* digital transformation—even if usage is

fully realised only through functioning electronic networks and the legal frameworks that manage permissions.

Nevertheless, during recent years digital scholarship within the ALM-sector has expanded the focus of digitisation activities towards different forms of investigations and explorations. In short, one can observe a scholarly driven progression within the institutional heritage domain from *preservation* via *access*—to *analyses*. Today all forms of digital heritage are computational—hence, how to enhance and increase the *research potential* of this material? If humanities and social science scholars traditionally were interested in the collections that archives and libraries had to offer deep down in their stacks and vaults, such archival driven humanities research has increasingly turned into *data driven research* due to the digitisation of heritage. And more data is better data (as Google would have it).

Today, governmental decrees for national libraries (and similar statutes for university libraries) usually stipulate that libraries are to provide a beneficial infrastructure for research. During centuries great book and manuscript collections at university libraries and national libraries played a pivotal role for the humanities and social sciences. They were envisioned as key infrastructures for scholarship. National libraries and deposit laws are, in fact, illustrative examples of how traditional knowledge structures were enacted through concrete and primarily humanistic infrastructures. They have essentially remained the same over centuries, but have during the last decade—due to repeated digitisation efforts—begun to alter. The long-term magnitude of this ongoing transformation is staggering—both for scholars and libraries. Within the library sector the gradual alteration effects the very foundation and principles of what libraries are—and should be at a time when 'the digital' is becoming default.

If the applications of digitisation within the ALM-sector initially had a preservation focus, novel ways of giving access and sustaining digital scholarship represent the other side of the same digital development. In short, mass digitisation combined with new media, technology and distribution networks has transformed the possibilities for libraries and their users. Emerging scholarly disciplines—from data science to the digital humanities—all take advantage of new computing tools and infrastructure, and provide different models for creating new forms of access to and analyses of library collections. Especially within digital humanities scholarship the systematic intertwining of research questions, digital materials, and tools have stressed the need to reformulate what an apt library and research infrastructure for the humanities (and social sciences) should pertain. Digitisation has thus in essence begun to transform the epistemic foundation of the library. The knowledge that can be deduced from collections in digital form is different—and foremost one of scale. So called *distant reading* of major textual corpora have even been envisioned as a new "condition of knowledge".

About the Report

About a year ago I was asked if I had an interest to examine, survey and evaluate in what ways a lab might—or could be—established at the National Library of Sweden. As a media studies professor at Umeå University, I have for a number of years worked and done research at the digital humanities center Humlab. I accepted the offer and applied for the position—a PM for a "pilot study" on a data lab at the National Library was drafted by library personnel Lars Björk and Peter Krantz, and additional funding was made available by Riksbankens jubileumsfond.

From January 2018 I have been working (part time) during nine months with this report and ways to prepare the ground for making my recommendations a reality. Lars Björk (at the National Library) has functioned as my co-worker. During winter, spring and summer 2018 we visited a number of scholarly environments, university libraries and research groups in Sweden with an interest in using a library lab. We have talked to many Swedish scholars and librarians with an interest in the matter; we established both a reference group and a steering committee for our work; we made a study trip to the British Library Labs and the Dutch KB Lab; we sent out a survey (with the help of Cecilia Ranemo, at the National Library) regarding available digital collections within the ALM-sector in Sweden (Appendix A), and we presented and discussed our work within the "Group for digitisation and digital access"—with me as chair and Björk as secretary—a group that is part of the "Forum for national library collaboration and development". I have also made a number of presentations of library lab ideas at Swedish universities, at the management board of the Swedish National Archives—who explicitly supported and endorsed the establishment of a library lab—at the Research board of the National Library, and at national and international conferences. Furthermore I organised a workshop on digital scholarship at the National Library (in April 2018) with some 25 scholars and librarians (funded by Riksbankens jubileumsfond). Our preparatory work, conversations and scholarly visits have thus been thorough.

This report is entitled *datalab.kb.se*—it is a term Björk and I suggest for naming an eventual datalab at KB, where the digital and Swedish connotation are obvious (including a necessary distinction and contrast to the Danish and Dutch KB Lab). The report is divided into three subsequent sections—"Library Labs" and "Digital Scholarship" (with subsections)—as well as a final part on "Recommendations". The first part sketches and maps the international terrain of current library labs, with a focus on different lab environments at national libraries. The second section puts novel forms of computational scholarship at the center of attention, with a particular emphasis on methods and (necessary) curation of datasets. In the final section on recommendations I suggest how a lab at the National Library could be organised, focusing both on actual tasks and workflow, as well as short job descriptions and required skill sets.

Library Labs

Digitally inclined research within the humanities and social sciences have during the last decade started to influence both national and university libraries to take advantage of the scholarly possibilities that arise when *documents as data* are sharable and networked, linkable and traceable, reusable and processable. The development and set up of library labs is one result of previous digitisation activities. The primary function of library labs are to deliver digital collections as data (or datasets) to researchers and other interested users, as well as to enable research or tinkering with the same data. Following the literal meaning of the term laboratory—"a room or building equipped for scientific experiments"—library labs are usually devoted to experimentation with provided datasets. "British Library Labs – experiment with our collections", as the slogan goes. Library labs can hence be envisioned as a scholarly, artistic or creative industries playground. The British Library Labs is, for example, an endeavor that supports and "inspires the public use of the British Library's digital collections and data in exciting and innovative ways." In a similar manner the Dutch KB Lab wants to be experimental; "we try out new techniques and tinker with tools to make our content as accessible as we can. Warning, that means stuff can be broken."

However, since library labs are also becoming more and more common, the focus on experimentation can become misleading. Providing datasets and working with these in different ways is nowadays hardly rocket science. The products from previous and ongoing digitisation activities simply allow—and to some extent makes it easy—for scholars to work with large scale datasets. Hence library labs can consequently be perceived as a *core service* that national libraries provide, with the lab (or its services) becoming an integrated part of a developed digital infrastructure.

Such perspectives were advocated at a recent conference at the British Library, *Building Library Labs* in mid September 2018. It brought some 40 libraries and partner institutions from North America, Europe, Asia and Africa—with no less than ten national libraries present. "Around the world, leading national, state, university and public libraries are creating 'digital lab type environments'", the conference program stated. The aim is often to develop novel forms of library usage, where library labs ensure that "digitised and born digital collections/data can be opened up and reused for creative, innovative and inspiring projects by everyone such as digital researchers, artists, entrepreneurs and educators."

The issue of library labs is hence timely. Presentations and discussions in London evolved around issues such as labs services and spaces, technical infrastructures, the values of a library lab, planning a lab and establishing it, as well as various funding models for labs. Usage, research and presentations of ongoing projects were also on the agenda. One result of the conference was a supportive network, another a forthcoming global report on library labs. Most libraries and institutions present did also participate in a library lab survey. The results are in no way conclusive, but give a tentative impression of how major libraries presently deal with lab issues (Appendix B). One thing to note from the survey is that library labs started to emerge between 2013 and 2015. Interestingly, this first wave of initiatives is now reinforced by a more general

‘lab trend’ (which this report is also part of). Following the survey some 20 libraries are about to launch a lab in 2019 or 2020.

According to the survey, most existing library labs today are aimed to serve academic research followed by internal staff, the general public or creative industries. The most common tasks are “facilitating access to data & digital collections at scale” and “creating new datasets & digital collections”, succeeded by “providing training in digital methods & tools” and “public engagement”. Half of the library labs (following the survey) provided access to restricted digital collections (through various forms of contracts and agreements), and (only) half of them offered a physical space in the library—thus for many library labs focus is mostly put on web based presence.

The conference *Building Library Labs* attest to the considerable international interest in library lab issues at the moment—and the prime reason why this report is written in English. Even if library labs are usually established with the purpose to enhance and amplify usage of digitised (or born digital) collections and datasets, they differ in approach, scope and orientation. Therefore, a brief description of some different types of library labs—with a focus on national library labs—can serve as a smorgasbord of how labs can be designed and organised, accustomed and staffed. I have not been able to find out in detail how labs are funded (including the size of budgets), but the amount of personnel gives a rough estimation of the dimension and proportions of library lab activities.

British Library Labs

The lab at the British Library was founded in 2013, with a major five year grant from the Andrew Mellon foundation. It is staffed with a manager, a “research software engineer” and a data curator—but the lab is also part of the department of Digital scholarship, hence resources can proliferate. The BL Labs hosts and supports a wide range of events, including workshops, seminars and presentations, which explore the uses of the British Library’s “digital collections and provide networking opportunities”. The work is also carried out by encouraging “researchers, developers, educators and artists to talk to us about collaborating on projects using the British Library’s digital content”. In order to raise awareness of its collections, BL Labs has done a number of “road shows” to universities in Britain, and arranged competitions and awards. The road shows includes both presentations and “practical hands-on workshops”, and a chance to explore and discuss what “you may do with some of the Library's data and for you to speak to and get feedback from experts”. Winners of “BL Labs Competition” become “researchers or artists in residence” at the library for a couple of months, where they are provided support to incubate and enhance ideas. Importantly, the BL Labs offers curatorial knowledge of digital collections, often in relation to physical collections and representativeness of digitised material. BL labs is also active within the internal “Digital Scholarship Training Programme” which creates opportunities for staff to

develop necessary skills and knowledge to support emerging areas of modern digital scholarship.

Dutch KB Lab

The KB Lab was founded in 2014 within the Research department of the National Library of the Netherlands. It is presently staffed with a manager/co-ordinator, two data curators and two research software engineers. The general purpose of the lab is to "showcase the tools that are built for and by users of our digital collections", and the wish of the lab "is to offer better access to our digital collections and promote the use of digitised content." On the one hand, KB Lab wants to develop novel scholarly ways to analyse digital and digitised collections at the National Library of the Netherlands—particularly the more than 10 million historical text pages of newspapers, books, journals and radio bulletins that are accessible on the so called Delpher platform. On the other hand, the aim is to supplement the library with innovative ways of, for example, searching or visualising the collections. Internal workflow and improvements are thus enhanced by activities performed within the lab environment—KB Lab is, in short, internally envisioned as an *innovation hub* for library development. Services and programmes at KB Lab are primarily geared towards facilitating digital humanities research through digital collections. Datasets are provided to external researchers via a general contract, an "Agreement and conditions governing the use of KB datasets" (Appendix C). Interestingly, KB Lab also offers access to *derived datasets*—that is, datasets that are the result of research activities, based on KB collections. In addition, the KB Lab also funds a "Researcher-in-residence programme" that invites early career digital humanities researchers to develop innovative research methods in close collaboration with KB data experts.

Library of Congress Labs

The American Library of Congress Labs (LC Labs) was founded in 2017, and is presently staffed with six "innovation specialists". The aim of LC Labs is to enable "transformational experiences" by connecting users with the library and its digital collections, "prototype ideas and build relationships with stakeholders" that might realise some parts of the library's digital strategy, and "strengthen our community by sharing our work for transparency, feedback and knowledge exchange." LC Labs is thus envisioned as a place to encourage innovation with digital collections, and not primarily geared towards digital scholarship. Rather the focus is on *public engagement*, innovation and creative industries. A current crowdsourcing pilot is, for example, about identifying "illustrations and provide captions in WWI-era newspapers". The lab also offers a position as "innovator-in-residence"; hence in contrast to both the British and Dutch labs, LC Labs has a more public profile. Online, LC Labs do offer APIs and "bulk downloads"—including "LC for Robots" in the form of "machine-readable access

to its digital collections"—but the focus is put foremost on "experiments" with "tools, art, applications, and visualizations we've made with our collections."

Europeana Labs

Europeana Labs started in 2014 and is staffed with one manager and additional personnel within the Europeana consortium in the Hague. Europeana Labs was "relaunched" in 2016 with an "improved product" and new visual design, including better API features and more re-usable datasets. Like the LC Labs, Europeana Labs tries to target a wider audience and the creative industries at large: "Europeana Labs is the go-to place for those who have the imagination, skills and desire to play with digital cultural content and use it in their experimental works or sustainable business projects." Europeana Labs runs hackathons and challenges to encourage large scale usage of Europeana's rich collections of digital heritage, and also offers a number of APIs and more than hundred different datasets. The latter represent "over a million" of Europena's directly accessible objects, "newspapers, books, photos, art, artefacts, audio clips". Most datasets are openly licensed and free to use "in any way you like." Europeana Labs also provides more than 150 examples of cultural heritage apps and tools, separated into three categories: "Showcase" examples of apps and games that use the Europeana API to make use of the collections, "Tools" that offer specific tools for working directly with the Europeana API, and "FLOSS"—open source tools relevant for digital cultural heritage developers.

Danish KB Lab

The Danish KB Lab was founded in 2016 as a subsection of the IT department at the Royal Danish Library. At present, the Danish KB Lab is thus more of an internal workforce. It seeks, however, to find new ways to "combine the library's digital cultural heritage collections and research, with the latest methods within machine learning." Since the lab is a part of an IT department, the focus lies within data science in general. Online, however, are also a number of experiments with the collections made that "visualize, engage or showcase the different materials or collections that we have available, to inspire and deepen the knowledge of what collections we actually have". These applications are considered "experimental projects", but the KB Lab also co-operates with a number of similar initiatives in Denmark, for example the KUB Datalabs at Copenhagen University Library, the DIGHUMLAB and the NetLab.

Austrian National Library (ÖNB-LAB)

The Austrian National Library lab promotes itself as a different and alternative type of library environment. If the Austrian National Library is usually perceived as a traditional (and somewhat old-fashioned) environment, ÖNB-LAB actively seeks to advertise itself as a computational counterpart—for example by using a modified

version of the library logotype and providing all information in English and German. ÖNB-LAB is staffed with a manager and one developer, and will foremost devote its activities to a web page with datasets and tools, including code and tutorials provided through Gitlab. In order to provide access to datasets (also restricted ones) ÖNB-LAB will divide users through a registration procedure (in three levels). Some collections will be offered online to everyone (without registration), some datasets to logged-in users (login), and all datasets to confirmed scholars (verified users). In order to stimulate usage of collections and datasets, ÖNB-LAB has prepared a number of examples (of lab experimentations) under the heading—"What can I do here?"—that will encourage and inspire researchers to tinker with the data. ÖNB-LAB will be launched in November 2018.

Apparent from these short descriptions, many national libraries are currently establishing library lab environments. They vary in scope and purpose—and examples from other library labs could have made the range even broader. A common denominator, however, are the focus on (1.) curation of digitised heritage material into machine readable datasets, and (2.) computational support, usually in the form of a dedicated developer. Most library labs also have a (full or part time) manager in charge of operations. For a number of library labs external funding has, in addition, been important—at least during the first phase of establishment. Many lab environments have, furthermore, either received a grant themselves to get going or benefited from being part of one (or more) major research project. In other words, determined scholarly interest is a prerequisite for the establishment of a library lab.

Digital Scholarship

During the last decade different forms of digital scholarship has emerged within the academy. Since distinct digital scholarly practices and methods have gradually become more common, digital scholarship has evolved into a rather miscellaneous field of knowledge. This type of research is conducted by scholars in totally different academic disciplines and thus vary in scope—but digital scholarship also shares some features. Digital scholarship is, for example, often embedded in digital methods that depart from digital evidence in the form of data or datasets. Since such data is often scalable, research often undergoes a fundamental change—or as the literary digital scholar Franco Moretti has stated: "when we work on 200,000 novels instead of 200, we are not doing the same thing … The new scale changes our relationship to our object, and in fact *it changes the object itself*."

Digital scholarship can also encompass scholarly communication *using* data or digital media. Furthermore, digital scholarship can be defined as a scholarly activity that explicitly applies the new possibilities opened up by the *affordances* of 'the digital', as for example new forms of collaboration, new methods for analysing and visualising data, or new forms of publication. When applying digital technology to scholarship—results simply change. Novel forms of scholarly data, presentations and dissemination thus represents a shift from traditional publishing, including the result of scholarship that has traditionally been collected and preserved at libraries.

Within a library setting, digital scholarship has often come to encompass ways that academics work with digitised (or born digital) collections. The Digital Scholarship Department at the British Library, for example, works to enable innovative research based on the library's digital collections through "collaborative projects", by offering "digital research support and guidance", or by digitising collections or making content in digital form researchable. Digital scholarship thus takes advantage of new ways to examine data—but it is also designated to critically examine sources in digital form. The concept of *digital hermeneutics* is sometimes used to accentuate such a critical and self-reflexive use of digital tools and technologies. A concrete example is the way in which poor OCR quality of digitised text has become a problem of—and within—digital scholarship. While we think we are searching original documents within the digital library, it has been argued, "we are actually searching markedly inaccurate representations of text."

As is apparent, digital scholarship covers many academic perspectives. The scope is indeed heterogeneous, but importantly digital scholarship also *unites* academic fields that otherwise lies miles apart. Data science, digital humanities or computational social science are, for example, diverse fields of knowledge—yet via 'the digital' they share both methods and interpretational patterns, and thus become similar in range and scope. Since proponents of a variety of scholarly disciplines might be interested in working with data and datasets provided by the National Library of Sweden in a lab environment, a brief description of rather different scholarly knowledge fields can serve as a test bed for the kinds of usage a library lab might cater to. I will hence in the

following briefly discuss data science, digital humanities and computational social science, with a particular emphasis on how these academic fields relate (or might relate) to a library lab environment. It should be stressed that a number of other scholarly disciplines—from internet studies, to statistics or digital history—could have been described in a similar manner. Then again, what counts in the following is that a library lab at the National Library of Sweden should be prepared to facilitate and assist research from a wide array of disciplines—and not only the traditional focus groups of the humanities and social sciences—a diversity that will all likely effect how a lab is organised, including its principle aim and purpose.

Data Science

Data science is often described as an interdisciplinary field of knowledge that analyses large amounts of data—both within the academy and the commercial sector. By combining programming skills, processes and algorithms, data science tries to *extract* information, knowledge and insights from data in various forms—often through text mining *Big Data*. As a subfield of computer science, *data mining* is the process of discovering patterns (or predictions) in large datasets that involves both machine learning methods, statistics and database systems. Since statistical concepts as probability, inference and regression are often used, data science, data mining and statistics frequently get mixed up. In addition, data mining and data science are not the same thing, but they are also often used as synonyms—especially within the commercial domain. "The ability to take data—to be able to understand it, to process it, to extract value from it, to visualize it, to communicate it—that's going to be a hugely important skill in the next decades", Google economist Hal Varian has stated.

There are many definitions of what data science is actually about. Following the statistical mathematician Chikio Hayashi claim from the late 1990s, "data science is not only a synthetic concept to unify statistics, data analysis and their related methods but also comprises its results. It includes three phases, design for data, collection of data, and analysis of data." For a library lab environment it is important to note that data science uses both structured and unstructured data—even if this distinction is relative. The former is usually characterised by a high degree of organisation (often in a predefined data schema as Excel), the latter contains disparate data as numbers, facts or dates. However, most data has some kind of structure; most text for example follows syntactic rules of language.

Regarding the relation between data science and library labs one should note that there exists a potential antagonism between data science—with its inclination towards the hard STEM sciences (science, technology, engineering and mathematics)—and knowledge fields as the digital humanities or computational social science. The British Library Labs, for example, encourage all forms of scholarship, but with a tendency to advocate humanities and social science perspectives. Therefore, a certain rivalry has developed with the Alan Turing Institute—also located at the British Library—a "national institute for data science and artificial intelligence". There exist co-operations

between these two institutions, for instance within the research project, *Living with machines*—a collaboration between historians, data scientists and curators that will harness digitised archives, analyse and model the effects of mechanisation on society—but a scientific and economic competitiveness also reigns, not the least in terms of funding opportunities.

Digital Humanities

During the last 15 years the digital humanities has emerged as a field of knowledge that unites different forms of humanistic scholarship, characterised by the systematic use of digital technology. Digital humanities scholarship can today be found in an array of disciplines—in archaeology (with its frequent usage of geographic information systems (GIS) tools), within comparative literature studies (with its distant reading of large text corpora), or within social media studies (and its usage of network visualisation tools as Gephi). The digital humanities is often defined as a new way of *doing* humanistic scholarship—a practice which involves both interdisciplinary collaboration, work with digital methods and tools, as well as computationally engaged scholarship, where developers and programmers become an integral part of the research process. Importantly, the digital humanities also involves *critical reflections* on these tools, methods and applications. There are many definitions of the digital humanities—the website whatisdigitalhumanities.com offers a new answer every time one refreshes the page (quotes are pulled from a database with 817 answers).

Regarding the relation to the library sector, some scholars have argued that there exists an "overlap" between the digital humanities and libraries in general—for example around management of data, digitisation and curation. In 2017 a special issue of the journal *College & Undergraduate Libraries* entitled, "The Digital Humanities: Implications for Librarians, Libraries, and Librarianship" reflected on some of the current "challenges that occupy librarians who are engaging the academic community in the digital humanities." Yet, already in 2010 the digital humanities scholar Patrik Svensson stated that humanities-based engagement with information technology had by then developed into "a rich multi-level interaction with the 'digital'" as a result of the "persuasiveness" of digital technology. Humanists were now increasingly "exploring differing modes of engagement, institutional models, technologies and discursive strategies. There is also a strategy-level push for the digital humanities which … affects university research strategies, external funding and recruitment"—and libraries one might add.

In fact, for a number of the library labs that have been established during recent years, the digital humanities have been *the* prioritised and targeted focus group. The library lab at Yale University Library, for example, has a distinct digital humanities agenda—it is nowadays even called, The Digital Humanities Lab—and seeks to help humanities "scholars in their own engagement with digital tools and methods in the pursuit of humanistic questions." The Dutch KB lab states that their datasets are "a veritable treasure trove for digital humanities researchers", and digital humanities library lab can today be found at many universities: Digital Humanities Innovation Lab at the Simon

Fraser University (Vancouver) or the UWM Libraries Digital Humanities Lab at the University of Wisconsin–Milwaukee.

The digital humanities uses digital tools and methods in humanities study—but similar tools and methods are also deployed within data science or computational social science. Even though a library lab at the National Library of Sweden might primarily cater to scholars within the humanities and social sciences—since these are by tradition the ones who use the library and its collections—there is no need for a National Library lab to have a distinct digital humanities profile. On the contrary, the aim of the lab should rather be to facilitate *any* type of research that deals with digitised heritage. However, given the rapid development of the digital humanities—including tutorials, tools, methods, datasets and scripts (on Github)—a library lab should also make sure to take advantage of this wealth of online resources, specialist skills, services and computational support that has flourished within the digital humanities.

Computational Social Science

Situated at the intersection of computer science, statistics and the social sciences, the emerging field of computational social science uses large-scale demographic, behavioral and network data to investigate human activity and its relationships. A major part of scholarship within computational social science departs from the fact that the integration of digital technology into our lives has created unprecedented volumes of data on everyday social behaviour. With such an increasing amount of data (what we buy, where we travel, whom we know), computational social science is able to *measure* and *model* human behavior with a precision that was believed to be impossible just a decade ago. In short, statistical and computational methods are deployed to understand society and human behaviour.

It is sometimes argued that computational social scientists (with a computer science background) mostly try to establish empirical regularities and 'social laws', whereas *analytical sociology*—via agent-based simulations, machine learning, and large-scale web experiments—tries to move from "mere descriptions and predictions to the explanation of social phenomena." Nevertheless, like data science, computational social science (or analytical sociology) also has a commercial potential and is of great interest to major IT players—or as Microsoft Research states online: "troves of detailed social data related to choices, affiliations, preferences and interests are now digitally archived by internet service providers, media companies, other private-sector firms, and governments. New computational approaches based on machine learning, agent-based modelling, natural language processing, and network science have made it possible to analyse these data in ways previously unimaginable."

Even if computational social science is mostly focused on contemporary society, collections and datasets at the National Library will be of immense interest, for example harvested web archiving collections (Kulturarw3), newspaper datasets or statistics around the actual usage of digital collections. From a library lab perspective, analysing

web archives, social media and other data to extend the knowledge about present "drivers of human behavior and the most common patterns of interaction" will, however, also cause problems—not the least in terms of providing access to more or less restricted (contemporary) datasets. Another thing to note is that while methods and datasets might be similar or equivalent within data science, the digital humanities and computational social science—research questions and results often differ substantially. Within data science, for example, evaluating the effectiveness and speed of a certain topic model is a perfectly valid research question—i.e. datasets from the National Library might thus be used to assess, validate and improve probabilistic machine learning methods (rather than explain the data).

In fact, a potential problem with a broad agenda to support research at a library lab at the National Library—encompassing a range of academic disciplines, and with the purpose to facilitate basically all forms of scholarly research—is that projects and results (within these contrasting academic disciplines) will be perceived through standards for research excellency that differ substantially. From a data science perspective extracting tens of thousands of images from 19th century newspapers—and algorithmically comparing these through a script that displays similarities between images—would be exciting for some scholars, but less for others. Thus, a broad agenda might become problematic, and could potentially affect how a library lab is run—including tricky decisions regarding where to put resources.

Digital Tools & Methods

The set up of a library lab today does not mean that a host institution needs to start from scratch. On the contrary, a new library lab—staffed with a skilled developer—can kick-start its undertakings by making use of a wide array of already existing digital tools and methods. Brand new appliances can naturally be developed, but it is pivotal to understand that a huge number of applicable digital tools and methods—many of whom are open source and free to use online without a cost—are already at hand. There are also a number of websites that explicitly invites researchers and developers to start working with digitised heritage, methods and tools, scripts and code. Github is most well known, with its millions of developers who work together, host and review code, manage projects and build software—Github has a number of entries on 'digital heritage'. The Programming Historian, is another example, a site that publishes "novice-friendly, peer-reviewed tutorials that help humanists learn a wide range of digital tools, techniques, and workflows to facilitate research". Some of the tutorials are more difficult than others; they span from illustrating "strategies for taking raw OCR output from a scanned text, parsing it to isolate and correct essential elements of metadata, and generating an ordered data set" to ways in which researchers "can document and structure their research data so as to ensure it remains useful in the future." Another similar site is the Library Carpentry, "a group of librarians, repository managers, metadata librarians, research data managers, and other information workers

who are committed to teaching and developing a range of lessons designed to help librarians develop skills around coding and data analysis."

Then again, the success of a library lab does not rely on existing tools or methods, but rather on the ways in which they are implemented by interested researchers—as well as the quality of provided datasets. Nevertheless, in a report on the establishment of a library lab it is still useful to state a few things regarding what methods digital scholarship (predominantly within the humanities and social sciences) today makes use of. Trying to present specific tools, however, would be too time consuming—the downloadable software applications and SDKs (software development kits) at the *Programming Historian* (as an example) includes hundreds of programs—but the most common methods are not that many. In the following I will hence briefly present some of the digital methods that a library lab at the National Library would all likely work with, including text and visual analyses, as well as spatial and network analyses. Most of these methods are centred on the media modality of text, so I will put first and foremost attention on textual methods.

During the last decade the term *distant reading* (Franco Moretti) has become a popular way of broadly describing the analyses of major textual corpora, usually in the form of loosely tied algorithmic text-mining approaches. As a computational method, text mining can include a number of applications, from tracking textual reuse and fluctuation of words to stylometry or topic modeling. Important for the method of distant reading is the notion of scale, and the consequent movement from studying only particular texts (within a canon) to the aggregation and analyses of massive quantities of textual data. It is, however, important to remember that such corpora are artificial objects created by researchers—or a library lab. There was, for example, never the intention of the National Library, that *all* Swedish Governmental Reports (SOU) should be treated as *one* single text, once digitised. Yet for a library lab today it is possible to gather (in a similar manner) many forms of textual documents or books into massive single textual datasets.

Analysing such major datasets can be done in many ways, and here I will briefly touch upon three different methods: *topic modeling*, *named entity recognition* (NER) and *word embedding models*. Topic modeling is a computational method to study themes in unstructured texts (with no computer-readable annotations) by accentuating words that tend to co-occur, and together create different topics—akin to themes or discourses. As a method of extracting clusters of words from massive sets of documents, topic models are based on the linguistic assumption that words which are similar in meaning often occur in similar contexts. Importantly, topic modeling is a computer generated and automated method for organising, managing, and delivering results, where the latter depend on the algorithm being used. Themes or discourses (topics) are thus *automatically* discovered by an algorithm that analyses a dataset in its entirety. In short, topic modeling provides "a suite of algorithms to discover hidden thematic structures" in major collections of texts. Topic modeling has therefore emerged as a potent research tool with the possibility to both infer latent semantic topics and asses *where* in a corpus topics are prevalent. Themes in a textual dataset can thus be traced, and through the

discovery of hidden semantic structures, a dataset can be rigorously explored by discovering underlying discourses within it.

Unlike latent semantic analysis as topic modeling, named entity recognition (NER) uses syntactic information in sentences to identify named entities, such as persons, organisations, locations or temporal expressions. In applying NER algorithms on textual corpora, it becomes possible to *classify* and sort entities into a range of pre-defined categories. NER is, for example, able to trace the semantics of space, and has proven to be a useful methodological approach regarding, for example, the geographical analyses of textual datasets. In short, by entity extraction NER makes it possible to single out geographical places and locations in a major text corpora, and mark these (as structured data). In a similar manner, expressions of time (as for example years or dates) and actual person names can also be gleaned as structured data for analyses.

A third textual method which of lately has become increasingly popular are so called word embedding models. Such methods compute latent semantics of individual words by representing each unique word as a numerical vector. Every word in a text thus becomes a mathematical number. A word embedding model, in short, embeds a (numerical) word in a space that represent semantic and syntactic relationships between words. Word embedding models—such as word2vec—are today part of neural network based language models being developed and used by Google and Facebook. In moving from semantics to mathematics, word2vec embedding models makes it possible to perform simple algebraic operations on word vectors that have a semantical meaning—including ways to study analogies or dichotomies.

Setting up a library lab one should be aware of the bias towards textual methods and materials. There are, for example, way fewer open source tools for *image analysis* than there are for text analysis. Nevertheless, prototyping, design and data visualisation are methods that are becoming increasingly popular as a scholarly way to present data in visual form. A library lab should hence take the opportunity and advantage of visualisation possibilities and techniques—not the least as a way to present and promote library lab activities. Visualisation of large-scale images databases is, for instance, a novel way of presenting and exploring a major image collection. The software PixPlot can visualise tens of thousands of images in a two-dimensional projection where similar images are clustered together. In general, visual analysis consists of using computational processes to study large collections of images. Software can thus be used to visualise similar images within a large image corpus, to study patterns in visual datasets, or to apply image recognition algorithms that can identify different (or particular) objects in images.

As stated named entity recognition (NER) is a useful method for a geographical analyses of textual datasets, but most methods within *spatial analyses* depart from different GIS applications. Geographic information systems (GIS) involve many technologies, processes and methods. Within a library lab setting, however, so called *georectification* is particularly promising. Georectification is a method that takes a digital image of an old map and applies GIS to it, so that the image of the map can be used as a computable layer in other maps. Old maps in this way become possible to

analyse via GIS tools. There are for example many applications that study changes in a geographic area over time by comparing data extracted from maps covering a range of years.

If textual, visual and spatial analyses offer three distinct methods for digital scholarship, *network analyses* is a fourth example that a library lab might devote resources to. Computational social science, for example, often uses large-scale network data to investigate relationships. Networks can offer a dynamic way to study and visualise 'nodes' and connections or links between them. Through visualisation and exploration software as Gephi, nodes and links can be encoded in color and size to discover and explore patterns. As digital methods, visual analyses and network analyses thus share a number of attributes. Online are also a number of large network dataset collections and methodological descriptions how to develop similar ones in a library lab setting. Stanford University, for example, offers both (anonymised) social networks, "email communication networks with edges representing communication" and "Amazon networks: nodes represent products and edges link commonly co-purchased products".

As is well known, the collections at the National Library of Sweden are vast. Even if digital scholarship has had a tendency to focus on textual sources, specific methods for other modalities as images, maps, handwritten documents, sound and moving images will also be important in a scholarly library lab environment. At present, the Transkribus platform, for example, promises cutting edge research in handwritten text recognition, especially for automated transcription and searching of historical documents. Hence, a library lab might not only focus on major textual datasets but also devote resources to handwritten records and collections since they will increasingly become machine readable as well.

In a similar manner, new forms of speech to text methods and technologies makes audiovisual content computable and searchable in novel ways. The goal of the (current) project application from the National Library, "Speech technology and the availability of the audiovisual collections of the Swedish National Library" is, for example, to use automatic speech recognition to make the audiovisual collections significantly more available then they are today. Via so called *video fingerprinting* methods—a technique in which software identifies, extracts and compresses characteristic components of a video, enabling particular footage to be uniquely identified—reuse of audiovisual heritage can also be tracked and traced. On Github are several applications for the computational analysis of moving images as, for instance, the "Distant Viewing Toolkit (DVT) for the Cultural Analysis of Moving Images". The point to be made is that major textual collections (in the form of curated datasets) will all likely be in focus for a library lab. Yet, at present other forms of heritage (in nontextual modalities) will also increasingly become both available (as datasets) and possible to analyse and research

through applications that are currently being developed for computational analyses of basically any form of media modality.

Curating Data Collections

Throughout history librarians and archivists have curated—that is, appraised and selected, arranged and described, catalogued and preserved—heritage materials and historic records. Curation in many ways lies at the heart of all professional librarianship, but digital curation of collections (in the form of datasets) differs from traditional curation since data can be assembled in quite *different* ways. It is well known that the process of digitisation alters the informative capacity of the original sources; bad OCR quality is the most apparent result. Yet, data curation (sometimes) goes one step further and deliberately transforms content into a new type of source.

A concrete example can suffice to make my point: I have for a few years been heading a research project together with the National Museum of Science and Technology in Stockholm where we have digitised all the museum's yearbooks *Daedalus*—in all some 85 volumes spanning approximately 15,000 pages. Within my research group we wanted to perform large scale textual analyses of all yearbooks bundled together as .txt-files in one single dataset. The OCR quality of the textual data was fine. However, it turned out to be necessary to *restructure* all texts to make the dataset machine readable for our research purposes. For instance, we needed to separate all articles in the yearbooks, and hence we inserted the symbols ### (between all articles) in the dataset in order for the machine to understand when articles started (and ended). Moreover, as we began to use named entity recognition-algorithms to extract geographical locations from the dataset, results appeared strange. In turned out that all *Daedalus* yearbooks started with a very long list of people (and their addresses) who were part of the membership organisation that supported the museum. When we erased all the names (and addresses) our NER-algorithms could be applied with splendid results.

The *Daedalus* example is a vivid illustration of the ways in which basically all digitised collections today need to go through some form of data curation to meet research criteria. Since the latter differs from one research question to another, so does curation, including data cleansing, filtering, enriching etcetera. The process of data curation—organising and integrating data collected from different sources, annotating and maintaining, assembling, presenting and preserving datasets for research—is today a central task for *all* library labs. From a scholarly perspective it has even been argued that collaboration with "memory institutions on [the] single issue of digital data curation could dramatically improve the quality of humanities research". As the episode with the *Daedalus* dataset attest to, data curation can however take many forms: it might mean to simply mark divisions between texts in a dataset, or chopping it into smaller textual chunks for better probabilistic topic modeling results. Importantly, a library lab should curate datasets and make them available—but also be prepared to *rework* such datasets.

If a software developer is important for a library lab, a data curator is even more significant.

At present, there are a number of competing terms used to describe the activity of managing digital materials for research: digital curation, digital stewardship, data curation or digital archiving. Data curation is, however, the term chosen by this report, stressing novel library practices as the production and assemblage of datasets, digital thematic research collections, building scholarly editions and working with data visualisations. The digital humanities scholar Alan Liu has over the years collected thousands of data collections and datasets at dhresourcesforprojectbuilding.com, and it should be stressed that datasets can come in many forms, as "demo corpora", as "image collections", as "linguistic corpora", as "map collections", or as "audiovisual data".

In a broad sense, data curation at libraries can today be defined as the active and ongoing management of digital assets that are of interest to and useful for scholarship. According to the data librarian Arjun Sabharwal data curation includes a number of activities and processes: "description (documenting the context and relationship of various forms and of research data); annotation (enhanced information on the data with more granularity and context); collection and aggregation (connecting data and teams); storage (maintaining a platform for stable and accessible data); and migration (to ensure continued access via emulation or preservation)."

Providing Datasets

Obtaining a dataset from a digitised collection might come across as a trivial task—yet data curation point to the fact that once a collection is digitised and OCR converted, then the real work begins. A library lab should on the one hand publish more generic datasets (based on its ongoing digitisation activities), but on the other hand also engage in dialogue with scholars and assemble more particular datasets in relation to actual research. Resources at a library lab are naturally limited, and selection criteria concerning which prioritised dataset to work with and publish will be an issue that a library lab will more or less constantly be preoccupied with.

In principle, there are two primary methods of publishing open data or datasets—as bulk data or with an Application Programming Interface (API). An API is a kind of software that communicates via the web (and foremost useful to programmers), whereas bulk data in common file formats—especially CSV (used to store tabular data as in Excel)—can be managed by almost anybody. A library lab environment should enable data transfer in both ways.

At present, the almost 30 open datasets available at data.kb.se offers an excellent start for a library lab at the National Library. Other collections of digitised material of interest to curate are, for example, (parts of) web archiving content (Kulturarw3) or digitised textual material originating from the previous research site filmarkivforskning.se. Given the support from the Swedish National Archives to endorse a library lab, digitised content and data collection from the archival domain

would naturally be of tremendous interest to curate into machine readable datasets. In addition, the survey regarding available digital collections within the ALM-sector in Sweden (Appendix A) is a forthcoming and challenging task for a data curator—i.e. to examine potential collections (with a CC-license), curate and include them as datasets at datalab.kb.se. A number of such digitised collections are mentioned in the survey, and the general idea is simply that a library lab might aggregate (some) already digitised content. The scope of such an undertaking would naturally have to be discussed—to accumulate all collections of interest would be too time consuming. But a library lab should aggregate at least some material of Swedish origin (digitised elsewhere, but present in the collections at the National Library). At KvinnSam (Gothenburg University Library) there are, for example, a number of women's periodicals (from around 1900) that have been digitised, as *Dagny* (1886-1912), *Idun* (1887-1926) or *Herta* (1914-1931) that would indisputably be of great interest to present as curated datasets.

In general, at datalab.kb.se it should be possible to acquire datasets in at least four different categories: (A.) as general datasets (bulk data), (B.) through APIs, (C.) as specific datasets (curated in co-operation with researchers), and (D.) as derived datasets (in the form of resulting research data). Importantly, curating all datasets should always involve a process of licensing, making clear if datasets are freely open or have a legal restriction in some sense. Data curation of general datasets and available APIs is a task that a skilled data curator at the National Library will be able to work with on her own, (with the help of a dedicated developer). Enabling access to specific and derived datasets, however, will require cooperation with (external) researchers. Hence, if data curation in general deals with the management of data, in a scholarly library lab setting it will often be governed by and geared towards a *particular* research question—which the process around 'cleaning' the *Daedalus* dataset testifies to. Data curation in relation to a specific research question will almost always involve some kind of *data cleaning*—the process of detecting and correcting (or removing) corrupt or inaccurate records. For a library lab it is then of paramount interest that such processing details are made clear in the form of a data disclaimer. All dataset on the Dutch CLARIAH platform (Common Lab Research Infrastructure for the Arts and Humanities), for example, come with a disclaimer stating that the "dataset has undergone processing before it was uploaded to this register." Examples of possible processing operations are "filter, transform, enrich, clean, interpret, combine or reconciliate."

The production of specific and derived datasets (C. and D.) might come across as strenuous or demanding. Yet, what they point to are the ways in which scholars, data curators and developers today work collaboratively—including the ways in which data curation can take many forms depending on the actual research. At the .txtLAB, for example, a laboratory for cultural analytics at McGill University (Canada) they use computational and quantitative approaches towards understanding literature and culture in both the past and present. The datasets provided by .txtLAB are a fine example of the different ways that digital scholarship today operates. First of all a research question is posed, secondly work with gathering digitised collections or statistics start, and finally a dataset is produced which will lay the computational foundation for trying to answer the initial research question. Online the .txtLAB hence offers quite specific datasets:

"Novel450—a collection of 450 novels in German, French, and English that span 1770 to 1930. Each language is represented by 150 novels with a roughly even distribution across time, length, and gender", or "Race and Film—this dataset contains character dialogue from 780 Hollywood movies produced between 1970 and 2014. Characters have been labeled by their racial and ethnic identity", or "Contemporary Novels—a collection of 1,211 novels published between 2000-2015. They are categorized by the following 6 groups: Bestsellers (BS), Prizewinners (PW), Novels reviewed in the New York Times (NYT), Mysteries (MYST), Romances (ROM), and Science Fiction (SCIFI)."

Thematic research collections or text corpuses can thus become novel products of digital research methods, and data curation in such a scholarly setting suggest a blend of both editing and archiving methods. Data curation then becomes a process where it remains of utter importance to state how the data was prepared—the choice of source (the edition or the specific copy that served as the basis for the digital object), calibration of instrumentation, methods of data capture, details of transcription, levels of quality assurance, the kinds of editorial oversight that have been exercised, and the details of any subsequent curatorial activity. Such derived datasets in many ways resemble *research data* (as it is presently being stored at for example Swedish National Data Service) but it also differs since research can also produce derived datasets. Exactly how such derived datasets—that is, datasets that have been edited by researchers (like the *Daedalus* case)—should be made available in a library lab environment needs to be decided case by case.

Recommendations

The foundation of a library lab at the National Library of Sweden would strengthened and support a commitment to cutting-edge data driven research, and also expand and update the library research infrastructure substantially. Currently, the setup of a library lab also comes at a particularly favorable time. Not only are other similar initiatives being taken within the library sector in general—and within the national library sector in particular—researchers, funding agencies and governmental research propositions are also increasingly pushing scholarship in a data intensive direction in order to promote digital scholarship (of various sorts). The Swedish Research Council will all likely in 2019 deliver yet a call on "digitisation and accessibility of cultural heritage collections", and Riksbankens jubileumsfond is currently preparing a research call devoted to data driven research with a focus on quantitative and qualitative methods. Other funding agencies are likely to follow suit. In addition, both the Swedish Ministry of Culture and the Ministry of Education and Research have (in different documents) underlined the importance of cultural heritage institutions making their collections available in digital format for the benefit of research and innovation.

Scholarly & Institutional Usage

As previous discussions in this report have made clear, digital scholarship is progressively expanding. The digital humanities is the best and most visible example, but more and more scholars (by necessity) work with digitised material and datasets. Scholarly digital humanities environment in Sweden, like the Humlabs at Umeå and Lund University as well as the Centre for Digital Humanities at the University of Gothenburg, are likely to be the most attentive users of a library lab. The same goes for data intensive humanities and social science environments like the Centre for Data Intensive Sciences and Applications at Linnaeus University or the Institute for Analytical Sociology at Linköping University. Moreover, within data science and artificial intelligence environment—like at RISE (Research Institutes of Sweden) and KTH (image analyses and sound-to-text applications)—one can envision a keen interest in working with major datasets.

A library lab environment will also be of interest *within* the ALM-sector. As stated, the Swedish National Archives supports the idea of a library lab, and co-operation with the platform Digisam (at the Swedish National Heritage Board) also has the potential of enhancing the library lab into a digital heritage infrastructure of national importance. Moreover, an envisioned library lab environment should principally devote resources to make datasets available for research. These datasets would predominantly originate from the collections at the National Library, and secondly via other digitisation activities within the ALM-sector. Importantly, however, there are also other partners that might show an interest in making datasets available in a robust library lab

environment in the form of various co-operations as for example with Statistics Sweden (SCB).

It is always difficult to speculate around potential scholarly and institutional usage of a library lab. However, at present there is a vivid interest in new forms of digital scholarship at many universities as well as within the ALM-sector. In the latest research survey on the humanities and social sciences from the Swedish Research Council, the digital humanities is, for instance, explicitly mentioned as a field with a current "forceful development"—and one can expect that funding will follow suite. Hence, a lab a the National Library at present has the potential to become a *national resource and facility* for both digital humanities research and data intensive social sciences. In fact, one can even envision that a library lab environment (in a few years time) could be prioritise by Swedish Research Council's Council for Research Infrastructure and be "upgraded" to a national infrastructure of "high scientific and strategic value."

Lab Objectives & Staff

From this report it is obvious that a national library needs a novel infrastructure to cope with new forms of digital scholarship. My suggestion is therefore that the National Library of Sweden should launch a library lab. The main purpose of the library lab—datalab.kb.se—is to support all forms of research on digitised heritage. The lab does not need to have a physical space in the library, but personnel should once or twice a week provide consultations during office hours on location at the National Library.

I suggest that datalab.kb.se should have five main objectives: (1.) to support digital scholarship—by showcasing applications or developing digital methods—and by helping scholars to articulate research question(s) and ways to answer them, where algorithmic approaches might be useful; (2.) to support digital scholarship by curating, assembling or aggregating datasets—mainly from collections that have been digitised by the National Library—and make them available (predominantly online); (3.) to support and actively participate in the co-development of research applications (predominantly around available or potential datasets) with the library lab as a dynamic partner; (4.) to support and supplement the National Library's digital development (in general)—from cataloguing via topic modeling, to new forms of searching and visualising the collections—where the lab is perceived as an internal innovation hub; (5.) to support and enhance digital knowledge of staff working at the National Library. Like at BL labs, one might even envision an internal "Digital Scholarship Training Programme" for library staff to develop and expand knowledge on digital scholarship. By supporting staff to instigate new digital skills, a lab could create an opportunity to delve into and explore all that digital content and new technologies have to offer in the digital research domain, including making staff familiar with foundational concepts, methods and tools of digital scholarship.

I furthermore suggest that datalab.kb.se should run as a pilot project spanning two plus two years—with a major assessment between the two phases. The principal workload

during the setup phase will be to establish a robust web presence at datalab.kb.se and enhance functionalities and information at data.kb.se, as well as creating and curating new datasets. Importantly, the lab should not envision itself as a local undertaking, but rather as national data infrastructure catering to many different forms of scholarship. However, during an initial phase academic outreach and dissemination will be of major importance in order for scholars to discover the digital collections and datasets on offer.

Rights restrictions will naturally pose an obstacle regarding access to datasets. I therefore suggest that datalab.kb.se should foremost devote its resources to collections and datasets that are freely available online and CC-0-licensed—a suggestion and guiding principle which is similar in scope at most library labs at national libraries. Where rights restrictions occur, I suggest that datalab.kb.se follows and mimic the legal European procedures that the Dutch National Library has developed, where external researchers can be provided with restricted datasets via a general contract: "Agreement and conditions governing the use of KB datasets" (Appendix C).

Furthermore, if a researcher requires the need for large scale high performance computing of major datasets at for example the Swedish National Infrastructure for Computing (SNIC), my suggestion is a similar agreement via a general contract. At SNIC sensitive personal data is handled more or less on a day to day basis, and legal frameworks are therefore already at hand. Moreover, since the National Library is currently working with different forms of remote access to collections, an equivalent type of access (via a university login) might also be feasible regarding retrieval of restricted datasets.

Given the tasks and workload above, I suggest that datalab.kb.se should be staffed with a minimum of three persons: a (part time) library lab manager, a data curator (or data librarian) and a developer. Since datalab.kb.se is a pilot project, I moreover suggest that within the National Library organisation, the lab might be embedded in one of the present departments on physical or digital collections, or within the department of management support—at least during the first initial two years, with a steering committee staffed with a number of heads of department. The evaluation after the initial two years of operation might consequently also address the future organisational belonging.

Regarding estimations for funding a library lab, major costs will naturally be devoted to personnel. Funding 2,5 persons requires a substantial amount of money. Hopefully, one (or some) of the research proposals that the National Library were part of in the 'digitisation call' from the Swedish Research Council will have a favorable outcome. Moreover, a library lab environment will in part all likely be able to use already existing IT infrastructure at the National Library. Funding for hardware (as a complimentary server and storage) is however needed, as well as for purchase of software and the setup of a dedicated webpage. An estimated 500 000 SEK (for IT costs) will probably be enough to get the lab started, with yearly operating costs of some 250 000 SEK—yet again depending on co-usage of existing IT infrastructure. Importantly, with an established lab one might anticipate a substantially increased scholarly interest seeking

funding for research projects in co-operation with the library lab—not the least since co-developing research applications is a lab objective.

The recruitment of library lab personnel will be somewhat tricky since they require specific qualifications. The library lab manager, for example, needs to have extensive experience of working with lab-related issues and research management. Since a successful library lab (in the long run) will be dependent on external funding, the lab manager also needs to be proficient in obtaining endowments, as well as functioning as an active link between academic research and the library lab environment. The data curator or data librarian, furthermore, needs to have experience in data management training, as well as experience in working with scholarly research processes. Of vital importance is also demonstrated experience in data curation and dataset production, including experience in cleaning data for different analytical tasks. Finally, the developer should preferably have experience from previous work within the library sector. He or she should have both programming and database skills (R and Python), and ideally also experience of working with machine learning and data visualisation.

Selected References

British Library Annual Report and Accounts 2017/18 –
https://www.bl.uk/aboutus/annrep/2017to2018/bl-annual-report-2017-18.pdf

Conway, Paul, "Digital transformations and the archival nature of surrogates" *Archival Science* no. 1, 2015.

Blei, David M., "Topic Modeling and Digital Humanities" nr. 1, 2012 –
http://journalofdigitalhumanities.org/2-1/topic-modeling-and-digital-humanities-by-david-m-blei/

Flanders, Julia & Trevor Muñoz "An Introduction to Humanities Data Curation" 2011 –
https://guide.dhcuration.org/contents/intro/

Library of Congress Digital Scholars Lab Pilot Project Report 2016 –
https://labs.loc.gov/portals/static/labs/meta/images/DChudnov-MGallinger_LCLabReport.pdf

Hayashi, Chikio, "What is Data Science ? Fundamental Concepts and a Heuristic Example", Data Science, Classification, and Related Methods (Springer 1998) –
https://link.springer.com/chapter/10.1007/978-4-431-65950-1_3

Hitchcock, Tim, "Confronting the digital" *Cultural and Social History* no. 1, 2013.

Jarlbrink, Johan & Pelle Snickars, "Cultural heritage as digital noise: nineteenth century newspapers in the digital archive" *Journal of Documentation* no. 6, 2017.

Keuschnigg, Marc, Niclas Lovsjö & Peter Hedström, "Analytical sociology and computational social science" *Journal of Computational Social Science* nr. 1, 2018 –
https://link.springer.com/article/10.1007/s42001-017-0006-5

Franco Moretti, "Patterns and Interpretation", Stanford Literary Lab Pamphlet 15, September 2017 –
https://litlab.stanford.edu/LiteraryLabPamphlet15.pdf

Sabharwal, Arjun, "Digital humanities and the emerging framework for digital curation" *College & Undergraduate Libraries* nr. 2-4, 2017

Sula, Chris Alen, "Digital Humanities and Libraries: A Conceptual Model" *Journal of Library Administration* nr. 1, 2013 –
http://chrisalensula.org/digital-humanities-and-libraries-a-conceptual-model/

Svensson, Patrik, “The Landscape of Digital Humanities” *Digital Humanities Quarterly* nr.1, 2010 – http://digitalhumanities.org/dhq/vol/4/1/000080/000080.html

Swedish Research Council, “Forskningsöversikt Humaniora och samhällsvetenskap 2019” – https://vr.se/download/18.30d42bf1659263a4d91e14/1537775497453/Forskningsöversikt%20humaniora%20och%20samhällsvetenskap%202019.pdf