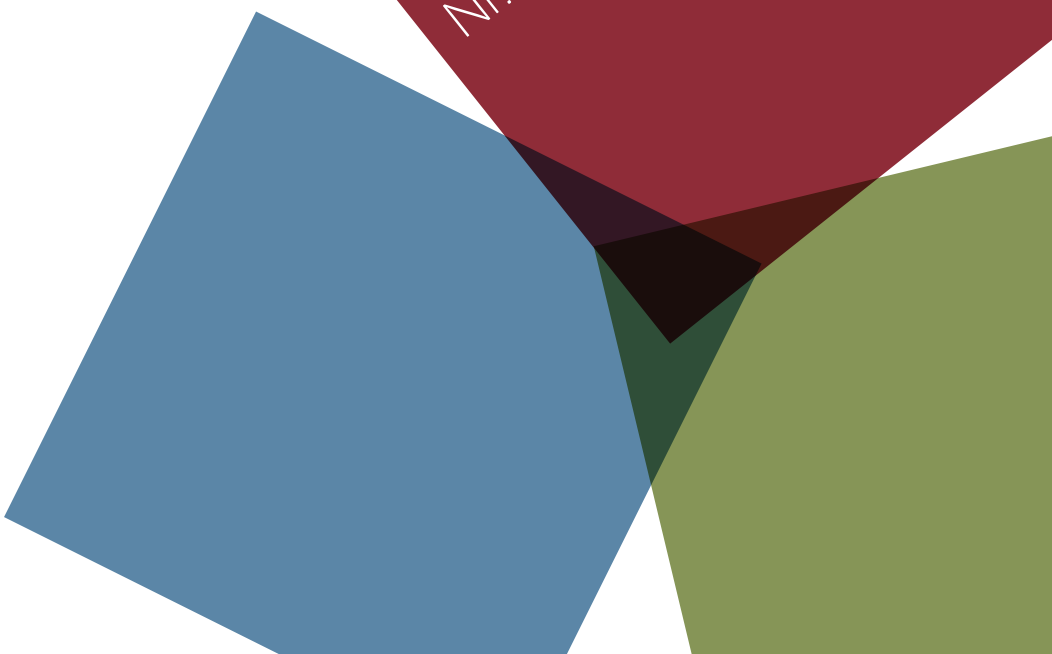
SFB 1199
Processes of Spatialization
under the Global Condition

Enock Seth Nyamador
Jana Moser
Philipp Meyer

Exploring school atlases: applying digital tools for visual data analysis and data management

Working paper series
des SFB 1199
an der Universität Leipzig
Nr. 36



Collaborative Research Centre (SFB) 1199
„Processes of Spatialization under the Global Condition“
at Leipzig University

Funded by

DFG Deutsche Forschungsgemeinschaft

Enock Seth Nyamador, Jana Moser, Philipp Meyer
Exploring school atlases: applying digital tools for visual data analysis and data management

This working paper is part of the Working Paper Series of the Collaborative Research Centre (SFB) 1199 “Processes of Spatialization under the Global Condition”. This working paper is also part of the Working Paper Series of ReCentGlobe, to which the SFB 1199 contributes since 2020.

10 / 2024

Vertrieb:
Leipziger Universitätsverlag GmbH, Oststraße 41, 04317 Leipzig
info@univerlag-leipzig.de

ISBN: 978-3-96023-538-5
ISSN: 2510-4845

Content

Abstract

Digital tools and computer programming are useful in easing and improving the speed and repeatability of outputs in social sciences and humanities research. Data visualisation plays an important role in getting insights into (large) datasets, communicating results and sharing knowledge amongst researchers. There exist several tools and software for data collection and visualisation but they are not always designed to fit all situations. In this rather technical working paper, we present some possibilities and advantages of using computer programming within the scope of a research project: (1) analysing quantitative datasets through means of visualisations produced within our research by (de)coding school atlases, and (2) data management for large sets of source-data, especially optimisation and embedding of coherent metadata in atlas scans to prepare for archiving and reuse. Together, we have developed an effective and efficient technical workflow for the processing, visualisation and management of our research data.

Keywords: *programming; digital humanities (DH) methods; visualisation; metadata*

1 Introduction

1.1 Aims and methods used in atlas analysis project

The project C05 "Maps and Atlases as Mediators and Producers of Spatial Knowledge under the Global Condition" examines maps as powerful media for imagining, establishing, reproducing and communicating spatial knowledge about the world and its parts in various scales. Being part of the Collaborative Research Centre (CRC) 1199 researching "Processes of Spatialisation under the Global Condition", we deal with a research frame and a defined terminology, such as spatial formats, spatial orders, or spatial semantics (Marung and Middell 2019), that are terms not used in our sources. These are geographical atlases for educational purposes, especially school and world atlases. We aim to find traces of how maps and atlases argue and communicate spatial concepts, if they visualise only concepts that are discussed in other disciplinary branches by spatial entrepreneurs (reproduction) or if mapmakers also included new ones, which we call spatial formats (production). To implement our research, we decided to collect and compare school atlases from mid-19th century until today while taking various world regions into account, such as French-speaking countries, the United States, Russia, China, and Germany. Our aim to at least rethink Eurocentric conditions using the method of reciprocal comparison (Austin 2007) is partly contradicted by focusing on regions producing maps and atlases. Even though school atlases are also used in African, South American and South-Asian countries, it is difficult to find atlases that were originally produced in these regions, not to mention series of atlas editions to trace changes within one region over a longer period.

Central to our research is data collected from analysing maps (Cherrier et al. 2019). In order to be able to compare a larger set of such atlases and maps, we developed a method to decode them, producing quantitative data that can be analysed using methods of digital humanities (DH). To achieve the research objectives of the CRC 1199, especially to apply concepts such as spatial formats or spatial orders and test how imaginations of space are visually communicated, we focused on the map content, map elements that determine the appearance of the maps such as projection, map orientation or central meridian, the graphical composition as well as map semiotics in the atlases. Using a semi-structured Excel sheet that has a comparatively low threshold for use, i.e. making it easy to use also for student assistants since there is no need for an elaborate introduction and familiarisation time. The map characteristics mentioned above are represented in columns in the Excel sheet. Each atlas map gets one line (row) and is (de)coded using in most cases only "0" for "not existing/no" and "1" for "existing/yes".

Despite the capabilities of Excel, the amount of data is large and not best explored and visualised within Excel. Based on these limitations and the need to create a reusable workflow, the project investigated options which lead to the decision to employ digital tools in analysis and visual exploration of the datasets.

Similarly, in order to keep records of some of the materials analysed, our project collected high-definition scans of maps for archival and reference purposes. Unfortunately, these high-definition scans are not well suited for regular on screen viewing as they are slower to open. They also do not have additional descriptive information (metadata). Therefore, despite applying a consistent file and folder naming system in the project, this is much prone to error since the files can be easily renamed defeating the purpose of storing referential information.

After introducing various definitions and tools used in section 1, we will describe in part 2 how we proceeded to solve our specific challenges in the C05 project.

1.2 Data processing

Data processing is the collection and manipulation of digital Data processing is the collection and manipulation of digital data to produce meaningful information (French 1996). Microsoft Excel, much like other bespoke software tools, can be used for data collection and presentation. As the data quickly grows, keeping track and ensuring repeatability of data presentation becomes tedious. It is in this regard that the C05 project, in which we developed a map coding scheme for school atlases, is employing computer programming for the creation of a reusable ecosystem for the management, processing, and presentation of data. In this technical paper, we describe using R, Python and command-line-based tools as utilised in our project for data processing and visualisations.

1.3 Data visualisation

Data visualisation in research can be likened to a powerful lens that allows researchers to view landscapes of data, uncovering patterns and insights that might otherwise remain hidden. By utilising a variety of visualisation tools, researchers can transform raw numbers into compelling narratives. According to Franconeri et al. (2021), effectively designed data visualisations allow viewers to use their powerful visual systems to recognise and understand patterns in data across several fields; science, education, health, and public policy. They also pointed out that ineffectively designed visualisations can cause confusion, misunderstanding, or even distrust—especially among viewers with low graphical literacy. But following Tufte, a statistician and a pioneer in the field of data visualisation, we want to highlight the aspects, that "often the most effective way to describe, explore, and summarize a set of numbers — even a very large set — is to look at pictures of those numbers". Tufte further argues that "all methods for analysing and communicating statistical information, well-designed data graphics are usually the simplest and at the same time the most powerful.[1] This in turn could empower researchers to not only convey their findings with clarity but also to foster a more profound appreciation of the underlying phenomena among their audience.

For us, data visualisation is a crucial part of the research process. It not only communicates findings but also acts as a catalyst for discovery and innovation (Takahashi et al. 2023). Hans Rosling, a pioneer in data visualisation, demonstrated the power of compelling visualisations in his 2006 TED talk "The Best Stats You've Ever Seen" (Rosling 2006). He believed that visualisations not only unveil existing knowledge but also inspire new hypotheses and further studies. By presenting data in an intuitive visual format, researchers can explore uncharted areas and uncover connections that would have otherwise remained unknown. Moreover, it allows researchers to re-visualise existing visualisations. The case of C05 is similar to the re-visualisation of maps, which reveals hidden information and trends that are not apparent by simple visual comparison of maps produced over time.

1.4 Metadata

The classical definition of metadata is "data about data". The National Information Standards Organization (NISO) in the United States defines metadata as structured information that describes, explains, locates, or otherwise makes it easier to retrieve, use, or manage an information resource. The organisation emphasises that metadata is a key to many systems from archival to large multinational businesses (NISO 2004).

1 Edward R Tufte (2001). "The visual display of quantitative information". In: vol. 2. Graphics press Cheshire, CT, p. 10.

Generally, metadata information is generated whenever a file is created or modified, but it is not always guaranteed to be complete enough to find or (re)use.[2] Furthermore, not all files created can have a meaningful set of metadata, i.e. in the case of photographed documents, metadata mostly will only contain information about the type of device used; software, date of creation/modification, and other information. In order to make the metadata of a digital file useful for the intended purpose with information such as title, author, etc. modification is sometimes required. However, we should also be aware that metadata on files can contain sensible data which may be used for unintended purposes. There are several standards for recording metadata, one example is Exif, which have been adopted in the case of this project.

The Exchangeable Image File Format (Exif), is a standard that specifies formats for images, sound, and ancillary tags used by digital cameras (including smartphones), scanners and other systems handling image and sound files recorded by digital cameras (Wikipedia contributors 2023).

1.5 Computer programming

Computer programming is the process of writing codes mostly with the aim of resulting in a fully functional software or program that performs a wide range of tasks. In the scope of this project, most of the programming related to this working paper is called *scripting*. Scripting involves writing codes in various programming languages that are meant to solve a specific task; they can be reusable or for one- time use. We have used a combination of different programming languages depending on the potential they have in completing various tasks.

1.5.1 R programming

R programming language provides a wide array of open-source packages tailored for data cleaning and wrangling, making them essential tools for researchers and data scientists. Some of these packages, including "dplyr" and "tidyr", simplify the process of transforming raw data into a structured format by offering functions for filtering, sorting, and reshaping data. They are particularly valuable for managing large and complex datasets where manual cleaning would be impractical. R's extensive library of packages covers nearly every aspect of data manipulation, enhancing its versatility. It also offers powerful data visualisation tools through packages like "ggplot2", which enables researchers to create customised and flexible visualisations such as scatter plots, bar charts, and heatmaps. Moreover, R seamlessly integrates with various data sources and software, allowing effortless data import from formats like XLSX, CSV, and SQL databases. This capability simplifies the integration of data collected from diverse sources into a unified analysis.

1.5.2 Python programming

Python programming offers a rich selection of open-source libraries[3] designed specifically for data cleaning and processing, making them indispensable resources for researchers and data scientists. Among these libraries, packages like "pandas" and "numpy" streamline the process of converting raw data into a structured format by providing functions for data filtering, sorting, and reshaping.

2 FAIR data are data which meet principles of findability, accessibility, interoperability, and reusability (FAIR), see www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/

3 In computer science, a library is a collection of read-only resources that is leveraged during software development to implement a computer program.

These tools prove particularly valuable when handling extensive and intricate datasets where manual cleaning would be unfeasible. Python is also a scripting language that can be used to automate processes and tasks that would have been manually performed.

The extensive ecosystem of Python's libraries covers virtually every aspect of data manipulation, enhancing its adaptability and utility. Additionally, it boasts of robust data visualisation capabilities via packages such as "matplotlib" and "seaborn", empowering researchers to create personalised and flexible visualisations, including scatter plots, bar charts, and heatmaps. Notwithstanding Python's capabilities for data processing visualisation as well, we resorted to R in our case as the team had a better leverage on R for data analysis and data visualisation.

1.5.3 Bash scripting

Bash, short for the Bourne Again Shell, is a powerful command line (is a means of interacting with a computer program by inputting lines of text called commands) programming language and environment commonly used in Unix-based operating systems. It enables users to automate tasks, manipulate files, and interact with the system through text-based commands and scripts. Bash scripts are sequences of commands that can perform a wide range of operations, from data processing and system administration to software automation and complex workflows (Perkel 2021). This versatility makes Bash scripting an essential tool for both beginners and experienced users, allowing them to enhance productivity, simplify repetitive tasks, and efficiently manage various aspects of computing and data processing.

1.6 Digital tools

Digital tools encompass a wide range of software programs and platforms designed to streamline collecting, analysing, and interpreting data. Example of these tools include programming languages like Python, R, and MATLAB, as well as specialized software for statistical analysis, data visualization, and simulation. By leveraging these digital tools, researchers can automate tasks, manipulate data, conduct experiments, and gain valuable insights to drive their research objectives forward.

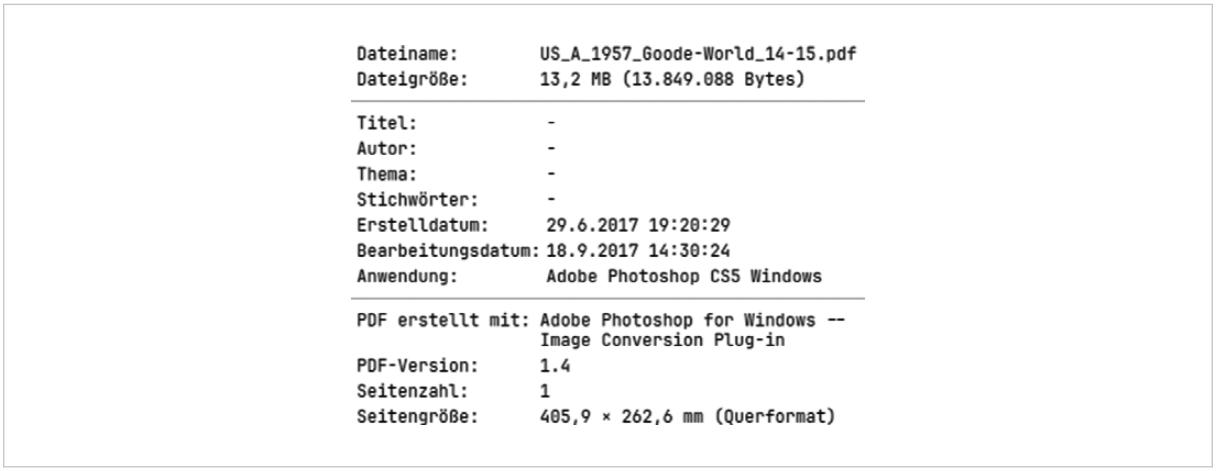
1.6.1 Ghostscript

Ghostscript is a widely used open-source software suite that interprets and renders PostScript and PDF files, making it a versatile tool for viewing and manipulating document formats. It enables the conversion of these files into other formats and provides extensive functionality for processing and manipulating document content programmatically.

1.6.2 Exif Tool

Exif Tool is a free and open-source software for reading, writing, and manipulating image, audio, video, and PDF metadata. Exif Tool is a cross-platform application and can be used on GNU/Linux, MacOS and Windows operating systems. The Exif Tool standalone command-line program enables users to use specific options known as *flags* much like other command-line based tools. These flags help in adding, modifying and deletion information from the metadata records. A basic command: `$ exiftool <filename>` where *filename* is the name of the file or location of file. This command displays all Exif key-value information embedded in the file including those that cannot be displayed in regular PDF

viewers i.e. viewing a PDF's metadata in a web browser will only display limited information similar to those in Figure 1.



Dateiname:	US_A_1957_Goode-World_14-15.pdf
Dateigröße:	13,2 MB (13.849.088 Bytes)
Titel:	-
Autor:	-
Thema:	-
Stichwörter:	-
Erstelldatum:	29.6.2017 19:20:29
Bearbeitungsdatum:	18.9.2017 14:30:24
Anwendung:	Adobe Photoshop CS5 Windows
PDF erstellt mit:	Adobe Photoshop for Windows -- Image Conversion Plug-in
PDF-Version:	1.4
Seitenzahl:	1
Seitengröße:	405,9 × 262,6 mm (Querformat)

Figure 1: Sample PDF file properties in a web browser

2 Applying digital tools in the C05 project

The C05 project used a combination of qualitative and quantitative methods. We started with a qualitative approach to understand the context of the atlases. This helped us gather quantitative data through decoding and analysis. We then used digital tools to process this quantitative data, which allowed us to identify interesting cases for more detailed analysis using quantitative techniques.

2.1 Using R for data processing and visualisation

The model proposed by Wickham, Çetinkaya-Rundel, and Grolemund (2023) as depicted in Figure 2 has been incorporated within the framework of this workflow.

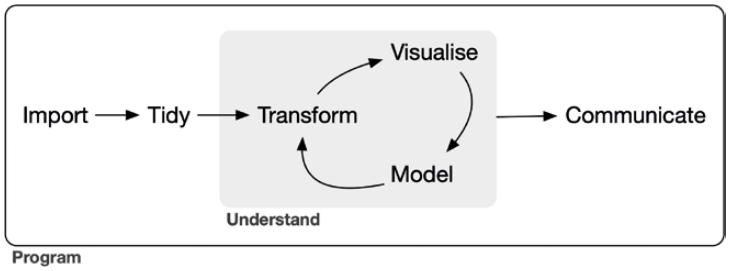


Figure 2: R for Data science model

In order to start working with our data, it needs to be imported. R can work with a wide range of data formats from single standalone files to data from relational databases (Wickham, Averick, et al. 2019). At the Leibniz Institute for Regional Geography (IfL) we have utilised Microsoft Excel for the collection of data from analysing graphical and cartographic information with maps using our developed coding scheme. Depending on the region of research, these spreadsheets can contains over 1500 records (rows) and 50 columns; there exist merged columns, which represent groupings for headers.

Cell merges make data "untidy". As shown in Figure 3, tidy data is a standard for structured data such that "each variable is a column, each observation is a row and each type of observational unit is a table" as described by (Wickham, Averick, et al. 2019).

2. Signs										
2.5 Thematic symbols										
2.5.1 Point symbols		2.5.2 Line Symbols								
2.5.1.A Symbol	2.5.1.B Diagram	2.5.2.A Territorial borders	2.5.2.B Flow lines (connector symbol)		2.5.2.C Arrows (connector symbol)		2.5.2.D Isolines	2.5.2.E Infra structure	2.5.2.F Boundary Lines	2.5.2.G Others
			2.5.2.B1 Simple	2.5.2.B2 Proportional	2.5.2.C1 Simple	2.5.2.C2 Proportional				

Figure 3: Initial table with cell column merges

After importing our spreadsheet file (`.xlsx`), each column is split, removing un-tidy data in form of merged cells that would hinder further processing. In order to obtain the affiliations of individual columns to their parent content, a reference table with correspondingly separated information was created in advance, as shown in Figure 4.

Code	Full Legend (as only use)	Sub-Heading (Title of Diagram)	Short Legend
2.5.1.A	Points: Symbols	Thematic Information: Points	Symbols
2.5.1.B	Points: Diagrams	Thematic Information: Points	Diagrams
2.5.2.A	Lines: National Borders	Thematic Information: Lines	National Borders
2.5.2.B.1	Lines: Flowlines Simple	Thematic Information: Lines	Flowlines Simple
2.5.2.B.2	Lines: Flowlines Proportional	Thematic Information: Lines	Flowlines Proportional
2.5.2.C.1	Lines: Arrows Simple	Thematic Information: Lines	Arrows Simple
2.5.2.C.2	Lines: Arrows Proportional	Thematic Information: Lines	Arrows Proportional
2.5.2.D	Lines: Isolines	Thematic Information: Lines	Isolines
2.5.2.E	Lines: Infrastructures	Thematic Information: Lines	Infrastructures
2.5.2.F	Lines: Non-National Boundary Lines	Thematic Information: Lines	Non-National Boundary Lines
2.5.2.G	Lines: Others	Thematic Information: Lines	Others

Figure 4: Coding reference table

The process continues to eliminate empty cells until our data is tidy enough. The data is transformed further by adding references to columns and removing texts from number only columns. Similarly, other data types such as dates can be converted into dates that R understands, which allows to easily compute time-based results. It should be noted that these processes and changes are made in close consultation and communication with the researchers only.

Some further transformations employed are aggregation, summation and grouping of fields specific to our needs as well as questions to be answered.

Finally, a combination of columns can be easily selected for visualisation based on the kind of visualisation best suited or attributes to be explored and compared. The following subsections showcases of stacked bar charts and line graphs. These charts and graphs are to be understood as functional visualisations for analysing the data and are not intended as a final visual representation of results.

Stacked bar charts

Stacked bar charts can be used for the representation of proportion of related values within maps over time. Leveraging the *ggplot2* package in R, the creation of a stacked bar chart typically begins with loading the necessary library and organising the data into a format compatible with the aesthetics of ggplot. Using the `ggplot()` function, we can specify the data, aesthetics, and the initial chart type.

In the case of stacked bar charts, the `geom_bar()` function is used, and the fill aesthetic is set to a categorical variable that defines the different segments. By adding the `geom_bar(stat = „identity“)` argument, the bars are stacked on top of each other based on the specified fill variable. Additional customization options, such as color, labels, and themes, can be incorporated to enhance the clarity and visual appeal of the chart. As an example, Figure 5 illustrates the proportion of four categories of classification for graphics as recorded from maps of the German Diercke atlases using our map coding scheme.

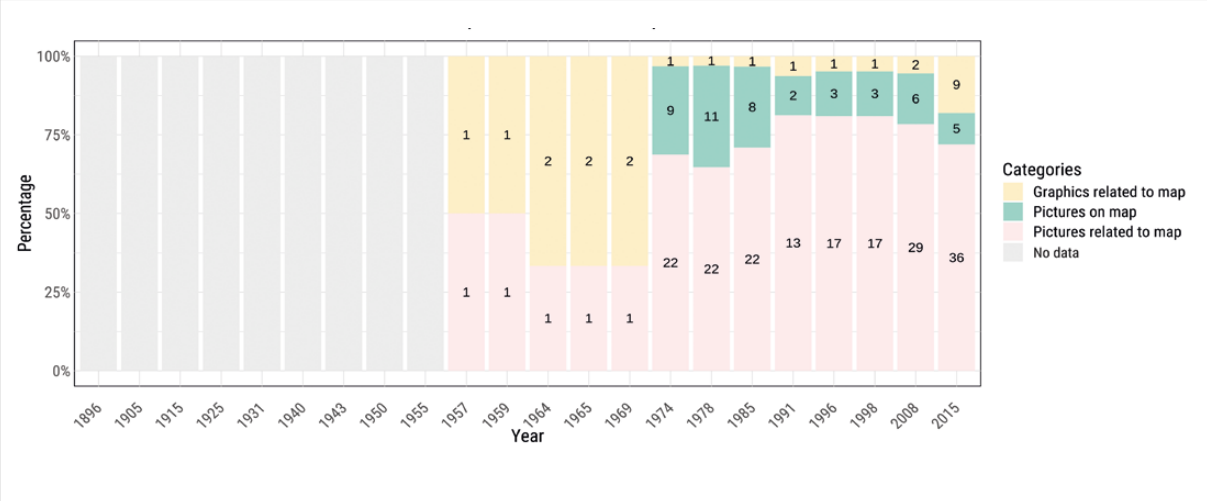


Figure 5: Stacked bar charts of pictures and graphics analysis. Before 1957 there were no graphics or pictures at all in Diercke's school atlases.

Line graphs

Line graphs are used to illustrate and compare the variations of selected variables across various time frames.

Similar to the stacked bar charts, line graphs are also based on the ggplot2 package. To create line charts with ggplot2 in R, the `geom_line()` function is utilized. Starting with the `ggplot()` function to set up the basic plot structure, users specify the data and aesthetics, including the x and y variables. The `geom_line()` function is then added to generate the line chart. The resulting plot connects data points with lines, providing a visual representation of the trends or relationships in the dataset. Additional features, such as color, line type, and labels, can be incorporated for further customization.

Figure 6 shows the variations of line symbols used in German Diercke school atlases over the years from 1896 to 2015.

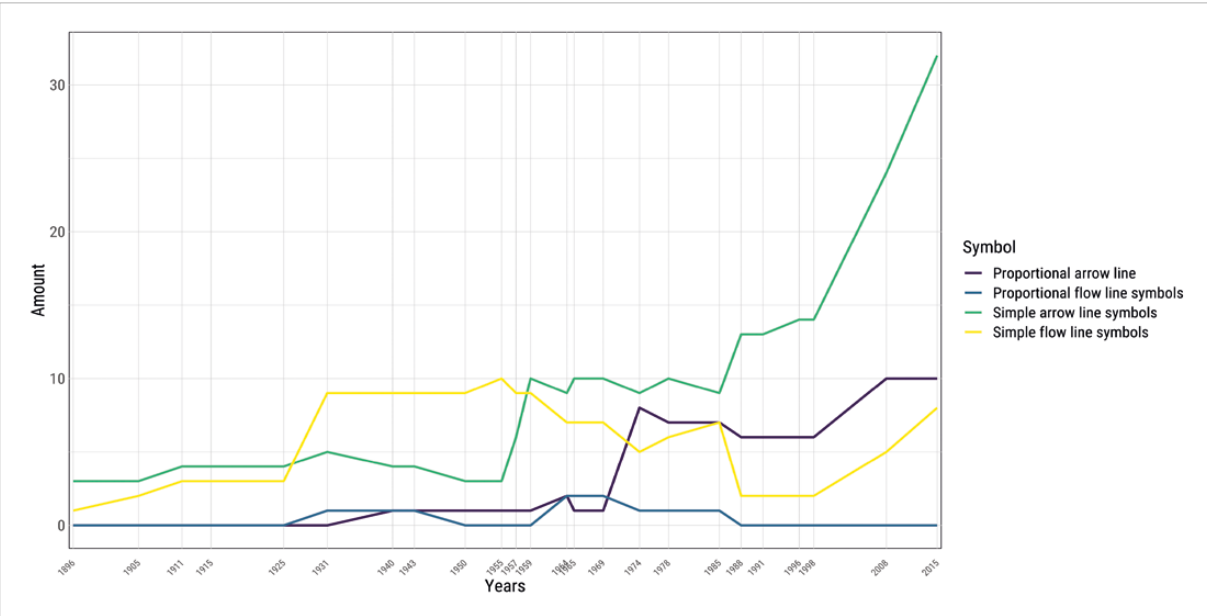
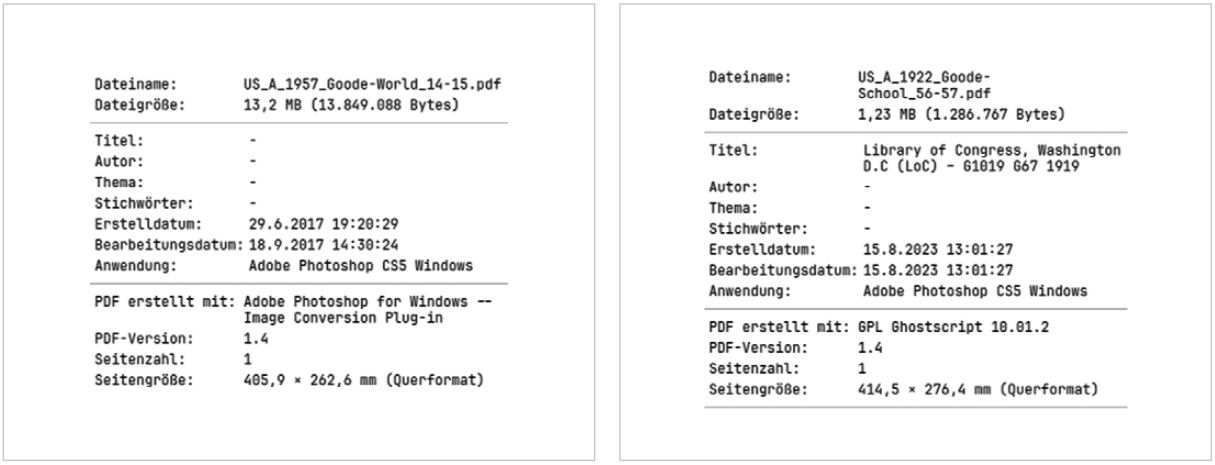


Figure 6: Aggregated line symbol dynamics for Diercke atlases by years

2.2 Optimising scanned documents

As part of the C05 project, series of maps were scanned or captured using a smartphone from different physical libraries. Scanned files with high dots per inch (DPI) for archival purposes can result in large amounts of storage space. On the other hand, compression of images can sometimes result in losing visual appearance.

We utilised a command-line tool "Ghostscript" to ensure good visual details whilst reducing size and at the same time improving screen display speed. Compression of large, high-definition scanned files to achieve unnoticeable visual difference in the content is useful for several reasons. Firstly, it solves the topmost problem of minimising storage space used, which is especially crucial in this era where digital documents are abound. Figure 7 shows the comparison of properties between the original scan and after compression as viewed from Mozilla Firefox web browser.



(a) Original scan metadata

(b) Compressed file metadata

Figure 7: Metadata before and after compression showing file sizes

Looking closer at the scanned files, for example those belonging to the Goode's School Atlas of 1922, a drastic reduction in the file sizes were noticed. For example, the scanned sized of *Page 94-95* was initially **108 MB**. However, after using Ghostscript from the terminal with option `-dPDFSETTINGS=/default`, the file's size reduced drastically to **1.1 MB** and improved opening time as well. Figure 8 shows a line graph which illustrates the file sizes before and after compression for files from 1922 Goode school atlas.

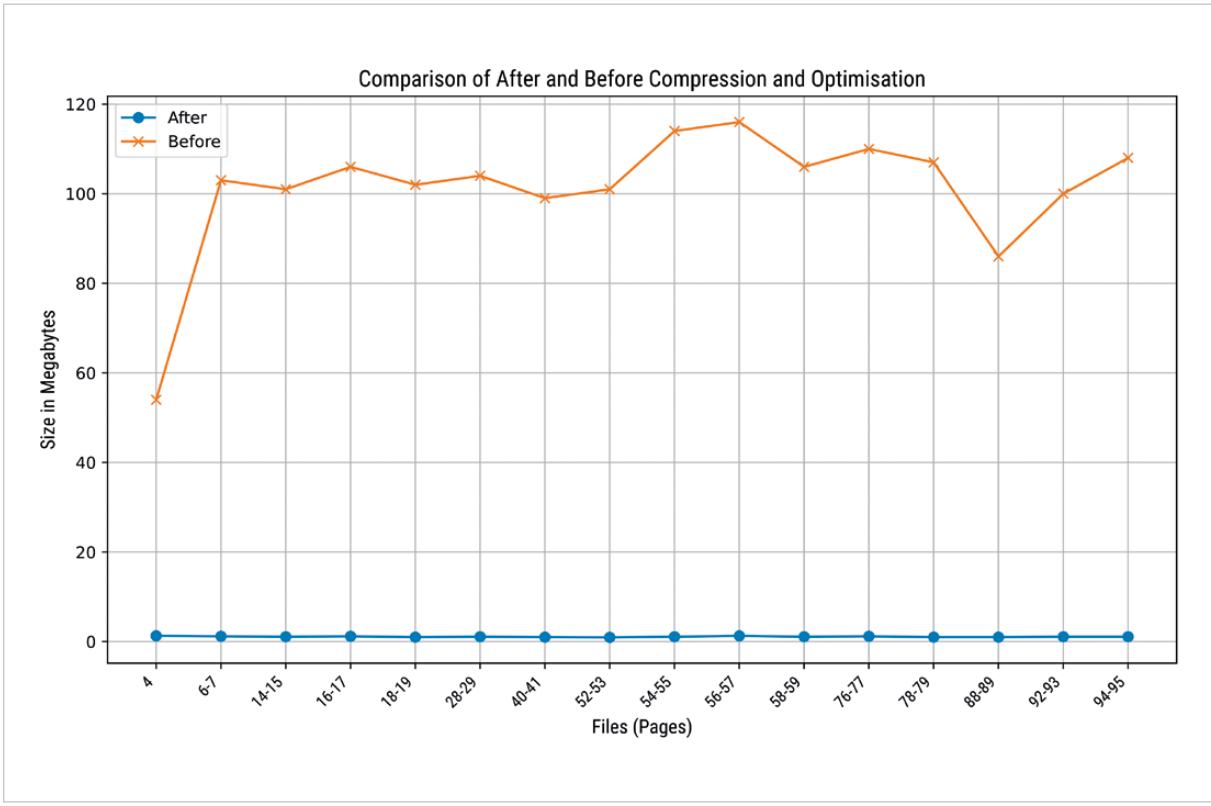


Figure 8: Graph of file sizes before and after compression & optimisation (example of Goode's School Atlas of 1922)

Secondly, compression also enables faster loading and viewing of files on various digital platforms and devices, which in turn enhances user experience, accessibility and file sharing. The scanned files were however kept in their original single-file states for backup and archival purposes. The optimised and compressed versions on the other hand will be used for on-demand research purposes. Comparing screenshots from the original uncompressed scans to its corresponding compressed version both at the same zoom level 400%, little or no visual differences could be identified as shown in Figure 9.

(a) Original scan

(b) Compressed and optimised file

Figure 9: Visual comparison of files

2.3 Metadata management

Metadata of files can be seen either by viewing their properties or from the application used in creating the file. In this case, we are dealing with PDF files. Hence, the metadata in question is accessible within a PDF reader application such as Adobe Reader, or a modern web browser such as Firefox, Chrome, Edge, etc.

Using metadata provides far better description than mere use of filenames since the information is embedded in the files and not dependent on the files naming. Additionally, more information can be added that helps referencing the single PDF files, such as atlas edition, year of printing, and the library with respective shelfmark, where the copy or photo was taken. This embedded information can be used and read by both humans and computers.

In order to add metadata, we have identified which fields within the metadata records is of interest to us. The project decided to use the *title* field in the metadata to store all our relevant information. The information stored includes (i) the name of the library, (ii) scanned items' collection ID or number i.e. "*Library of Congress, Washington D.C (LoC) – G1019 G67 1919*". The processing is accomplished with a *Python* script (*see Annex I*) that goes through each folder and looks for defined patterns in the filenames. Together with the Exif Tool, the title field of the metadata is updated based on the matching key i.e. pattern value from a dictionary of predefined values. Figure 10a shows initial metadata for the file *US_A_1957_GoodeWorld_1415.pdf* compared to Figure 10b which represents the information after running the script.

Dateiname:	US_A_1957_Goode-World_14-15.pdf
Dateigröße:	13,2 MB (13.849.088 Bytes)
Titel:	-
Autor:	-
Thema:	-
Stichwörter:	-
Erstelldatum:	29.6.2017 19:20:29
Bearbeitungsdatum:	18.9.2017 14:30:24
Anwendung:	Adobe Photoshop CS5 Windows
PDF erstellt mit:	Adobe Photoshop for Windows -- Image Conversion Plug-in
PDF-Version:	1.4
Seitenzahl:	1
Seitengröße:	405,9 × 262,6 mm (Querformat)

(a) Before

Dateiname:	US_A_1957_Goode-World_14-15.pdf
Dateigröße:	13,2 MB (13.868.359 Bytes)
Titel:	Library of Congress, Washington D.C (LoC) - G1019 G67 1957
Autor:	-
Thema:	-
Stichwörter:	-
Erstelldatum:	29.6.2017 19:20:29
Bearbeitungsdatum:	18.9.2017 14:30:24
Anwendung:	Adobe Photoshop CS5 Windows
PDF erstellt mit:	Adobe Photoshop for Windows -- Image Conversion Plug-in
PDF-Version:	1.4
Seitenzahl:	1
Seitengröße:	405,9 × 262,6 mm (Querformat)

(b) After

Figure 10: File information

3 Discussion and Conclusions

R, Python programming and other digital tools that have been described in this working paper are just a few of the many resources available that can be used by researchers for data processing and visualisation. Data visualisation is not just a means of presenting findings; it is a dynamic tool that advances the scientific process forward, throwing light on the hidden domains of data, especially larger datasets, and pushing research into new frontiers. Command-line-based tools offer great power and flexibility in manipulating and updating files.

Despite or perhaps because of the ease of use, we would like to emphasise that the use of such digital possibilities requires constant reflection on the part of the researchers. This includes, for example, constantly scrutinising the process: What does the data contain and what results do I expect? For example, are unexpected results meaningful or do they result from incorrect initial data or even incorrect or unintentional correlations of data? Does the visualisation suggest causalities that make no sense in terms of content? Digital tools, just like visual media, are tempting to be used quickly, even when communicating research results to third parties. Precisely because such results can be interpreted very differently depending on the perspective, researchers have a special duty to take care and scrutinise the results.

However, programming and the use of command-line tools require some technical skills and motivation for its incorporation into a research project. If collaboration with professional programmers is not an option, there are several resources for learning these digital skills created by both individuals and institutions in various formats. Programming Historians[4] is one of such unique initiatives by historians for historians offering peer-reviewed and quality digital skills articles in a tutorial format.

The programming in the C05 project is still in its early stages. It is our goal to publish the source codes via the GlobeData[5] repository provided by the Research Centre Global Dynamics at Leipzig University where also the map coding scheme will be shared in the near future. Notwithstanding, this is a proof of concept in use and can be applied in other projects.

Acknowledgements

This working paper is part of the Collaborative Research Centre (SFB) 1199, project C05 "Maps and Atlases as Mediators and Producers of Space (Knowledge) under the Global Condition", funded by the German Research Foundation (DFG). We thank Ihor Doroshenko for conceiving the initial R programming that have served as foundation for our further developement and use. We also thank Gina-Loreen Seydler for critically reading and commenting on a first version of this text.

4 https://programminghistorian.org/en/about
5 https://globedata.uni-leipzig.de/

Bibliography

Austin, Gareth (2007). "Reciprocal comparison and African history: tackling conceptual Eurocentrism in the study of Africa's economic past". In: *African Studies Review* 50.3, pp. 1–28.

Cherrier, Pierre et al. (2019). *Raumformate und Kartensprachen erkennen: Vorschlag einer Methodik zur Analyse von Karten und (Schul) Atlanten als Vermittlern von Weltbildern unter Globalisierungsprozessen*. Leipziger Universitätsverlag, SFB 1199 Working Paper Series.

Franconeri, Steven L. et al. (2021). "The Science of Visual Data Communication: What Works". In: *Psychological Science in the Public Interest* 22.3. PMID: 34907835, pp. 110–161. doi: 10.1177/15291006211051956. eprint: https://doi.org/10.1177/15291006211051956. url: https://doi.org/10.1177/15291006211051956.

French, Carl (1996). *Data processing and information technology*. Cengage Learning EMEA.

Marung, S and M Middell (2019). "The respatialization of the world as one of the driving dialectics under the global condition". In: *Spatial formats under the global condition: Dialectics of the global* 1, pp. 1–11.

NISO (2004). "Understanding metadata". In: *Washington DC, United States: National Information Standards Organization*, p. 1., url: www.lter.uaf.edu/metadata_files/UnderstandingMetadata.pdf.

Perkel, J. M. (2021). "Five reasons why researchers should learn to love the command line". In: *Nature* 590, p. 173.

Rosling, H (2006). "The best stats you've ever seen. TED talk". In: *Retrieved December 2023* 30, p. 2019.

Takahashi, Keisuke et al. (2023). "Catalysts informatics: paradigm shift towards data-driven catalyst design". In: *Chemical Communications* 59.16, pp. 2222–2238.

Tufte, Edward R (2001). "The visual display of quantitative information". In: vol. 2. Graphics press Cheshire, CT, p. 10.

Wickham, Hadley, Mara Averick, et al. (2019). "Welcome to the Tidyverse". In: *Journal of Open Source Software* 4.43, p. 1686. doi: 10.21105/joss.01686. url: https://doi.org/10.21105/joss. 01686.

Wickham, Hadley, Mine Çetinkaya-Rundel, and Garrett Grolemund (2023). *R for data science*. „O'Reilly Media, Inc.".

Wikipedia contributors (2023). *Exif — Wikipedia, The Free Encyclopedia*. [Online; accessed 1-October-2023]. url: https://en.wikipedia.org/w/index.php?title=Exif&oldid=1168568253.

Annex I

```
#!/ usr / bin / env
python import os
import re
import subprocess
from rich . progress import Progress

# Sample dictionary with patterns and corresponding titles
title_map = {
# ....... Reduced list
    ' US_A_ 1932 ': ' Library of Congress ( LoC ), Washington D. C - G1019
G67 1932 ' # ........
}

def count_files ( root_directory ):
    count = 0
    for foldername , _, filenames in os. walk (
        root_directory ): count += len ( filenames )
    return count

def update_pdf_titles ( root_directory , overwrite ):
    total_files = count_files ( root_directory )
    processed_files = 0

    with Progress () as progress :
        task = progress . add_task ("[ cyan ] Processing files ... ", total =
        total_files ) for foldername , subfolders , filenames in os. walk (
        root_directory ):
            for filename in filenames :
                filepath = os. path . join ( foldername , filename )

                # Check if the filename contains a pattern from the dictionary
                for pattern , title in title_map . items ():
                    if re. search ( pattern , filename ):
                        # Use exiftool to update the PDF
                        title cmd = f' exiftool - Title ="{
                        title }" '
                        if overwrite :
                            cmd += ' - overwrite_original '

                        cmd += f' "{ filepath }" '
                        subprocess . run ( cmd , shell =True , stdout = subprocess . PIPE ,
                            stderr = subprocess . PIPE )
                processed_files += 1
                progress . update ( task , completed = processed_files )

if __name_ == ' __main
    ': import     click
    @ click . command ()
    @ click . argument (' root_directory ', type = click . Path ( exists = True ),
    default =os. getcwd ()) @ click . option ('-- overwrite ', is_flag =True , help ='
    Enable to overwrite original files .') def main ( root_directory , overwrite ):
        update_pdf_titles ( root_directory , overwrite

    ) main ()
```

Listing 1: Python script for adding 'title field' in metadata

Leipzig Research Centre Global Dynamics

Working paper series des SFB 1199 an der Universität Leipzig No. 36

ISBN: 978-3-96023-538-5
ISSN: 2510-4845

Universität Leipzig
SFB 1199

E-Mail: sfb1199@uni-leipzig.de

http://research.uni-leipzig.de/~sfb1199

UNIVERSITÄT LEIPZIG
Leipzig Research Centre Global Dynamics

Leibniz-Institut für Länderkunde

GWZO Leibniz-Institut für Geschichte und Kultur des östlichen Europa

TECHNISCHE UNIVERSITÄT DRESDEN

Funded by
DFG Deutsche Forschungsgemeinschaft