Data management for historians

ECTS	5 EC
Level	M (Master)
Language	English
Venue	Online
Period	Semester 2 (February- April 2021)
Lecturer	Multiple guest lecturers • Wouter Beek • Albert Meroño-Peñuela • Rick Mourits • Björn Quanjer • Auke Rijpma • Ruben Schalk • Richard Zijdeman
Email	r.j.mourits@ru.nl

Table of contents:

1. General information

Course Description

Researchers in economic and social history regularly work with historical records, such as prices, wages, or demographic data. These historical data are studied to produce tables, graphs, and statistical analyses to get insight in historical processes. An often invisible, but crucial part of the quantitative research cycle is the transformation of rudimentary data into intelligible information. Data management and data manipulation are required to explore, clean, and operationalise otherwise unintelligible data. For example, counting the number of spelling variations in historical occupations, cleaning year entries, or grouping information from different sources into one new variable.

To allow master students within economic and social history to process their own rudimentary data and enable reuse according to the FAIR data principles[1], this course instructs them on the basics of quantitative data management. It introduces the quantitative research cycle, programming in R and SPARQL, and how to report on data questions. Combined, these skills will allow the participants to understand quantitative research better, make efficient and reproducible enquiries on the data, and clearly report on data questions. As such, the course prepares participants for quantitative research in general.

Learning Objectives

The aim of this course is to provide research-oriented historians with a thorough introduction into data management and introduce them to R and Linked Data. As such, the course provides essential background for historians who work with quantitative data and allows participants to understand, organise, and report on data management processes.

To do so, the course concentrates on four objectives:

- Understanding the quantitative research cycle
- Critical evaluation of quantitative data
- Basic understanding of programming in R and SPARQL
- Being able to report on data management

In addition, the course emphasizes transparent management by using platforms like GitHub and GRLC to share data management scripts.

Course Format

The course takes a hands-on-approach: the weekly meetings will be part lecture, part tutorial, and part self-study hours. The lectures will deal with the course literature and the weekly topics. Furthermore, the lectures will be used to emphasize the connections between the weekly lectures.

[1] www.go-fair.org/fair-principles/

Participants are required to prepare the literature in advance to facilitate active participation during the lectures. There is one weekly lecture of two hours (two times 45 minutes).

Literature

The required reading per lecture is listed in the course schedule. Overall, we will predominantly draw from open access books and articles.

Ducharme, B. (2013). *Learning SPARQL: Querying and updating with SPARQL 1.1.* Newton Massachusetts: O'Reilly Media. (Chapter 1-2, 5, 13)

Grolemund, R. & Wickham, H. (2017). *R for Data Science*. Newton Massachusetts: O'Reilly Media. http://r4ds.had.co.nz (Chapter 1-3).

Heath, T. & Bizer, C. (2011). *Linked Data: evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology.* Williston (VT): Morgan & Claypool Publishers. (Chapter 1-2) http://linkeddatabook.com/editions/1.0/

Quanjer, B. & Kok, J. (2020). Drafting the Dutch: Selecting biases in Dutch conscript records in the second half of the 19th century, *Social Science History, 44*(3), 501-524. doi: 10.1017/ssh.2020.13

Van den Boomen, N. & Ekamper, P. (2015). Denied their 'natural nourishment': religion, causes of death and infant mortality in the Netherlands, 1875-1899, *the History of the Family, 20*(3), 391-419. doi: 10.1080/1081602X.2015.1022199

Wickham, H., Chang, W., Henry, L., Lin Pedersen, T., Takahashi, K., Wilke, C., Woo, K., Yutani, H. & Dunnington, D. (2020). *Create elegant data visualisations using the grammar of graphics: ggplot2*. https://ggplot2.tidyverse.org/index.html

Wilkinson, M.D. et al (2016) The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data, 3*(160018). doi: 10.1038/sdata.2016.18

Attendance and participation

Attending and preparing for lectures is mandatory. Participants are expected to be active participants. If you cannot attend a lecture, make sure to inform the lecturer, preferably before the meeting. Due to illness or other types of force majeure you may unfortunately not be able to attend one or more meetings, but after you miss two lectures the course coordinator will email you with a cc to the Posthumus programme director. The study advisor will check if you had sufficient reason to be absent, determine whether leniency should apply, and advise the course coordinator accordingly. Using that advice, the course coordinator will be entitled to take appropriate measures, such as giving an extra assignment, disallowing a retake of the exam, or entirely excluding the participant from the course.

2. Assessment

The performance in the course will be assessed on the basis of two assignments:

Mid-term assignment

In the first 4 weeks, participants will learn about the basics of data manipulation and data reporting. After taking their first steps in R, participants will have to create a data report in which they structure data and present their output in a systematic way. This data report has to follow the guidelines in the data report manual, which is available to the participants on Surfdrive:

1. Picking an example dataset.
2. Transform data using R.
3. Visualise the data.
4. Summarise your conclusions from the enquiry.

The mid-term assignment counts for 30% of your course grade

Final assignment

At the end of the course, participants will have learned how to structure, visualise, and report on their data exploits. The final assignment will require the participants to apply all the acquired skills from the course in a brief, data-focused paper, which shows that they understand the Quantitative Research Cycle.

Participants will choose one of the available datasets provided and use this dataset to:

1. Formulate a research question which can be answered with the data.
2. Explain why this research question is relevant.
3. Discuss the data processing, by:
    a. Applying source criticism
    b. Showing how they transformed the data
4. Visualise the data
5. Write a short conclusion (1,000 words max) answering the research question, explaining conclusions, and discussing the pros and cons of the data. Special attention should be given to the question how others can make use of your findings / scripts.

The final assignment counts for 70% of your course grade

3. Course schedule

Overview

In this course, we will take our first steps in the world of data management. Over the course of 9 weeks, we will get acquainted with the quantitative research cycle, write research reports, and learn to program in two state-of-the-art environments: R and Linked Data.

As data management, manipulation, and visualization are arts that can only be mastered with time, we will take a hands-on approach in these issues. The course is structured in two blocks of four weeks.

- *In the first lecture we introduce the principles of data management, so that you can understand, structure, and critically reflect on each step of the quantitative research cycle.*
- *The second and third lecture focus on data manipulation to get hands-on experience with the R and Linked Data environment.*
- *The final lecture focuses on visualising and reporting data, so that you can report your findings elegantly and efficiently to your readers and/or supervisors.*

Course overview

	Date	**Topic**	**Teacher**
Block 1	4 February 11 February 18 February 25 February	Week 1: General introduction Week 2: Introduction to data management in R, 1 Week 3: Introduction to data management in R, 2 Week 4: Presenting your data	Ruben Schalk Rick Mourits Rick Mourits Björn Quanjer
Ass. 1	5 March	Mid-term assignment (30% of course grade)	
Block 2	11 March 18 March 25 March 1 April	Week 5: Quantitative Research Cycle Week 6: Introduction to Linked Data, 1 Week 7: Introduction to Linked Data, 2 Week 8: Presenting your Linked Data	Auke Rijpma Wouter Beek Albert Meroño-Peñuela Richard Zijdeman
Ass. 2	9 April	Week 9: Working on final assignment	-
	TBD	Final assignment (70% of course grade)	2-3 teachers

Week 1: General introduction

During the course, we will perform R and Linked Data assignments on this data to get acquainted with the challenges of quantitative research. In this meeting, we introduce the quantitative research cycle which helps to structure analytical enquiries. In the lecture, we will discuss the research paper by Van den Boomen & Ekamper as a practical example of how quantitative data explorations can take place in a historical context.

Teacher:

Ruben Schalk

Required reading:

Wilkinson, M.D. et al (2016) The FAIR Guiding Principles for scientific data management and stewardship, *Scientific Data, 3*(160018). doi: 10.1038/sdata.2016.18.

Van den Boomen, N. & Ekamper, P. (2015). Denied their 'natural nourishment': religion, causes of death and infant mortality in the Netherlands, 1875-1899, *the History of the Family, 20*(3), 391-419. Doi: 10.1080/1081602X.2015.1022199

Topics of this week:

- Introduction to data management
- Data sharing and reproducibility
- Moving from qualitative data/research to quantitative data/research
- Source criticism as part of quantitative data management

Learning goals:

- Get acquainted with the quantitative research cycle
- Learn which questions belong where in the quantitative research cycle
- Learn the appropriate outlets for help
- Participants are able to formulate their own data problem

Week 2: Introduction to data management in R, 1

This week, we start with programming in R. However, before we start, we reflect on last week's lecture. Participants are required to bring a short note—maximum 300 words—in which they briefly outline: which exact data questions they have (and/or expect to have) and how these questions fit in the quantitative research cycle. Furthermore, we focus on the basics of programming and the challenges of importing and exporting data, which causes way more errors inside and outside academia than you might think.

Teacher:

Rick Mourits

Required reading:

Grolemund, R. & Wickham, H. (2017). *R for Data Science*. Newton Massachusetts: O'Reilly. http://r4ds.had.co.nz (Chapters 1.1 & 2).

Additional literature:

Venables, W.N., Smith, D.M. et al (2020). *An introduction to R: A programming environment for data analysis and graphics*. https://cran.r-project.org/doc/manuals/r-release/R-intro.pdf

Topics of this week:

- Discussion of assignments
- First introduction to R

Learning goals:

- Participants can apply the quantitative research cycle to their own data questions
- Participants learn to critically reflect on their own data problem
- Participants are introduced to the basics of R
- Participants learn how to import and export data into the R environment

Week 3: Introduction to data management in R, 2

In week 3, we focus more on R and discuss data manipulation, relational tables, and how to join tables. However, our most important lesson this week is that you are not alone! Most of the problems you will have to deal with, others have had to solve as well. Therefore, we introduce you to the online platform of Stack Overflow, which is your best friend when you are in an R pickle.

Teacher:

Rick Mourits

Required reading:

Grolemund, R. & Wickham, H. (2017). *R for Data Science*. Newton Massachusetts: O'Reilly. http://r4ds.had.co.nz (Chapter 5 & 13).

Topics of this week:

- Importing and exporting data in R
- Programming in R

Learning goals:

- Participants are able to do data manipulation in R
- Participants can use the proper online platforms for feedback

Week 4: Presenting your data (data visualisation in R)

Data manipulation is generally not a goal in itself, but required to answer research questions. One of the hardest things in data science is to report results of data manipulation, so that you can discuss them with others, for example your supervisor or a colleague. This week we will show you how to write a data report: a short overview of (1) your research question, (2) one table or figure to answer your question, and (3) a short conclusion. This format will allow you to have in-depth discussion on your work with someone who is not as deeply invested in the data as you are.

Teacher:

Björn Quanjer

Required reading:

#TidyTuesday https://twitter.com/tidypod

Data report guide

Grolemund, R. & Wickham, H. (2017). *R for Data Science*. Newton Massachusetts: O'Reilly. http://r4ds.had.co.nz (Chapter 3).

Additional literature:

Wickham, H., Chang, W., Henry, L., Lin Pedersen, T., Takahashi, K., Wilke, C., Woo, K., Yutani, H. & Dunnington, D. (2020). *Create elegant data visualisations using the grammar of graphics: ggplot2*. https://ggplot2.tidyverse.org/index.html

Topics of this week:

- Writing data reports
- Visualising data in R

Learning goals:

- Participants are able to structure their research
- Participants can present their output in a systematic way
- Participants are able to link research questions to specific research output

Mid-term assignment

In the first 4 weeks, participants will learn about the basics of data manipulation and data reporting. After taking their first steps in R, participants will have to create a data report in which they structure data and present their output in a systematic way. This data report has to follow the guidelines in the data report manual, which is available to the participants on Surfdrive:

1. Picking an example dataset.
2. Transform data using R.
3. Visualise the data.
4. Summarise your conclusions from the enquiry.

The mid-term assignment counts for 30% of your course grade

Due date: 5 April

Week 5: Quantitative Research Cycle (QRC)

Now that participants learned how to manipulate data in R, it is time to look at the bigger picture. Data manipulation is often but a part of the research process. This week we will learn how to structure your data in a logical order. This is called the pipeline or - as we like to call it - the quantitative research cycle. Understanding the research cycle will not only help you perform your research more efficiently, but also helps to interpret and critically reflect on the research of others. In the lecture, we will discuss the research paper by Quanjer & Kok as a practical example of how quantitative data explorations can take place in a historical context.

Teacher:

Auke Rijpma

Required reading:

Quanjer, B. & Kok, J. (2020). Drafting the Dutch: Selecting biases in Dutch conscript records in the second half of the 19th century, *Social Science History, 44*(3), 501-524. doi: 10.1017/ssh.2020.13

Topics of this week:

- Quantitative research cycle and historical research

Learning goals:

- Participants can apply the QRC to publications and other forms of quantitative research.
- Participants are able to place source criticism in the quantitative research cycle.

- Participants are able to perform source criticism on quantitative research.

Week 6: Introduction to Linked Data, 1
In week 6, we take a look at Linked Data. This format focuses on standardising and connecting different sources of data, such as the tabular data sets that we previously worked with when we were using R, but also data that lends itself far less to the tabular form. Structuring and manipulating data allows researchers to extract more information from their data. However, it is often also a lot of work, which has already been done by other researchers that have worked with similar or relevant data. These efforts are not always shared and - if they are - often hard to find. Moreover, links between data sets generally need to be established anew for each new study, which costs time and may cause errors. Linked Data is a data format that deals with these problems.

Teacher

Wouter Beek van Triply

Required reading:

Ducharme, B. (2013). *Learning SPARQL: Querying and updating with SPARQL 1.1.* Newton Massachusetts: O'Reilly Media. (Chapter 2 + 1)

https://www.youtube.com/watch?v=4x_xzT5eF5Q

Topics of this week:

- Principles of Linked Data
- SPARQL

Learning goals:

- Participants understand the principles of Linked Data.
- Participants are able to perform basic queries on Linked Data.
- Participants can identify all 151 original Pokemon.

Week 7: Introduction to Linked Data, 2
Now that we understand what Linked Data is used for, and why it is so useful, it is time to actually create Linked Data of our own. In this week, participants will start by converting data, using standardised variable descriptions also known as vocabularies, and finally publishing it online.

Teacher:

Albert Meroño-Peñuela

Required reading:

Heath, T. & Bizer, C. (2011). *Linked Data: evolving the web into a global data space. Synthesis lectures on the semantic web: theory and technology.* Williston (VT): Morgan & Claypool Publishers. (Chapter 1-2) http://linkeddatabook.com/editions/1.0/

Cow manual: https://csvw-converter.readthedocs.io/

Topics of this week:

- Making your own Linked Data from your CSVs
- Querying/Enriching Linked Data

Learning goals:

- Understand the basic syntax of an RDF file
- Learn the challenges of creating your own Linked Data
- Understand the value of reusing existing vocabularies
- Learn how to combine Linked Data from different sources

Week 8: Presenting your Linked Data

Just like "normal" data, Linked Data is very hard to understand for the uninformed outsider. Often, a picture says more than a thousand words. Therefore, a Linked Data query is not complete without some form of visualisation. In the final week of the data management course, we discuss 'the good and evil of eye candy'. How can you summarise your research in one breathtaking image that immediately captivates the reader?

Teacher:

Richard Zijdeman

Recommended reading:

https://stories.datalegend.net/

Topics of this week:

- Visualising Linked Data
- Sharing Linked Data visualizations through GRLC

Learning goals:

- Participants can present their Linked Data output in a systematic way
- Participants are able to link research questions to specific research output
- Participants can share their Linked Data queries

Week 9: Working on assignment

In the final week, you have the time to work on your assignment. There might be no class, but our mailboxes are open for questions.

Final assignment

At the end of the course, participants will have learned how to structure, visualise, and report on their data exploits. The final assignment will require the participants to apply all the acquired skills from the course in a brief, data-focused paper of ***max 2,000 words***, which shows that they understand the Quantitative Research Cycle.

Participants will choose one of the available datasets provided and use this dataset to:

1. Formulate a research question which can be answered with the data.
2. Explain why this research question is relevant.
3. Discuss the data processing, by:
    a. Applying source criticism
    b. Showing how they manipulated the data
4. Visualise the data
5. Write a short conclusion (***1,000 words max***) answering the research question, explaining conclusions, and discussing the pros and cons of the data. Special attention should be given to the question how others can make use of your findings / scripts.

The final assignment counts for 70% of your course grade

Due date: TBD