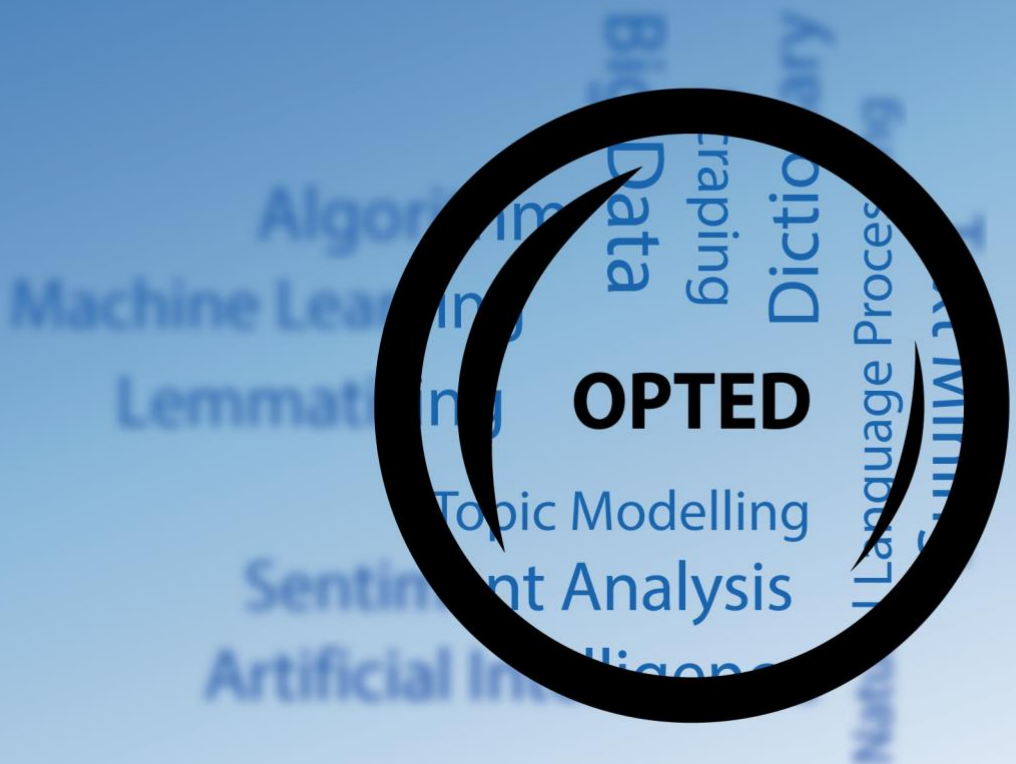
OPTED

Inventory of Concepts & Measurements

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

Report

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

Inventory of Concepts & Measurements

Deliverable D8.2

Shaun Bevan[1]

[1] University of Edinburgh

Due date: September 2021

Executive Summary

Based on Work Package 8's (WP8) focus on enabling the linkage of political text datasets to further their study and usefulness, this deliverable builds on the efforts of D8.1 through a focus on coded concepts and measurements. Specifically, this deliverable uses the inventory of datasets from D8.1 and its analysis of data linkages to unpack the concepts and measurements included in that inventory. It accomplishes this through the introduction and application of a framework for the categorization of coded variables on the D8.1 inventory of data sources. This deliverable both demonstrates the theoretical and practical complexity of variables coded from political text, as well as introducing additional considerations for WP8 and OPTED more broadly moving forward.

1 Introduction

The analysis of political text tends to favor the theoretical and/or practical elements of what is said or written in politics through the identification of concepts with terms like platform, frame, issue, and other common political terms. The definitions and bounding of these concepts can vary a great deal both within fields and especially between fields and subfields of social science. However, measurement tends to be far more consistent, with many common rules for human coding (e.g., issues or frames) or the use of the same or very similar tools for the analytical coding of (e.g., topics or sentiment) text. Subfields often use and measure these concepts consistently, but here some important differences can lead to confusion. For example, the difference between what is a priority and what is a preference when considering political opinions has led to considerable confusion in opinion research with debates over findings often being driven by terminology rather than differences in findings (see Jennings and Wlezien 2011; Bevan et al. 2016). The concept of framing is another example, where debates over what frames are and how they should be operationalized has muddled the progress of research in this area (see Cacciatore et al. 2016). Another example is the measurement of valence or sentiment where terminology is more consistent, but where different methods (e.g., dictionary-based and hand coding) tend to produce quite different results (see Boukes et al. 2020; van Atteveldt et al. 2021). Therefore, understanding the concepts and measurements used by work on political texts is essential to validly link and harmonize datasets, and existing research outputs. Beyond these linkages, setting standards for concept and measurement naming within the project can also help influence the wider community of political text researchers promoting good, transparent scientific practice.

2 Approach

The starting point of the Inventory of Concepts & Measurements (D8.2) detailed in this report is the Inventory of Data Sources and Linkage (D8.1) also created by WP8. The distinction between these two deliverables is their core focus. D8.1 samples coded political datasets detailing the metadata and identification variables included in these datasets to provide an overview of what is available while exploring the possible linkages between these and other similar datasets. This report (D8.2) focuses on the coded variables included in these datasets. Coded variable is a broad term which captures any variable that aims at measuring a concept gained from an analysis of part or all of the text in an observation. For example, a topic assigned through hand coding, a dictionary, or computational methods is a coded variable (in some fields the term annotations is more commonly used, although this deliverable uses the term coded variable throughout). This is different from an identification variable, such as a subject area, that is part of the data and not defined by the researchers analyzing the text. In short, coded variables are the information added by researchers to datasets as a measure of a concept.

This effort (D8.2) categorizes coded variables from D8.1 and deconstructs them by unpacking each coded variable based on Measurement, Method, Text Type, Stated Concept and most importantly General Concept. This deliverable (D8.2) will serve as a guide for understanding concepts and measurements across text datasets more broadly, including data created as part of OPTED and associated research in the future. Table 1 details the five components of coded variables.

Table 1: Components of Coded Variables	
Component	Description
General Concept	Theoretical Ideal of the Variable
Stated Concept	Definition and Operationalization
Text Type	Written, Spoken, and the medium of the text
Method	e.g., Hand coding and automated content analysis including supervised/unsupervised methods. Examples include machine learning, clustering methods including topic modelling, word embeddings, dictionary-based approaches, etc.
Measurement	Number, Title, Code, Nested Coding (e.g. CAP Codes), etc.

The **General Concept** of a coded variable is the theoretical ideal the variable intends to measure. For example, economic prosperity is a General Concept used in a great deal of social science research. While the General Concept is often explicitly stated, in many instances the General Concept is not concrete, as the definition and supporting research/literature is still developing or is debated. Additionally, many datasets focus on operationalized concepts or methodological improvements to these concepts' measurement, rather than starting from base theories. Continuing the economic example, a new way of measuring a Stated Concept like unemployment with no consideration of the General Concept of economic prosperity is a fairly common approach in social science research.

A coded variable's **Stated Concept** is the definition and operationalization of the variable chosen by the researcher. The Stated Concept in a piece of research focused on economic prosperity could, for instance, be GDP, GDP per capita, inflation, job creation, the misery index, unemployment, or even public opinion on the economy developed through the analysis of political text just to name a few. In some instances, the Stated Concept has the same name as the General Concept, but it is important to remember that the two can still differ despite sharing a name. For example, individual preferences is a General Concept that interests many social scientists, that refers to an individual's thoughts and desires. When operationalized into a Stated Concept, "individual preferences" comes to refer to a *measurement* of those thoughts through a survey, interview, or a text analysis of an individual's statements.

Text Type plays an important role with many coded variables. Spoken word tends to have more errors, uses language differently, and can be less direct. The written word is generally more formal and often longer, can include multiple authors even when attributed to only one, and is more often edited or revised before being made public. The line between these is often blurred in text like social media posts where colloquial language, informality, and errors are common and vary widely between authors. The use of different words, ways of communicating, and the potential involvement of multiple, often hidden actors, affect how researchers and their methods understand and code the data. Considerations of Text Type can also extend to the part of text considered, such as the whole document, a summary, or even just the title that is used for coding. Language can play another important role where the choice to code in the original language versus using translated text often makes little difference for stated concepts like issues, but can make all the difference for e.g., sentiment, where linguistic nuance plays more of a role.

The **Method** used to create a coded variable is generally quite clearly stated in publicly available datasets. However, a mixture of different methods, such as hand coding in support of supervised topic modelling, can be used with the exact method or methods used to create the coded variable for each observation often obscured in the dataset, through the coding process itself, and/or due to the quality of coding. More importantly, the method can introduce different biases or artifacts in the coded variable based on several factors that affect their reliability, precision, and recall. These can include unclear coding rules, improperly cleaned or pre-processed data, the use of too many or too few measurements of the coded variable, and more.

Coded variables are presented as a **Measurement**, and this can take many forms. It can be as simple as a number, code, or word. It can also be nested in or grouped with related values being a subset of a larger whole, different parts or regions of some value, or coordinates on some scale.

Text research often starts with different components, such as the raw data (Text Type) or a Stated Concept, and it may leave out any consideration of the General Concept. This is frequently true in the case of inductive research, focused on finding patterns in data, or when new methods or data are analyzed with a well-established Stated Concept.

3 Categorizing Coded Variables

Table 2 is based on D8.1's inventory and includes an overview of the coded variables contained in each of the datasets listed in that deliverable as of September 2021. For each coded variable the five components listed in Table 1 are listed based on an analysis of the dataset and its associated codebooks.

TABLE 2: COMPONENTS OF CODED VARIABLES FROM D8.1

Dataset	Variable	General Concept	Stated Concept	Text Type	Method	Measurement
Comparative Manifesto Project	Issue	Policy Preferences	Policy Area and Direction	Written (including "manifesto like" documents)	Hand Coding	Policy Codes Annotated as Positive or Negative
Content analysis of European issue salience	Actors, Issues, and Focus	Party Communication	Actors, Issues, and Focus	Written News Stories and Press Releases, Spoken TV Spots	Hand Coding	Categorical Codes
Issue Competition Comparative Project (ICCP)	Issue	Party Priorities	Issue Priorities	Written Tweets by Political Parties	Hand Coding	Counts and shares of attention
Social Media and Political Agenda Setting	Issue Salience	Agenda Setting Power	Political Salience	Written News Articles, Party Tweets, and Politician's Tweets	Supervised Machine Learning (multiple classifiers)	Categorical Codes
INTEREURO	Frames and Positions	Lobbying Goals	Frames and Positions	Written New Stories and Lobbying Proposals	Hand Coding	Dummy and Categorical Variables
Comparative Agendas Project	Major topic and Subtopic	Issue Attention	Policy Topics	Written, Transcripts, and Spoken	Hand Coding, Supervised Topic Modelling, Dictionaries	Subtopic Codes Nested in Major Topics Codes
DICEU	Ideal Points	Government Positions	Approval and Ideal Points	Video of Council Deliberations and Written Transcripts	Human Coding and Scaling	Scaled Ideal Points

For example, the CMP (Comparative Manifestos Project) is focused on policy preferences, in other words, the policy area and what policy is proposed as a General Concept. The Stated Concept is often also called policy preferences, however as operationalized the Stated Concept is more accurately a combination of policy area and if the mention is positive or negative for that area. In many analyses using the data, the positive or

negative coding is either ignored by combining both sides into a measure of policy area alone or the Stated Concept is split into two variables. The Text Type used are quasi-sentence mentions in party manifestos, with quasi-sentences representing complete policy mentions if not complete sentences (Volkens 2002) in the political party's native language using country-level experts and a common codebook. In instances where party manifestos do not exist, similar data on party policy preferences is used. For the majority of publications using CMP data, the Method used to produce it has been hand coding following a common codebook. More recent data has experimented with computational and other methods, although this is not yet widely used. Finally, the CMP produces a Measurement of policy codes annotated as positive or negative that are often grouped, such as the economy-focused coding for protectionism (positive 406; negative 407) and economic goals (408). Additionally, while the Measurement is inherently directional most codes are not in fact directional like economic goals (408).

Moving on to another example, CAP (Comparative Agendas Project) datasets focus on the General Concept of issue attention for different political and politically relevant texts such as laws, legislative debates, media outputs, public opinion and more. The Stated Concept for CAP data is policy topic, the functional area or means for policy making such as housing, health, or civil rights. This moves the data away from pure issue attention, to how government addresses issues using different tools. The data covered by the CAP is generally written or written transcripts of spoken data that have been edited. In some instances, the data are based on raw spoken text and need pre-processing or special coding rules based on Text Type to make all data comparable. In addition, the CAP includes data from over 30 nations and nearly as many languages, further complicating this component of the coded variable. The Methods used for CAP coding have historically favored hand coding, and an element of hand coding is used in the coding process for almost every dataset. More recently, supervised topic modelling has been used based on gold-standard hand coded data - especially for larger datasets and for newer or smaller project teams. Finally, in some rare instances, dictionaries have been used for coding budget data where large volumes of data, but with strict naming policies, exist. The Measurement produced by the CAP is a subtopic nested within a major topic. For example, insurance (302) is a subtopic of the major topic, health (3).

By deconstructing coded variables within datasets, it is possible to look not only for potential linkages between datasets, but also for potential problems for the creation and usage of the data. On the surface, the coded data from a source like the CAP is a straightforward classification of issue attention, but in reality, the varying Text Types, languages, and the focus on how governments structure policy make the data different than an effort at coding issues that started inductively or that started by transcribing raw data into English or some other language. In the case of the CAP this has been addressed through a common Master Codebook (see Bevan 2019), specific country-level codebooks, and large hand coded, gold-standard datasets over time and across many languages. This helps ensure that the Stated Concept of policy topics can be applied to different Text Types and captured using different Methods, all to produce the same Measurement (CAP codes) intended to reflect the same General Concept (issue attention). Only by understanding the components of coded variables can researchers code data in a way that produces their intended outcomes. Additionally, examining the components of coded variables helps with identifying opportunities for linking the coded variables between datasets something WP8 will continue to investigate.

4 Future Considerations

The process of developing a framework for the categorization of coded variables has left a few open considerations for future WP8 deliverables and OPTED more broadly.

- Would the categorization of text analysis methods by theoretical approach and intent alongside process (e.g., hand coding, (un)supervised, and algorithm (or codebook)) be useful?
- Are coded variables internally valid and how can that be assessed? In other words, do the components of coded variables line up with one another so that measurements are reasonable reflections of General and/or Stated Concepts?
- How replicable are coded variables based on the details provided within datasets and their associated documents?

Within OPTED and in the work that follows from it these considerations should be discussed and addressed through careful planning. This would include thorough theoretical consideration of coded variables,

a more open discussion of the validity not only of measures, but how closely coded variables match concepts, and most importantly careful descriptions of processes through detailed documentation allowing for replication within and outside of the project. Combined these elements will help lead to more standardization of how concepts are coded as well as less confusion within and between fields.

References

Bevan, S., Jennings, W., & Wlezien, C. (2016). An Analysis of the Public's Personal, National and EU Priorities. *Journal of European Public Policy*, 26(6): 871-887.

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

Boukes, M., van de Velde, B., Araujo, T., & Vliegenthart, R. (2020). What's the tone? Easy doesn't do it: Analyzing performance and agreement between off-the-shelf sentiment analysis tools. *Communication Methods and Measures*, 14(2), 83-104.

Cacciatore, M. A., Scheufele, D. A., & Iyengar, S. (2016). The end of framing as we know it… and the future of media effects. *Mass Communication and Society*, *19*(1), 7-23.

Jennings, W., & Wlezien, C. (2011). Distinguishing between Most Important Problems and Issues? *Public Opinion Quarterly*, 75(3): 545-555.

van Atteveldt, W., van der Velden, M. A., & Boukes, M. (2021). The Validity of Sentiment Analysis: Comparing Manual Annotation, Crowd-Coding, Dictionary Approaches, and Machine Learning Algorithms. *Communication Methods and Measures*, 15(2), 121-140.

Volkens, A. (2002). Manifesto coding instructions. Discussion Paper FS III 02-201. *Berlin: WZB*.