OID Library Dataset Documentation

Metadata on the research gathered and analysed by the Observatory on Information and Democracy First Research Cycle

OBSERVATORY ON INFORMATION AND DEMOCRACY

April 11, 2025

1 Description

Published by the Observatory on Information and Democracy (OID), the **OID Library Dataset** contains data for **2,739 academic sources** gathered by the Observatory during its first research cycle. Originally used to synthesise available evidence and produce a critical state-of-the-art on information and democracy, the data contains 22 variables that can be used to study the academic literature on the topic. Special focus in the dataset is devoted to sources dealing with the **crosscutting topic of mis- and disinformation** through the perspectives of:

1. **News Media**;
2. **Artificial Intelligence**;
3. **Data Governance**.

The report from the Observatory's first research cycle, titled Information Ecosystems and Troubled Democracy, has been published in January 2025.

1.1 Variables Description

The dataset contains **22 variables** explained here below:

- `Key` (str): unique key associated to the source.
- `Item Type` (str - categorical): type of document detected by Zotero.
- `Publication Type` (str - categorical): the type of source reworked in the analysis - takes values `Book Chapter, Book, Journal Article, Report, Conference Paper, Other`.
- `Author` (str - list): list of authors separated by "`;`".
- `Editor` (str - list): for book chapers and edited works, the editor(s) separated by "`;`".
- `Title` (str): title of the source.
- `Abstract Note` (str): if available, source abstract.
- `Publication Year` (int): year of publication of the source.
- `Publication Title` (str): for journal articles, the journal where the article was published.
- `Journal Abbreviation` (str): for journal articles, the abbreviation of the journal where the article was published.
- `Publisher` (str): for books, reports, and others, the name of the publishing company or website.
- `ISSN` (str).
- `ISBN` (str).
- `DOI` (str).
- `Url` (str).
- `Automatic Tags` (str): if available, the tags automatically detected by Zotero.
- `in_report` (str - categorical): title of the Observatory's report where the source was cited.
- `in_chapter` (int - list): chapter(s) of the Observatory's report in which the source was cited.
- `region` (str - categorical): region studied by the source (available for sources cited in the report).
- `main_topic` (str - categorical): the macro-topic that the source mainly deals with.
- `sub_topics` (str - list): sub-topics included in at least one of the source's paragraphs.

2 Data Collection

The data was collected throughout the Observatory on Information and Democracy's first research cycle. Such search was conducted using a combination of (i) a collaborative network consultation approach; (ii) keyword based search of academic databases; and (iii) call for papers targeted at gathering sources from the global majority world. Gathered sources were then analysed for quality and pertinence to the thematic priorities outlined in section 1 by the Observatory's Scientific Director, the report rapporteurs, and the OID Research Assessment Panels.

More information about the data collection methodology can be found at the report's methodology section.

3 Data Analysis

To develop additional metadata which can be used for research, the OID data team has conducted quantitative and qualitative analyses on the raw text of the sources gathered throughout the first research cycle. We make these variables available in our dataset. Here below we outline the methodology used to derived them :

- **Geographical variable**: the variable `region` (str) encodes information regarding which region the study refers to. It is the result of a manual coding conducted by the Observatory's Scientific Director, Prof. Dr. Robin Mansell and the first cycle lead rapporteur Prof. Dr. Rob Procter. It can take the values `global north, global majority, global`
- **Thematic variables** (`main_topic`, str & `sub_topics`, str): these provide information regarding what topics the source deals with. To identify this we used a NLP based method involving the following steps:
    1. **Chunking step**: we first chunked the raw text files of each source into paragraphs using LangChain's RecursiveTextSplitter;
    2. **Embedding step**: we then embedded the paragraphs using sentence-transformer's `all-MiniLM-L6-v2` model;

3. **Dimensionality reduction**: we projected the embedding space onto two dimensions using Uniform Manifold Approximation and Projection – or UMAP (McInnes & Healy, 2018)
4. **Clustering step**: we then run HDBSCAN, a density-based clustering algorithm based on hierarchical density by Campello et al. (2013), to identify clusters in the 2D space.
5. **Interpretation step**: finally, we interpreted the clusters combining a qualitative approach analysing the content of random samples from each cluster, and a quantitative approach combining BERTopic and LDA topic modelling. This yielded specific topics for each cluster. In a second interpretation step, we qualitatively abstracted sub topics to wider categories.

The computational work was conducted by the Observatory's data analyst Giovanni Maggi and the interpretation in cooperation with Iris Boyer, Head of the Observatory, and Prof. Dr. Robin Mansell, Scientific Director. The results of this analysis are also available as an interactive visualisation and our code is open source here.

Acknowledgements and Contribution

The **data collection** was conducted by the OID first research cycle – led by Iris Boyer, Head of the Observatory, Emma Gruden, Project officer, and under the guidance of Prof. Dr. Robin Mansell, Scientific Director and the Observatory's steering committee. The data collection benefited from numerous contributions across the OID network.

The quantitative **data analysis** and computation of topics was conducted by Giovanni Maggi, OID Data Officer. The qualitative regional coding has been conducted by Prof. Dr. Robin Mansell and Prof. Dr. Rob procter.

The OID Library Dataset was formatted, documented, and compiled by Giovanni Maggi, and benefited from the advice of Dr. Anselm Küsters and Dr. Paul Bouchaud.

Cite As

Observatory on Information and Democracy (2025). *OID Library Dataset*. Paris: Observatory on Information and Democracy.

Correspondence

All questions and enquiries concerning the dataset can be addressed to giovanni.maggi45@gmail.com

The Observatory on Information and Democracy is an Initiative of the Forum on Information and Democracy. Learn more at observatory.informationdemocracy.org