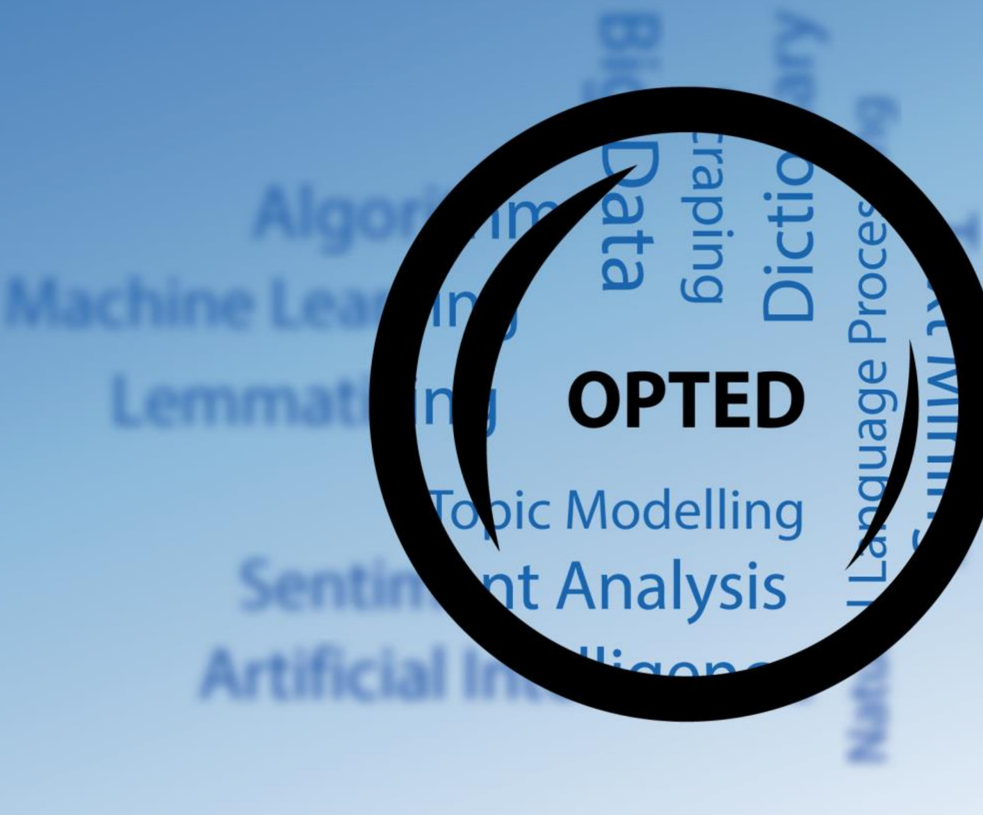
OPTED

Deliverable 8.1 Inventory of data sources and opportunities for linkage

Christoffer Green-Pedersen & Matt W. Loftis



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

Report

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

Inventory of data sources and opportunities for linkage

Deliverable D8.1

Authors: Christoffer Green-Pedersen[1] & Matt W. Loftis[1]

[1] Aarhus University

Due date: September 2021

Executive Summary

Work Package 8's (WP8) long-term goal is to ease and enable linking political text data together to unlock opportunities to answer new social scientific questions. This deliverable has two elements, namely first building an inventory of databases (datasets and data sources), and second, it describes our progress on investigating opportunities for data linkages in terms of aggregation across time and linking databases with different levels of time aggregation. We further introduce related, ongoing steps toward our long-term goal, including establishing standards for data interoperability and systematizing a shared vocabulary that supports researchers in data-linking efforts.

Future work will build on the progress described here. The inventory of databases also serves as a foundation for further work toward linking databases in terms of key concepts, while the metadata model and our developing standards for data interoperability provide theoretical and conceptual groundwork for our data-linkage workflow and technical tools.

1 Building the inventory

The inventory draws on the work of WP 2-5., and especially deliverable 4.2 and 5.1. We have included in this inventory, finished data sets compiled for research purposes from original full-text resources. These data sets should either include the original full text in the data set or should be easy to connect to the original, publicly available full text. All of these data sets include measurements produced by researchers and applied in published political science research. The inventory has been assembled based on the following considerations.

The main purpose is to generate an overview of how databases typically are set up in terms of especially:

- What is the unit of analysis: party, country, organization
- What is the time granularity: Yearly, monthly, etc. (and its exact format)
- Which key concepts have been coded: frame, issue, tone, actors, etc.

The way in which data sets are structured with regard to these characteristics is often a reflection of the nature of the original data source. For example, party manifesto data sets typically have parties as the unit of analysis and election period as their time granularity—media data typically have news articles as the unit and day as time granularity (combined with precise time in the case of on-line media). Conclusions by the European Council have conclusions as the unit and date of meeting as time granularity.

To investigate questions about linkages, it was important for us to include data stemming from a wide variety of original sources. One thing is to link different data produced by the same political actors—e.g. press releases, TV ads, party manifestos from political parties. Another thing is to link these data sets to data sets containing data compiled from citizens (tweets or Facebooks posts), institutions (bills, conclusions), other political actors (press releases, manifestos), and news media (news articles)

Therefore, the focus in building the inventory has been on having a *broad* scope in terms of the type of data sources covered rather than having a *deep* coverage of many data sets drawing on the same types of data sources. We have further favored data sets focused on the European context, and we have excluded data sets that are highly unstructured or not in a tabular format, such as web data that have not been parsed and organized. The inventory is far from exhaustive, but it has been crafted to provide a broad overview of the challenges involved in linking databases.

The inventory has been focused on data sets rather than original corpora or "raw" data sources because questions about linking databases typically relate to the steps taken after moving from raw political text to datasets of measured variables that researchers eventually analyze. Thus, the ParlLawSpeech database built in WP5 is less relevant here since it is a database of original texts and not data immediately ready for analysis, complete with measurements, labels, etc. for concepts such as issues, frames, positions or the like. In principle, of course, the linkage discussion here applies to raw data if it structured in a standard tabular format and the discussion of standardization and interoperability are relevant for raw data for which the researcher can legally share the data.

Further, our focus has been on data sets with a broad cross-national coverage. The reason for this is that these data sets typically serve as de facto standards in social science research applying text as data. The design choices, formatting systems, and measurements found in these data sets have accreted to other data sets by virtue of their wide application. The practices used in the data sets are therefore central to discuss from a linkage perspective. We plan to expand the inventory with further national data sets as time goes on, to ensure we can provide a better overview of issues that arise in the process of linking data sets.

1.1 Data sets in the inventory

Since the inventory is intended to be a living document, receiving running updates, we have opted to maintain it in a Google Sheets document. This can be accessed openly at the following link.

We provide here a brief example of linking two of the data sets in the inventory. The example helps to clarify the sorts of questions, decisions, and concrete steps involved in carrying out a data linkage. In Section 2, we will return to clarify several of these points and provide an introduction to the tools WP8 is developing to ease this process. Finally, in Section 3 we will see how the work carried out for this deliverable connects to the broader set of tools and resources we are developing to address the challenges of linking data sets for social science text analysis.

1.1.1 Example: Linking policy agendas data to party manifestos

To illustrate the process of linking data sets and to foreshadow the issues the present deliverable addresses, we provide an example of linking two quite different data sets. Suppose a researcher's analytical goals required linking data measuring the policy agenda in national legislatures to the policy preferences expressed by major national political parties at elections. This can be achieved by using two of the data sets included in the inventory presented below. Namely, the Comparative Agendas Project (CAP) provides data sets with measurements of the national policy agenda based on issue topic codes applied to national legislation in multiple countries, and the Comparative Manifesto Project (CMP) provides data on the policy preferences parties express in their election manifestos—the documents or statements parties release around election time to present their platforms to voters.

We begin by summarizing the basic features of the relevant data sets. The CMP data are organized by country, party, and election period. Each observation in the data is a summary of measurements describing the stance of one party, in a particular country, during a particular election campaign toward multiple different political issues. CMP includes data from more than 1,200 parties in 61 countries. The CAP data are provided in separate data sets by country. Each country data set on legislative policy agendas is organized into observations of individual bills—i.e. draft legislative instruments that come before parliament for discussion or a vote. Each bill includes information on the date it was proposed in parliament, and all bills are classified into one of around 200 subtopics, according to which specific political issues they address.

The first and most fundamental decision facing the researcher is the shape of the linked data set to be produced. Several options are possible, since the output data could take on the shape of the CAP data, the CMP data, or it could take on some other shape. Since the CAP data identify the date each bill is proposed, the researcher could join every bill to the policy positions of major political parties at the most recently conducted election. The researcher could, instead, aggregate up all bills in the CAP data within election periods, summarizing the policy agenda as the proportion of bills devoted to each policy topic during each election period. The researcher could then join this aggregated CAP data to the CMP data, producing an output data set organized the same as the CMP data, but annotating each country-party-election observation with summaries of the national legislative policy agenda.

The possibilities for organizing the output data are endless, therefore we illustrate the steps involved in executing a quite standard data linkage in political science, organizing output data by country, year, and party. Each observation in the output data set will describe a political party in a particular country, during a particular year. From the CMP data, we will include a measurement of the party's expressed policy preferences, and from the CAP data we will include one summary measure of the national legislative policy agenda, the percentage of legislative attention directed to the economy (major topic 1).

In the following subsections, we provide short discussions of the steps involved in executing a data linkage and what obstacles a researcher must overcome to do so:

1.1.1.1 Reshape and combine the two CAP data sets

We begin with the CAP data. Since CAP data are organized into separate data sets by country, it is most efficient to first reshape the country-level policy agendas data sets and then to combine the country-specific data into a single data set to join to the CMP data. We take as examples, two countries with long time series CAP data on legislative bills: France (1974-2013) and Denmark (1953-2016). To begin with, we reshape both data sets to the annual level, 64 observations for Denmark and 40 for France, and we add to each data set a variable calculating the percentage of bills in each year that focused on the economy. At this point, we can simply stack the two country-level data sets on top of each other to combine them into a single CAP data set.

1.1.1.2 Harmonize country reference variable

Upon combining the two CAP data sets, the researcher encounters an obstacle. CAP data sets do not include a variable for country; therefore, this must be added to both the France and Denmark data before combining them. However, in order to execute the later data linkage with the CMP data, the country variable we add to the CAP data must match the CMP's method for labelling countries. CMP identifies country with two variables. The first is a unique number for each country, while the second is the name of the country spelled out in English. Both variables are non-standardized data formats, unique to the CMP. The country numbers, for example, label France as 31 and Denmark as 13. The ordering is not alphabetic, since Sweden comes first at number 11, and the numbering system makes occasional large leaps, for example assigning Mexico to number 171. The CMP country variable featuring the English name of each country works well in this example, since it uses the standard spelling for both countries we consider. However, a larger data set would need to verify the spellings and usage of several special cases. For instance, in the CMP country variable, the United Kingdom is separated into Great Britain and Northern Ireland.

1.1.1.3 Prepare the CMP data for joining

At this point, the CAP data has been reshaped into 104 country-year observations and includes three variables: country name, year, and the percentage of attention to the economy in legislative bills in the respective year. To join this to the CMP to create output data organized by party, country, and year, the CMP data must be reshaped to match. This reshaping forces the researcher to make a substantively important decision about how to allocate information from elections to the years between elections. In the CAP case, this decision was less substantive because the researcher is aggregating information from days (when bills are introduced) to years. There are multiple ways to aggregate this information, but the researcher is working with information at hand. In the CMP case, we are required to disaggregate information from election periods (typically once every several years) to years.

Information about parties' annual policy preferences is not represented in the CMP data, so the researcher must make assumptions about how preferences measured at election time are best assigned to the intervening years. For example, one option is to assume that parties are tied to their expressed policy preferences from the previous election until they produce a new manifesto. In that case, in the expanded CMP data set, the researcher would repeat observations of the most recent expressed preferences for every year until the next election. On the other hand, the researcher might prefer to assume that parties' preferences evolve smoothly from one election period to the next, interpolating all measurements of party preferences over the years between one election and the next.

Whichever choice the researcher makes, once the CMP data has been expanded, it will be organized in the same way as the targeted output data: by country, party, and year.

1.1.1.4 Final merge

With the two input data sets in hand and reshaped appropriately, the researcher is prepared to link them to produce the output data. Executing the linkage itself is a smoothly automated process in all modern statistical software packages. However, the researcher will have to keep a careful accounting of which variables contain the variables that structure the merge. For example, in the CMP data the variable containing the English names of each county is titled "countryname." These names of the variable structuring the merge must either match exactly in both data sets, or the researcher must note for the software precisely which variables in each data set structure the final merge. Finally, the researcher must keep notes or well-commented software to ensure that their work to execute the data linkage is transparent and reproducible for themselves and others in the future.

Using this example as a launching point, the remainder of this deliverable provides vocabulary and background for discussing and addressing the types of challenges and frictions that arose here, and we present our solutions and work in progress to both ease the process of linking data sets and to strengthen the tools available to researchers to ensure data linkages are transparent and reproducible. Specifically, our work addresses the need for shared standards and vocabulary in several areas, including: the choice of how to structure output data, variable naming conventions, and standardized references (e.g. for country names, times and dates, etc.). We also describe a system for automatically annotating linked output data sets with metadata describing the sources of input data, the nature of the linkages made, and any (dis)aggregations performed to the time scale of the data.

2 Data linkages and time aggregation

Most features of social science data resources are represented in the data's columns, or variables. Two crucial features of these data, that are often only implicitly represented in the data itself, are the *unit of analysis* and the data's *time granularity*. Observations in social science data have a particular unit of analysis—i.e. observations represent data on single entities or events *at a particular time.* The unit of time denominating the chronology of individual observations is the data's time granularity. Examples of a unit of analysis and its time granularity include: a legislator in a legislative session, a country in a year, or a news article published in a particular minute. Data for analyzing political texts frequently have both a unit of analysis that covers many individual units (e.g. many legislators or many news articles) and a time granularity that covers a chronological range (e.g. many legislative sessions or many days). Data sets with this structure are often termed *time-series cross-sectional* data (TSCS) or *pooled time-series* data.

In principle, data can be linked so long as the researcher harmonizes the unit of analysis in the parent data and selects an appropriate, harmonized time granularity for the linked data. When joining two data sets with different time granularity, three scenarios are possible. The output data may either preserve the courser time scale, the finer time scale, or aggregate both data resources to a still coarser time scale. For example, suppose an analyst joins daily news media data to data on monthly meetings in a parliamentary committee. Output data at the *monthly* level would aggregate media data to the coarser time granularity of months. Output data at the *daily* level would spread the monthly data to the finer granularity of days, repeating values of the appropriate month for each daily observation. Alternatively, the output data could aggregate observations to the level of a still *coarser time scale*, like three-month quarters, years, etc. Disaggregating a data resource's time granularity without adding additional information or assumptions (e.g. interpolating missing values) to the data is not possible, as we saw with the CAP-CMP example in Section 1.1.1. Although there is no natural time granularity for the output of data linkages, the difficulty of justifying assumptions required for disaggregating coarse time scale data argues in favor of defaulting to aggregating data to the time granularity of the coarser input data when linking.

We list below, in Table 2.1, data sources from inventories in deliverables 4.2, 5.1, and 8.1, classified by unit of analysis and time granularity. We include only data sources identified as existing.

Table 2.1: General data sources by unit of analysis and time granularity

Deliv.	Name/category	Unit of analysis	Time granularity
4.2	Parties / internal communication	Congress speech/motion	Annual or election cycle
4.2	Parties / internal communication	Statutes	Day
4.2	Parties / external communication	Press releases	Day
4.2	Parties / external communication	Website	Arbitrary (minute)
4.2	Parties / external communication	Manifestos	Election cycle
4.2	Parties / external communication	Coalition agreement	Event /government formation
4.2	Parties / external communication	Social media posts	Arbitrary (minute)
4.2	Interest groups / internal comm.	Newsletter, magazine, etc.	Publication cycle
4.2	Interest groups / internal comm.	Statutes, internal deliberations	Irregular (day)

4.2	Interest groups / external comm.	Press releases / Position papers	Day
4.2	Interest groups / external comm.	Website	Arbitrary (minute)
4.2	Interest groups / external comm.	Consultation submission	Day / event
4.2	Interest groups / external comm.	Public speeches	Day
4.2	Interest groups / external comm.	Social media posts	Arbitrary (minute)
5.1	Legislative speeches	Speech/question/interpellation	Day / event
5.1	Legislative documents	Law/bill/amendment	Day / event
8.1	Comparative Manifesto Project	Party manifestos	Election cycle
8.1	Content analysis of European issue salience	Press release	Day
8.1	Issue Competition Comparative Project (ICCP)	Party	Single election cycle
8.1	Social Media and Political Agenda Setting	Press releases	Day
8.1	Social Media and Political Agenda Setting	Social media posts	Arbitrary (minute)
8.1	Social Media and Political Agenda Setting	Newspaper articles	Day
8.1	INTEREURO	Interest group consultations	Event
8.1	Comparative Agendas Project	Party manifestos	Election cycle
8.1	Comparative Agendas Project	Newspaper articles	Day
8.1	Comparative Agendas Project	Executive speeches	Year/event
8.1	Comparative Agendas Project	Bills	Day
8.1	Comparative Agendas Project	Executive orders (decree)	Day
8.1	Comparative Agendas Project	Parliamentary questions	Day
8.1	Comparative Agendas Project	Motions, interpellations, etc.	Day
8.1	Comparative Agendas Project	Coalition agreement	Event/government formation
8.1	Comparative Agendas Project	European Council conclusions	Day
8.1	Comparative Agendas Project	Constitutional/Supreme court rulings	Event/day
8.1	Comparative Agendas Project	Government communications	Day
8.1	Comparative Agendas Project	Congressional hearings	Day
8.1	DICEU	European Council debates	Day

Note: "Event" refers to irregular time granularities that can only be denominated by the opportunity for an observation to occur—e.g. interest groups can only consult on a policy when a policy or decision is under consideration and open for consultation. "Arbitrary" refers to events that can be measured with arbitrary precision—e.g. the timing of website updates or social media posts can, in principle, be measured to the second or finer.

Table 2.1 excludes mentioning several features of these data sources, including: their use of standardized identifiers for features like time or country, their naming conventions for common variables, or their organization beyond unit of analysis and time granularity. This is due to the great variety of standards across these data sources. As evidenced by this sample of prominent, publicly available data sets for social science researchers, the use of standardized identifiers and names is not widespread in the field, and data sets tend to be organized in various ways that suit individual researchers' specific tasks and tastes. For this reason, the types of challenges encountered in the example in Section 1.1.1 are quite common. Although social science researchers working with text data frequently have strong skills in the technical aspects of linking data, the field itself lacks shared best practices and organizational standards for data and has failed to coordinate on standardization methods that can bring tremendous efficiency gains. Thus, individual researchers often spend inordinate amounts of time performing and redoing the tedious work required to make it possible to link two existing data sets. This situation has become widely understood as a basic feature of the field. In fact, the most prominent current tool for easing researchers' process of linking data on political parties, PartyFacts, is simply a collection of metadata that translate the ad hoc variables and identifiers in one data set to the ad hoc variables and identifiers in every other (see: https://partyfacts.herokuapp.com/). Although this is an invaluable contribution to the field, it clearly demonstrates the problem. Rather than bring clarity to the babel of data management standards, researchers have resorted to developing new software to overcome the field's lack of coordination on basic data standards.

Given the existing circumstances, we begin our work of easing and standardizing data linkage at the end of the process, with documenting linkages. At the documentation stage, the lack of strong shared standards has serious consequences for research transparency and reproducibility. Section 2.1 presents our proposal for a consistent, transparent standard for generating metadata that documents how data sets were joined or aggregated. Usefully, our standard can be fully automated, and therefore it can be integrated by default into the tools WP8 will produce. Following this, we move on to describe several ongoing contributions to setting and improving standards for data management.

2.1 Convention for annotating data linkages

Information about time granularity of the original data resources is not automatically preserved when linking data. Therefore, a system for preserving this information and recording operations that alter data's unit of analysis and time granularity represents a needed advancement for transparency and reproducibility in social science research and a practical benefit for individual researchers. We address this challenge by creating a metadata model that can be applied to annotate output data resulting from joining data resources.

We are developing a standardized metadata model based on adapting the logic of the *resource description framework* (RDF) metadata model (Schreiber & Raimond, 2014). In the most general terms, the RDF standard structures metadata expressing statements about resources (e.g. documents, people, concepts, etc.). RDF's structure is described as *subject-predicate-object*. The subject and object are two resources, and the predicate expresses their relationship. Our metadata model annotates a joined data set with metadata on the two parent data sets and the nature of the linkage that produced the joined data set. As we will see below, in our model the description of the nature of the linkage means that annotations can be expressed with either data listed in the position of the subject or object. Furthermore, our model can be extended to joins of multiple data sets by concatenating annotations describing additional joins.

Several first principles informed our design of this data linkage metadata standard:

- Time granularity is a key feature of data for social science research.
- Data linkage operations come in many types and often specify a hierarchy for retaining or dropping observations from one or the other of the joined data resources. Therefore, retaining this information in metadata is also crucial to transparency and reproducibility.
- Non-join aggregations of time granularity must also be preserved in metadata for reasons of transparency and reproducibility.
- Metadata annotations must identify the identities of the parent data resources joined to produce the output data.
- The metadata model must be systematized to allow annotations to be automated as part of the joining/aggregation process.
- Metadata annotations must be extensible, so that further joins or aggregations operating on already-joined data resources may be appended to existing annotations.

We describe briefly our system respecting each of these first principles. Our metadata model prescribes annotating joined data resources with an expression of the following form:

```
parent_data_A %join_description% parent_data_B
```

The '%' notation is applied to distinguish predicates from resources. Aggregations are further noted with their unit of analysis in parenthesis:

```
parent_data_A %join_description% parent_data_B %aggregate(year)%
```

We specify identifying parent data sets by a digital object identifier (DOI) (see: International DOI Foundation). Preferably, the user provides a DOI that points to the parent data set directly. In the absence of a DOI for the data set, users can specify a DOI for an article or other digital object introducing, describing, or first applying the data.

We further specify complete descriptions for join operations. The range of join operations for social science data can be expressed with the vocabulary of relational algebra as implemented, most notably, in structured query language (SQL). The SQL expressions for data joins (e.g. inner join, natural left outer join, etc.) elegantly convey the hierarchy and rules applied to preserving observations in the joined data (Silberschatz, et al., 2020; Pratt & Last 2015).

Aggregate operations, finally, require description in terms of the target time granularity of the aggregation (i.e. minute, day, etc.) or in terms of unique events (e.g. election, parliamentary session, etc.)

With that, an example metadata annotation would resemble the following:

```
doi:0.000/0 %left_outer% doi:0.00/00 %aggregate(election)%
```

In this example, the "left_outer" describes a join in which all observations in the dataset to the left of the predicate are preserved in the output dataset, but only observations in the dataset to the right of the predicate that match observations in the left dataset are preserved. We recommend including join metadata integrated directly into the joined data for maximum transparency. For small data sets, the join metadata can be included as a column, with values repeated over rows, in the joined data set. In the case of very large data sets, including an extra column in the data can be memory intensive, therefore in these cases we recommend that users clearly indicate the join metadata in a codebook or readme or other file describing the joined data set.

Finally, note that this system is extensible. Additional joins can be appended to the end of the existing data join metadata. Furthermore, consistent application of the recommended join operation descriptors ensures that unique full metadata can be resolved when joining two already-joined data resources.

3 Prerequisite interoperability standards for metadata annotation

We preview here two ongoing processes we have launched in support our future work and to address practical aspects of the application of our metadata model. The adoption of a standardized metadata model to enable tracing time aggregations and the parent data sets of linked data sets will be most impactful in a field with strong norms around data management and shared vocabulary. We briefly discuss ongoing projects to support both and plans for integrating these projects into later outputs.

3.1 Norms around data management

Consistent use of metadata is one of several practices that can be grouped under the heading of data management norms. Unlocking the fullest value from WP8's efforts to ease and enable data linkages will require the participation of the community of scholars working in social science text analysis in adopting these practices and many of the broader norms of good data management developed by the wider scientific community. To support this, we are collaborating with our colleagues in OPTED and other stakeholders to compile a set of minimum recommended standard data management practices. These will feature in deliverables 8.3 and 8.5, and the tool in deliverable 8.4 will support and enforce basic standards.

These include, for example, use of standardized formats for dates (e.g., ISO 8601), countries, etc. Under this heading we also include naming conventions and formats for variable or feature names in data resources.

Finally, we also include standards for data structure here. Joining data resources relies on an assumed shared structure between those resources. We will provide guidance and recommendations on how researchers can apply widely accepted principles for good data structure (i.e., Cobb, 1990; Wickham, 2014). The integration of these standards into our tools for automating data linkages will support researchers in the execution of time-consuming, error-prone tasks, while at the same time increasing transparency and reproducibility. Furthermore, it will encourage best practices in the scientific community by providing researchers with a useful incentive to adopt basic standards for data interoperability.

To illustrate, briefly, how data structure interacts with data linkage, consider a hypothetical example of two ways of organizing the same data. A typical, well-structured data set contains data along its rows—one observation of one unit of analysis per row—and its columns denote features of those observations. The example of the CMP data above is an example of a well-structured data set: the unit of analysis is a party in a country at an election, and the data contain columns for party, country, and election alongside measurements of party preferences in the respective manifesto. In practice, however, researchers can mix the unit of analysis in the definitions of columns, for example by performing "long-to-wide" transformations. For example, if we restricted the CMP data to have columns for country and election, alongside columns labelled something like "party_1_preference" and "party_2_preference." The latter data structure is a denser structure, but by mixing the unit of analysis with the definition of columns it makes linking data sets much more difficult by adding several steps of reshaping data to the process of completing a merge. Although automatic transformations can be done straightforwardly in most statistical software, this sort of data structure is something good data management practices would discourage because it obscures the unit of analysis and adds unnecessary steps to the data-management process.

3.2 Shared vocabulary for social science text analysis

Political texts in European democracies are produced by a variety of actors (e.g. political parties, interest groups, media organizations, government institutions, etc.) across many countries and refer to many concepts. Furthermore, social scientists employ a variety of measurements and methods to study these texts. Linking data resources requires, first, that the features in these data resources enjoy a baseline interoperability. Furthermore, however, researchers must share a vocabulary or taxonomy that supports interoperability. *Controlled vocabularies* are a widely employed tool by which communities of experts rationalize and standardize taxonomies (Francart, et al., 2019; Harpring, 2010; Lancaster, 1986).

We have launched the process of constructing a controlled vocabulary to standardize the language used to name variables or features in social science text analysis data resources. This process will proceed in several stages and its progress will feature in the deliverables 8.2, 8.3, 8.4, and 8.5.

The first stage in constructing a controlled vocabulary is typically a broad exploration of terms applied currently by the community of experts (e.g. Mundie & McIntire, 2013; Moine, et al., 2014). This is always followed by several rounds of refinements, consultations with the wider expert community, and consensus building. We are in the first stage of this process now, consulting codebooks of the data resources in the present inventory, published research, and colleagues to establish a broad-based starting point for a taxonomy of entities, measurements, concepts, and events relevant to social science text analysis.

Below we provide a brief excerpt from the controlled vocabulary working draft. Initial categories in the typology have been identified by reviewing the codebooks of several of the data sets included in the inventory, providing us a starting point for the set of concepts, labels, measurements, and other features that the controlled vocabulary must address. Since this is a working draft, it will be updated regularly as this work proceeds.

Our current controlled vocabulary model aims to allow for consistent variable names across data sets of different types by constructing variable names from a combination of prefix, root word, and suffix. The set of prefixes, root words, and suffixes in the controlled vocabulary will have consistently definitions and scope conditions to enable consistent application.

Level 1 (variable name prefixes, excerpt) – denotes data type of variable

- ID: Unique numeric entity identifier
- IND: Binary indicator variables (i.e. 0 / 1 values only)
- N: Count variables (i.e. non-negative integers)

- DT: Dates always formatted as YYYY-MM-DD (indifferent to punctuation)
- TM: Time stamps formatted as YYYY-MM-DD HH:MM:SS (again, indifferent to punctuation)

Level 2 (variable name root words, excerpt) – denotes entity that is the object of measurement

- Entities

o PARTY: political party
o GOV: government institution (ministry, agency, etc.)
o GROUP: organized interest group or organization (not a firm, party, media organization, or government institution)

- Events

o ELEC: election
o PUB: publication
o HRG: Hearing

Level 3 (variable name suffixes – can stack multiple) – denotes details of measurement

- Identifier / standard

o Reference to a standard for entity ID/name, e.g.
□ Party IDs, including CMP, ParlGov, etc.
□ Country IDs, including ISO 3166-1 alpha-2, ISO 3166-1 alpha-3, or ISO 3166-1 numeric
□ Interest group or media outlet names/IDs from Wikidata
o Reference to a methodology, as modifier to a measurement, e.g.
□ Ideology/preference scaling method, including Wordfish, DW-NOMINATE, Wordshoal, etc.
□ Class label estimation method, including Topic models

Applying this controlled vocabulary results in variable names of the following variety:

- ID_PARTY_CMP – unique codes identifying political parties using the CMP system
- N_HRG – count of hearing events
- DT_ELEC – date of election

A process of testing and consultation lies ahead in building out and refining this controlled vocabulary. This excerpt provides a brief look at the logic we have adopted from similar systems constructed by other groups of researchers to bring consistency and clarify language in their own fields. The ambition for this process is that completing this project will both achieve a greater clarity of language and ease the process of data linkage across the field. Consistent language and variable names will help illuminate which linkages are possible and desirable across data sets, it will help make those merges mechanically simpler, and it will contribute to making the process of data linkage more easily and smoothly automated. Ultimately, these advantages will accumulate to strengthen the field's ability to answer substantive social science research questions.

References

Codd, E.F. (1990). *The Relational Model for Database Management: Version 2*. Addison-Wesley Longman Publishing.

Francart, T., Dann, J., Pappalardo, R., Malagon, C. & Pellegrino, M. (2019). The European Legislation Identifier. *Knowledge of the Law in the Big Data Age.* Peruginelli, G. & Faro, S. eds. IOS Press.

Harpring, P. (2010). *Introduction to Controlled Vocabularies: Terminology for Art, Architecture, and Other Cultural Works*. Getty Publications.

International DOI Foundation. *DOI Handbook.* https://www.doi.org/hb.html.

Lancaster, F. W. (1986). *Vocabulary control for information retrieval*. Information Resources Press.

Moine, M.-P., Valcke, S., Lawrence, B. N., Pascoe, C., Ford, R. W., Alias, A., Balaji, V., Bentley, P., Devine, G., Callaghan, S. A., & Guilyardi, E. (2014). Development and exploitation of a controlled vocabulary in support of climate modelling. *Geosci. Model Dev, N. 7.* 479-493.

Mundie, D. A. & McIntire, D. M. (2013) *The MAL: A Malware Analysis Lexicon*. Technical Note, CMU/SEI-2013-TN-010. Software Engineering Institute.

Pratt, P.J. & Last, M.Z. (2015). A *Guide to SQL.* 9th ed. Cengage Learning.

Schreiber, G., & Raimond, Y. (eds.) (2014). *RDF 1.1 Primer: W3C Working Group Note 24 June 2014*. https://www.w3.org/TR/rdf11-primer/.

Silberschatz, A., Korth, H. F., & Sudarshan, S. (2020). *Database System Concepts*. 7th ed. McGraw Hill Education.

Wickham, H. (2014). Tidy Data. *Journal of Statistical Software, N. 54:10*, http://dx.doi.org/10.18637/jss.v059.i10