JuDDGES (Judicial Decision Data Gathering, Encoding and Sharing) DMP CHIST-ERA

Candice Fillaud, Chérifa Boukacem-Zeghmouri

JuDDGES (Judicial Decision Data Gathering, Encoding and Sharing) DMP CHIST-ERA

Version 1

Description

This document corresponds to the Data Mangement Plan for the JuDDGES (Judicial Decision Data Gathering, Encoding and Sharing) project and aims to apply FAIR principles to data and software produced. It will be updated regularly, at least every semester until the end of the project.

The overall goal of the JuDDGES project is to harness state-of-the-art Natural Language Processing & Human-In-The-Loop technologies to provide legal researchers with new Open software and tools that enable extensive, flexible and on-going meta-annotation capability (both automated and employing domain experts in-the-loop). This capability is applied to legal records/judgments from criminal courts across jurisdictions with varied legal constitutions (Poland, England & Wales).

The project is organized into 4 WPs :

Website link coming soon

 DOI: - 17/07/2024

WP4: Open science practices & engaging early career researchers

WP3: NLP and HITL Machine Learning Methodological Development

WP2: Gathering and Human Encoding of Judicial Decision Data

WP1: Project management

Funder

CHIST-ERA||CHIST-ERA

Grant

Judicial Decision Data Gathering, Encoding and Sharing/ No ANR-23-CHRO-0001

Researchers

Krzysztof Kaminski (orcid:0000-0003-2103-371X), Mandeep Dhami (orcid:0000-0001-6157-3142), Chérifa Boukacem-Zeghmouri (orcid:0000-0002-0201-6159), David Windridge (orcid:0000-0001-5507-8516), Łukasz Augustyniak (orcid:0000-0002-4090-4480), Tomasz Kajdanowicz (orcid:0000-0002-8417-1012), Santosh Tirunagari (orcid:0000-0002-9064-1965), Michał Bernaczyk (orcid:0000-0001-7683-8852)

Organizations

Wrocław University of Science and Technology, Middlesex University, Université Claude Bernard Lyon1

1. Basic Information

Title: JuDDGES (Judicial Decision Data Gathering, Encoding and Sharing) DMP CHIST-ERA

Description:

This document corresponds to the Data Mangement Plan for the JuDDGES (Judicial Decision Data Gathering, Encoding and Sharing) project and aims to apply FAIR principles to data and software produced. It will be updated regularly, at least every semester until the end of the project.

The overall goal of the JuDDGES project is to harness state-of-the-art Natural Language Processing & Human-In-The-Loop technologies to provide legal researchers with new Open software and tools that enable extensive, flexible and on-going meta-annotation capability (both automated and employing domain experts in-the-loop). This capability is applied to legal records/judgments from criminal courts across jurisdictions with varied legal constitutions (Poland, England & Wales).

The project is organized into 4 WPs :

- WP1: Project management
- WP2: Gathering and Human Encoding of Judicial Decision Data
- WP3: NLP and HITL Machine Learning Methodological Development
- WP4: Open science practices & engaging early career researchers

Website link coming soon

Researchers:
Krzysztof Kaminski (orcid:0000-0003-2103-371X)
Mandeep Dhami (orcid:0000-0001-6157-3142)
Chérifa Boukacem-Zeghmouri (orcid:0000-0002-0201-6159)
David Windridge (orcid:0000-0001-5507-8516)
Łukasz Augustyniak (orcid:0000-0002-4090-4480)
Tomasz Kajdanowicz (orcid:0000-0002-8417-1012)
Santosh Tirunagari (orcid:0000-0002-9064-1965)
Michał Bernaczyk (orcid:0000-0001-7683-8852)

Affiliations:
Wrocław University of Science and Technology
Middlesex University
Université Claude Bernard Lyon1

Contact: Candice FILLAUD

2. License

3. Funding

Funder: CHIST-ERA||CHIST-ERA

Grant: Judicial Decision Data Gathering, Encoding and Sharing

4. Data Management

Descriptions

WP2_Raw & Instruct Datasets

The main aim of WP2 is to establish the data to develop and test the tool. The objectives will be to (1) collate/gather openly sourced legal case records and judgments for coding; (2) develop a coding scheme for extracting data from legal case records and judgments; (3) train human coders to reliably code data from legal case records and judgments using the specially designed coding scheme; (4) make human only coded data available for use by WP3, (5) facilitate human-in-loop coding for WP3, and (6) enable WP4 to make data Open and re-usable by others beyond the project team. It also aims to allow the WP compliant to the standard of Open Publishing.

To date, the datasets produced are those corresponding to objective (1). Here are the 4 datasets produced and available:

- **JuDDGES/pl-court-raw**: Raw data from Polish court decisions
- **JuDDGES/pl-court-instruct**: Annotated and structured data from Polish court decisions
- **JuDDGES/pl-court-graph**: Graphical representations of Polish courts judgements
- **JuDDGES/en-court-raw**: Raw data from England and Wales Appeal Court judgements

It is planned to possibly modify or version them by adding new metadata or new judgments rather than creating new datasets.

Template: CHIST-ERA Data Management

Type: Dataset

1.1 What data (for example the kind, formats, and volumes), will be collected or produced?

1.1.1 Give details on the kind of data

a. • Derived or compiled (e.g.

• text mining

• 3D models)

b. Other

The collected datasets come from digital databases (particularly government databases); they consist of primary digital data (raw data) and secondary data.

- **JuDDGES/pl-court-raw**: The dataset consists of Polish Court judgements available at https://orzeczenia.ms.gov.pl/, containing full content of the judgements along with metadata sourced from official API and extracted from the judgement contents. This dataset contains raw data.
- **JuDDGES/pl-court-instruct**: The dataset consists of Polish Court judgements available at https://orzeczenia.ms.gov.pl/, containing full content of the judgements along with metadata sourced from official API and extracted from

the judgement contents. This dataset is designed for fine-tuning large language models (LLMs) for information extraction tasks and is formatted as instructions.

- **JuDDGES/pl-court-grap**h: This dataset is primarily based on the JuDDGES/pl-court-raw. The dataset consists of nodes representing either judgments or legal bases, and edges connecting judgments to the legal bases they refer to. Also, the graph was cleaned from small disconnected components, leaving single giant component.

- **JuDDGES/en-court-raw**: The dataset consists of England and Wales Appeal Court judgements available at https://caselaw.nationalarchives.gov.uk/judgments/advanced_search?court=ewca/crim/, containing full content of the judgements from official website. This dataset contains raw data.

1.1.2 Give details on the data format

The data consists of primary digital data (raw data) and secondary data (for analysis) extracted via API or HTML parser.

The formats are non-proprietary and open source, such as parquet, JSON, xml and PyG formats.

1.1.3 Justify the use of certain formats

widely supported format across disciplines

The .parquet format is a commonly used format in the discipline. Given that the data is on the HuggingFace platform, it is the most appropriate format (more information here: "*Parquet is a columnar storage format optimized for querying and processing large datasets. Parquet is a popular choice for big data*

processing and analytics and is widely used for data processing and machine learning.").

For this dataset: JuDDGES/pl-court-graph, the JSON format is intended for analysis and contains most of the attributes available in JuDDGES/pl-court-raw. We excluded some less-useful attributes and text content, which can be easily retrieved from the raw dataset and added to the graph as needed. The PyG format is designed for machine learning applications, such as link prediction on graphs, and is fully compatible with the Pytorch Geometric framework (more information here).

1.1.4 Give details on the volumes

GB (gigabyte)

It depends on the dataset:

- **JuDDGES/pl-court-raw**: 10.3 GB
- **JuDDGES/pl-court-instruct**: 2.94 GB
- **JuDDGES/pl-court-graph:** 1.57 GB (PyG format)/ 250MB (JSON)
- **JuDDGES/en-court-raw**: 86.6 MB

1.2 How will new data be collected or produced?

1.2.1 Explain which methodologies or software will be used if new data are collected or produced or if third party data are used

Polish data are extracted via API from official websites including the Polish Court judgements available at https://orzeczenia.ms.gov.pl/.

For the UK data, an HTML parser has been developed to automatically browse the search results and download the XML files for each judgment. The data has been extracted from the England and Wales Appeal Court judgements available at https://caselaw.nationalarchives.gov.uk/judgments/advanced_search?court=ewca/crim/ and can be downloaded as XML or PDF files under the Crown Copyright license.

1.2.2 Explain how data provenance will be documented

For now, the WP2 datasets are published on HuggingFace platform and accompanied by enriched documentation within the 'dataset card': dataset cards help users understand the contents of the dataset and give context for how the dataset should be used. You can also add dataset metadata to your card. The metadata describes important information about a dataset such as its license, language, and size. We are considering creating a datasheet template for the JuDDGES project data. This to provide consolidation to our approach in terms of data sharing, but also in terms of enhancing open practices within the project.

2.1.1 What metadata and documentation will accompany the data?

2.1.1.1 Indicate which metadata will be provided to help others identify and discover the data

- Structural
- Administrative
- Descriptive
- Technical

2.1.1.2 Indicate which metadata standards will be used

OpenAIRE Guidelines

The data is documented through dataset cards. We do not use specific standards as we create our schema based on use cases and annotation guidelines.

2.1.1.3 Indicate how the data will be organised during the project

For now, the dataset names follow the naming convention 'project name/country-data type', for example: JuDDGES/pl-court-raw.

2.1.1.4 Consider what other documentation is needed to enable re-use

All the documentation needed is available within the datasets cards.

- https://huggingface.co/datasets/JuDDGES/pl-court-raw
- https://huggingface.co/datasets/JuDDGES/pl-court-instruct
- https://huggingface.co/datasets/JuDDGES/pl-court-graph
- https://huggingface.co/datasets/JuDDGES/en-court-raw

What the dataset cards contain:

- **Description**: An overview of the dataset, including its purpose, potential applications, and history.
- **Composition**: Details about dataset features such as the number of rows, data types, specific column characteristics, etc.
- **Usage**: Information on how the dataset can be used, including code examples, specific tasks it is suited for, and recommended configurations.
- **Creation**: Information on how the dataset was created, including data sources, collection process, cleaning methods, and authors or contributors.
- **Ethical Considerations**: Discussions on ethical considerations related to dataset use, such as potential biases, privacy impacts, and precautions to take.
- **License**: Information on usage rights and the license under which the dataset is published.
- **References**: Citations and references to related works or articles associated with the dataset.

The data will also be directly linked to the publications; the detailed methodology in the publications will help complete the metadata.

2.1.1.5 Consider how this information will be captured and where it will be recorded

Dataset cards, datasheets for dataset, readme text file

2.1.2 What data quality control measures will be used?

2.1.2.1 Explain how the consistency and quality of data collection will be controlled and documented

Data entry validation

POLISH DATASETS

In the first stage of ensuring data quality control for our Polish dataset, we utilized manually created metadata sourced from https://orzeczenia.ms.gov.pl/, a reputable repository of legal documents and

judgments. This initial step was crucial in establishing a reliable foundation for subsequent analyses.

To further enhance the consistency and quality of data collection, we will implement two key measures:

- **Schema Structure Validation:** We will employ JSON Schema validation tools (e.g., Python's jsonschema library available at https://python-jsonschema.readthedocs.io/en/stable/validate/) to verify that each data entry adheres to our predefined schema structure. This process involves checking whether the data conforms to specific formats, contains required fields, and meets all constraints outlined in our schema definition. By doing so, we can detect any discrepancies or anomalies early on and ensure that only well-structured data is included in our dataset.
- **Automated Data Quality Checks:** Following schema validation, automated scripts will be developed to perform routine checks on the dataset. These scripts will look for common issues such as missing values, duplicate records, outliers that deviate significantly from expected patterns, and inconsistencies across related fields. Any identified problems will be logged systematically for review by the research team.
- Both measures are designed to maintain high standards of data integrity and meticulously document every step taken during the validation process. The documentation generated through these controls provides transparency into how the dataset has been curated and maintained over time—essential information when sharing results with an academic audience or using this resource as a basis for further research endeavors.

By implementing these robust quality control mechanisms early in our workflow pipeline—and continuously refining them based on feedback—we aim to establish a trustworthy database capable of supporting rigorous scientific inquiry within Polish legal texts analysis.

UK dataset

An HTML parser has been developed to automatically traverse the search results and download the XML files for each judgment. The dataset includes the full content of 6,154 XML files with judgments published no later than 2024-05-15. This HTML parser is available in scripts/github.

3.1 Reused Data

3.1.2 Where can re-used data be found?

No datasets will be reused.

3.1.3 Which data will be re-used?

No datasets will be reused.

3.1.4 State any constraints on re-use of existing data if there are any

Not relevant

3.1.5 Briefly state the reasons if the re-use of any existing data sources has been considered but discarded

Not relevant

4.1.1 How will data and metadata be stored and backed up during the research?

4.1.1.1 Describe where the data will be stored and backed up during research activities and how often the backup will be performed

Daily

Polish dataset

During our research activities, we must ensure that data and metadata are stored securely and backed up regularly to prevent any loss or corruption. To this end, we have chosen MongoDB Atlas as our primary storage solution.

MongoDB Atlas Database Storage: MongoDB Atlas provides a fully managed cloud database service that offers high availability, scalability, and compliance with the most stringent data security standards. Our datasets and corresponding metadata will be hosted on MongoDB Atlas clusters, which are configured for optimal performance and reliability.

Backup Strategy: The backup protocol within MongoDB Atlas is robust and automated. We will perform backups of our entire dataset every 6 hours without fail. This frequent backup schedule ensures that in the event of an unexpected system failure or data corruption incident, so we can quickly restore the most recent version of our data with minimal loss.

The backups themselves are snapshots of the database state at consistent points in time. These snapshots include all collections, documents, indexes, and configurations present when each backup is taken. They are encrypted using advanced encryption methods to protect sensitive information from unauthorized access during transit and at rest.

In addition to our rigorous backup strategy within MongoDB Atlas, we will periodically create new versions of datasets for integration with the Hugging Face platform. This process involves exporting curated collections from our MongoDB database and transforming them into formats suitable for machine learning models hosted on Hugging Face.

Dataset Versioning for Hugging Face: We deliberately create dataset versions to track changes over time and provide researchers with access to historical and up-to-date data. By versioning our datasets, we facilitate reproducibility in research by enabling comparisons across different iterations of the same dataset.

Significant updates or enhancements made to the existing collections in MongoDB Atlas will determine the frequency at which these new versions are created. Each version released on Hugging Face will be thoroughly documented, detailing any modifications or additions since the previous release. This ensures transparency and provides users with clear insights into the evolution of the dataset.

By maintaining an active presence on platforms like Huggling Face, we contribute valuable resources to a broader community and invite collaboration and feedback that can lead to further improvements in our work. Our commitment extends beyond just storing and backing up data; it encompasses sharing knowledge and fostering advancements within the field through accessible, high-quality datasets.

Currently the UK datasets are stored locally on the data manager’s machine. They will gradually be backed up to MongoDB.

4.1.2 How will data security and protection of sensitive data be taken care of during the research?

4.1.2.1 Explain how the data will be recovered in the event of an incident

The platforms allow data retrieval online, knowing that they are open source, regarding MongoDB, see the previous question.

4.1.2.2 Explain who will have access to the data during the research and how access to data is controlled, especially in collaborative partnerships

- Santosh Tirunagari (orcid: 0000-0002-9064-1965)
- Łukasz Augustyniak (orcid: 0000-0002-4090-4480)
- Tomasz Kajdanowicz (orcid: 0000-0002-8417-1012)

4.1.2.3 Describe the main risks and how these will be managed

Pseudonymization or anonymization will be applied if necessary. For Polish datasets, the judgments are already pseudonymized in the databases. The backups in the database are encrypted to protect sensitive data.

4.1.2.4 Explain which institutional data protection policies are in place

a. Middlesex University London

b. Wroclaw University of Technology

c. Université Claude Bernard Lyon1

5.1.1 Personal data

5.1.1.1 Are there any personal data to be formulated?

No

5.1.1.2 Explain whether there is a managed access procedure in place for authorised users of personal data

Not at this stage of the project within this WP.

5.1.2.1 Data ownership and accessibility

5.1.2.1.1 Who will be the owner(s) of the data?

a. Wroclaw University of Technology

Łukasz Augustyniak (orcid:0000-0002-4090-4480)

b. Middlesex University

Santosh Tirunagari (orcid:0000-0002-9064-1965)

5.1.2.1.2 Explain what access will apply to the data?

Open

The data will be open to the Hugging Face community, which is an aknowledged open platform.

5.1.2.2 Intellectual property rights

5.1.2.2.1 Explain which intellectual property and how will they be dealt with

Other

Not relevant

5.1.2.3 Third-party data restrictions

5.1.2.3.1 Are there any restrictions on the re-use of third-party data?

No

5.1.3 Ethical issues

5.1.3.1 What ethical issues and codes of conduct are there, and how will they be taken into account?

Ethical issues surrounding personal data

A letter of approval regarding the ethical considerations of the project was issued on March 8, 2024 for all the documents such as Informed Consent Form, Data Protection Declaration, Participant Recruitment Information.

1. As mentioned in the project proposal:

'*The research will first undergo ethics review from the Department of Psychology Research Ethics Committee at Middlesex University London, followed by reviews from Wroclaw and Lyon where necessary and applicable. The University has a strict data protection and declaration procedure which adheres to 7 principles, namely, that the information is Fairly and lawfully processed, processed for specified and lawful purposes, adequate, relevant and not excessive, accurate and kept up date where necessary, not kept for longer than is necessary, kept secure, and necessary to actively demonstrate compliance with all of the above principles. The University abides by GDPR rules and exemptions. DHAMI has previously been granted ethics approval for student projects involving human encoding of court records/legal judgments including from the BAILII data resource and so we do not foresee any ethical issues.*'

6.1.1 How and when will data be shared? Are there possible restrictions to data sharing or embargo reasons?

6.1.1.1 Explain how the data will be discoverable and shared

• Deposit in a trustworthy data repository

• Indexed in a catalogue

6.1.1.2 Outline the plan for data preservation and give information on how long the data will be retained

To be decided and updated.

6.1.1.3 Explain when the data will be made available

Data are already on Hugging Face and GitHub.

6.1.1.5 Will exclusive use of the data be claimed?

No

6.1.1.7 Indicate whether data sharing will be embargoed or restricted

Embargoed

No restriction or embargoed.

6.1.1.8 Indicate who will be able to use the data

• Researchers

• Research communities

• Decision makers

• Education

• Economy

• The public

6.1.1.9 Is it necessary to restrict access to certain communities or to apply a data sharing agreement?

No

6.1.2 How will data for preservation be selected, and where data will be preserved long-term?

6.1.2.1 Indicate what data must be retained or destroyed for contractual, legal, or regulatory purposes

a. Retained

To be decided and updated.

6.1.2.2 Indicate how it will be decided what data to keep

To be decided and updated.

6.1.2.3 Describe the data to be preserved long-term

Other

To be decided and updated.

6.1.2.4 Explain the foreseeable research uses (and/ or users) for the data

In making new software, tools and data resources open on a public repository, researchers will be empowered to develop and empirically test theories of judicial decision making and address judicial policy and practice-relevant questions. Researchers from public (legal) institutions can also reuse the data for their purposes.

6.1.2.5 Indicate where the data will be deposited

Repository of University of Wroclaw

It has not been decided yet, but we are considering Zenodo.

6.1.2.6 Demonstrate that the data can be curated effectively beyond the lifetime of the grant

Zenodo is running from the CERN data center, whose purpose is long term and persistent preservation of digital objects.

No storage space limit per researcher, a DOI will be assigned to each dataset, the costs related to data management are included in the project within WP4.

6.1.3 What methods or software tools are needed to access and use data?

6.1.3.1 Indicate how the data will be shared

Repository

6.1.3.2 Indicate whether potential users need specific tools to access and (re-)use the ata.

There is no specific software required to reuse the data. The data is available as open access and in an open format.

6.1.4 How will the application of a unique and persistent identifier to each data set be ensured?

6.1.4.1 Select how the data might be re-used in other contexts

• To obtain information

• To share information

• To make informed decisions

6.1.4.2 What type of persistent identifier (PID) will be used?

DOI

7.1.1 Who will be responsible for data management?

7.1.1.1 Outline the roles and responsibilities for data management/stewardship activities

a. Łukasz Augustyniak (orcid:0000-0002-4090-4480)

Project Data Manager and Supervisor (PL datasets)

b. Chérifa Boukacem-Zeghmouri (orcid:0000-0002-0201-6159)

Coordinator and author of the data management plan

c. Candice FILLAUD (orcid:0009-0005-2084-3384)

Coordinator and author of the data management plan

d. Santosh Tirunagari (orcid:0000-0002-9064-1965)

Data Manager and Supervisor (UK datasets)

7.1.1.2 Explain the co-ordination of data management responsibilities across partners

Each partner is responsible for the collection and management of their own language data. Weekly exchange meetings are scheduled to discuss guidelines for annotators and schemas for information extraction and share practices. The teams also share on a common GitHub repository dedicated to the project.

7.1.2 What resources will be dedicated to data management and ensuring that data will be FAIR (Findable, Accessible, Interoperable, Re-usable)?

7.1.2.1 Explain how the necessary resources to prepare the data for sharing/preservation have been costed in

Infrastructure Grant

7.1.2.2 Indicate whether additional resources will be needed to prepare data for deposit or to meet any charges from data repositories

No

/

7.1.2.3 Explain how much cost provisionally is needed

/

At this stage, we do not have the exact amount, but a dedicated budget is allocated for the management and opening of the project data. This is included in WP4: 'Open science practices & engaging early career researchers', objective (1) provide open access to publication, data and software in accordance with the Open Science policy of the call.

- The storage of Polish datasets in MongoDB represents 100-200 USD monthly, the subscription will increase based on the size of the datasets.
- The digital infrastructures (HuggingFace and GitHub) are free and open source, which does not represent a cost to the project. However, within WP2 and the data management, the teams responsible for data management represent a human cost (collection, analysis, management, publication on digital platforms), which can be quantified in terms of working hours or salaries.

5. Software Management

Powered by