# Digitalization and the Organization of Work in the Digital Government Research Library v17.5: A Structural Topic Model

Andres Aguilera Castillo

26 April 2022

## Abstract

Digital Government is a growing and vibrant multidisciplinary field of research, the fast increase in research output has challenged researchers to explore and use novel computational ways and methods to perform evidence synthesis on the extant literature and be able to map a scientific discipline, explore the thematic evolution over time and identify potential avenues for further research. The exploration of the linkages between digitalization and the organization of work remains relatively unexplored in a public sector context. Topic modeling has emerged as a powerful computational technique that contributes to the examination of large amounts of text data. This chapter aims to perform a 'smart literature review' on a subset of the Digital Government Reference Library (DGRL) version 17.5, by using text as data approach and a topic modeling technique known as Latent Dirichlet Allocation (LDA). To our best knowledge, this is the first attempt to use unsupervised machine learning techniques with this data set. This effort may contribute to creating a map of the field, identify the evolving themes in the literature and help to identify promising areas of research.

## Introduction

Recent trends in global scientific output demonstrate a rapid and sustained increase in the production of vast amounts of unstructured data in the form of digitized text. This bounty in research content is challenging researchers to explore and pursue novel methodological approaches and techniques to examine massive volumes of scientific publications. Essentially, the ever-growing amounts of bibliographic information available in almost any field of research exceed human capacity making it necessary to explore computational-assisted approaches for research synthesis. Studies in the history of science have identified a relatively sustained growth pattern in scientific publications over time, this exponential growth rate means a doubling in scientific output every 17 years approximately (Bornmann, Haunschild, and Mutz 2021). This level of growth might be attributed to the increased resources dedicated to the global scientific endeavor and consequently the communication of science via publications. However, it may also be due to what has been dubbed "salami sliced publishing" or the multiple publications of a single research study (Bornmann and Daniel 2007; Bornmann and Mutz 2015).

Digital Government Research (DGR) as a multidisciplinary research field is experiencing rapid growth in its research output. Contributions to this research domain come from established disciplines such as information science, computer science, organization science, sociology, public administration, and political science (Scholl 2021). The diversity in scope and methods of these disciplines converge in the field of Digital Government enriching it, but at the same time, raising questions as regards the lack of native theoretical developments, thus relying upon frameworks, theories, and conceptualizations from related disciplines (Bannister and Connolly 2015).

The advent of computerization and digitalization has had broad impacts in most aspects of contemporary life, including scientific research. Digitalization has influenced how research is designed and conducted, allowing

for the creation and increased availability of ever-growing data sets that require powerful computational methods and enhanced tools to handle abundant information (Meyer and Schroeder 2015). Case in point is unsupervised machine learning techniques for text analysis, this research technology can be used in a wide range of disciplines to examine databases, repositories and corpora thus expanding the methodological repertoire of scholars and opening an opportunity to explore large troves of data.

Research synthesis is part of the literature review process in which the extant scientific knowledge in each academic field is examined to help scholars understand the conceptual structure, themes, and debates to identify trends in the literature and potential areas for further research. This crucial task is labor-intensive, time-consuming, and restricted to a limited number of documents if conducted by traditional "manual" methods (Antons et al. 2020; Asmussen and Møller 2019). Still, computer-assisted text analysis does not substitute human intervention, instead it "augments our reading ability" (Grimmer, Roberts, and Stewart 2022), human judgement is deemed necessary for the evaluation and validation of the outcome of these models (Barberá et al. 2021).

This chapter explores the contents of The Digital Government Research Library in its version 17.5 (Scholl, 2021a), using a text as data approach and an unsupervised machine learning technique known as Latent Dirichlet Allocation (LDA). In its raw form, version 17.5 of the data set contains 16531 references related to the Digital Government Research (DGR) domain, the most prevalent types of documents are conference papers (33%) and journal articles (50%). Previous explorations of this reference library using bibliometric and scientometric approaches have revealed the thematic evolution (Alcaide–Muñoz et al., 2017), and identified the most influential journals, conferences and leading scholars in the field (Scholl, 2021b). To our best knowledge this is the first attempt to run a topic model for a corpus in the field of Digital Government Research.

Scholars exploring labor-saving technologies have applied similar techniques expanding the methodological repertoire available and inspiring a similar pursuit for the exploration of the impact of technological change in a public sector context (Arduini and Zanfei 2014; Montobbio et al. 2022).

RQ

Process

# Literature Review // Conceptual framework

You can also embed plots, for example:

## Methods and Data

The Digital Government Research Library version 17.5 is a large curated repository of publications contributing to the field of Digital Government Research (DGR), it contains more than 16500 references among its records. The Library can be downloaded from the website DGRL. The download package contains three types of bibliographic files BibTeX, RIS, and ENL (EndNote) that can be explored with diverse reference management software and or imported to R Studio for data exploration and data cleaning. By exploring the BiBTeX file, RIS file and ENL file, it can be observed that there are some information that is available in one of the file types mentioned above but missing in the other. For this exercise, the following variables have been deemed of interest for the analysis: type of reference (conference paper or journal article), year of publication, author, document title, publication title and the presence of abstract. Text is a type of unstructured data that requires processing before using it. For replicability purposes, the script for data wrangling, cleaning and processing is available in a GitHub repository.

The first criterion for keeping the reference for further analysis was the presence of the Digital Object Identifier (DOI) number, a unique number that was used as the value used to join the data sets and to avoid duplicates.

Abstracts, our main unit of analysis are present in some of the references in the RIS file, year of publication, a key covariate for the structural topic model

The file formats exploratioThe BIB file For this analysis, an initial exploration of the files indicated important differences in the was downloaded and fields of related metadata, the lion's share being journal articles n=8278 and conference papers n=5493. The unit of analysis for this analysis is the abstract of these two types of documents: conference papers and journal articles.

Table 1: Table 1. Contents of DGRL v17.5 without pre-processing

| Document Type | Number of Documents |
|---|---|
| journalArticle | 8278 |
| conferencePaper | 5493 |
| bookSection | 2084 |
| book | 636 |
| report | 33 |
| thesis | 3 |
| magazineArticle | 2 |
| manuscript | 1 |
| webpage | 1 |

In its raw and unprocessed form, the data set has a large proportion of missing values 80.75%, mostly clustered in metadata not considered relevant for the analysis. Out of the 87 columns just 8 are deemed pertinent and may render some insights. The following visualizations explore the missing values in a subset of the DGRL containing journal articles and conference papers.

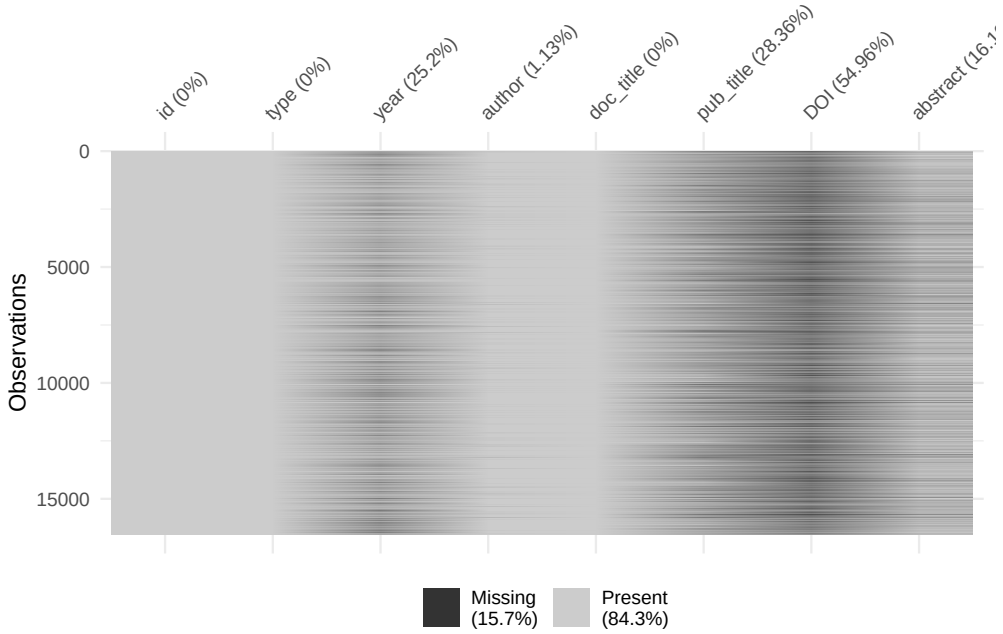
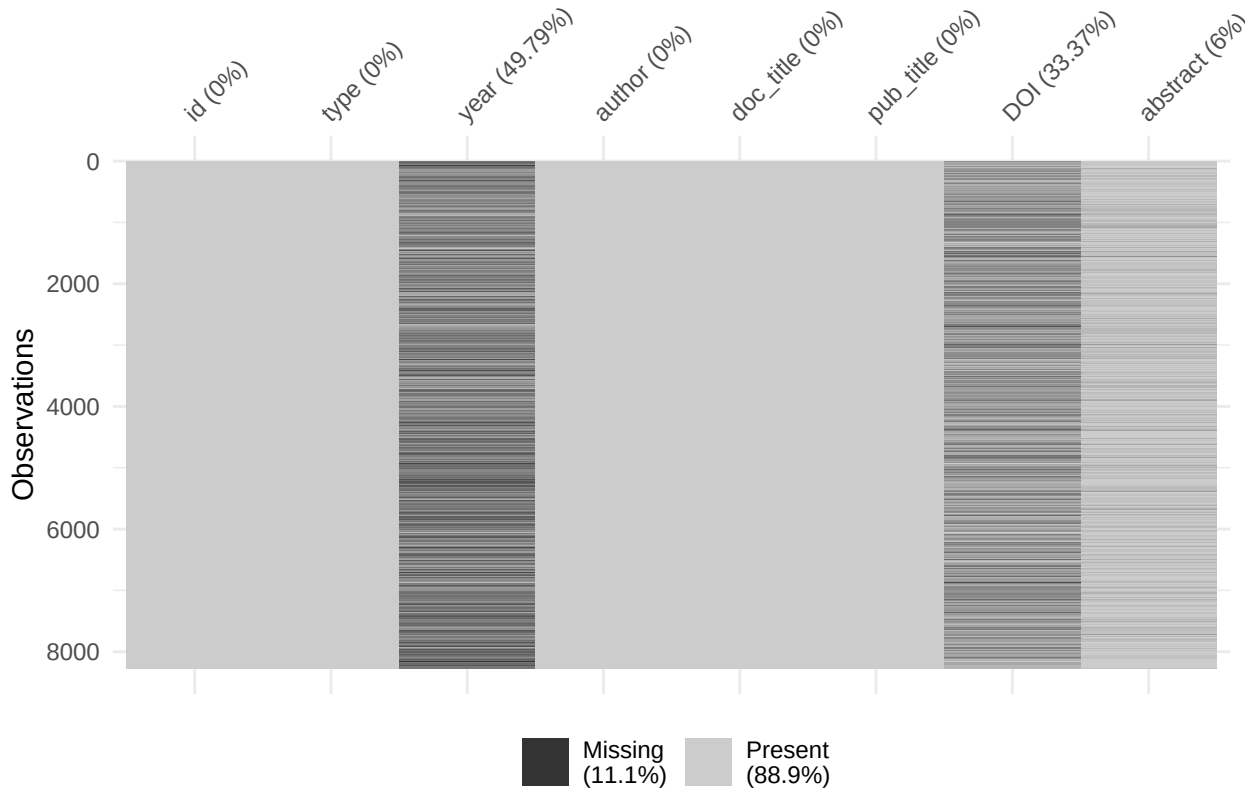


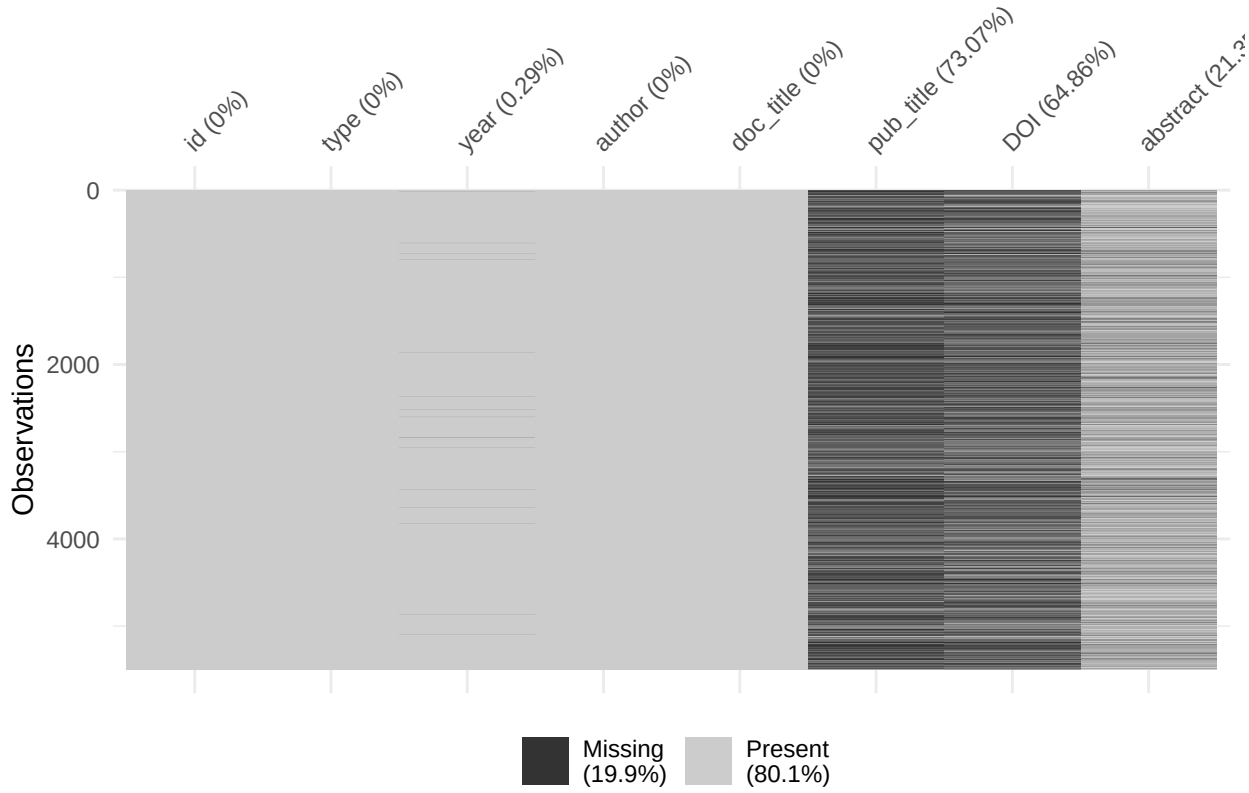

Figure 1: Percentage of missing values in the variables of interest

By visually exploring the contents of the data set, it is evident that a large amount of relevant information is missing. Important covariates for the analysis have a substantial proportion of missing information like year, publication title, also 16.16% of abstracts (our main unit of analysis) is missing. However, the

amount of missing data is very different from conference papers and journal articles as seen in the following visualizations.




Almost 50% of the metadata related to the year of publication, a key covariate for this analysis, is missing for the journal articles' subsample. In addition, there is missing a third of the Digital Object Identifier (DOI) information. To solve this problem and not to lose this information deemed relevant, we will describe the use a Zotero add-on DOI Manager to search for the missing information by using the available DOI numbers.




The conference paper subsample visualization tells a different story, most of the missing information is for the publication title column 73%, followed by the proportion of missing DOI information 64% and 21% of missing information for abstracts.

Zotero Search for Metadata and Retract Watch

The workflow for topic modeling includes text pre-processing meaning data cleaning and data transformation. Some variables of interest are missing and there must be some data preparation before running creating a corpus. In particular, we would like to explore if the covariate year, missing in journal articles could be imputed in a reproducible way by publishing the code used to process the data.

Our main unit of analysis is the abstract of the documents that have the covariates deemed of interest for our exploration. Text is a type of unstructured data, these type of data can be structured for processing using the bag of words approach or the splitting of documents, abstracts in our case, into separate word units or terms and every occurrence of a term is defined as a token. The creation of a bag of words is known as tokenizing. The bag of words approach deliberately ignores the syntax or structure of the text, additional treatment of text include the elimination of punctuation, transform each word to lowercase and in some cases stemming which is a way to reduce a word to its stem or root.

Stemming algorithms: Porter / Snowball / Lancaster

Stop words

## Results

## Discussion

## Conclusions

## References


Antons, David, Eduard Grünwald, Patrick Cichy, and Torsten Oliver Salge. 2020. "The Application of Text Mining Methods in Innovation Research: Current State, Evolution Patterns, and Development Priorities." *R&D Management* 50 (3): 329–51. https://doi.org/10.1111/RADM.12408.

Arduini, Davide, and Antonello Zanfei. 2014. "An Overview of Scholarly Research on Public e-Services? A Meta-Analysis of the Literature." *Telecommunications Policy* 38 (5-6): 476–95. https://doi.org/10.1016/j.telpol.2013.10.007.

Asmussen, Claus Boye, and Charles Møller. 2019. "Smart Literature Review: A Practical Topic Modelling Approach to Exploratory Literature Review." *Journal of Big Data* 6 (1): 1–18. https://doi.org/10.1186/S40537-019-0255-7/TABLES/6.

Bannister, Frank, and Regina Connolly. 2015. "The Great Theory Hunt: Does e-Government Really Have a Problem?" *Government Information Quarterly* 32 (1): 1–11. https://doi.org/10.1016/J.GIQ.2014.10.003.

Barberá, Pablo, Amber E. Boydstun, Suzanna Linn, Ryan McMahon, and Jonathan Nagler. 2021. "Automated Text Classification of News Articles: A Practical Guide." *Political Analysis* 29 (1): 19–42. https://doi.org/10.1017/PAN.2020.8.

Bornmann, Lutz, and Hans-Dieter Daniel. 2007. "Multiple Publication on a Single Research Study: Does It Pay? The Influence of Number of Research Articles on Total Citation Counts in Biomedicine." *Journal of the American Society for Information Science and Technology* 58 (8): 1100–1107. https://doi.org/10.1002/ASI.20531.

Bornmann, Lutz, Robin Haunschild, and Rüdiger Mutz. 2021. "Growth Rates of Modern Science: A Latent Piecewise Growth Curve Approach to Model Publication Numbers from Established and New Literature Databases." *Humanities and Social Sciences Communications 2021 8:1* 8 (1): 1–15. https://doi.org/10.1057/s41599-021-00903-w.

Bornmann, Lutz, and Rüdiger Mutz. 2015. "Growth Rates of Modern Science: A Bibliometric Analysis Based on the Number of Publications and Cited References." *Journal of the Association for Information Science and Technology* 66 (11): 2215–22. https://doi.org/10.1002/asi.23329.

Grimmer, Justin, Margaret E. Roberts, and Brandon M. Stewart. 2022. *Text as data: a new framework for machine learning and the social sciences*. Princeton, New Jersey Oxford: Princeton University Press.

Meyer, Eric T., and Ralph Schroeder. 2015. *Knowledge Machines: Digital Transformations of the Sciences and Humanities*. Infrastructures. Cambridge, Massachusetts: The MIT Press.

Montobbio, Fabio, Jacopo Staccioli, Maria Enrica Virgillito, and Marco Vivarelli. 2022. "Robots and the Origin of Their Labour-Saving Impact." *Technological Forecasting and Social Change* 174 (January): 121122. https://doi.org/10.1016/J.TECHFORE.2021.121122.

Scholl, Hans J. 2021. "The Digital Government Reference Library (DGRL) and Its Potential Formative Impact on Digital Government Research (DGR)." *Government Information Quarterly*, July, 101613. https://doi.org/10.1016/J.GIQ.2021.101613.