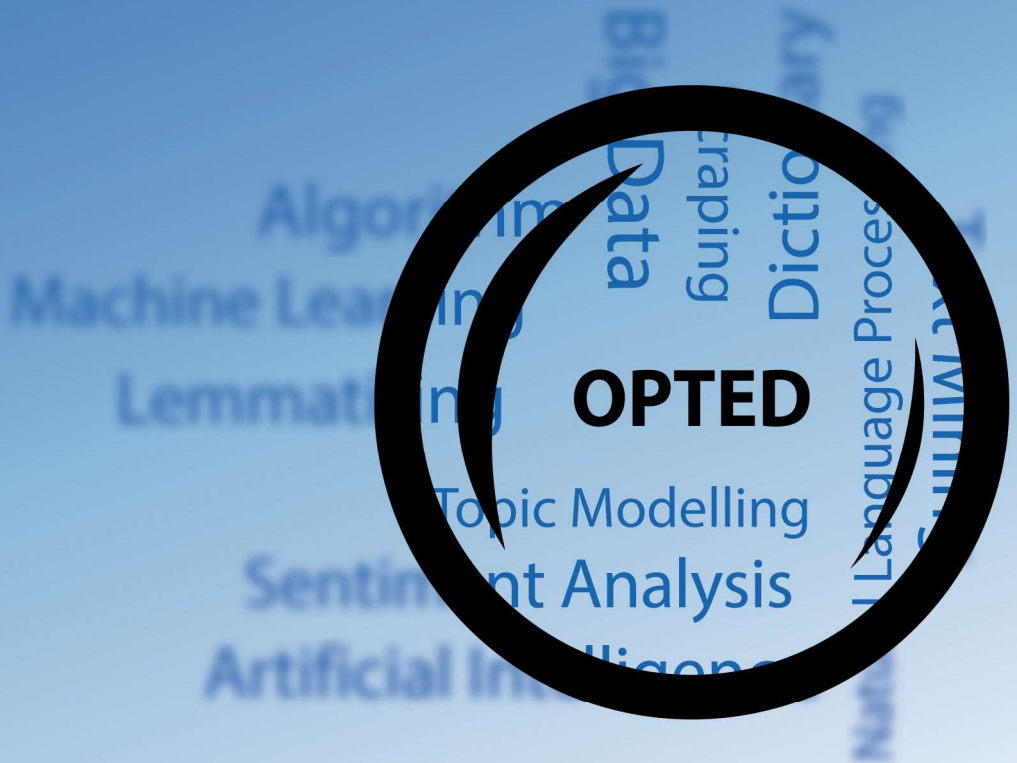
OPTED

Stepwise Workflow for Linking Different Types of Data Sources

Shaun Bevan



Disclaimer

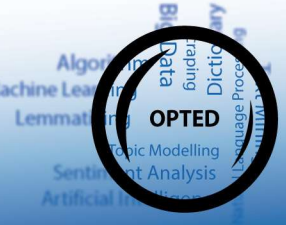
This project has received funding from the European Union's Horizon 2020 research & innovation programme under grant agreement No 951832. The document reflects only the authors' views. The European Union is not liable for any use that may be made of the information contained herein.

Dissemination level

Public

Type

Workflow



OPTED
Observatory for Political Texts in European Democracies:
A European research infrastructure

Stepwise Workflow for Linking Different Types of Data Sources

Deliverable D8.3

Shaun Bevan[1]

[1] University of Edinburgh

Due date: March 2023

Executive Summary

The process of linking different types of data sources is a difficult and often time-consuming activity. Furthermore, despite the very common need to link different data sources in order to fully test theoretical expectations in most empirical research, linking is often ad-hoc with experience gained through trial-and-error attempts rather than the application of a clear and consistent workflow. This document introduces and explains a stepwise process for linking data sources based on a set of five steps. The steps are identifying linkage points; renaming, recoding, and cleaning linkage points; linking processes; sanity checking; and archiving. It also introduces a simplified workflow chart and describes the process through an example linking four data sources including two existing political text datasets, public opinion data, and related national statistics. The document closes with a discussion of ongoing concerns related to the linking of social science data in general.

1 Introduction

The linking of data is often one of the biggest time sinks within research projects, at least those projects that are not collecting or creating original data. Even then, making original data work with additional outside sources can be a challenging task under the best of circumstances with no clear standards for metadata, data structures, coding offering a clear starting point. Moreover, minor errors or undocumented changes within data sources can create substantial issues with linking and are often hard to diagnose or even notice. While we cannot remedy these issues, we can make it easier to address them through a common process for linking data by suggesting a Stepwise Workflow for Linking Different Types of Data Sources (from this point forward called workflow) which sets out the principles and the process for the linking of political text and associated data sources within the social sciences. This is intended as a guide which lays out best practice for linking data for theory testing, common concerns and best practices at each stage of the process, and emphasizes the value of establishing and promoting norms for newly created datasets that could be of tremendous value for research communities.

2 Approach

The starting point of the workflow (D8.3) detailed here are the two previous inventories (D8.1 and D8.2), created by WP8, which considered existing data sources, opportunities for linkages between them, as well as the core concepts and measurements contained within political text data. However, while these are the starting point for this workflow and their content is used as examples, the workflow outlined in this deliverable is intended to apply to other use cases in political text data and social science data more broadly. This workflow also lends support to the future deliverable D8.4, a tool for data linkage focused on common attributes and norms within existing data aimed at partially or fully automating linkages between OPTED data sources and several other commonly used data sources.

The remainder of this document starts by introducing and discussing the steps of the workflow. It next offers a brief version of the workflow as a chart for easy reference followed by an example of how to link four data sources. Finally, this deliverable concludes by discussing some ongoing goals and issues related to linking different types of data sources.

3 Discussing the Steps

The steps of the workflow are presented in Table 1. Each step is discussed in detail in this section. The next section offers a brief outline of this process through a workflow chart as well as an example of linking datasets following this process.

Table 1: Linking Process Steps	
Step	Short Description
1	Identifying Linkage Points
2	Renaming, recoding, and cleaning linkage points
3	Linking processes
4	Sanity checking
5	Archiving

3.1 Identifying Linkage points

The first step is the identification of linkage points or opportunities. Potential links between data can exist in any part of a data source. Oftentimes multiple variables are needed to create linkages such as combinations of date, location, and/or actor variables. While the reasons for linking data should be based on the needs of theory testing, links are not always theoretical and can be found between metadata, coded variables, and/or other variables such as demographic, dummy, or institutional variables. However, linkages can also be driven by theoretical or conceptual reasons. For instance, coded data on a policy area like the environment could be linked to a dataset containing one of more variables related to environmental conditions, like carbon emissions, the percentage of electric vehicles sold, renewable investments, etc.

The most common way to link data is through metadata either on its own or in combination with coded data and conceptual matches. Metadata can and often does vary from dataset to dataset, not just in regards to what is included, but also if it is considered Metadata (see D8.1). For example, while an official database may list the political party of a speaker, an analysis or debate text may identify political party through topic modelling or named entity recognition and this could be considered coded or enriched data. Regardless, of whether an individual dataset considers a variable as metadata or some other form of data, metadata is distinct in that it represents a substantive fact about each observation. For example, when something happened (e.g. a year or a time), where it happened (e.g. a country or an office), what was it (e.g. a law or headline), who did it (e.g. a prime minister or a member of the public), etc. It does not include concepts, theories, or arguments, unlike coded data.

Metadata examples:

- ID variables
- Dates
- Location (e.g. country, region)
- Types (e.g. speech, question, law, pledge)
- Political Party

In ideal circumstances linked data will have perfectly matched coded variables when they measure the same concept such as a policy area, frame, or sentiment. In many cases, researchers might not even consider well-linked data as different data sources, when functionally they are. For example, the CMP contains many different data sources across countries and through time, but these are often joined at the point of download: all sharing a common coding scheme with different sources identified through a combination of country and party id variables. CAP data follows a common Master Codebook (see Bevan 2019), but the data included in the CAP comes from many funded and unfunded projects from different creators even within the same nation, sub-national, or supernational unit. Most open access data can be compared through the CAP Trends Tool, but that data may need to be recoded and/or restructured into different data formats such as cross-sectional data to be linked.

Oftentimes data sources will touch on the same concept, but with different measurements for that concept. This mismatch can be relatively simple to address such as in the use of different scales. It could be more difficult when variables contain part of a concept covered in each data source. For example, CAP and CMP data both focus on policy areas/issues/topics, but their coded issues are not the same. In this example, CMP data covers both issues and direction, while CAP data only covers issues (topics), but often includes multiple topics for each CMP issue (this will be further explored in the future deliverable D8.4). In other cases, data may be truly unique between data sets such as a set of different economic variables or the sentiment of different actors in each data set, but this data could still be linked conceptually with the help of other, linkage points. Linking data based on these conceptual matches requires additional work discussed in the next subsection (3.2).

Coded Examples:

- Common or overlapping schemes (e.g. scales or common codebooks)
- Conceptual matches such as topics or entities

Depending on the data sources that a researcher intends to link to, no, all or just some variables may be unique. These can include identifiers that are only appropriate or included in one of the data sources, as well as coded, uncoded, and even meta data variables that are unique within the comparison. While these variables do not help with linkages and can create contradictions in linked data, they are often the key reason for linking data sources. As such the coverage and retention of these variables through the linkage process steps is often essential.

Unmatched / Unmatchable Examples

- Unique identifiers (e.g. demographics, data source specific indicators)
- Uncommon conceptual variables. Specifically, some concepts may not apply to or be worth coding for all variables as not all data sources cover concepts like sentiment, frames, or even topics

3.2 Renaming, recoding, and cleaning linkage points

The second step is the renaming, recoding, and/or cleaning of linkage points. Even the simplest data linkage will generally involve the renaming of variables. This is so that the same variable such as a date can be linked even if it is listed as “Date” in one dataset and “Day” in another despite being the same and having the same format. Extending the date example, dates will often be in different formats and require recoding to be linked. For example, 1 January 2022 could be listed in one dataset with another listing 1/1/22 and still another listing a number, based on the logical rules of the software package the dataset was stored or created in. Other examples include country names vs country abbreviations, and responses to public opinion questions listed on a scale from 1 to 5 or from strongly agree to strongly disagree.

It is also important to remember that datasets vary significantly in quality, and that even high-quality datasets of a sufficiently large size are likely to have errors or typos. These could include simple, but easy to miss errors like the transposition of some dates from a European (day/month/year) to an American format (month/day/year) for some observations. They can also include larger errors like the mistaken inclusion of a 55 in in a scale from 1 to 5 or the dropping of the word agree in a different scale. What is a single mistake in one dataset can easily perpetuate when linking data as a single row of data in one dataset, may end up as several rows of data in the linked data when the unit of analysis differs.

3.2.1 Unit of Analysis

A choice also needs to be made concerning the unit of analysis. In the majority of cases linked data will take on the same unit of analysis as one of the linked datasets with choosing the unit largely dependent on the intended research or use case for the linked data. In some instances, data on regions or countries such as the unemployment rate may be linked to individual level data that encompasses multiple regions of countries. In

other cases, the intended research might call for the aggregation of individual level data. For example, individual or group level data might be aggregated and averaged to create national level data. Differences in the unit of analysis may also come down to theoretical or practical approach. For example, many text datasets will choose to separate observations by sentence or by quasi-sentence intended to measure the main or all elements of a given sentence (Däubler et al 2012). While in practice the differences between these two approaches are minimal (ibid), the unit of analysis differs between them and observations cannot be matched without some form of aggregation or pre-processing.

3.3 Linking processes

The third step is linking, which should generally be done through joining, pivots, searches, and/or query type operations depending on data format as well as the software being used for data management and/or analyses. There is no best software or programming language for linking, but suitable tools should follow replicable processes that can be shared as code and/or through a step-by-step guide in order to create a replicable process. Linking by hand, copying and pasting, or other bespoke means are prone to errors and these errors can be hard if not impossible to spot. Idiosyncratic means of linking data sources also make these processes hard to replicate, complicating Open Access goals, documentation, and the ability to revisit parts of the research making data management processes more difficult in the very least (see Bevan 2022; Corti et al 2019; Kruse and Thestrup 2017).

3.4 Sanity checking

The fourth step is checking linked data against expected or known links. For example, if the first data source contains a set of cases completely contained within a second data source the linked data should encompass the full set of cases in the first data source. This is more complicated with missing observations, variations in the length, number, or circumstances of links, but such complexity makes it even more important to check the data as errors would be less obvious. Many tools that can be used to link data can/will provide one or more check variables noting if linking variables agreed providing a useful shortcut. Crosstabs and visualizations are also useful tools for checking linked data against the unlinked sources and the expected qualities of the linked data.

Extending the checks of known links and general patterns to both random and targeted checks of the linked data in comparison to the unlinked data sources is another important part of the process. This further helps make sure observations were matched correctly and that variables were transcribed or updated appropriately. While random checks help to assess whether the linkage was successful in general terms, targeted checks can be used to investigate areas where earlier issues were seen, like cases or variables that needed to be recoded, or cases that are essential to the research being conducted.

3.5 Archiving and sharing

The fifth and final step is the archiving and sharing of the linkage process. This often-overlooked step is essential to Open Access research. While the sharing of linkage processes is often not specifically stated as a requirement for funded or published research, in many cases it is necessary for the aspirational goals of Open Access to be possible. Noting the sources used to conduct a piece of research is informative, but as the previous steps in this process have made clear, linking data is often complicated and time consuming. The process itself can have a number of decision points and include many small changes to existing data sources to make the linkage work. Other researchers linking the data can reasonably make different choices leading to slight or significant differences in the linked data.

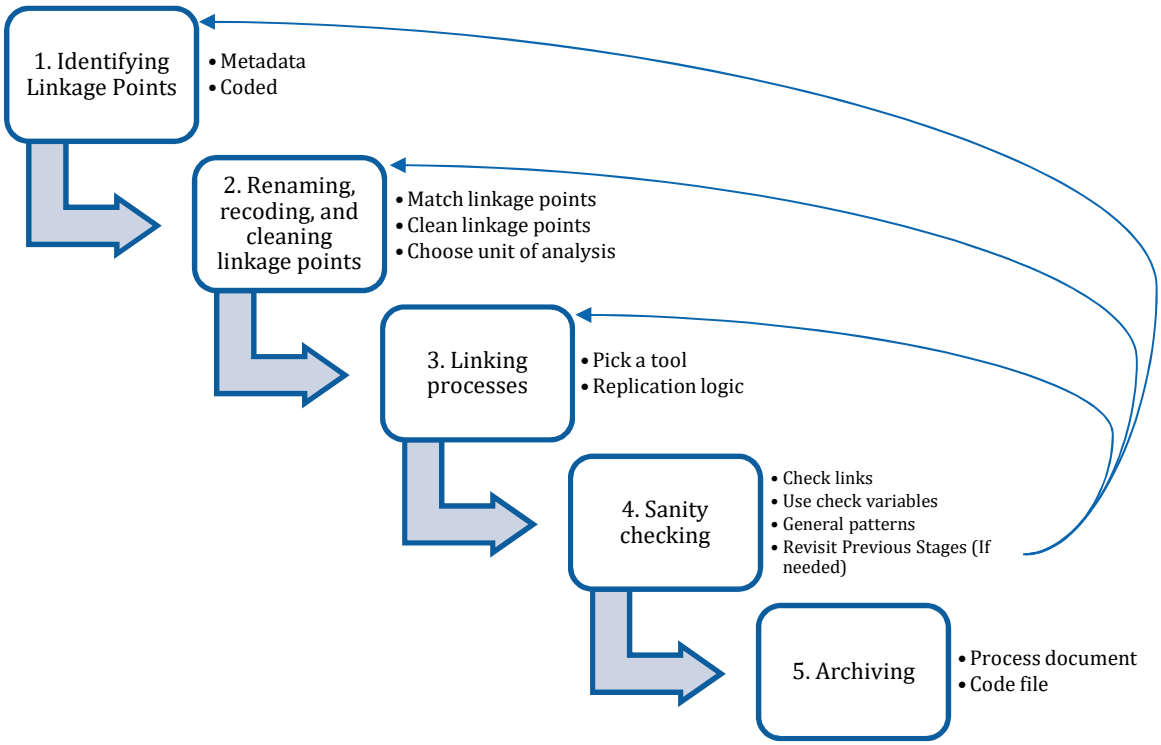
In part linkage processes are not commonly shared because of the lack of a formal requirement, but they are also often unshared due to the norms and expectations for sharing data sources. However, sharing the linked

data is often impossible due to copyright and ownership concerns over some or in the case of purely secondary research all of the linked data. No ownership issues exist when producing a detailed process document or code file allowing data sources to be linked in the same way by other researchers in the future.

Perhaps more importantly than promoting good practice and even making research more accessible by extending the goals of Open Access is that the archiving of the linkage process formalizes institutional memory. Without a detailed explanation of processes, it becomes difficult to replicate and extend research (even internally) to projects and, with every additional researcher involvement increasing the hurdles. The push to maintain good records included in Open Access goals, FAIR guidelines, and good data management practices is, at least in the long run, a boon and not a burden for researchers (see Bevan 2022; Corti et al 2019; Kruse and Thestrup 2017, Wilkinson et al 2016).

4 Stepwise Workflow

Figure 1 STEPWISE WORKFLOW CHART



4.1 Example

The following example discusses a workflow for the linking of Comparative Agendas Project (CAP), Comparative Manifestos Project (CMP), Public Opinion Data, and official statistics following the Stepwise Workflow presented in Figure 1.

4.1.1 CAP and CMP

The linking of CAP and CMP data is a useful and realistic example within political science due to the considerable overlap in those interested in the policies that governments (CAP) and political parties (CMP) pursue. The first step in linking CAP and CMP data is the identification of linkage points. As previous work has shown, this data can be linked based on a combination of metadata level variables including time and country, as well as coded variables, specifically topics (CAP) and issues (CMP) (e.g. Green-Pedersen 2019). Other variables like political party variables can also be used in some cases, but not all CAP data contains political party variables. In this kind of case multiple observations from the CMP could be linked to a single CAP observation to capture the focus on different political parties' issues from the CMP.

The second step includes the renaming and recoding of data. In this case CMP data would generally need to be extended through time to represent issues from the previous election to the next. Coded variables would also need to be aggregated, recoded, and/or matched between CMP issues and CAP topics. For example, CMP issues often include two directional, left/right codes and those need to be recoded to capture overall attention to match CAP topics. Cleaning is not a major concern given the general quality of these datasets, but some errors may still be present such as typos or missing data and these should be identified. The use of summary statistics and visualisations of linkage variables early on in the process can help identify problems before the linkage process.

The linkage process itself can be done in many ways, but it is advisable to use a software or code-based solution given the likely size of these datasets. This process should either be recorded in a process codebook or through the documentation of code from statistical software, programming languages, or other tools. Ideally, whatever solution is chosen will produce summary variables and/or statistics indicating successful and failed linkages to be used in the next step.

Sanity checking is the fourth step. The combination of CMP and CAP data will very likely produce missing linkages. The main reason for this is the absence of data in both datasets such as data not covering the full time period, certain political parties, or some countries. Other issues are possible though. If renaming and recoding was unsuccessful,a expected links may be missing or incomplete. In some instances, the tools used to link data will produce an error noting the issue or even preventing the tool from running. In others, the process of sanity checking includes comparing the coverage (such as countries and time periods) in the linked data to the expected coverage. This is where summary variables can be particularly helpful. If the process fails at this stage, the previous stages should be revisited.

Once the data has going through the linkage process and passed its sanity check the final step is archiving the process. Often first an internal step, a complete record of what was done to link the data is essential for rigorous research and replication, even internally to the project. Additionally, the detailed and archived linkage process allows other scholars to replicate and extend research. Oftentimes linked data sources cannot be shared as they make use of data that is not owned by the researcher. An archived linkage process is one way to get around this limitation and represents best practice for Open Access (Bevan 2022).

4.1.2 Public Opinion

The linking of public opinion data to already linked CAP and CMP data will follow the same general steps as above. However, there are some additional elements to consider when it comes to linking data that is not the same type or aiming to capture the same concept, like party/government data compared to public opinion data.

First, the identification of linkage points is a conceptual/theoretical exercise. Public opinion can measure many different things, from topic salience, to preferences, to popularity, to competency. Each of these measures can theoretically affect the combined CAP/CMP data. Often, data like public opinion is aimed at capturing a related concept to topics such as the level of importance, how much spending should change (or remain the same), or which political party is viewed as the most competent on the topic. The difficulty is in how these sometimes more specific and sometimes narrower sources of public opinion data can be matched. Continuing the example of CAP/CMP data, in some cases this requires aggregation of data before linking it to public opinion in order to match very broad topics. In other cases, it may require a narrower selection of data using other variables or additional original or recoded data to match a specific public opinion response.

Given the theoretical element of linking, the process of recoding and cleaning is essential, but in practice faces the same issues as linking the previous two data sources.

The linkage process should also follow the same process and ideally use the same tools to avoid errors and make the process easier to revisit or replicate.

Sanity checking is likely to be more involved with theoretically driven linkage points. While the same methods for checking coverage can be used, the fit of the data to a known or anticipated reality is important in this case. Opinion data may not cleanly fit a concept like a CAP topic. For example, public opinion rarely distinguishes between international affairs and defense as unique topics even if governments tend to have separate departments or ministries to address each. In this kind of case, issues may need to be combined in the CAP/CMP dataset, as well as in the opinion data. This would lead to a repeat of the process from step 2.

Like with the first linkage, the linkage with public opinion should also be archived. Here, it can oftentimes be more important to lay out the steps of the process given that most opinion data is generally gathered for commercial purposes and the ability to reproduce the data directly can be extremely limited.

4.1.3 Official Statistics

Official statistics such as unemployment, prescription drug costs, train delays, and thousands of other variables captured for one reason or another including the evaluation of existing policies and programs can be linked based on different approaches to theory as well.

Like with public opinion data, identifying the linkage points with other types of data is at first theoretical. Does the measure affect multiple or potentially all topics in the current example like with a headline economic measure such as the gross domestic product (GDP)? Is it closely tied or related to a single topic (like drug costs) or could it affect multiple topics, but not all such as unemployment which affects headline economic performance, labor issues, and business concerns. Regardless, theory should drive the establishment of these linkages.

The remainder of the process should follow the same process as public opinion. However, there may be some temptation to skip the final step of archiving as most official statistics do allow for reproduction of the data. This would be a mistake though, as the linkage will still depend on the previous linkages and a set of more theoretical and practical choices that may be impossible to replicate or understand without a clear, archived guide to the process.

5 Ongoing considerations

There is no perfect or universal way to join political text data or any type of social science data, at least without a common and strictly adhered to naming and structure standard for data. Standards are the goal of research communities like OPTED internally, but the diversity of disciplines involved in the study of political text inevitably means that many datasets will continue to be created outside those standards and this deliverable (8.3) has provided a workflow on how to approach linking both text and other forms of social science data. Deliverable 8.4 will introduce a standard for naming and structuring OPTED data as well as how to transform many of the commonly used political text and related datasets to that standard. It will demonstrate the value of standardization through the automated joining for many datasets commonly used in and in combination with political text data.

Despite the clear advantages of standardization for end users, we do accept that this would be a difficult and perhaps intractable goal within the social sciences without strong existing norms changing best practices. Even then, efforts at standardization take considerable time and support namely from institutional buy-in. The scale of the journey to Open Access research and data being the norm for funded academic work in general has been a long one and one that is still not complete with a great deal of work in the social sciences being unfunded and not having to adhere to the norm. Many funders, universities, research groups, journals and publishers have made that goal more likely, but it is still not fully realized with important outputs like books still having no clear solution. The higher standards in Open Access introduced through the goal of making data FAIR (Findable, Accessible, Interoperable and Reusable; Wilkinson et al. 2016) have a strong uptake in some fields but FAIR has yet to catch on in the social sciences with findable and accessible data still representing mostly aspirational goals. Reusability and interoperability can include standards allowing for easier data linking but creating standards of interoperability in the social sciences where how data is used is so widely varied creates a greater problem. While many research areas have inherent standards due to the common tools and models used, the social sciences tend to collect models almost as much as data and scholars constantly work with new and different tools from outside fields as a means of innovation putting different requirements on data structures and forms.

References

Bevan, S. (2019). "Gone fishing: The creation of the comparative agendas project master codebook." In *Comparative Policy Agendas* (pp. 17-34). Oxford University Press.

Bevan, S. (2022). "How to Create a Data Management Plan for an Online Research Project." *SAGE Research Methods: Doing Research Online*. (DOI: 10.4135/9781529608953).

Corti, L., Van den Eynden, V., Bishop, L., & Woollard, M. (2019). *Managing and sharing research data: A guide to good practice*. Sage.

Däubler, T., Benoit, K., Mikhaylov, S., & Laver, M. (2012). Natural sentences as valid units for coded political texts. *British Journal of Political Science*, 42(4), 937-951.

Green-Pedersen, C. (2019). The reshaping of West European party politics: agenda-setting and party competition in comparative perspective. Oxford University Press.

Kruse, F., & Thestrup, J. B. (Eds.). (2017). *Research data management-A European perspective*. Berlin, Boston: De Gruyter

Wilkinson, M., Dumontier, M., Aalbersberg, I. et al. (2016) "The FAIR Guiding Principles for scientific data management and stewardship." *Sci Data 3*, 160018.