

Introduction to Digital History

<https://wissen-ist-acht.github.io/digitalhistory.intro/>

Ina Serif

Last compiled on 05.06.2026

Table of contents

Welcome	4
I Introduction	5
1 What is Digital History?	6
2 Teaching and Research	7
2.1 Digitised Sources, Digital Sources	8
2.2 Digital Tools for Analysis	9
2.3 Digital Tools for Communication	11
2.4 Projects and Resources for Teaching and Research	12
2.4.1 Ancient History	12
2.4.2 Middle Ages and Early Modern History	12
2.4.3 Modern and Contemporary History	13
2.4.4 Jewish History	14
2.4.5 African History	14
2.4.6 Eastern European History	15
2.4.7 Supra-national and Trans-epochal History	15
3 Digital Literacy, Digital Criticism	16
3.1 Digital Literacy, Data Literacy	16
3.2 Digital Criticism, Data Criticism	16
3.3 Artificial Intelligence and Algorithm Criticism	22
4 Data Gathering, Data Processing, Data Analysis	24
4.1 Data Gathering	25
4.2 Data Processing	28
4.3 Data Analysis	29
4.4 Data Storage	31
5 FAIR, CARE and LOUD	32
5.1 FAIR Data	32
5.2 CARE Principles	33
5.3 LO(U)D	33

II Practical Section	35
Hands on	36
6 ‘Der Sturm’ – Edited Correspondence	37
7 Through the Front Door	40
8 Through the Back Door	41
8.1 Shell 101	42
8.2 Ctrl-F 2.0	43
8.3 First Steps	46
9 Perspectives	52
Appendices	57
A Glossary	57
Glossary	57
B Literature, Tools, Tutorials	58
B.1 What is digital history?	58
B.2 Introductions and Guides	58
B.3 Digital History Tools (free/open source)	59
B.3.1 General	59
B.3.2 Data Analysis	59
B.3.3 Maps	59
B.3.4 Network Analysis	60
B.3.5 Text/Corpus Analysis	60
B.3.6 Storytelling	60
B.3.7 Visualisation	60
B.4 Digital Literacy, Digital Criticism	61
B.5 Programming Languages & Editors	61
B.6 Regular Expressions	61
B.7 Terminal/Command Line/Shell	62
B.8 XML	62

Welcome

The following guide, created during the autumn semester of 2022, is intended for first-year students of [History at the University of Basel](#) and gives first insights into the subject of Digital History. It is a living document that will be regularly updated and it means to take into account the different epochal and areal foci of the different introductory courses by adding more and more Digital History projects from throughout the field over time. For the participants of the introductory course, the guide will be accompanied by an in-person meeting, but it can hopefully also be used independently. I am happy to receive [comments, suggestions or complaints](#).

The guide is made up of two parts: Chapters one to five give an initial overview of Digital History and the changes that digital methods have brought to the field of historical studies. The ensuing practical part uses a concrete example in order to show the application of different techniques that can be used by (not only) historians working with source material. This part has two aims. Firstly it is intended to reduce people's shyness when using the computer as something other than simply an electronic typewriter. Secondly it wants to create a basic understanding of what possibilities computer assisted analyses have to offer and how to implement these in historical research.

The overview is as brief as possible – there are numerous detailed handbooks, for which reason many themes are only touched upon briefly and I give references to further literature. The same goes for the practical part: links are given to tutorials, online courses and further instructions when applicable. These references are by no means complete and I am happy to include your suggestions for further online resources.

Part I

Introduction

1 What is Digital History?

There is endless possibility for discussing what the right answer to the question of what the term Digital History can encompass is. Since it is a part of the Digital Humanities, the current pragmatic definition by Blaney et. al (2021) can be useful:

Digital humanities, in our view, is a question of approach: if you are actively and critically using digital tools to aid your work in researching, teaching or learning, you are probably doing digital humanities. We would encourage anyone to learn to program if they are interested in doing so, but we do not see it as a defining characteristic of work in digital humanities. [1, S. 6]

“Digital tools” can mean a whole lot of different things, and you will hardly be able to find anyone who completely abstains from using any of these in their studies, in research or in teaching. We are all as historians living in a digital age and as such need to develop new competencies. We use word processing programmes as a matter of course for our seminar papers or essays, research via search engines and online catalogues or lexica are part of the daily business of historians. But in addition to that we can decide to use methods and techniques for a research project that go beyond the tradition tools of the historians – the analysis and interpretation of sources through [close reading](#) – and let ourselves be helped by the computer. Whether we do this with the help of existing software or write our own programmes, understanding ourselves not only as historians of a digital age but also as digital historians may for some be a matter of principle; an inclusionary attitude towards this question however seems to me to be only beneficiary.[Quite opposed to a frequently cited opinion of Emmanuel Le Roy Ladurie’s (*1929) the historian of tomorrow shall either be a programmer or not be at all: “l’historien de demain sera un programmeur ou il ne sera pas.” 5]

For a first idea of how one can answer historiographical questions with the help of digital methods and how diverse digitally assisted research projects can look like, take a look at the article [“State of the Field: Digital History”](#).^[9]

A growing list of examples from different epochs and thematic fields can be found in chapter two under Projects and Resources.

This guide is about an active, critical and reflected use of digital methods in teaching and research, with a focus on the uses for historians. Further reading on the question of what Digital History is as a whole can be found under Literature, Tools, Tutorials

2 Teaching and Research

The way that, in all walks of life, things are becoming more and more digitised has a number of different consequences for historical studies, on many levels: when it comes to research and the work with sources, using digital methods not only for analysis but also for communication, and finally for teaching in higher education.



Figure 2.1: Randall Munroe, History Department, xkcd.com (17.12.2018).

2.1 Digitised Sources, Digital Sources

As historians, sources are central to our analyses. That means that access and availability of documents has a great influence on what questions we are able to answer or which analyses we can undertake. Restrictions as to access, which can influence the size and quality of our corpus, can be set by institutions such as museums, archives and libraries, for instance when contemporary documents are protected for a period of time, or an object is too fragile for use. It can also be difficult for financial or organisational reasons to access archives further away, in order to be able to include more documents in one's analysis. Large-scale digitisation project

in libraries and archives thus offer the possibility of finding additional material not just as a catalogue entry but as actual digitised copies that one can download onto one's own computer. Especially for valuable historical objects – such as antique papyri, early medieval manuscripts, early printed documents etc – this offers a possibility of making the objects available to a far larger audience without the strain of continuous usage, and without the users having to take long journeys to see them. For medieval manuscripts and early modern manuscripts and prints for instance there are nowadays many (mostly national) portals that allow a central search of all collections; you will find a selection in [Section 2.4.2](#).

Next to the digitisation of existing sources (retro digitisation) we have the unstoppable production of new sources purely in digital form (born digital data). The relative scarcity of sources often bemoaned by scholars of premodern history is in opposition to the surfeit of contemporary material. Both situations – too little, too incomplete and too much, too confusing – come with methodological difficulties: How does one put together a corpus (a collection) of sources that contains sufficient documents to allow one to answer questions, underline theses and gain new insights, but still remains manageable? Historians need to acquire new competencies in order to deal with such questions: in addition to the classic source critical analysis we now need digital source criticism, in addition to the ability to read and understand analogue sources, the ability to do the same for the digital realm. More about digital literacy and digital criticism in [chapter 3](#).

2.2 Digital Tools for Analysis

The definition already cited [here](#) that it is the active critical use of digital tools in teaching, research and study which makes up the Digital Humanities, begs the question of what exactly we mean by digital tools and to what ends we use them. Even reading this guide without digital methods is not possible – it does not exist in printed form. Reading at the screen does not make you a *digital humanist*, but one does not need to learn programming to be able to use the computer for one's own work and to achieve results that would not have been possible in the same measure with classical methods – in history mainly paper-based *close reading* of sources and scholarly texts.

Research that uses digital methods is usually scalable – if one is using software that counts the frequency of terms in one document, it should not make a difference to the software whether it does this with one or with one hundred documents. If you do this by hand however, the amount of analysis grows in proportion to the amount of documents. So digital tools allow us, amongst other things, to ask the same question of a far larger corpus of sources. It also allows us to ask different questions of this large corpus than would be possible with a small one. In the main, historical sources are text-based. Written by hand, carved, or printed – and with the possibility of changing this text via text recognition software into data that is readable to

a computer, sources can be turned into data that is able to be examined and analysed with the help of quantitative methods.¹

For Literary studies, for instance, one important field of use is to check for authorship. Whether an anonymous text can be attributed to a known author can either be judged by a literary scholar through *close reading* of the text, or through the search for patterns, for quantifiable attributes of a text, such as the frequency of function words, particles, punctuation etc. The detective novel *The Cuckoo's Calling*, published under the pseudonym of Robert Galbraith, could be attributed to Joanne K. Rowling by the relevant software – and this in thirty minutes, about as much time as it takes to read twenty pages of the novel. [Here](#) you can read an article that deals with this case and situates it within the field of linguistic forensics, which investigates offenders with the help of quantitative text analysis.

A video about the development and use of software for authorship attribution can be found [here](#). The software used, JGGAP,² can obviously also be used for historical analysis – think of regimes with strict censorship and numerous authors not publishing under their real names. By identifying anonymous writers different themes connected with censorship can be studied: which authors were publicly recognised, who at the same time published anonymously and under their real names, which authors wrote in exile, what networks can be reconstructed, etc. Having a programme that *can* do the quantitative analysis and thus the main job of identification (on how to think carefully about data and algorithms, see [chapter 3](#) –, means that there is more time remaining for the qualitative analysis. At the same time the analysis is based on a significant amount of data instead of looking just at individual cases.

The idea is not to play off qualitative and quantitative methods against each other, rather show that both methods have positive and negative aspects and that in the best case they can be used to good advantage in combination. Quantitative analysis just for the sake of it and without specific historical question is hardly ever of added value.

Depending on the data base, analytical purpose and research question, there are different tools you can use. For most of the research you will be doing during your studies, existing software should be sufficient, be it for statistical analysis, network analysis, geomapping or visualisation. You can find a selection of tools (all free of charge / open source) under the heading Literature, Tools, Tutorials. For certain analyses it can be helpful to learn some basic programming skills – being able to write your own scripts (smaller programmes) means having full control over the way in which data is read, prepared, enriched, analysed and visualised; with repeated processes that take a lot of time when doing them by hand, a lot of time can be saved.

For projects in the humanities there are at the moment two programming languages mainly being used, [R](#) and [Python](#). Since both are popular in the humanities you can nowadays find numerous packages that make data and text mining (large-scale data and text analyses) very easy. Such packages for programming languages are like plug-ins for programmes, for instance

¹This is easier to do for contemporary history, as many texts are already extant in digital form, for those periods or fields that are generally less rich in sources, generating data from text can nevertheless be even more interesting.

²Java Graphical Authorship Attribution Program, <http://evllabs.github.io/JGAAP/>.

like an ad blocker for the browser. These were not originally intended by the developers, but someone wanted to block advertisements and wrote a programme to do so, making it free for the general public to use. The difference in a package is that a package offers numerous different functions that the users can then choose and implement. Who has not had any contact with programming in school and university may possibly have some trepidation at first, but once again: you do not need to be able to programme in order to do quantitative work. There is a project called “[The Programming Historian](#)”, which has offered numerous tutorials since 2008 for historians without previous programming experience, introducing different tools, techniques and workflows for historical research and teaching.

2.3 Digital Tools for Communication

Even without quantitative/computer based analyses there are various options in historical research to use digital methods to assist in communicating results. These are on the one hand different types of digital publishing and on the other the use of digital tools to visualise results. Printed academic journals and publishers have certain criteria for accepting texts for publication – content quality is important for any academic work, but formal criteria such as length/imaging etc. are less central for online formats. In this way first results from a new research project or even a seminar paper can be offered up to an interested audience in the form of a blog article without much organisational preparation or regard for a publisher’s workload.

The number of academic blogs has been rising continuously in recent years, so that there are a number of suitable publication options available for every possible subject. These are edited, meaning that they are in the hands of academics who see themselves as responsible for the quality of the contributions. PhD projects are also often accompanied by blogs – giving visibility to the research and offering a space for themes that have no room in the actual dissertation, but are still noteworthy.

[Hypothèses](#) has established itself as an important platform that hosts a number of academic blogs. You can find a catalogue of all the blogs [here](#), and you can filter according to subject and period. The following blogs are the result of research and teaching at the History Department of the University of Basel:

- [Exilland Schweiz](#). Kulturschaffende und Intellektuelle im Schweizer Exil, 2021–
- [Global Health Africa](#). Circulating Knowledge and Innovations, 2019–
- [Materialized Histories](#). Materielle Kultur und digitale Forschung, 2021–
- [Materialized Identities](#). Objects, Affects and Effects in Early Modern Culture (1450–1750), 2016–2021.

- [Stadt.Geschichte.Basel](#). Blog zum Forschungsprojekt, 2019–
- [The Color Line](#). Race Relations in Schlüsseltexten amerikanischer Autor:innen (1881–1953), 2022–

A relatively recent format are Data Stories – narratives that describe a situation on the basis of (quantitative or qualitative) data and analyses. This type of (data) publication is used especially in journalism, and helps to ensure interactivity and actuality of data. There are various tools that can be used to create such data stories, some of them include publication possibilities; there is a selection under Literature, Tools, Tutorials.

One example that allows the users to add content themselves is [Darüber spricht der Bundestag](#), a Data Story by the newspaper “DIE ZEIT”, which allows analysis of all the speeches given in the German Bundestag since 1949. An example for a map-centric depiction that embeds various different media is [Arya’s Journey](#) from *Game of Thrones*. An example for the use of public census data is by the Schweizer Bundesamt für Statistik, [Die Schweiz \(er\)zählen](#).

2.4 Projects and Resources for Teaching and Research

2.4.1 Ancient History

Projects:

- [D-Scribes](#): Project recognising ancient authors of Greek and Coptic papyri

Resources/portals:

- [Papyrus Portal](#): digital collection of papyry
- [The Arabic Papyrology Database](#): Data base of premodern arabic texts written on papyri, parchment or paper from the 7th to the 16th century

2.4.2 Middle Ages and Early Modern History

Projects:

- [Burchards Dekret Digital](#): Digital edition analysing the manuscript transmission of the Decretum Burchardi
- [République des Lettres](#): Edition and research platform with numerous collections of texts by scholars between 1700 and 1850, connected to structural data
- [Printed Markets](#): Project digitising and commenting on the early modern “Avisblatt” from Basel (1729–1844)

- [Repertorium Academicum](#): Project compiling data on european scholars between 1250 and 1550

Resources/portals:

- [dMGH](#): Monumenta Germaniae Historica online (Beta-Version)
- [e-codices](#): Virtual Manuscript library of Switzerland
- [Fragmentarium](#): Laboratory for Medieval Manuscript Fragments
- [Handschriftenportal](#): Central national register of book manuscripts in German libraries and in German language (in development)
- [Innovating Knowledge](#): Data base and digital edition of Isidor of Sevilla’s “Etymologiae”
- [e-manuscripta](#): Digitised manuscripts from Swiss libraries and archives
- [e-rara](#): Platform for digitised prints from Swiss Institutions
- [Gallica](#): Digitised sources from French libraries
- [Stapfer Enquête](#): Edition of a Swiss school census from 1799
- [swisscollections](#): search platform for historical Swiss collections
- [transcriptiones](#): platform for the creation, sharing and use of transcriptions of historical manuscripts

2.4.3 Modern and Contemporary History

Projects:

- [impresso. Media Monitoring of the Past](#): project for the processing, semantic editing, representation, exploration and research of data in historical media (newspapers and radio), across time, languages and national borders
- [Living with Machines](#): research project on the impact of the mechanisation of work during the industrialisation.
- [Refugee History](#): academic blog and interactive network on the current debates around “refugees”

Resources/portals:

- [Datenbank Bild + Ton](#): on the history of Swiss social movements
- [Dodis](#): historical editions of documents on Swiss foreign affairs
- [Gallica](#): digitised sources from French libraries

- e-newspaperarchives.ch: Schweizer Zeitungen online
- e-periodica: Swiss newspapers online
- [Historische Statistik der Schweiz](http://HistorischeStatistikderSchweiz) (HSSO): historical statistic of Switzerland
- histat: historical statistics

2.4.4 Jewish History

Projects:

- [Digital Jewish Studies Online](http://DigitalJewishStudiesOnline), Stroum Center for Jewish Studies, University of Washington

Resources/portals:

- [Blavatnik Archive](http://BlavatnikArchive): archive for the preservation of material on jewish history of the 20th century with a focus on the world wars and soviet Russia
- Menny, Anna; Rürup, Miriam; Siegel, Björn: Jüdische Geschichte im deutschsprachigen Raum, in: Busse, Laura u. a. (Hg.): Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften, Berlin 2018, S. E.2-1–E.2-56. Online: <https://doi.org/10.18452/19244>.

2.4.5 African History

Projects:

- Emandulo: Digital archive that brings together and curates archival/museal collections and presentations on precolonial South African history
- [Legacies of British Slavery](http://LegaciesofBritishSlavery): research project on the British slave trade and ownership

Resources/portals:

- FHYA: Experimental digital research platform on precolonial South African history
- [Legacies of British Slavery](http://LegaciesofBritishSlavery): data base on British Slave trade and ownership
- [Slave Voyages](http://SlaveVoyages): data bases on the transatlantic and interamerican slave trade with a data base of individuals

2.4.6 Eastern European History

Projects:

- [Gulag: Many Days, Many Lives](#): Archive and presentation on sovjet gulags
- [Gulag Online](#): virtual museum with presentations and sources on life in the gulag
- [Seventeen Moments in Soviet History](#): multimedia online archive with selected sources on events in soviet history based around 17 different years between 1917 and 1991
- [The Imperia Project](#): a spatial history of the Russian Empire

Resources/portals:

- [Blavatnik Archive](#): archive for the preservation of material on jewish history of the 20th century with a focus on the world wars and soviet Russia
- [The Other Side](#): Web archive with interviews of ex-Ostarbeiter, POWs and inmates of German camps. Publication platform

2.4.7 Supra-national and Trans-epochal History

Projects:

- [Lord of the Rings Project](#): interactive analysis of the work of JRR Tolkien

Resources/portals:

- [Around DH in 80 days](#): portal presenting 80 different Digital-Humanities- Projects world-wide and from various disciplines
- [Internet Archive](#): digital library archiving books, images, films, software, music and websites

3 Digital Literacy, Digital Criticism

3.1 Digital Literacy, Data Literacy

What we understand by Data Literacy is the competence to collect, manage, evaluate and use data,[7, p. 8] a competence everyone should develop for themselves in order to deal with all the various forms of data we encounter in everyday life nowadays. Depending on the discipline there are differences in the specific competencies necessary – a humanities student may not think that algorithm criticism is the most important thing to learn when starting to study.¹ But you do not need to understand every detail of machine learning software’s source code in order to have a basic understanding of the way such applications function, and to be able to reflect on their usage. This kind of digital or data literacy is relevant especially when it comes to interpreting results that appear to be objective, or to have been produced without bias. A good example are the results of a query in a search engine. Depending on which engine you are using, different circumstances will play into generating your results list. For instance your own search history, meaning that *search neutrality* is no longer a given.² Go to [Google Image Search](#) and search for “historian”. What do you see?

If I did not know anything about historians, the results of my search would lead me to believe that a historian is an old, white man with glasses, a beard, and a large bookshelf. If you look around at the History Department in Basel, you might get a rather different impression. The results of search engines, that use algorithms in order to function, are biased: They are based on previous searches, preferences, geographical location – and on metadata offered by humans, data with information about other data. It is part of working in a digitised world that we are aware of this, and know to question data.

3.2 Digital Criticism, Data Criticism

Digitised Sources, just like purely digital ones, need advanced source criticism – in the introductory course at the University of Basel you will learn the basics of classic source criticism:

¹Different organisations and interest groups have thought about this, and the German “Gesellschaft für Informatik” has published a white paper on Data Literacy and Data Science Education as a digital competence in higher education: https://gi.de/fileadmin/GI/Hauptseite/Aktuelles/Aktionen/Data_Literacy/GI_Data_Science_2018-04-20_FINAL.pdf.

²There are alternative search engines, such as [Startpage](#) or [DuckDuckGo](#).

Where does a source come from, who has created it under what circumstances? What objectives could have been behind the creation, what biases could be in the text due to the circumstances of its creation? What can you read in a high-medieval king's chronicle, if the author was directly dependent on the person commissioning the work? How do we read the witness statements of an inquisition into witchcraft, if they were gathered under threat of torture? How carefully do we need to analyse the contents of a diary that was written with a view to later publication?

Apart from the internal criticism, working with sources always also concerns the question of the creation of corpora: what constitutes a significant corpus of sources that can be used for a specific historical question, but at the same time can be analysed in a manageable amount of time? Additionally, different types of sources bring different difficulties: With analogue sources that are also available in digitised form, there is the danger that a subject, area or aspect is neglected if only the immediately accessible, digitised collections are used. If for instance you are interested in the Swiss historian and women's rights advocate [Meta von Salis](#) (1855–1929) and her correspondence – Friedrich Nietzsche was one of her correspondents – and you go on to the catalogue [swisscollections](#), the search engine for Swiss historical collections, to look for the relevant documents in national libraries and archives, you will get 361 hits. :: {layout="[49,-2,49]"}

← Einfache Suche

Erweiterte Suche

Mit

<input type="text" value="meta von salis"/>	<input type="text" value="Alle Felder"/>	×
<input type="text"/>	<input type="text" value="Alle Felder"/>	×
<input type="text"/>	<input type="text" value="Alle Felder"/>	×

+ Suchfeld hinzufügen

+ Suchgruppe hinzufügen

Suchen

Suchtipps

[Hilfe zur erweiterten Suche](#)

[Hilfe zu den Suchoperatoren](#)

Suche verfeinern

Sprache <input type="text" value="Sprachen suchen"/>	Jahr von: <input type="text" value="300"/> bis: <input type="text" value="2023"/> <input type="range"/>
Format <input type="text" value="Brief"/> ×	Treffer pro Seite <input type="text" value="20"/>
Bibliothek <input type="text" value="Bibliotheken suchen"/>	

Suchen

[Stichwortsuche](#) [Indextsuche](#) [Durchstöbern](#) [Archivtechnik](#) [Bibliographien](#)

swisscollections

[Erweiterte Suche bearbeiten](#) [Neue erweiterte Suche starten](#) [Neue einfache Suche starten](#)

Format Brief ×

361 Suchergebnisse für: (Alle Felder:meta von salis)

Seite: ← 1 von 19 →

Sortieren nach Relevanz ▾

Digitalisiert verfügbar



Brief an Meta von Salis-Marschlins
von Louise von Schwerin
 Schwerin, Louise von
 1909.06.22
 Basel, UB, UBH NL 61 : M. IX 25



Brief an Meta von Salis-Marschlins
von Emmy von Strössenreuter
 Strössenreuter, Emmy von
 1918.12.27
 Basel, UB, UBH NL 61 : S. III IV 188



Brief an Meta von Salis-Marschlins
von Bertha von Cramer
 Cramer, Bertha von
 1867.04.19
 Basel, UB, UBH NL 61 : M I A 2a



Brief an Meta von Salis-Marschlins
von Anton von Salis
 Salis, Anton von
 1915.01.23
 Basel, UB, UBH NL 61 : S. III IV 31

Ressourcentyp ^

- + Archivmaterial (361) ×
- + Handschrift (361) ×

Bibliothek ^

- Basel, UB (355) ×
- Zürich, ZB, Handschriftenabteilung (6) ×

Person / Körperschaft ^

- + Salis-Marschlins, Meta von ... (356) ×
- + Nietzsche, Friedrich (1844-19... (15) ×
- + Kym, Hedwig (1860-1949) (5) ×
- + Eschmann, Marie (3) ×
- + Hauser, Johann (1853-1921) (3) ×

mehr

Sprache ^

- Deutsch (356) ×

Schlagwörter ^

- Ravenna, Bürgermeisteramt (1) ×

⋮

Of these, only three entries were available digitally in October 2022, of which the first was one letter of Nietzsche's to Meta von Salis, the second seven letters by Caroline Farner, and the third entry is neither by nor to Meta von Salis, but simply mentions her:

Stichwortsuche Indexsuche Durchstöbern Archivtechnik Bibliographien

swisscollections Erweiterte Suche bearbeiten Neue erweiterte Suche starten Neue einfache Suche starten

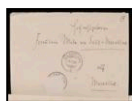
Digitalisat verfügbar × Format Brief ×

3 Suchergebnisse für: (Alle Felder:meta von salis)

Seite: ← 1 von 1 →

Sortieren nach Relevanz ▾

Digitalisat verfügbar



Brief an Meta von Salis-Marschlins
von Friedrich Nietzsche
 Nietzsche, Friedrich
 1887.09.01-14
 Basel, UB, UBH NL 61 : N. II 10 IV



7 Briefe an Meta von Salis-Marschlins
von Caroline Farner
 Farner, Karoline
 1893.05.19-1895.04.07
 Basel, UB, UBH NL 61 : S. II C. 2 1a-1b u.a.



Brief an Prof. Dr. Richard Oehler
von der Universitätsbibliothek Basel
 Universitätsbibliothek Basel
 25. Mai 1937
 Dokument=Item=Pièce
 Basel, UB, UBH NL 53 : B III 1, Beil. 3

Seite: ← 1 von 1 →

Ressourcentyp ^

- + Archivmaterial (3) ×
- + Handschrift (3) ×

Bibliothek ^

- Basel, UB (3) ×

Person / Körperschaft ^

- + Salis-Marschlins, Meta von (18... (2) ×
- + Farner, Karoline (1842-1913) (1) ×
- + Nietzsche, Friedrich (1844-1900) (1) ×
- + Oehler, Richard (1878-1948) (1) ×
- + Universitätsbibliothek Basel (1) ×

Sprache ^

- Deutsch (3) ×

Ort ^

- Basel (1) ×

Figure 3.1: Search results for “Meta von Salis” + “Brief” + “Digitalisat verfügbar”

So had you attempted to create your source corpus from the comfort of your desk, only taking into account digitised sources, you would have missed most of the relevant material and the results of your analysis would have been badly skewed, had you tried to make statistical pronouncements: Meta von Salis corresponded with one man and one woman, so the gender

ratio would have been equal – but more women on average wrote to Meta von Salis than men. Looking at all the search results, your analysis would have been quite different, so it would be a good idea to eliminate this bias from your data basis.

In addition to this there is always of course the basic problem when looking for sources: swisscollections and similar portals can always only show you what their cooperational partners have made available. If a library has letters by Meta von Salis in their collections, but have not yet recorded them as a data set, you probably would not even realise – contrary to the above example – that you have missed something, that there is a bias in your corpus.

Similar caution in order to avoid bias in your data basis is necessary with purely virtual data, for instance when evaluating data sets from questionnaires. If you stand in front of the University Library in Basel on the 27th of October 2022, and for a whole day, with the help of a short questionnaire and a spreadsheet, record how satisfied those you question are with the food in the university cafeteria, you will get a data set in which probably over 80% of those questioned would like better and cheaper food – which would make a good headline for the BZ, which could then claim to be reporting based on the newest scientific research. If you repeat the experiment a week later, during the autumn fair, the results will probably look very different. The probability that the cafeteria will have revised their menu and lowered their prices within days as a reaction to the headlines in the BZ is presumably less likely than that your sample, your choice of data points, meaning the people you questioned, have changed significantly due to the fair: You will in the second time round not have met mainly students and university staff, but also people visiting the fair on the Petersplatz. Here too, we have evidence of skewed results, similar to the last example with the correspondence: If out of a total only a specific subquantity with one common trait is regarded – digitised sources or visitors of the university library – the data basis and with it your research result is biased. In order not to transmit existing biases that may be inherent in the data you are using, you need to practice data criticism.

There is a good interview with Roopika Risam (2020) on the fact that data is not simply “given” (lat. dare, datum: give, given), but created, and for that reason needs to be interpreted;[8] on the cementing of clichés by translation algorithms there is an article in the Republik by Marie-José Kolly and Simon Schmid (2021);[4] and on the power of data science and the potential for change through data feminism there is an entire book by Catherine D’Ignazio and Lauren F. Klein (2020).[3]

On the question of how the digital turn influences source criticism, watch this short video from the [Ranke.2 project – Quellenkritik im digitalen Zeitalter](#):³

A handout concerning the use of digitised and digital data that has been developed in the same project can be found [here](#).

³You find a self-study course on the subject here: <https://ranke2.uni.lu/u/archival-digital-turn/>.

3.3 Artificial Intelligence and Algorithm Criticism

With the release of ChatGPT-3 by the firm Open AI in the winter of 2022, the use of artificial intelligence (AI) all of a sudden stood open to a very large number of people. A chatbot can, after an initial registration via an [URL](#), answer all kinds of different questions, can explain Fermat's last theorem, produce sonnets in the style of Shakespeare about an Emmental cheese being in love with a paper clip, or can summarise the so-called Historiker:innenstreit – a dispute about the singularity of the holocaust that took place in universities in the 1980s in Germany. Technically, ChatGPT is a generative language model (generative pre-trained transformer), a so-called large language model (LLM), that has been trained with a very large data set – around one trillion words from six million Wikipedia articles, millions of websites and digitised books, all in all about 570 GB, in order to produce new text in answer to user requests.⁴ If you ask ChatGPT itself to explain the concept of a generative language model and its own algorithm, you get the following answer:

Hello! I am happy to explain the concept of generative language models and my algorithm. A generative language model is an artificial intelligence-algorithm which is trained to create texts in a specific language or style. The model uses a statistical method called probability distribution in order to judge the probability as to where a specific word or phrase will be placed in a sentence. The aim of the algorithm is to generate texts with natural speech structures and patterns, in order to be understandable and legible to humans. A generative language model can use different approaches, such as the use of recurrent neuronal networks (RNN) or transformer models. As ChatGPT, I am based on the GPT-3.5 architecture and also use a transformer model. My algorithm uses a large amount of text data in order to learn speech patterns and structures and then be able to generate texts that are similar to that language. In order to generate text, my algorithm uses a so-called autoregressive method in which it uses the previously generated word or phrase as input in order to predict the next word or phrase. In this way text is generated step by step, each step based on the previous. Through perpetual adjustment of the emphasis in the model, depending on how good it is as predicting the subsequent text, the generative language model becomes better with time at generating text that sounds good and is realistic.⁵

This means that the answers are especially good, or make most sense, for subjects for which there are large amounts of training data. This also means, that everything that went online after the completion of the training in November 2021 has not found its way into the model

⁴You can find detailed and visual explanation by Helmut Linde here: <https://www.golem.de/news/kuenstliche-intelligenz-so-funktioniert-chatgpt-2302-171644.html>, an introductory lecture with a view to the consequences for higher education by Thomas Arnold here: <https://www.youtube.com/watch?v=-c8ogAwX6KI>.

⁵ChatGPT-3.5 result, request „Wie funktioniert ein generatives Sprachmodell und der dazugehörige Algorithmus?“, 11. November 2022, German answer translated into English.

3.5; there are differences depending on release and model, but events such as the Russia–Ukraine war that started in 2022 were not known to ChatGPT in 2024.⁶ It is also not always easy to judge the quality and balanced nature of the answers – by creating a programming function you can directly test the chatbot’s suggestion. With questions that are formulated more openly, LLMs offer convincing sounding answers – though they are of course guided by the training material, the model and the training method, which are not always transparent.⁷ ChatGPT-4o answers the prompt “name ten medieval philosophers” with ten men of which seven are from Western/Central Europe.⁸ Only when told that the list is very male-biased and white does it offer a greater selection of Arabic and Persian philosophers as well as (western) women. Another query will elicit Indian and Chinese names. The large part of the texts that the model was trained with are in English, and thus from the Anglo-American sphere. A bias towards knowledge production and opinions from this sphere is thus invariable and must be taken into account.

Looking at the above reply, beautifully formulated as it is, it would seem obvious to use LLMs for all kinds of text work, and in (higher) education it caused some alarm⁹ – especially also since the programmes offered freely invented facts with great confidence, or referred to non-existent references. If people uncritically adopt results and sell it as their own text it does not only suggest capabilities that are not actually there, but the likelihood that false information is picked up and circulated, for instance in presentation, increases. As a reaction, so-called codes of conduct have been established at most universities; Basel has a new form that students have to sign, declaring that any passages created with the help of AI have been marked as such. There is of course also software able to identify texts generated by AI.¹⁰

There are also discussions about helpful uses of generative language models as learning assistants,¹¹ as writing coaches, or as research assistants, as well as successful examples of how ChatGPT can be used in teaching.¹² Central to an informed and productive use is definitely a general understanding of how the algorithms work and the preparedness to critically check the received answers, and be prepared to dismiss them if necessary.

⁶By May 2026, the newest version of ChatGPT, GPT-5.5, has training data until June 2024; [perplexity](#) has October 2023 as cut-off date, Google’s [Gemini](#) gives January 2025 as cut-off date. Newer online articles are incorporated into answers, and models can access online search, but depending on the result there is relevant information missing from the time between the completion of training and an online-search months later.

⁷The “Open” in OpenAI, the firm behind ChatGPT is thus not programmatic. The complete technical report, which nevertheless leaves out many relevant parameters, is available [here](#). For criticism of OpenAI’s methods, see for instance [this article by Leigh Mc Gowran](#). [dolly](#) is an example for an open source alternative with open training set and model, as well as Google’s [Gemma](#). The LLM [perplexity](#) advertises itself as referencing its online sources and making its results more transparent.

⁸ChatGPT-4o-result, prompt „Name ten medieval philosophers“, September 3, 2024.

⁹See [Basel jetzt](#), 19.01.2023; [NDR](#), 09.02.2023; [bajour](#), 20.03.2023; [20 minuten](#), 24.03.2023.

¹⁰The Vizerektorat Lehre has created new [guidelines](#) “Citing AI tools”.

¹¹See for instance [Pedagogical Educational Tutor \(PET\)](#), developed at the University of Hohenheim, with an [introductory video](#).

¹²See for example those by [Andrea Klein](#) or [Ethan Mollick](#).

4 Data Gathering, Data Processing, Data Analysis

Every kind of research is reliant on data, be it gathered by questioning people, medical measurements, web scraping or the interpretative analyses of texts. On the basis of data, research questions can be answered, theses postulated, hypotheses refuted, narratives strengthened. Analyses that rely on a small set of sources, or data, often present their results in the form of syntheses, which result from a previous interpretation of the documents. Via the source bibliography and references in the text, the source basis becomes comprehensible; that a certain passage, sentence or word can be interpreted in a certain way though is also due to the influence of the researchers themselves – a literature scholar for example, who wrote their PhD on male figures in Joanne K. Rowling’s work, will read *The Cuckoo’s Calling* (see Section 2.2) differently from a long-time Harry Potter fan, with much reading experience but different, less formal education. In the discussion around the text’s authorship, each will have different arguments for or against Rowling’s authorship, both will give well-founded arguments, both will refer to their experience and engagement with Rowling’s work and both will use individual passages or sentences in order to strengthen their argument, which a third person might use to argue exactly the opposite. The data basis is the same, and understandable, but the data evaluation, or the strategy by which it is evaluated is not, and thus the results gained, which again are research data, are not either.

Computer based analyses have the aspiration that all individual steps should be comprehensible, and thus that they produce reusable data: not just the source basis, i.e. the generation of data and the creation of a data set, but all the steps from the enrichment and refinement of data and the methods or programmes used for analysis to data storage and archiving should also be transparent, well-documented and comprehensible. For one, in order for the results and statements based on them to be credible, for another to allow the data to be accessible for further use, free of cost. We shall discuss the principles one should follow when working with data in [chapter 5](#). For the moment we are talking about the individual steps that are frequently needed in digital history projects: data gathering, data processing, data analysis and data storage.

4.1 Data Gathering

There are different methods you can use to gather data for historical research, we shall just mention a few in the following.

For periods of time in which there are relatively few sources and we have no serial data, **digitisation of texts** and their subsequent analysis is useful. Digitisation does not simply mean the transformation of a physical source into an image, but also enriching the picture with layout and text: only by marking those areas that are present in the text is it possible in a second step to recognise the text as such, and make it readable by a machine, and thus analyseable. This kind of transformation from image to text is possible both for modern texts, for which there is a typescript, and for premodern manuscripts or prints, in Latin script as well as in Arabic, Chinese or Japanese. There are fee-based programmes such as [Abbyy FineReader](#), but also open-source tools with or without a graphical user interface (GUI). Quite well known is [Transkribus](#), which compiles a number of functionalities; but the text recognition is liable to a fee after a certain number of pages, though some student's projects can be funded on request. Programmes that are run via the command line, are entirely free of charge and also offer a number of functionalities, are for instance [eScriptorium](#), [Kraken](#), [OCR4all](#), [OCRopus](#) or [Calamari](#).

For the **extraction of data** from digital/digitised texts there are a number of possibilities: you can use small command line programmes (these tend to be tricky to use and difficult to read) or packages for programming languages, for the humanities usually R or Python (see Section [B.3](#)). With these you can, for example, extract entities from digitised telephone books (person names, street names, employment) or from old theatre programmes (actors, directors, etc) and use them as data sets.¹

The initial effort it takes to create automated data extraction and the steep learning curve for the use of some of the necessary programmes can be daunting at first. If you only want to analyse one theatre programme, you are sure to be quicker if you just enter the relevant data into a table. But if you have a larger corpus of sources with similar internal structure, such as a telephone book or a series of theatre programmes, it hardly makes a difference whether you are analysing ten or a hundred with the help of a script. And in addition to that you can let others use your script, or use it for similarly structured data in another project.

If you want to work with already digitised collections from public institutions such as galleries, museums or archives, (so-called **GLAMs**: **G**alleries, **L**ibraries, **A**rchives, **M**useums), there is frequently the possibility of downloading data via an **interface**.² Such **A**pplication **P**rogramming **I**nterfaces (**API**) allow communication between two computers without a graphic

¹A useful tutorial for the extraction of datasets from telephone books has been created by [Lindsey Wieck](#) for a DH course at St. Mary's University in San Antonio: https://lindseywieck.com/fall_2016_sf/gatheringdatatutorial.html. [Derek Miller](#) works on Broadway shows: [Visualizing Broadway](#). The project is described [here](#) and [here](#).

²At openglam.ch you can find information about Swiss GLAM institutions that offer open data.

interface being necessary. Instead of, for example, searching for objects or documents via the search mask of the [Staatliche Museen zu Berlin](#) using different catchphrases, and then downloading the results individually, your computer can directly communicate with the museum's API and, with quite simple commands, download long lists of results for you to work with. For this kind of request you can use the command line or a programming language. You basically just need one line, as here in the programming language R:

```
library(jsonlite)
cats <- fromJSON("https://smb.museum-digital.de/json/objects?&s=katze")
```

If you want to comprehend the individual steps, you can download R [here](#). When you open the programme you first need to install the `jsonlite` package: `install.packages("jsonlite")` “Enter” installs the package. Then you can type in the two lines from above and execute them by again pressing “enter”. The results are shown with `cats + “Enter”`.

The result for your query of “katze” will be saved in the variable `cats` which can then be exported for further use as a table: `write.csv(cats, "docs/cats_smb.csv")`

The function `write.csv` saves the contents of the variable `cats` as a csv file³ under the file path “docs/cats_smb.csv” on your hard drive.

	A	B	C	D	E	F	G	H	I
1		objekt_id	objekt_name	objekt_inventarnr	objekt_erfasst_am	institution_id	institution_name	image	total
2	1	456	Statuette der Göttin Bastet in Gestalt einer sitzenden Katze	AM 2598	2021-11-02 21:15:59	9	Ägyptisches Museum und Papyrussammlung	data/smb/resources/images/201806/200w_21081501496.jpg	134
3	2	52029	Shinto-Götterschrein mit heiliger Katze	ID 17657 a,b	2021-11-02 21:15:59	11	Ethnologisches Museum	data/smb/resources/images/201807/200w_06173101822.jpg	134
4	3	108895	Katze auf Rädern zum Ziehen	N (35 F) 917/1995,a	2021-11-02 21:15:59		Museum Europäischer Kulturen	data/smb/resources/images/201808/200w_04160418311.jpg	134
5	4	256381	Einseitig bemaltes Ostrakon mit Darstellung einer Katze vor einem Opferisch mit Gans (Tefnut-Legende)	AM 3317	2021-11-02 21:15:59		Ägyptisches Museum und Papyrussammlung	data/smb/resources/images/202009/200w_5f5700958cb2.jpg	134
6	5	589	Figur der Göttin Bastet in Gestalt einer sitzenden Katze	AM 11385	2021-11-02 21:15:59		Ägyptisches Museum und Papyrussammlung	data/smb/resources/images/201806/200w_21081927984.jpg	134
7	6	6962	"Hälfte eines breiten Rings, darauf eine gelagerte säugende Hündin" (tatsächlich Teil eines ägyptischen Sitstrums mit säugender Katze)	Misc. 8482	2021-11-02 21:15:59		10 Antikensammlung	data/smb/resources/images/201806/200w_27194138039.jpg	134
8	7	7012	Vierfüßiges Tier. Katze? (es handelt sich um einen Amoretten mit Löwen)	Misc. 7899	2021-11-02 21:15:59		10 Antikensammlung	data/smb/resources/images/201806/200w_27194155897.jpg	134
9	8	230034	Amorette mit Katze	1928107	2021-11-02 21:15:59		14 Kunstgewerbemuseum	data/smb/resources/images/202009/200w_5f5ea0d3c91bf.jpg	134
10	9	63364	Mädchen mit Katze	60025-04.415	2021-11-02 21:15:59		Museum für Asiatische Kunst	data/smb/resources/images/201807/200w_15173931440.jpg	134
11	10	106633	Geliebte Katze	N (47 B) 3/2017,35	2021-11-02 21:15:59		Museum Europäischer Kulturen	data/smb/resources/images/201808/200w_04153954180.jpg	134
12	11	50739	Katze	ID 51881	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06165932553.jpg	134
13	12	51123	Katze mit Schellenbaum	ID 51948	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06171126839.jpg	134
14	13	51234	Katze 猫	ID 50252	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06171645114.jpg	134
15	14	51441	Erna: Katze	ID 52073	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06171906607.jpg	134
16	15	51467	Erna: Katze	ID 52049	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06171932133.jpg	134
17	16	51562	Erna: Katze	ID 52125	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06172105136.jpg	134
18	17	51736	Katzen	ID 52252 a,b	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06172711680.jpg	134
19	18	51794	Katzen	ID 52251 a-c	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06172750863.jpg	134
20	19	51890	Katze	ID 52290	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06172843317.jpg	134
21	20	51904	Nikko "Nemuri-no-neko" "die schlafende Katze" vor dem Eingang zur Cryptomerien Allee, die zur Grabstätte des Ieyasu führt. Nach Hidori Gingoro	VIII D 12502	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06172852239.jpg	134
22	21	50605	Kauernde Katze 猫 Neko	ID 51748	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06165645371.jpg	134
23	22	50966	Katze	ID 51834	2021-11-02 21:15:59		11 Ethnologisches Museum	data/smb/resources/images/201807/200w_06170156251.jpg	134
24	23	83570	Fränzi mit Katze auf einer Decke liegend	KdZ 26603	2021-11-02 21:15:59		15 Kupferstichkabinett	data/smb/resources/images/201807/200w_28193302235.jpg	134
25	24	85574	St. Goar. Blick auf St. Goarshausen und Burg Katz	SM 9.13	2021-11-02 21:15:59		15 Kupferstichkabinett	data/smb/resources/images/201807/200w_28195922385.jpg	134

Figure 4.1: First rows of the results for the query “katze” using the API of the Staatliche Museen zu Berlin

³comma separated value is a format in which individual *values* can be clearly distinguished by specific delimiters, usually *commas*, and can thus be displayed in a tabular format, where each value is stored in a separate cell. Spreadsheet software like Excel, Google Sheets, or Numbers can open csv files.

In order to avoid requests that will overload their servers, most APIs have built in either an authentication or a number of maximum results per query. With the above example this means that you are not given the entire results list (134, as you can see in the column ‘total’), but only the first 24 – these settings have been implemented by the developers of the API. If you want all results, you need to read their documentation and modify your request. For those who are interested, you can find details here.⁴

⁴The API from the example is configured in such a way, that with results over 24 hits you will only be given the first 24; that is somewhat unusual, but we can deal with it by setting the maximum result to 10. That is not too high and a number with which it is easy to do sums. You can set the parameters for maximum results with `&breitenat=10`. The starting point can be changed with the parameter `&startwert=`. Thus in order to get all hits, you can ask for the results in steps of ten, and add them together. So that that doesn’t develop into a copy & paste exercise, you need to use a slightly more comprehensive variable, or a number of variables. That has the benefit that you can then search for any term.

```
base_URL <- "https://smb.museum-digital.de/json/objects?&s=katze"
cats <- fromJSON(base_URL)
start <- 0
breite <- 10
iterations <- cats$total[1]%/%10 + 1
endsize <- cats$total[1]-(iterations-1) * 10
cat_list <- data.frame()
for (i in 1:iterations){
  if(i < iterations){
cat_list <- rbind(cat_list, fromJSON(paste(base_URL, "&gbreitenat=10&startwert=", start , sep="")))
  } else {
cat_list <- rbind(cat_list,fromJSON(paste(base_URL, "&gbreitenat=",
                                         endsize, "&startwert=", start, sep="")))
  }
  start <- start + 10
write.csv(cat_list, "Desktop/cat_list.csv")
}
```

First we clean the code and save the major part of the URL in `base_URL`:

```
base_URL <- "https://smb.museum-digital.de/json/objects?&s=katze"
```

The results of the query are again stored in the object `cats`:

```
cats <- fromJSON(base_URL)
```

The number of times you do this for one request is counted by the number of total results/10 + 1; the number of hits can be taken from the column “total” in the object `cats`. In R you write this as follows:

```
cats$total[1]
```

For the cat example, this gives you 134 hits, so: (134/10 without carryover)+ 1, so 14 iterations:

```
iterations <- cats$total[1]%/%10 + 1
```

Then you set the start value to 0:

```
start <- 0
```

And the maximum hits to 10:

```
breite <- 10
```

The last iteration need not retrieve the next ten hits, but only the remaining 4 (the last after 130):

```
endsize <- cats$total[1]-(iterations-1) * 10
```

Then we create an empty table, a data frame that we can gradually fill with our results. (With smaller quantities of data, the function `rbind` can be used to combine single tables; with larger quantities the iterative extension of data frames is not recommended.):

If websites do not offer APIs there is the possibility of reaching your goal with **Web Scraping**. Depending on the website or its contents however, the legal situation is not always entirely clear. To download websites with Python there is [a course in the Programming Historian](#) by William J. Turkel and Adam Crymble. A further tutorial on data acquisition by Zach Coble, Liz Rodrigues, Erin Pappas, Chelcie Rowell, and Yasmeen Shorish can be found [here](#).

4.2 Data Processing⁵

Working with data sets, be they your own or collected by a third party, it is often the case that there is information missing or collected irregularly, making a later analysis more difficult.

If in a survey among students concerning their studies, their subject is entered into a table without having first defined values for the category, you might find instead of “History” and “German” the variants “Historical studies”, “Hist.”, “Hitsory”, “German studies”, “German language and literature”, “German studies”, “Germanistics”. Instead of the two values for two subjects, you now have nine – without there actually more subjects being studied. In the best case, variants like these are prevented from the outset, by giving a set list of values. If, however, you receive a data set with different variants for one and the same word, you will need to combine these in order to have a useful data base. You could use **Ctrl-R** to try and find different variants and replace them; in programmes like Excel, Open Office or Google sheets you can have unique variants displayed in one column, and can then combine them to one base value; most helpful and easiest to use – even for large data sets – is the software

```
cat_list <- data.frame()
```

Once we have set these variables, we can build a loop that implements different actions under certain conditions: :

```
for (i in 1:iterations){
```

If the last iteration has not been reached yet, the request will be answered in steps of ten, each iteration moving the starting value of ten further, and the results are filled into the `cat_list`.

```
if(i < iterations){
```

```
cat_list <- rbind(cat_list, fromJSON(paste(base_URL, "&gbreitenat=10&startwert=", start ,
sep="")))
} else {
```

As soon as the last iteration has been reached, not ten but the number of hits saved in `endsize` are requested, which in our example is 4:

```
cat_list <- rbind(cat_list,fromJSON(paste(base_URL, "&gbreitenat=", endsize, "&startwert=",
start, sep="")))
}
```

```
}
```

```
start <- start + 10
```

In the end, so after 14 iterations, the table is written to a file::

```
write.csv(cat_list, "Desktop/cat_list.csv")
}
```

⁵It is often said that data editing/preprocessing takes up 80% of your time leaving only 20% for analysis and interpretation. Leigh Dodds looked at these numbers in a [blog article](#) from 2020, and the numbers are not really as dramatic as that.

[OpenRefine](#), with which you can extract,⁶ clean up/normalise⁷ and enrich⁸ data in order to have a good data basis for your research question and the necessary analysis.

For text files a number of steps are needed, depending on which method or software you want to use. For most analyses it is sensible to work with so-called stop word lists. [Stop words](#) are words that have been removed from a corpus before analysis in order to get more significant results, especially when purely quantitative methods are being used for content analysis. Words with grammatical functions are present in great number in documents, but carry little individual meaning. If you analyse the raw text of this guide according to word frequency (here with [Voyant-Tools](#)), you can only guess at the subject matter. “digital” is only on 12th place, articles and prepositions are far more frequent. Using a stop word list that removes the commonest words not carrying meaning from the text, the contents are clearer:

Further steps could be [tokenisation](#), segmenting the text on word-level, and [lemmatisation](#), reducing various word forms to their one base form – “is”, “was”, “are” are turned into “be”. Just like with the variants in subjects earlier, the different variants have no added value for your research question and can be merged for further analysis. There are software and packages for these steps too, so that you need not do it all yourself. Especially for the more widely spoken modern languages, see also Section [B.3](#). For non-standardised languages or language forms (dialects or premodern texts) it is more difficult. There are programmes for this too, but how precise they are must be judged individually.

4.3 Data Analysis

If you have a data set, either from your own data or someone else’s, and have done preprocessing for your own use, you are (finally) ready to analyse it. What method or software you use depends not only on the type and amount of data you are working with, as well as the format your data is in, it also depends on your research question. If you have got a data base with correspondents (authors and addresses) of whom you know where they live but have forgotten to note the dates of the individual letters, you can only illustrate a spatial distribution, but hardly one across time and space.⁹ If you are only interested in the spatial distribution of female and male authors but are not interested in the question of when, then it is hardly necessary to record this information. Before you decide on your method, you need to ask

⁶Evan Peter Williamson: Fetching and Parsing Data from the Web with OpenRefine, Programming Historian 6 (2017), <https://doi.org/10.46430/phen0065>.

⁷Seth van Hooland, Ruben Verborgh, Max De Wilde: Cleaning Data with OpenRefine, Programming Historian 2 (2013), <https://doi.org/10.46430/phen0023>.

⁸Karen Li-Lun Hwang: Enriching Reconciled Data with OpenRefine, The Bytegeist Blog 2018, <https://medium.com/the-bytegeist-blog/enriching-reconciled-data-with-openrefine-89b885dcadb>

⁹A large project of Stanford University, “[Mapping the Republic of Letters](#)”, has modeled the correspondence network of 18th century scholars through their letters. One example is that of Voltaire, in different visualisations: <http://republicofletters.stanford.edu/publications/voltaire/letters/>. Dan Edelstein. Interactive Visualization for Voltaire’s Correspondence Network. Letters in Voltaire’s Network [Created using Palladio, <http://hdlab.stanford.edu/palladio>].

		Term	Count
<input type="checkbox"/>	1	und	191
<input type="checkbox"/>	2	die	170
<input type="checkbox"/>	3	https	150
<input type="checkbox"/>	4	in	119
<input type="checkbox"/>	5	der	101
<input type="checkbox"/>	6	sie	96
<input type="checkbox"/>	7	für	89
<input type="checkbox"/>	8	von	87
<input type="checkbox"/>	9	zu	83
<input type="checkbox"/>	10	mit	74
<input type="checkbox"/>	11	ist	72
<input type="checkbox"/>	12	digital	66
<input type="checkbox"/>	13	sich	64
<input type="checkbox"/>	14	data	61
<input type="checkbox"/>	15	oder	50
<input type="checkbox"/>	16	zur	49
<input type="checkbox"/>	17	eine	49
<input type="checkbox"/>	18	daten	49
<input type="checkbox"/>	19	ein	47
<input type="checkbox"/>	20	das	46
<input type="checkbox"/>	21	es	44
<input type="checkbox"/>	22	werden	42
<input type="checkbox"/>	23	den	37
<input type="checkbox"/>	24	auf	37
<input type="checkbox"/>	25	um	36

(a) Word frequencies raw text

		Term	Count
<input type="checkbox"/>	1	https	150
<input type="checkbox"/>	2	digital	66
<input type="checkbox"/>	3	data	61
<input type="checkbox"/>	4	daten	49
<input type="checkbox"/>	5	history	35
<input type="checkbox"/>	6	wiki	32
<input type="checkbox"/>	7	doi.org	29
<input type="checkbox"/>	8	tools	24
<input type="checkbox"/>	9	online	23
<input type="checkbox"/>	10	en.wikipedia.org	23
<input type="checkbox"/>	11	quellen	21
<input type="checkbox"/>	12	command	21
<input type="checkbox"/>	13	chapter	21
<input type="checkbox"/>	14	shell	18
<input type="checkbox"/>	15	digitale	18
<input type="checkbox"/>	16	text	17
<input type="checkbox"/>	17	arbeit	17
<input type="checkbox"/>	18	analyse	16
<input type="checkbox"/>	19	terminal	15
<input type="checkbox"/>	20	line	15
<input type="checkbox"/>	21	interface	14
<input type="checkbox"/>	22	geschichte	14
<input type="checkbox"/>	23	forschung	14
<input type="checkbox"/>	24	ressourcen	13
<input type="checkbox"/>	25	literacy	13

(a) Word frequencies with stop words excluded

yourself how and to what end you want to use your data set and which questions you would like answered. In a next step we should think about the specific type of analysis that is possible with the existing data. Among the many possibilities for working with **structural data**, the methods most commonly used in history are [network analysis](#) or [regression analysis](#). For **textual data** there are also a number of forms of analysis, for instance word frequency count as part of [stylometry](#)/attribution of authorship (see Section 2.2), [topic modeling](#) as statistical method for identifying recurring themes in larger corpora, or [sentiment analysis](#) for extracting emotions, feelings and values from text passages. If you have **georeferenced data**, you can analyse and visualise your data in various ways with the help of [GIS](#) (Geographic Information System).

Whether you use your own script for topic modeling or use existing software, whether you do your regression analysis yourself or via websites, remains your own decision. It is often a good idea to use existing online possibilities for initial short analyses in order to decide whether the intended method really can give the hoped-for results. For larger projects, in which you will perform complex analyses over a longer period of time, it could be a good idea to work with a programming language simply because you can then adjust the functions to your own needs and have full control over your data. A list of tools frequently used for historical analysis can be found in Section B.3.

4.4 Data Storage

In [chapter 5](#) we will be looking at how to store your research data sustainably. At this point I shall just mention that beside storing your data, it is sensible to use version control and provide a detailed documentation. **Data versioning** has the advantage that you can redo individual steps, save data at various different stages and for later use, and attribute them to different team members. Additional version control is more than the regular back-up done by back-up programmes or clouds such as dropbox or Switchdrive, and for collaborative work in academia as well as industry, the use of [git](#) has established itself, often in combination with data/code repositories on [GitHub](#). Most of you will probably not have personal GitHub repositories, but will at some point use the system, mainly by downloading data that has been provided. The text data for this guide is also on a [GitHub repository](#). Finally, the **documentation** of stored data comprises information about the development of the data set. How and by whom was the data collected? How was it annotated? In what format is it available? What software was used where? What does it depict? Storing data in different places, for instance on your local hard drive, in a cloud and on a flash-drive, can save you from losing it. Documentation and storing it on a repository, a long-term storage space for data, gives it additional visibility and the possibility for reuse. Repositories for the humanities are for instance [DARIAH-DE](#) or the [DaSCH](#); there are very specialised repositories or ones that are open to all disciplines such as [Zenodo](#) (run by Cern). You can deposit your data there for free, have your authorship recognised, and give the data/publication a Digital Object Identifier (DOI), an individual and enduring digital identifier, which makes it permanently citable.

5 FAIR, CARE and LOUD

Already at the beginning of a project, be it a seminar paper or a large collaborative project, you should ask yourself questions about storage, compatibility and reusability of research data. Too often projects end without the data that has been produced being made usable for later research, either because nobody thought of looking for options for permanent storage, or because the data was collected and saved in a form that makes future reuse difficult or even impossible. At the latest when you visit the archives – be it analogue or digitally – and are going to transcribe a document, you will probably ask yourself whether that has not already been done by somebody else, and you could save yourself the job, or if you will want to offer your own transcriptions to other people. The question is simply where, and how.

At the beginning of your studies, such questions are presumably not yet central; but still some questions surrounding securing, storing, and reusing data and data formats shall be looked at briefly here, in order to create awareness, but also because they influence the [data gathering](#).

5.1 FAIR Data

The principles of FAIR data were defined in 2016 by a consortium of academics and organisations as follows:[10] **F**indability, **A**ccessibility, **I**nteroperability, **R**euse of digital assets.

Data should be **findable** and **accessible**, **interoperable** (usable with different systems) and **reusable**. If you take photographs of ten wills from the 18th century in the Basel State Archives for a seminar paper and subsequently transcribe them, identify the objects mentioned, compare the testators and hand in your results on paper to your tutor, your data is the exact opposite. Nobody knows that you have collected the data, it cannot be found via general search tools but only via personal contacts, and if your tutor would like to put it at the disposal of other students in order to encourage further research, they need to do it by copying your printed paper. Paper copies are neither interoperable nor sensibly reusable – you need to type them into a computer in order to make them readable by a machine and to be able to work with them. If, however, you publish your transcribed texts and identified objects – in a standardised format and with an open licence – on a repository, you are not only making important parts of your work visible, but also making further research easier.¹ You can also ensure that the work does not have to be done twice.²

¹For different repositories see Section 4.4.

²The platform [transcriptiones](#) allows transcriptions to be shared easily.

5.2 CARE Principles

Following the FAIR principles, the Global Indigenous Data Alliance formulated the CARE principles for dealing with indigenous data in 2019:[2] **C**ollective Benefit, **A**uthority to Control, **R**esponsibility, **E**thics.

The focus is not just on propagating open data and sharing of data, but also to consider the people involved and the goals in order not to reinforce existing power imbalance between different actors. Indigenous data should serve **collective use**, there should be a **right to control**, **responsibility** should be taken for the use of the data, and **ethical principles** taken into account. Even though these guidelines are specially worked out for the use of indigenous data, they are an addition to the data centered approach formulated in the FAIR principles that takes into account the origin of the data and urges a reflective use.

5.3 LO(U)D

Tim Berners-Lee, the inventor of the World Wide Web, early on promoted standardised digital data could be linked, and thus the development of a **Semantic Web**, in which data is human and machine readable:

The Semantic Web isn't just about putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find other, related, data.³

In order not simply to be able to find data but to be able to reuse and combine it, principles have also been formulated for **Linked Open Data**, which should be adhered to when creating or publishing data; well-known examples for LOD data sets are **Wikibase**, **Wikidata** or **GeoNames**.

Berners-Lee suggested a five-star system to classify open data sets, wherein five stars is the equivalent of Linked Open Data.

1 Star: A data set is openly accessible, in a random format, for instance as a pdf.

2 stars: a data set is openly accessible in a structured format, for instance Microsoft Excel (.xls).

3 stars: a data set is openly accessible in a non-proprietary format, for instance as comma separated values (.csv).

4 stars: a data set follows the standards of the World Wide Web Consortium (**W3C**), such as using Resource Description Frameworks (**RDF**) and adopting Uniform Resource Identifiers (**URIs**).

5 stars: a data set fulfills all the above criteria and in addition it contains links to other Linked Open Data.

³Tim Berners-Lee: Linked Data, 2009. Online: <https://www.w3.org/DesignIssues/LinkedData.html>.

Going even further, **Linked Open Usable Data** principles aim at not only providing data in open and linked form, but also to structure them in a comprehensible way and to document them in order to secure their reusability.

It is unlikely that you will have to think about things such as the W3C, RDF and URIs during your studies, but be aware that your research data is more visible and reusable if you choose a file format that does not only work on one operating system or needs a programme that is not freely available.

Part II

Practical Section

Hands on

In this part different steps surrounding the **collection of data** – how do I get from sources to (structured) data and what is that anyway? –, the **preprocessing of data** – how can or must I edit the data for my own use? – and the **analysis of data** – what is the structured data good for and what do I do with it? – will be exemplified. We will only touch many of the practices and concepts that can be delved into deeper with further Literature and Tutorials. You might miss some contents – that is unavoidable, but comments are always welcome.

As source we are using a correspondence that has been compiled as part of the digital edition “Der Sturm” at the Akademie der Wissenschaften in Mainz.⁴ The project which edits letters from individuals belonging to the international Avantgarde surrounding the Journal “Der Sturm”, had different users in view and offers the data in different formats:

On the Website of the project those [letters already edited](#) can be read, and there is an index of the named entities in the text, such as [persons](#), [places](#) and [works](#); in addition to that there is the possibility of downloading the sources or index data via an [API](#). This means we can approach the letters via the front door, the website, or the back door, the command line, and compare the two approaches. Basic concepts for working with data, the necessary steps to get from a source to a data set, will be touched on.

Letters are a common source genre in many different epochs and areas and they can be used both for textual analysis and for the gathering of structural data. The example is therefore aimed at showing processes that can be very useful for actual historical research.

⁴DER STURM. Digitale Quellenedition zur Geschichte der internationalen Avantgarde, erarbeitet und herausgegeben von Marjam Trautmann und Torsten Schrade. Mainz, Akademie der Wissenschaften und der Literatur, 2018. Online: <https://sturm-edition.de/>, licence: [CC-BY-4.0](#).

6 ‘Der Sturm’ – Edited Correspondence

The website <https://sturm-edition.de/> serves as a portal for the work with sources concerning the avantgardist “STURM” enterprise that began with the founding of the journal of the same name in Berlin in 1910, by [Herwarth Walden](#), and the subsequent founding of an art gallery, a theatre, and a publishing company that achieved international importance. Besides Walden other actors were a part of the enterprise and by summer 2026 179 letters by three different artists written to Walden were available in the portal. In addition to that, the journal, exhibition catalogues, yearbooks, publisher’s documents and further sources such as posters, photographs and invitations have been digitised.

Go to the [homepage](#) and read the brief description of the letters, people and places, to get a first impression of the material. Then click on the letter section and look at letter number 8, from [Franz Marc](#) to Herwarth Walden.

As you can see, the letter has not only been digitised, i.e. transformed into a digital image which you can look at via an external viewer, but also edited in historical-critical mode and encoded in XML following the TEI P5 guidelines.

What does that mean and why is it important for our work as historians?

In an historical-critical edition you do not produce a simple rendition for accessible reading as you are used to, for instance, from the way text is presented in a novel. Additional material, relevant for better understanding the text, is consulted and given as information in a so-called critical apparatus. For example, sources that the author has used as templates for certain passages are mentioned, or events that influenced the text are referenced. Named entities such as persons, places, works etc. are explained and commented on. The text itself is reproduced as close to the original as possible, orthographical or grammatical mistakes are not corrected, and usually extratextual elements such as crossed-out or underlined text is marked in some way.

[1914-04-14 / Sindelsdorf]

➤ [185v]

Folio number, i.e. the page in the source corpus at the Staatsbibliothek zu Berlin

[Empfänger]
Herr
Herw. Walden
Verlag „Sturm“
Berlin W. 9.
Potsdamerstr. 134/a

➤ [185r] L. W,

bestätige mit vielem Dank 60 Mk für 2 Holzschnitte. Auch diese kleinen Verkäufe nutzen mir momentan sehr. Ich habe wohl neun Holzschnitte in Arbeit, aber alle für die Illustration an der Genesis gedacht, die ich keinesfalls vorher veröffentlichen darf. Aber ich werde noch suchen, nebenbei für den Sturm ➤ (i) ein paar neue zu schneiden. Hat Arnold eigentlich die Bilder aus Breslau (3 kleine Kompositionen u. Wasserfall) Ihnen zugesandt? Ich habe von dort nichts gehört. Vollmacht hab ich geschickt. Ist Klugen ganz vom Erdboden verschwunden. Ich hegte noch immer stille Hoffnungen auf ihn. Auf die komische

➤ [185v] Sache bin ich neugierig. Daß Aug. Macke in der | Sezession ausgestellt haben soll, kann ich schwer glauben, wenigstens nicht von sich aus. C. wird die Sachen von anderer Seite haben. Oder ist es vielleicht Hellmuth Macke? August ist momentan mit Klee u. Moillet in Tunis!

Representation of extratextual elements such as underlines

Hrzl.
Ihr FMarc.

1. Die Tunisreise von Klee, Macke und Moillet wird in der Kunstgeschichte als bedeutendes Ereignis für die moderne Kunst behandelt. Vgl. Güse, Ernst-Gerhard: Die Tunisreise. Klee, Macke, Moillet. Stuttgart 1982. »

FAKSIMILES DIESER QUELLE



Bestandshaltende Institution: Staatsbibliothek zu Berlin - Preussischer Kulturbesitz. Lizenz: Public Domain.

Digitised letter

Click on „DFG Viewer“ leads to a representation in an external viewer

Critical apparatus containing explanations about the trip to Tunis that is mentioned in the text

Figure 6.1: Letter no 8 from Franz Marc to Herwarth Walden.¹ Red text contains a link, usually to entities such as persons or places.

The precise way in which the editors proceeded – this is different from edition to edition – is laid out in the editorial principles. For the sources in the STURM project there are different principles for the different genres of sources. Those for the letter edition can be found [here](#).

An encoding in XML means that a text file has been encoded with [Extensible Markup Language](#), meaning that structures within the text have been marked with previously defined signs, so that they can be interpreted by both humans and machines. Somewhat commoner is HTML, [Hypertext Markup Language](#), a language for the tagging of electronic documents. The principles are similar. Textual elements are marked with an opening and a closing tag,

¹Online unter <https://sturm-edition.de/quellen/briefe/fma/Q.01.19140414.FMA.01.html>

for instance

`<salute>Hrzl.</salute>`,

in order to signalise that the salutation in a letter (as in letter 8, see above) is exactly that, a salutation. That has the benefit, that you do not have to search for specific terms if you are interested in different forms of salutations, but can search for the appropriate element `<salute>`, and not only you, but the machine too. We will come back to that. First we'll look at the letters on the website.

7 Through the Front Door

Imagine you would like to study the letters of Franz Marc to Herwarth Walden for a research project. You are mainly interested in what salutations Marc uses in his letters to Walden, in order to possibly draw inferences about the relationship between the two. How would you go about getting an overview? Would you use the download options of the website or read the letters online? What steps would you take to save the letters on your computer? Please note your answers to all questions in a document.

There are a number of paths you could take and none are better or worse, but some might be more efficient than others, meaning that you save time you can use for other things, be that in studying or at the swimming pool.

Whatever method you chose, continue with the following:

1. Create an overview of all the salutations in the 45 letters by Franz Marc to Herwarth Walden. Which format you choose, analogue or digital, or which file type you use, is up to you.
2. Take letter no 8 from Marc to Walden and divide it into its constitutive structural parts. Again it is your choice whether you print the letter and mark the elements with a pen or work on your computer. What you define as being a structural element is part of the exercise. (In a poem, possible elements could be title and verse)
3. Take the same letter and mark all entities, such as persons, places etc. Create a spreadsheet (with Excel, Google Sheets, Open Office or suchlike), and enter the entities.

After you have done this, read “The Ten Commandments of Inputting Data” in chapter 3 of the book “Quantitative Methods in the Humanities”.[6, pp. 57–60] You will find the chapter in the ADAM-Workspace of your introductory course – the whole thing is worth reading, but the “Ten Commandments” are enough to start with. You can find a more practically focused online version [here](#) in the blog accompanying the book.

After reading the “Ten Commandments”, would you do something different in step 3 of the exercise? Write down everything you would do differently, or any new insights, and bring your results to the scheduled meeting in the introductory course, be it in analogue or digital form.

8 Through the Back Door

Before we take a closer look at the corpus of the letter edition, we'll take a quick look at the possibilities of interacting with the computer, and how this can be useful for our work as historians, either for collecting, editing or analysing data.

There are two ways of interacting with (or using) a computer: via a **Graphical User Interface** (GUI), so by using the mouse and clicking on objects, or, somewhat more directly, via the **command line**.¹ If you want to delete the file “letter1.txt” in the folder “letters” via GUI, you open the Finder (Mac), the Explorer (Windows) or the file browser of your choice (Linux), and click your way to the folder “letters”, click right on the file you want to delete (“letter1.txt”), click “move to bin”, or drag it there directly with the mouse. The same action can be written as a command: You open the Terminal (Linux or Mac; open the Finder and enter “terminal” in the search window, then open the programme) or a Power Shell (Windows; click right on the start symbol, then choose “Windows Power Shell”), navigate to the relevant folder by entering a command, for instance `cd Documents/letters` + ‘Enter’ (Mac und Linux) or `cd ./Documents/letters` (Windows) and enter the command `rm letter1.txt`, which is executed by pressing ‘Enter’.

```
(base) serina00@dg-19-mac-02 ~ % cd Documents/letters
```

```
(base) serina00@dg-19-mac-02 letters % rm letter1.txt
```

How to proceed in the command line/in the terminal using MacOS

The two methods differ in three points:

1. The command `rm` is final, the file is deleted without grace period in the bin.
2. The command is relatively simple to use for a number of documents at once, whereby quite different conditions can be taken into account, and it can be combined with other commands.
3. Terminal looks k3wl.

¹Command Line, Bash, Shell or Prompt are often found as synonymously used terms for command line interfaces. On UNIX-based operating systems like Mac OS and Linux the terminal is common as interface; details on: https://en.wikipedia.org/wiki/Command-line_interface#History. Windows users should be able to work quite well using the PowerShell, but might want to install [Cygwin](#) or [MinGW](#), in order to be able to work with a UNIX-based interface.

Before we take a look at the second and for us most helpful difference, a brief note on the command line.

8.1 Shell 101

In a terminal/shell (see footnote for differentiation) commands or programmes can be executed that occur on the *structural* level – for instance deleting a file, `rm filename.xyz` (`rm` for *remove*), or creating a folder, `mkdir NewFolder` (`mkdir` for *make directory*). Equally possible are operations on a *content* level – such as searching for a certain term in a text file, `grep 'term' textfile.txt` (Mac/Linux) or `Select-String -Path textfile.txt -Pattern 'term'` (Windows), or counting terms and saving the results in a new file, `grep -Ec '(term1|term2)' textfile.txt | wc -l > results.txt` (Mac/Linux) or `Select-String -Path textfile.txt -Pattern '(term1|term2)'.Matches.Count > results.txt` (Windows) – the commands will explained more fully below.

But how does your Shell know what it is supposed to do when you type `rm` or `grep/String-Select`? There are all sorts of Shell programmes that are already installed on your system and with which you can do a lot. Open your Shell, type in `date` and press ‘Enter’: You will see the current date and time appear. (Your Shell looks for the first argument, the command `date`, in the filesystem of your computer, and if it is successful, implements the action according to the given parameters.)

i Note

tmi: If you type in `echo $PATH` into the Terminal (Mac/Linux) or `$env:PATH` (Windows), you will see a list of the places in which a command is searched for. Type `which date` and press ‘Enter’ to see where the programme ‘date’ is on your computer.

If you type a command that does not exist, or for which there is no programme installed on your computer, you will get an error notification – you can’t do any damage.

```
(base) serina00@dg-19-mac-02 ~ % nonsense
```

```
command not found: nonsense
```

The windows output is a bit more extensive:

```
nonsense: The term 'nonsense' is not recognized as a name of a cmdlet, function, script file, or executable program. Check the spelling of the name, or if a path was included, verify that the path is correct and that the path points to the file you want.
```

The current date is probably also shown in your toolbar, and you can create a new folder with a right click – you don't really need the Terminal for that. To find a term in a text file you can open the document, press **Ctrl-F**, type in the term and see the result. If you want to search for more terms, you need to repeat the same action with **Ctrl-F**, term2. And if you want to search more than one file, perhaps to find out how often the salutation “Mit herzlichem Gruß” appears in a letter collection, you have to repeat the search in every file. If you then need to look for the variant “Mit herzlichem Grüßen” or even “Herz. Gruß”, your work will be multiplied.

You can also do the same in the Terminal and use some of the built-in programmes to save yourself time and work.

8.2 Ctrl-F 2.0

As in the previous chapter, we are working with a part of the source corpus of the edition “Der Sturm”, with those letters written by [Franz Marc](#).

To fully understand the following steps, please download the folder “letters_Der_Sturm”. You can either download the complete [GitHub repository of this guide](#) as a zip file, and find the folder “letters_Der_Sturm” in the folder “docs”.

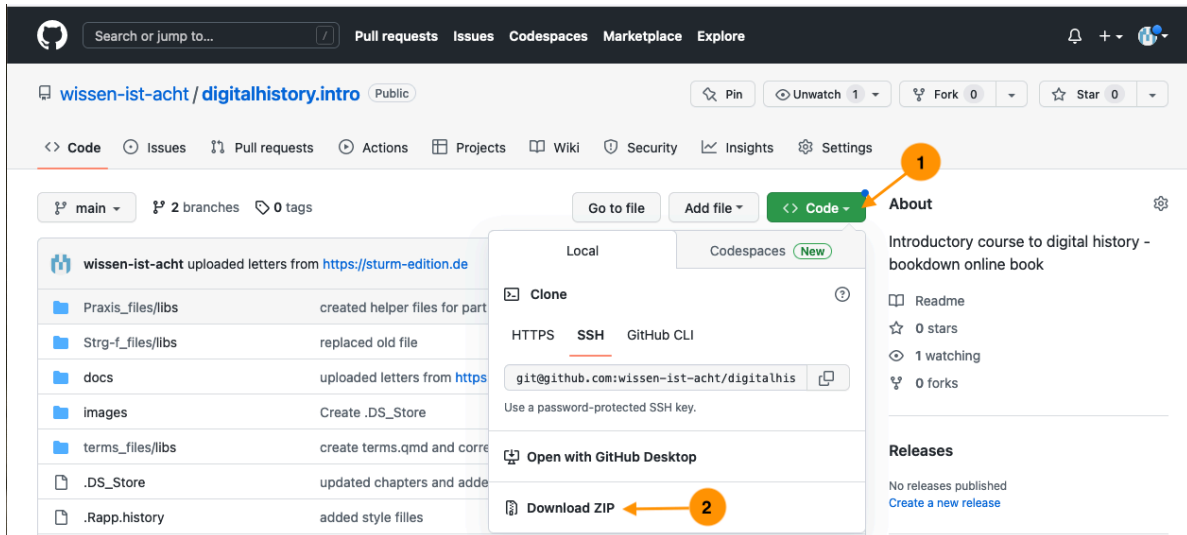


Figure 8.1: GitHub repository with source code, download via “Code” and “Download ZIP”.

You can clone the repository via the command line

```
git clone https://github.com/wissen-ist-acht/digitalhistory.intro.git
```

Or for a lazy option use this direct [link](#).

If you want to use the [website's API](#) you only need a few commands to get at the files.

Download using the API

With the first command we create a file, “letters_marc.xml”, with the file names of all the letters written by Franz Marc – from the register on the website we know that he has the person ID P.0000003; the URL of the API is in the [documentation](#):
Mac/Linux:

```
curl https://sturm-edition.de/api/persons/P.0000003 --output letters_marc.xml
```

Windows:

```
Invoke-WebRequest -Uri "https://sturm-edition.de/api/persons/P.0000003" -OutFile "letters_marc.xml"
```

If you open the file with an editor that can read XML files, you can see that next to the file names following “target=”, there are lots of things we do not need:

```
<person xmlns="http://www.tei-c.org/ns/1.0" source="http://d-nb.info/gnd/11857745X" xml:id="P.0000003">
  <persName type="pref">Marc, Franz</persName>
  <persName type="fn">Franz Marc</persName>
  <linkGrp type="files">
    <ptr n="B1.375" target="Q.01.19191212.JVH.01.xml"/>
    <ptr n="B1.377" target="Q.01.19200114.JVH.01.xml"/>
    <ptr n="B1.219" target="Q.01.19160128.FMA.01.xml"/>
    <ptr n="B1.222" target="Q.01.19160205.FMA.01.xml"/>
    <ptr n="B1.223" target="Q.01.19160302.FMA.01.xml"/>
    <ptr n="B1.218" target="Q.01.19160101.FMA.01.xml"/>
    <ptr n="B1.221" target="Q.01.19160122.FMA.01.xml"/>
    <ptr n="B1.220" target="Q.01.19160115.FMA.01.xml"/>
    <ptr n="B1.207" target="Q.01.19150703.FMA.01.xml"/>
    ...
  </linkGrp>
</person>
```

We only really need the file names to download the files with the relevant command. We can also see that it is not only letters with the abbreviation “FMA” for Franz Marc that are listed, but also nine with “JVH”, Jacoba van Heemskerck.

If you open the relevant files, you can see that they are ones in which Franz Marc is mentioned and is tagged in TEI-XML with `persName key="P.0000003"` which is why they are part of the results listed. With a second command we then create a new file, in

which the individual extracted file names are listed without those letters of Jacoba van Heemskerck, combined with the download command `curl` and supplemented with the relevant URL for the download:

Mac/Linux:

```
cat letters_marc.xml | grep -o 'Q.*FMA.*.xml\b' | perl -nle 'print "curl -o $_ https://sturm-edition.de/api/files/"
```

Windows:

```
Select-String -Path "letters_marc.xml" -Pattern 'Q.*FMA.*\.xml\b' | ForEach-Object {  
    $match = $_.Matches[0].Value  
    "curl -o $match https://sturm-edition.de/api/files/$match"  
} > filenames_letters_marc.txt
```

The file “`filenames_letters_marc.txt`” looks like this:

```
curl -o Q.01.19160128.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19160128.FMA.01.xml  
curl -o Q.01.19160205.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19160205.FMA.01.xml  
curl -o Q.01.19160302.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19160302.FMA.01.xml  
curl -o Q.01.19160101.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19160101.FMA.01.xml  
curl -o Q.01.19160122.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19160122.FMA.01.xml  
curl -o Q.01.19160115.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19160115.FMA.01.xml  
curl -o Q.01.19150703.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19150703.FMA.01.xml  
curl -o Q.01.19150417.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19150417.FMA.01.xml  
curl -o Q.01.19151106.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19151106.FMA.01.xml  
curl -o Q.01.19150918.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19150918.FMA.01.xml  
curl -o Q.01.19150827.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19150827.FMA.01.xml  
curl -o Q.01.19150303.FMA.01.xml https://sturm-edition.de/api/files/Q.01.19150303.FMA.01.xml  
...
```

`cat letters_marc.xml` (Mac/Linux)/`-Path "letters_marc.xml"` (Windows) passes the content of the file to the Terminal; `grep -o 'Q.*FMA.01.xml\b'` (Mac/Linux) or `Select-String -Pattern 'Q.*FMA.*\.xml\b'` (Windows) finds all character strings between “Q” and “FMA.01.xml” in the list, with `\b` after “xml” marking the end of the pattern; the 45 discovered strings are then each written in a new line, with `curl -o $_` giving the command `curl -o` and with `$_` (Mac/Linux) or `$match` (Windows) as placeholder for the character string (i.e. the file name), followed by `https://sturm-edition.de/api/files/$_` (Mac/Linux) or `https://sturm-edition.de/api/files/$match`(Windows) and again with `$_` or `$match` the character string (again the file name). With a third command, `bash`, you execute the created file, meaning that the commands in the file will be executed and the letters downloaded via the command `curl` (Client **URL**).

```
bash filenames_letters_marc.txt
```

However you have decided to download the files, you should now have 45 letters in XML format on your computer. Now open the Terminal (Mac/Linux) or the PowerShell (Windows) and with `cd`, **change directory**, navigate to the folder in which your text files are saved.

In my case this is in Documents/GitHub/digital_history_intro/docs/letters_Der_Sturm.

```
(base) serina00@dg-19-mac-02 ~ % cd Documents/GitHub/digital_history_intro/docs/letters_Der_S
```

For most of you, it will probably be in “Downloads” – give it a try.

(In order to check what is in a folder, you can enter `ls` (for **list**) in the Terminal, or `dir` (for **directory**) in the PowerShell):

```
ls
```

```
Q.01.19140115.FMA.01.xml    Q.01.19150315.FMA.02.xml
Q.01.19140119.FMA.01.xml    Q.01.19150327.FMA.01.xml
Q.01.19140121.FMA.01.xml    Q.01.19150417.FMA.01.xml
Q.01.19140124.FMA.01.xml    Q.01.19150501.FMA.01.xml
Q.01.19140125.FMA.01.xml    Q.01.19150615.FMA.01.xml
Q.01.19140125.FMA.02.xml    Q.01.19150703.FMA.01.xml
Q.01.19140409.FMA.01.xml    Q.01.19150710.FMA.01.xml
Q.01.19140414.FMA.01.xml    Q.01.19140421.FMA.01.xml
Q.01.19150827.FMA.01.xml    Q.01.19140507.FMA.01.xml
...
```

8.3 First Steps

Once you have navigated to the folder that has your letter files in them, instead of opening every single file as you would in a text editor, and searching with `Ctrl-F`, you can search for salutations with one single command. In the Terminal/in the PowerShell, with the programme `grep` (Global Regular Expression Print, Mac/Linux) or `Select-String` (Windows), you can search all the letters in the folder by including all files ending on “.xml” in the search. You can see the results – in this search for the salutations “Mit herzlichem Gruss” or “Mit herzlichen Grüßen” one hit per file – in your Terminal/PowerShell.

Mac/Linux:

```
grep -E -i '(Mit herzlichem Gruß|Mit herzlichen Grüßen)' *.xml
```

Windows:

```
Select-String -Path *.xml -Pattern "(Mit herzlichem Gruß|Mit herzlichen Grüßen)"
```

Output:

```
Q.01.19160115.FMA.01.xml:          <salute>Mit herzlichen Grüßen für Sie beide</salute>
```

So the formulation “Mit herzlichen Grüßen” appears once in your corpus, in the document Q.01.19160115.FMA.01.xml.

You can also get the word count of the number of hits found on line level, `-l`, with `wc -l` (Mac/Linux), or `Matches.Count` (Windows) and write them into a new file with `>` (that will be created when you execute your command).

Mac/Linux:

```
grep -E -i '(Mit herzlichem Gruß|Mit herzlichen Grüßen)' *.xml | wc -l > count_greetings.txt
```

Windows:

```
(Select-String -Path *.xml -Pattern "(Mit herzlichem Gruß|Mit herzlichen Grüßen)").Matches.C
```

When you open the newly created file `count_greetings.txt`, which will be in the same folder as the letters, you should find that it says “1”, since our search gave us one hit.

The command `grep` (Mac/Linux) was given the additional parameter `E` in the above command, and the command `Select.String` (Windows) the parameter `-Pattern`, meaning that we are not looking for an exact string, but are using the option of pattern searches. These are formulated as so-called **Extended Regular Expressions** (that is where you get the `E`). We have not only searched for “Mit herzlichem Gruß”, but also for “Mit herzlichen Grüßen”, which we formulated with the symbol “|”, which here means “or”. With the help of regular expressions we can extend our search further and search for different variants at once.

i Note

Regular expressions have different *flavours* – depending on the programming language things are formulated in different ways, and some default settings differ. In our case, `grep` needs the parameter `-i`, in order to ignore upper/lower case. `Select.String` ignores it

by default and does not need additional parameters. These details are important to know about when working with regular expressions, but is something you will learn on the go.

Mac/Linux:

```
grep -E -i '(Mit herzlichem Gru(ß|ss)|Mit herzlichen Grü(ß|ss)en|H(e|.?)rzl. Gru(ß|ss))' *.xml
```

Windows:

```
(Select-String -Path *.xml -Pattern "(Mit herzlichem Gru(ß|ss)|Mit herzlichen Grü(ß|ss)en|H(e|.?)rzl. Gru(ß|ss))")
```

With this formulation you will find 17 hits for a salutation, with the possible variants of “Mit herzlichem Gruß”, “Mit herzlichem Gruss”, “Mit herzlichen Grüßen”, “Mit herzlichen Grüssen”, “Herzl. Gruß”, “Herzl. Gruss”, “Hrzl. Gruß”, “Hrzl. Gruss”.

If you would like to find out whether greetings were sent *herzlich*, *hrzl.* or *freundlich* you can modify the search and form of output:

Mac/Linux:

```
grep -E -i 'Gr(u|ü)(ß|ss)' *.xml
```

Windows:

```
Select-String -Path *.xml -Pattern "Gr(u|ü)(ß|ss)"
```

Output:

```
Q.01.19140115.FMA.01.xml:      stets sofort antworte; es muß verloren gegangen
Q.01.19140119.FMA.01.xml:      <salute>Hrzl. Gruß</salute> <signed>Ihr <persName
Q.01.19140125.FMA.02.xml:      <salute>Hrzl. Gruß</salute>
Q.01.19140421.FMA.01.xml:      <closer>Gute Besserung <persName key="P.000
Q.01.19140507.FMA.01.xml:      <salute>besten Gruß</salute>
Q.01.19140831.FMA.01.xml:      <salute>Hrzl. Gruß von Eurem Freund in Waffen</s
Q.01.19141113.FMA.01.xml:      <salute>Hrzl. Gruß 1 x 2</salute> <signed>Ihr <p
Q.01.19150112.FMA.01.xml:      <salute>Hrzl. Gruß Ihnen beiden</salute>
Q.01.19150116.FMA.01.xml:      <salute>Mit herzl. Gruß Ihnen beiden</salute> <
Q.01.19150121.FMA.01.xml:      <salute>Herzl. Gruß</salute> <signed>Ihr <persName
...
```

With this command you are searching the text for the pattern `Gr(u|ü)(ß|ss)`, so a phrase that starts with `Gr` or `gr`, followed by either a `u` or a `ü`, then followed by either a `ß` or `ss`. As we have not marked a word-end (you would do that with `\b`), you will also find “Grüße” or “Grüssen”.

If you have clicked your way through the letters while reading the previous chapter, you will have noticed that not all letters end on “Gruss” or “Grüssen”. In the output of the queries in the Terminal you can see that all salutations are surrounded by a tag-pair:

`<salute>` marks the beginning of the salutation, `</salute>` the end. Open one of the letter files and search for “salute”. (If you do not have an XML capable editor on your computer, just open the file in your browser.)

```

112 ▾ <opener>
113 <salute>Lieber <persName key="P.0000001" ref="http://d-
nb.info/gnd/118770950">Walden</persName>,</salute>
114 </opener>
115 ▾ <p>ich vermute, daß <persName key="P.0000058" ref="http://d-
nb.info/gnd/118870645">Cassirer</persName> es auf den Proceß ankommen läßt; ich
rate nicht dazu;
116 Sie wi<hi rend="underline">sse</hi>n ja meine Bedingung, daß ich auf gar
keinen
117 Fall irgend welche Kosten von der Sache haben möchte. Das ist mir weder
118 <persName key="P.0000058" ref="http://d-
nb.info/gnd/118870645">Cassirer</persName> noch <term type="journal"
key="W.0000070" ref="http://d-nb.info/gnd/4127687-5">die
119 Aktion</term> wert.</p>
120 <p>Von <persName key="P.0000057" ref="http://d-
nb.info/gnd/143669230">Reiche</persName> resp. <persName key="P.0000059"
ref="http://d-nb.info/gnd/2097840-6">Arnold</persName> hab ich illustr. Katalog
erhalten; Sie werden ihn wohl
121 auch inzwischen gesehen haben; sonst kann ich Ihnen den meinen schicken.</p>
122 ▾ <p>Von
123 <persName key="P.0000033" ref="http://d-
nb.info/gnd/118969161">Filla</persName> hab ich die <term type="artwork"
key="W.0000011">
124 <hi rend="underline">Häringe</hi>
125 </term> gewählt, von
126 <persName key="P.0000060" ref="P.0000060">Beneé</persName> ein größeres
Aquarell, ich glaube mit blau (hell,
127 geometrisch) kein, hochformat.<pb xml:id="S.178v.02" n="178v"
facs="http://resolver.staatsbibliothek-berlin.de/SBB0000DAA400000001"/> Auf
beide Bilder haben sie
128 auf der Rückseite mit Blei „Marc“ geschrieben. Sie werden es schon finden.
129 Sobald ich die Sachen habe, sende ich 2 Aquarelle; geben Sie mir bitte die
130 Adressen.</p>
131 ▾ <closer>
132 <salute>Hrzl. Gruß</salute> <signed>Ihr <persName key="P.0000003"
ref="http://d-nb.info/gnd/11857745X">F. Marc</persName>
133 </signed>
134 </closer>

```

Figure 8.2: Extract from letter no 1 from Franz Marc to Herwarth Walden

As you can see, you twice have the tag-pair `<salute>-</salute>`, once framed by the tag-pair `<opener>-</opener>`, once by `<closer>-</closer>`. The address is marked with the first, the salutation with the second tag-pair. So when we are working with files that have been

marked up according to fixed rules, we can search for the element of the salutation without first having to look at the texts and formulate various queries. Now we change the formulation of our query and search for a string of characters beginning with `<closer>`, followed by from zero to however many `(.*)` characters of the class `cntrl`, so invisible characters such as tabs, page breaks, or line breaks. Then follows `<salute>`, again followed by from zero to however many `(.*)` characters, followed by from zero to however many `(.*)` characters of class `cntrl` and again followed by from zero to however many `(.*)` characters, until you get to the beginning of the closing tag `</salute>`. Like this, you can cover all possible cases in the letters in which you find text or line breaks between `<closer>` and `<salute>` or not, and in which there is text, no text or a line break between `<salute>` and `</salute>`.

Mac/Linux:

```
grep -E -zo '<closer>[:cntrl:]*<salute>.*[:cntrl:]*<' *.xml
```

Output:

```
Q.01.19140115.FMA.01.xml:<closer>
    <salute>Hrzl.<
Q.01.19140119.FMA.01.xml:<closer>
    <salute>Hrzl. Gruß</salute> <signed>Ihr <persName key="P.0000003" ref="http://
Q.01.19140121.FMA.01.xml:<closer>
    <salute>Hrzl.</salute> <signed>Ihr <persName key="P.0000003" ref="http://
Q.01.19140125.FMA.02.xml:<closer>
    <salute>Hrzl. Gruß<
Q.01.19140409.FMA.01.xml:<closer>
    <salute>Herztl.<
Q.01.19140414.FMA.01.xml:<closer>
    <salute>Hrzl.</salute> <signed>Ihr <persName key="P.0000003" ref="http://
Q.01.19140507.FMA.01.xml:<closer>
    <salute>besten Gruß<
Q.01.19140512.FMA.01.xml:<closer>
    <salute>Hrzl.</salute> <signed>Ihr <persName key="P.0000003" ref="http://
Q.01.19140606.FMA.01.xml:<closer>
    <salute>Hrzl.</salute> <signed>Ihr <persName key="P.0000003" ref="http://
Q.01.19140608.FMA.01.xml:<closer>
    <salute>hrzl.</salute> <signed>Ihr <persName key="P.0000003" ref="http://
...
```

If you want to write the results directly into a file you can of course also do that:

Mac/Linux:

```
grep -E -zo '<closer>[[:cntrl:]].*<salute>.*[[:cntrl:]].*<' *.xml > salutations.txt
```

But this is the moment, at the latest, to change your tools. You can do all sorts of things with the Terminal/Shell and there are any number of programmes you can install in addition – to parse XML files, to work on image files or to download Youtube videos. But it starts getting less and less neat and for the analysis of structural and textual data you can use programming languages like R or Python (as mentioned in [?@sec-digitaletools](#) which are more practical.

To continue working with the full text of the letters, be it for close reading or quantitative text analysis, you could, for example, strip the XML tags to achieve a better reading experience and to facilitate further analysis.

9 Perspectives

If you now compare the way you structured letter no 8 from Franz Marc to Herwarth Walden and the way in which you identified entities with the way in which it is done in the XML files of the STURM editors, you will probably find some differences. One difference will most certainly be that the editors followed a model in their structuring, which is the standard for text editions. It is called [TEI](#), and using it can have a number of benefits. It means for instance, that you can follow through with the queries shown in the previous chapter without wondering whether the tag name is going to change halfway through. And if you decide during your research to concentrate on the address rather than the salutation, or to include the corpus of letters by [Jacoba van Heemskerck](#), you can do this with only very few changes to your queries, as these too are encoded with a unified tag. If you only read the texts via the browser and counted things out by hand, you would have to start all over again with the addition of new documents.

The creation of standardised data according to a model or fixed guidelines also means that you can compare different data sets with each other or enrich them with additional data. Looking at the letters, on the website or in the XML file, you can see that entities such as persons or places are not only marked as such and linked internally within the project, but are also linked to other authoritative data, for instance with the relevant entry in the [GND](#), the Gemeinsame Normdatei der Deutschen Nationalbibliothek, or in [Geonames](#), a data base for geographical data.

```

112 ▾      <opener>
113      <salute>Lieber <persName key="P.0000001" ref="http://d-
nb.info/gnd/118770950">Walden</persName>,</salute>
114      </opener>
115 ▾      <p>ich habe nicht das geringste vom Anwalt erhalten; Sie kennen mich ja, daß ich
116 ▾      stets sofort antworte; es muß verloren gegangen sein. Grüßen Sie bitte D<hi rend="super">
117      <hi rend="underline">r</hi>
118      </hi>
119      <persName key="P.0000056" ref="P.0000056">Feige</persName> und sagen Sie ihm
120      das.</p>
121 ▾      <p>In einer niederrheinischen Zeitung soll von einer bevorstehenden großen
122      Ausstellung in <placeName key="0.0000045"
ref="http://sws.geonames.org/2952539">Barmen</placeName> von <persName key="P.0000009"
ref="http://d-nb.info/gnd/118559737">Kandinsky</persName> mir u.s.w.
123      berichtet worden sein; ich kann mir nur denken, daß die <persName key="P.0000057"
ref="http://d-nb.info/gnd/143669230">Reiche</persName>-<placeName key="0.0000032"
ref="http://sws.geonames.org/2935022">Dresden</placeName> Collection dahin
124      kommt;<pb xml:id="S.177v.02" n="177v" facs="http://resolver.staatsbibliothek-
berlin.de/SBB0000DAA300000001"/> ich
125      bin natürlich einverstanden, nur soll die frei<hi rend="underline">e
126      Rück</hi>fracht <placeName key="0.0000045"
ref="http://sws.geonames.org/2952539">Barmen</placeName> - <placeName key="0.0000002"
ref="http://sws.geonames.org/2950159">Berlin</placeName> gesichert
127 ▾      sein.<note>Im Jahr 1913 fand in der Stadt Barmen eine STURM-Ausstellung mit
128      Werken der Künstlergruppe „Der Blaue Reiter“ statt. In der Literatur gilt
129      diese Ausstellung als nicht ausreichend nachgewiesen; dieser Brief gibt
130      endgültigen Aufschluss über das tatsächliche Stattfinden der Ausstellung.
131      Vgl. Enders, Rainer: Ausstellungen außerhalb der Berliner Galerie. URL: <ref
target="https://www.arthistoricum.net/themen/portale/sturm/ausstellungen/">https://www.ar
thistoricum.net/themen/portale/sturm/ausstellungen/</ref>
132      (Aufruf 06.04.2017).</note> Aber es ist natürlich nur Vermutung von mir. Von
133      <persName key="P.0000057" ref="http://d-nb.info/gnd/143669230">Reiche</persName> höre ich
gar nichts.</p>
134 ▾      <closer>
135      <salute>Hrzl.</salute>
136 ▾      <signed>Ihr <persName key="P.0000003" ref="http://d-nb.info/gnd/11857745X">F.
137      Marc</persName>
138      </signed>
139      </closer>

```

Figure 9.1: Extract from letter no 1 from Franz Marc to Herwarth Walden, authoritative data offset in yellow.

If you click on the link to [Kandinski](#) or [Berlin](#), you will find additional information for the person or place on the website of GND or Geonames, such as dates and geocoordinates. If you now for example would like to know which places are named in the letters of Franz Marc, you can extract these not only with the tag pair `<placeName>-</placeName>`, but also with the added geocoordinates, and you can then visualise them on a map.

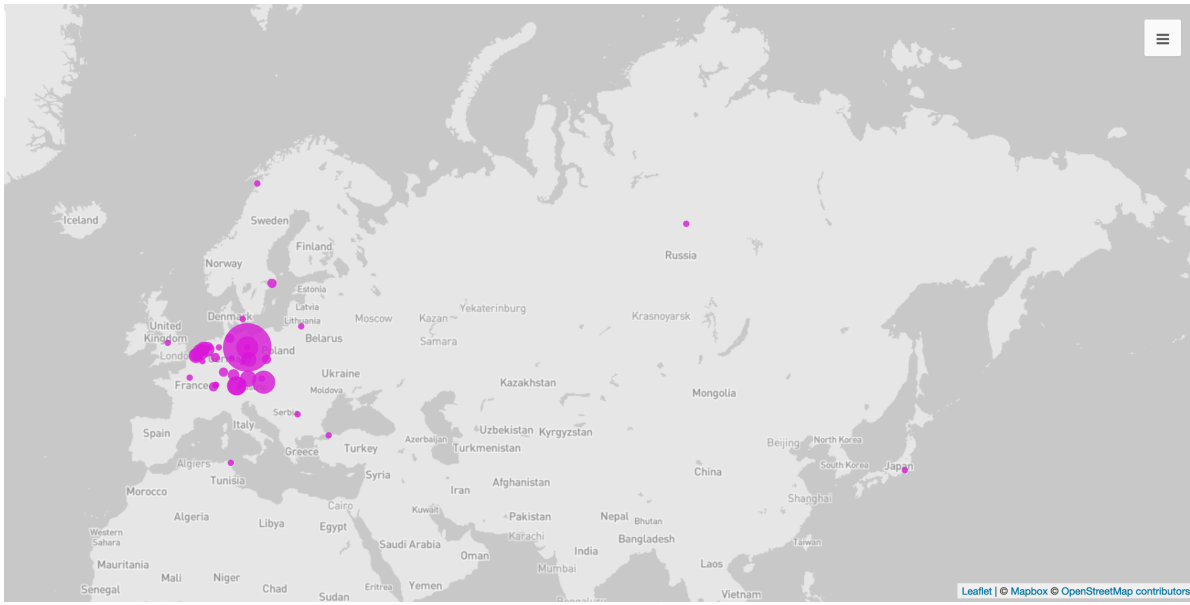


Figure 9.2: Places mentioned in the corpus of letters by Franz Marc to Herwarth Walden (created with Palladio).

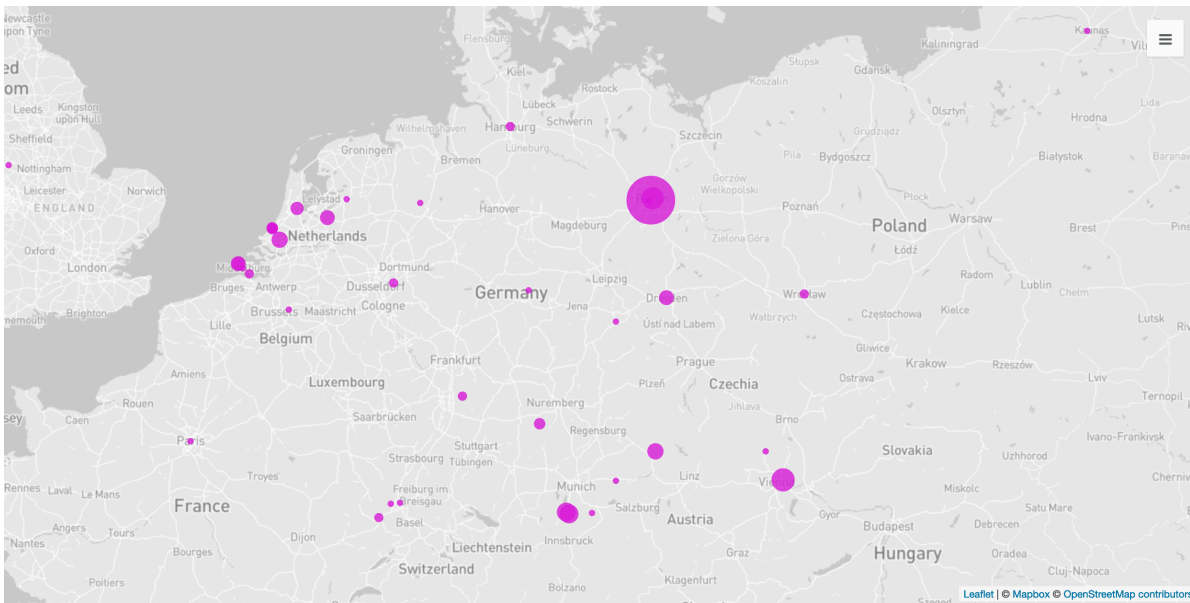


Figure 9.3: places mentioned in the corpus of letters by Franz Marc to Herwarth Walden, detail (created with Palladio).

For such processes you need a brief (though not necessarily briefly created) [script](#), that you

can change according to your aim or desired style of visualisation.¹ This can then be used with relatively little additional work for other documents – whether you are only interested in extracting and visualising places from the letters of Franz Marc or also those of Jacoba van Heemskerck, has no impact on the script’s computation time.

So, creating data sets according to specific principles, formal ones but also with a view to the [FAIR principles](#), has many benefits for both your own work – you need not formulate new data models by yourself – and for that of others – important basic information can be transferred, leaving more time for actual research.

A nice example for the reuse of data is the project [quoteSalute](#), created by students, a website on which you can generate historical salutations, in case you are getting bored with current standard forms. The project uses a number of XML encoded letter corpora, that are all online as openly usable data, for which they extracted and enriched the salutations. You can find the exact description [here](#).

The screenshot shows the homepage of the quoteSalute website. At the top left, there is an orange button labeled "New Greeting" and a "quote me!" button. Below these, a historical salutation is displayed: "»Adieu, Herr Graf, verzeihen Sie die Kritzelei und die Unordnung in meinem Briefe und denken Sie bisweilen an die Abwesenden.«". Underneath the salutation is a link to a document: "Lebenswelten, Erfahrungsräume und politische Horizonte der ostpreußischen Adelsfamilie Lehndorff vom 18. bis in das 20. Jahrhundert" by Prinzessin Amalie von Preußen. Below the document information is a "Filter Greetings" section with four columns: "From", "To", "Politeness", and "Language". Each column has radio button options. The "From" column has options for Female, Male, and Neutral. The "To" column has the same options. The "Politeness" column has options for Formal and Informal. The "Language" column has options for German, English, Spanish, Italian, French, Greek, and Latin. At the bottom of the filter section are "Apply Filters" and "Clear Filters" buttons. On the right side of the page, there are two informational boxes. The top one is titled "Creative Correspondence" and explains the website's purpose. The bottom one is titled "Contribute" and asks for user input.

Figure 9.4: Homepage of quoteSalute

As you can see, you can not only generate salutations, but also filter the search, for instance according to the gender of the sender or according to language. This is possible because

¹In this case the [csv file](#) was created with the help of the script in the link (written in the programming language R), and was [refined](#), with OpenRefine in order to add geocoordinates to the addresses. This [enriched spreadsheet](#) was uploaded into the online tool [Palladio](#) gwhich can place coordinates on a map.

the existing annotation of the text files has been extended in the project, for example with information on the gender of a person, if apparent. The code used is openly accessible on a [GitHub repository](#), and new corpora are welcome.

What this guide wanted to show is that computer-assisted and computer-based work can make the processes we can use for our historical research easier, quicker, sometimes even at all possible. There are nowadays countless programmes you can use via a graphic interface that are perfectly adequate for most requirements we have in the humanities. But some analyses require very specific steps or a lot of computing capacity for which it can be worth learning a programming language. Both for formulating an interesting research question and for the choice of data, as well as for the interpretation of the results given by the machine, you need informed expertise – a programme that should identify the author of an unknown text needs an incredibly long time if it wants to compare all existing texts with the unknown one. You need an expert, a literary scholar, to choose the relevant corpus it should refer to. Historical research is always also the analysis of the individual case, the particular, it is a close reading; the possibility of widening one's gaze with the assistance of the computer can in most cases be an interesting option.

A Glossary

Glossary

- **API: Application Programming Interface:** a facility offered by a web resource which allows search queries independent of a **GUI**, often performed using scripts
- **bash:** default program that runs in the **command line**
- **big data:** huge amount of data, identifiable through repeated freezing of your standard program when opening a file
- **born digital data:** data which originated in a digital form
- **CLI: Command Line Interface,** text interface that allows interaction with the computer; see also **bash**
- **CMS: Content Management System**
- **Console:** See **CLI**
- **Crowdsourcing:** projects that include the active participation of the public to generate content, transcribe sources etc.
- **csv: comma separated values,** a structured text format, using commas as separators between columns
- **distant reading:** quantitative approach to huge amounts of texts, using computational methods to search for interpretable patterns
- **GUI: Graphical User Interface**
- **HTML: Hypertext Markup Language,** a structured text format, like the format this guide is written in, to render documents in a browser
- **Jupyter notebook:** web application/interactive coding environment that runs in a browser; let's you create and share code (<https://jupyter.org>)
- **machine readable:** transformation of, for example, text into a data format that is processable by a computer
- **OCR: Optical Character Recognition,** process of transforming text on an image into a data format
- **OS: Operating System**
- **OSS: Open Source Software**
- **Regular Expression:** syntax for search and replace text using patterns (instead of exact matches)
- **terminal:** See **CLI**

B Literature, Tools, Tutorials

B.1 What is digital history?

- Antenhofer, Christina, Kühberger, Christoph, Strohmeier, Arno (Hrsg.): Digital Humanities in den Geschichtswissenschaften. Wien 2023.
- Brennan, Sheila A.: Digital History, in: The Inclusive Historian’s Handbook, <https://inclusivehistorian.com/digital-history/>, 04.06.2019.
- Hohls, Rüdiger: Digital Humanities und digitale Geschichtswissenschaften, in: Busse, Laura et al. (ed.): Clio-Guide. Ein Handbuch zu digitalen Ressourcen für die Geschichtswissenschaften, Berlin 2018, pp. A.1-1–B.1-34. Online: <https://doi.org/10.18452/19244>.
- Romein, C. Annemieke et al.: State of the Field: Digital History, in: History 105 (365), 04.2020, pp. 291—312. Online: <https://doi.org/10.1111/1468-229X.12969>.
- Winters, Jane: Digital History, in: Tamm, Marek; Burke, Peter (Hg.): Debating New Approaches to History, London 2019, pp. 277–300.
- Art. “Digital history”, in: Wikipedia, 07.09.2022. Online: https://en.wikipedia.org/w/index.php?title=Digital_history&oldid=1220662831, last updated: 25.04.2024.

B.2 Introductions and Guides

- Battershill, Claire; Ross, Shawna: Using Digital Humanities in the Classroom. A Practical Introduction for Teachers, Lecturers, and Students, London et al. 2022.
- Blaney, Jonathan et al.: Doing Digital History. A Beginner’s Guide to Working with Text as Data, Manchester 2021.
- Cohen, Daniel J.; Rosenzweig, Roy: Digital History. A Guide to Gathering, Preserving, and Presenting the Past on the Web, Philadelphia 2006. Online: <https://chnm.gmu.edu/digitalhistory/>.
- Döring, Karoline et al. (ed.): Digital History. Konzepte, Methoden und Kritiken Digitaler Geschichtswissenschaft, Boston 2022, Online: <https://doi.org/10.1515/9783110757101>.

- Dougherty, Jack; Nawrotzki, Kristen (Hg.): Writing History in the Digital Age, 2013. Online: <https://doi.org/10.3998/dh.12230987.0001.001>.
- Graham, Shawn et al.: Exploring Big Historical Data. The Historian's Macroscope, 2022. Online: <https://doi.org/10.1142/12435>.
- Lemerrier, Claire; Zalc, Claire: Quantitative Methods in the Humanities. An Introduction, Charlottesville 2019.
- Peroni, Silvio: [The CTP Book](#). A book for teaching Computational Thinking and Programming skills to people with a background in the Humanities.

B.3 Digital History Tools (free/open source)

B.3.1 General

- Programming Historian: [Tutorials for different tools and methods for historical teaching and research](#)

B.3.2 Data Analysis

- [Datasette](#): interactive tool for data exploration and visualisation
- [Heurist](#): web based research environment for data collection, exploration and visualisation
- [Nodegoat](#): web based research environment for data collection, exploration and visualisation

B.3.3 Maps

- [Dariah GeoBrowser](#): web based tool for the depiction of spatiotemporal data
- [QGis](#): software for creating, processing, depicting and analysing spatial information
- [Map in Seconds](#): web based tool for creating maps (optimised for Chrome)
- [Palladio](#): web based tool for visualising structural data, a.o. geodata

B.3.4 Network Analysis

- [Gephi](#): programme for analysing and visualising graphs and networks
- [Nodegoat](#): web based research environment for data collection, exploration and visualisation, a.o. network analysis
- [Palladio](#): web based tool for visualising structural data, a.o. network analysis

B.3.5 Text/Corpus Analysis

- AntConc: [tool kit for corpus analysis](#)
- Lemmatisation: [collection of the FID Romanistik](#)
- Natural Language Toolkit, package for Python for Tokenisation, Lemmatisation etc.: [NLTK](#)
- Tokenisation: [tutorial by fortext for NLTK](#)
- Voyant-Tools: [collection of tools for text analysis, browser based or standalone](#)

B.3.6 Storytelling

- [Omeka](#): web based platform for the curation of digital exhibitions, for exhibiting and publishing images/objects
- [StoryLine JS](#): tool for creating interactive diagrams
- [StoryMap JS](#): tool for creating interactive maps/images with various elements
- [Timeline JS](#): tool for creating interactive timelines

B.3.7 Visualisation

- Bostock, Michael; Heer, Jeffrey; Ogievetsky, Vadim: A Tour through the Visualization Zoo. A Survey of Powerful Visualization Techniques, from the Obvious to the Obscure, in: Queue 8, No. 5 (2010). Online: <https://queue.acm.org/detail.cfm?id=1805128>
- [Data Visualisation Catalogue](#): a guide to choosing forms of visualisation
- [Datasette](#): interactive tool for data exploration and visualisation
- FID Romanistik: [collection of tools for visualisation](#)
- [Palladio](#): web based tool for visualising structural data
- [RAWGraphs](#): tool for the visualisation of tabular data (.tsv, .csv, .dsv or .json-files)

B.4 Digital Literacy, Digital Criticism

- Ekström, Andreas: The Moral Bias behind your Search Results, TED talk 7.12.2015 (9:18), Online: https://www.youtube.com/watch?v=_vBggxCNNno.
- Gibbs, Frederick W.: New Forms of History: Critiquing Data and Its Representations, in: The American Historian, February 2016. Online: <http://tah.oah.org/february-2016/new-forms-of-history-critiquing-data-and-its-representations/>.
- Tavani, Herman; Zimmer, Michael: Search Engines and Ethics, in: Edward N. Zalta (Hg.): The Stanford Encyclopedia of Philosophy (Fall 2020 Edition), Online: <https://plato.stanford.edu/archives/fall2020/entries/ethics-search/>, ch. 3.1.

B.5 Programming Languages & Editors

- Python:
 - Althage, Melanie et al.: Python für Historiker:innen. Ein anwendungsorientierter und interaktiver Einstieg, 20.07.2022, [Jupyter Book, \(v1.0\)](#)
 - Karsdorp, Folgert: [Python Programming for the Humanities. Interactive Tutorial and Introduction to Programming with Python for the Humanities.](#)
 - Mattingly, William : [PythonHumanities.com. A Free Course for the Digital Humanities.](#)
 - Editor e.g. [PyCharm](#)
- R:
 - Mullen, Lincoln A.: [Computational Historical Thinking With Applications in R](#)
 - Phillips, Nathaniel D.: [YaRrr! The Pirate's Guide to R.](#)
 - Wickham, Hadley; Golemund, Garrett: [R for Data Science.](#)
 - Editor e.g. [RStudio](#)
- Editors: [Anaconda](#); [Sublime Text](#); [Visual Studio Code](#)

B.6 Regular Expressions

- Knox, Doug: Understanding Regular Expressions, Programming Historian 2 (2013), <https://doi.org/10.46430/phen0033>. (self-learning lesson)
- RegexOne: [Learn Regular Expressions with simple, interactive exercises.](#) (interactive self-learning tutorial)

B.7 Terminal/Command Line/Shell

- Dawson, Ted: Introduction to the Windows Command Line with PowerShell, Programming Historian 5 (2016), <https://doi.org/10.46430/phen0054>. (self-learning lesson)
- MIT Computer Science Department: [1-hour-lecture on the Shell](#) (video)
- Milligan, Ian; Baker, James: Introduction to the Bash Command Line, Programming Historian 3 (2014), <https://doi.org/10.46430/phen0037> (self-learning lesson)
- datacamp course: [Introduction to Shell](#) (interactive self-learning lesson)
- Jeroen Janssens: [Data Science at the command line](#) (book)

B.8 XML

- Latex Ninja Blog: [A shamelessly short intro to XML for DH beginners \(includes TEI\)](#) (blog post)

References

- [1] Jonathan Blaney et al. *Doing digital history: a beginner's guide to working with text as data*. eng. IHR research guides. Manchester: Manchester University Press, 2021. ISBN: 978-1-5261-3268-0.
- [2] Stephanie Russo Carroll et al. “The CARE Principles for Indigenous Data Governance”. en. In: *Data Science Journal* 19 (Nov. 2020), p. 43. ISSN: 1683-1470. DOI: [10.5334/dsj-2020-043](https://doi.org/10.5334/dsj-2020-043). URL: <http://datascience.codata.org/articles/10.5334/dsj-2020-043/> (visited on 11/28/2022).
- [3] Catherine D’Ignazio and Lauren F. Klein. *Data feminism*. English. OCLC: 1157171756. 2020. ISBN: 978-0-262-35852-1. URL: <https://direct.mit.edu/books/book/4660/Data-Feminism>.
- [4] Marie-José Kolly and Simon Schmid. “Sie ist hübsch. Er ist stark. Er ist Lehrer. Sie ist Kindergärtnerin”. de. In: *Republik* (Apr. 2021). URL: <https://www.republik.ch/2021/04/19/sie-ist-huebsch-er-ist-stark-er-ist-lehrer-sie-ist-kindergaertnerin> (visited on 08/23/2022).
- [5] Emmanuel Le Roy Ladurie. “La fin des érudits”. In: *Le Nouvel Observateur* (Aug. 1968).
- [6] Claire Lemercier and Claire Zalc. *Quantitative Methods in the Humanities. An Introduction*. Charlottesville: University of Virginia Press, 2019. ISBN: 978-0-8139-4270-4.
- [7] Chantel Ridsdale et al. *Strategies and Best Practices for Data Literacy Education Knowledge Synthesis Report*. Jan. 2015. DOI: [10.13140/RG.2.1.1922.5044](https://doi.org/10.13140/RG.2.1.1922.5044).
- [8] Roopika Risam. “It’s Data, Not Reality”: On Situated Data With Jill Walker Rettberg. en. June 2020. URL: <https://medium.com/nightingale/its-data-not-reality-on-situated-data-with-jill-walker-rettberg-d27c71b0b451> (visited on 08/16/2022).
- [9] C. Annemieke Romein et al. “State of the Field: Digital History”. en. In: *History* 105.365 (Apr. 2020), pp. 291–312. ISSN: 0018-2648, 1468-229X. DOI: [10.1111/1468-229X.12969](https://doi.org/10.1111/1468-229X.12969). URL: <https://onlinelibrary.wiley.com/doi/10.1111/1468-229X.12969> (visited on 09/15/2022).
- [10] Mark D. Wilkinson et al. “The FAIR Guiding Principles for scientific data management and stewardship”. en. In: *Scientific Data* 3.1 (Mar. 2016). Number: 1 Publisher: Nature Publishing Group, p. 160018. ISSN: 2052-4463. DOI: [10.1038/sdata.2016.18](https://doi.org/10.1038/sdata.2016.18). URL: <https://www.nature.com/articles/sdata201618> (visited on 11/09/2022).