Publications Office
of the European Union

Initiative for Semantic Analysis of the Official Texts on COVID-19 measures in the European Union

Towards creating the Euro-COVID19 dataset for natural language processing

Disclaimer

Document metadata

Reference	WP 1.5: Towards creating the Euro-COVID19 dataset for natural language processing
Corporate Author	Publications Office of the European Union
Author	Eugeniu Costetchi
Reviewers	Nataliya Rozbroj Jasinskaja, Carlos Perez
Contractor	Eugeniu Costetchi
Legal framework	CFE 2019/S 041-092027 / BDC 36370
Work package	WP 1.5
Delivery date	5^{th} May 2021
Suggested readers	data scientists, information architects, business analysts,

Abstract

This document aims at describing how the Euro-COVID19 dataset was created. It comprises texts and metadata on COVID19 crisis measures taken by the European Union (EU) and the European Member States (MS). The creation of this dataset belongs to the SemCovid project financed by the Open Data Portal unit from the European Publications Office.

We provide, in this document, some of the underlying assumptions, a summary of methodological considerations, a description of the collected data and the technical stack used to perform the data processing.

Contents

1 Introduction

This document aims at describing how was created the *Euro-COVID19* dataset comprising texts and metadata on COVID19 crisis measures taken by the European Union (EU) and the European Member States (MS). The creation of this dataset belongs to the SemCovid project financed by the Open Data Portal unit from the European Publications Office, which is described elsewhere.

In the Euro-COVID19 dataset, data are collected from various sources, each providing content with different characteristics and available metadata. What is fetched from a selected source we also call a dataset due to its internal homogeneity and difference to the content available from other sources. Thus we say that the *Euro-COVID19 dataset* is a composition of different datasets.

Further in the document, we provide some of the underlying assumptions, a summary of methodological considerations, a description of the collected data and the technical stack used to perform the data processing.

A dataset shall fulfil a clearly defined goal. That is why, before diving into the *hows* of its creation, we first iterate through its whys.

2 Project goals: the whys

The current project goals originate from two stakeholders, whose interests differ in a complementary manner. The EU Open Data Portal team has an investigative novelty-seeking research orientation, while the EuroFound team is pragmatically oriented to satisfy the business needs of the European policymakers.

From a research perspective, this project sets out to *investigate and establish the semantic mapping of the European Union (EU) and Member States (MS) response to the COVID19 crisis in the area of living and working conditions.* This goal can be further elaborated in terms of the following research questions:

- How do the measures compare between Member States collectively and the EU regarding the issues they address, various types of categorisation and degree of content similarity?

- How do the measures compare between the Member States individually and also to the EU measures?
- What are the emerging topics in the dataset(s), and how did they evolve over time?

After discussing the business needs of policymakers with the EuroFound team, a set of prototypical business interests emerged that is best expressed as the following questions:

- Who has done what on which issue? What acting bodies are involved, and which categories of responses are used for classification?
- Who pays for the measures and how much? What sources of financing are employed, and what amounts are allocated for each measure or issue?
- Who benefits from the measure outcomes? What are the target groups for each measure?
- Where is the measure applicable? What is the territorial coverage of the measure?
- When is the measure executed? When is the measure adopted, and how long does it last?

3 Business features: the whats

To answer the above business question, we further ask ourselves what sort of properties (or features) shall be available in the datasets to enable answer computation. Having articulated these features explicitly serves as a compass in the data collection and data processing processes. And where the automation has reached its limitation, the same compass will indicate what sort of data shall be collected/produced and published by the data providers in the future.

To give you an example of how the business questions related to the business features, lets take BF8: the target group feature in Table 1. The dataset needs to provide explicitly in a dedicated field the information about the target groups of the measure to answer the business question “Who benefits from the measure

outcomes?" which is relevant to the policymakers.

ID	Business feature	Description
BF1	adopting entity	The organisation that adopts the COVID measure.
BF2	categories	A classification of the measure following a well-defined scheme manually assigned by the data provider.
BF3	issue date	The date when the measure is published.
BF4	temporal coverage	The beginning and the end dates of the measure applicability or execution leading to a duration definition.
BF5	spatial coverage	The definition of spatial coverage where the measure is applicable. Usually denoted by a codified reference to a territorial unit (country/region).
BF6	sources of financing	The mention of the financing source(s): usually denoted by a reference to an EU programme, special national or international fund, or a generic label such as "own funds" or "national budget" .
BF7	funding amounts	The total amount of money allocated to or spend for a particular source of financing.
BF8	recipients & beneficiaries (target groups)	The groups which shall benefit from the measure. Beneficiaries targeted by the action. The beneficiaries may be expressed either as groups of entities (e.g. SMEs, self-employed, etc.), demographically defined groups (e.g. elderly over 65, single-parent families, etc.) or functionally defined roles (e.g. doctors, policemen etc.).

continued on next page

continued from previous page

BF9	semantic similarity	In comparative text analysis studies, the semantic similarity represents how close is the meaning of two texts
BF10	textual description	The textual description of the measure.
BF11	topics	Topics of the measure automatically discovered by the machine learning techniques.

Table 1: Business features that necessary to answer the business questions

4 Building a dataset: the how

A dataset is a remarkable thing, not so much because it is a collection of data records, but because of the properties that it acquires if it is well-designed and carefully constructed. The guiding principles for conceiving a good dataset cannot be strictly defined but rely heavily on the good sense and clear thinking of the people involved and *feedback from a consensus of users*. For this, deciding upfront, "W*hat is the dataset for?*" , "*How will it be used?*" and "*What are the project goals?*" plays a crucial role in its design.

In addition, if the dataset involves the organisation of textual content (linguistic data), then it is important to understand and take into consideration *how language in general works*. One needs to pay particular attention to the *communicative functions*[1] and the *community* in which the texts arise. However unsteady is the notion of *representativeness*[2], it is an unavoidable one in the dataset design, and others such as *sample* and *balance* need to be faced as well. Therefore the *linguistic dataset* builders should strive to make the dataset as representative as possible of the language and purpose for which it is chosen.

[1]Metafunction is a systematic cluster, which groups semantic systems that make meanings of a related kind.

[2]Representativeness heuristic is simply described as assessing similarity of objects and organizing them based around the category prototype.

Any selection must be made on some criteria and the first major step in linguistic dataset building is the determination of the criteria on which the sources and the text that form the dataset will be selected.

1. The *text register*[3], that is a variety of language used for a particular purpose or in a particular communicative situation. For example, when speaking officially or in a public setting, an English speaker may be more likely to follow prescriptive norms for formal usage than in a casual setting.
2. The *(semantic) domain of the text*[4], which is a specific place that shares a set of meanings, or a language that holds its meaning, within the given context of the place. In lexicography a semantic domain or semantic field is defined as "*an area of meaning and the words used to talk about it*" . Many sports have specific semantic domains that entail terminology that is specific to that particular sport. In order to understand the meanings of these terms one would need to understand the context and domain of that sport.
3. The location of the text, for example Spain, Ireland or European institutions (as a pseudo location).
4. The date of the text, because language evolves over time or can be bound to certain events.

The criteria for determining the structure of a linguistic dataset should be small in number, clearly separate from each other, and efficient as a group in delineating a dataset that is representative of the language or variety under examination. The selected criteria and assumptions about them we address in the sections below, starting from delineating the dataset domain.

[3]Text register
[4]Text domain

5 Domain delineation: what is a COVID19 measure?

A pandemic such as COVID19 is so pervasive that it reaches deep into all aspects of human life. It serves as a thematic context and a broad discourse domain of the dataset language. In addition, however, we need to make more fine grained distinctions. This drive comes, especially from the EuroFound stakeholder, who emphasised interest in responses and not reactions or preventive actions to COVID19 pandemics.

In this section we establish the conceptual framework for making the above conceptual distinctions, first by human beings, and then, hypothetically, similar distincts could be done to a certain degree by the machines. In interaction with our stakeholders several terms were frequently used to refer to actions taken by public and private organisations in relation to the COVID19 pandemics: "*reaction*", "*response*", "*measure*", "*action*". When asked there seems to be no obvious official definition of these terms used in the context of COVID19 pandemic. So, there seems to be missing an official definition for each of these terms.

We take a generic crisis management framework, and attempt at simple definitions derived from the interactions with the stakeholders, which we slightly extended to complete the picture. The crisis management point of view offers four types of actions: *assessment*, *prevention*, *reaction* and *response*, which are depicted in Figure 1.

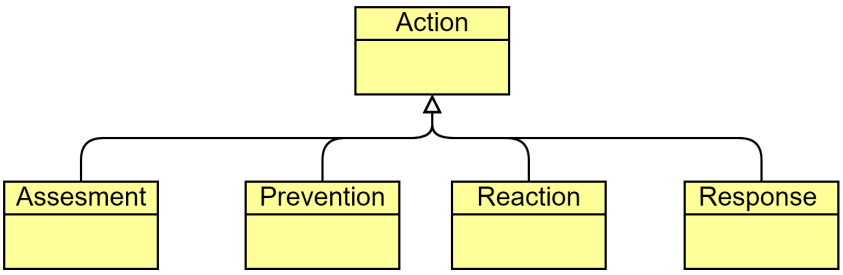

Figure 1: Types of actions in a crisis

"Action" is a broad term covering all sorts of activities. Under it are depicted four classes of actions: the first two (on the left side of the diagram) are notably occurring before the crisis event happens. The other two (on the right side of the chart) apply to the aftermath. The depiction of the types of actions ordered sequentially along the temporal axis is depicted in Figure 2 where the crisis event takes a central place.

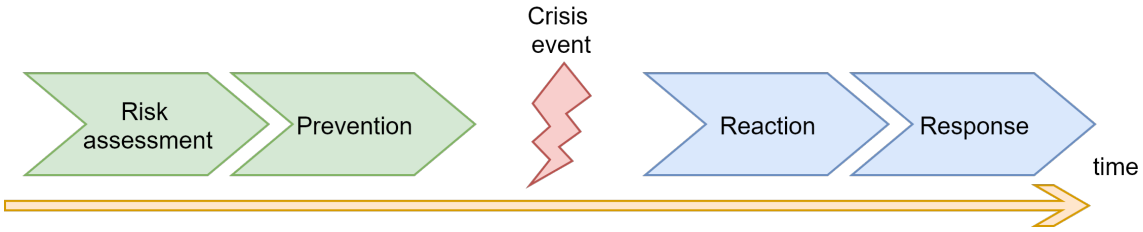

Figure 2: Phases in crisis management

The *risk assessment* actions deal with risk identification, estimation of probability for each risk factor to occur. Also, at this stage is investigated the impact from the environmental, economic, social and public health perspectives, including who is impacted, how, and to what degree.

The *prevention* actions focus on what could be done to mitigate and alleviate the crisis impact in case it occurs.

The *crisis event* (or period) leads to an unstable and dangerous situation affecting an individual, group, or all of society.

The *reaction* actions deal with immediate mitigation while the *responses* with medium to long-term mitigate the crisis consequences. These two action types: responses and reactions, which we collectively call *measures*, and in our case *measures to the COVID19 crisis*, are the primary focus of the dataset. So, next, we define them in this context specifically.

The *COVID19 reaction* is what an organisation plans to do to minimise the impact of the pandemics, such as including extra protective equipment for health workers, procurement of additional medical equipment, the introduction of social distancing, more rigorous border controls or their temporary closure. The reactions are actions related to the pandemic poses health threats: "a virus makes people

sick, and something must be done about that" . In addition to the pandemic consequences, the reactions have economic and social effects that can be directly observed or indirectly deduced. And so the reaction of the authorities to the virus affects the people too: consider, for example, the social distancing or the lockdown measures. We can say that the reactions are, in fact, to the health threat and that the main topics are anything related to health, pandemics, virus, infection etc.

The *COVID19 responses* are measures set in place to compensate for the economic and social consequences of the pandemic crisis and diminish the effects of the authorities' reactions to the threat. Therefore the temporal horizon of the responses is much longer, and the topics they deal with are much broader and challenging to pinpoint. They could be anything related to social, economic, welfare, child care, telework, income support, increase in social protection and many other topics.

Now, that the dataset domain is set to the COVID19 reaction and response measures, with a predominant preference for the COVID19 responses, we describe other criteria and provide some of our assumptions.

6 Criteria and assumptions about the data

- We investigate possible sources and collect text and metadata describing and classifying responses and reactions to the COVID19 crisis.
- If possible, we distinguish between reaction and response measures because the latter is of particular importance to the EuroFound stakeholder.
- We assume that the official documents are a reliable source of measure descriptions.
- The primary focus is on collecting official texts of various registers (administrative, journalistic, legal, etc.) to enable further exercises to train models that can identify and distinguish a COVID19 measure from other texts and, if possible, differences between the reactions and responses.
- Selecting the texts with these properties leads to a relatively scarce set of available resources. Therefore no (data) sampling was considered or any tests

for the (data) balancing. Mainly we aimed at covering selected initial sources given the apparent degree of (data) representativeness to the semantic domain, which remains to be tested and analysed.

- We could identify two text registers (or genres) in the available sources, which is a valuable distinction in the further analysis. First is the journalistic or reportage style of language, which summarises and describes COVID19 measures. Second is the legal style of language that characterises the source texts from where COVID19 actions summaries are extracted. Possibly more genres could be identified in a detailed analysis.
- Geographically, we select texts produced in the European Union, both at the EU level, by the institutions and at the member state level, by the national governments.
- We select texts written in English as a first step and as an established limit to the scope, and later, in future work, this can be extended to texts written in the official EU 24 languages.
- We consider only texts issued after the beginning of 2020 when the COVID19 pandemic started.
- A basic assumption is that Policy Watch Database (PWDB) is a suitable set of summarised descriptions of what a COVID19 measure looks like. They cover broader economic and social issues and intentionally exclude COVID19 reactions, focusing primarily on public health-related issues.

7 Established data scope: how much is enough?

Following an investigation and a series of discussions with the project stakeholders, it was decided to set the initial scope to collect data. This scope was set on pragmatic grounds with an awareness of possible extensions in future work.

This section presents the four datasets extracted from another source and constitute the current project dataset, Euro-COVID19. The detailed description for each dataset is provided in a dedicated section below. Below is an outline of these

datasets, along with their subcomponents. Note that the subcomponents are noteworthy at the level of this document to facilitate a deeper understanding of the data, but the dataset distributions merge them into usable wholes. We also provide a codified identifier for each dataset to allow for precise referencing.

- *[ds_pwdb]* Policy Watch Dataset
  - *[ds_pwdb_core]* core: the content from EuroFound website[5]
  - *[ds_pwdb_ext]* extended: the transformed and enriched version of the core
- *[ds_eu_cellar]* EU Cellar COVID19 Dataset
  - *[ds_eu_cellar_core]* core: EurLex[6] documents in Cellar[7] marked with a special "COVID19" tag
  - *[ds_eu_cellar_ext]* extended: Cellar documents about selected EuroVoc[8] concepts (tightly related to COVID19) dated after the 1st of January 2020.
- *[ds_eu_timeline]* EU Action Timeline Dataset
  - *[ds_eu_timeline_core]* core: documents crawled from the EU action timeline website[9]
  - *[ds_eu_timeline_ext]* extended: documents from the European Commission Press corner[10] crawled by searching for selected EuroVoc concepts dated after 1st of January 2020.
- *[ds_ireland_timeline]* Ireland COVID19 Timeline Dataset (optional)
  - *[ds_ireland_timeline]* timeline crawled from the Irish Government press corner by searching for selected EuroVoc concepts in the press releases dated after 1st of January 2020.

The datasets listed above differ in terms of their internal structure and in terms of the content they cover. Each dataset will be presented in the sections below. A summary of the dataset content classification is provided in Table 2.

[5]COVID19 EU Policy Watch
[6]EurLex
[7]The semantic repository of the Publications Office
[8]EuroVoc The EU's Multilingual Thesaurus
[9]Timeline of EU action
[10]Press material from the Commission Spokesperson's Service

@en	EU	MS (collectively)	MS (individual)
Journalistic text	EU actions timeline	Policy Watch DB	Ireland timeline
Legal text	EU Cellar COVID19	-	-

Table 2: Classification of the dataset content

In terms of content coverage, we distinguish datasets that cover: the EU measures, the MS measures for all EU states in a single collection, the MS measures segregated for each country.

In terms of language register[11] two major ones are identified that characterise existent texts: (a) *Journalistic or reportage text register* and (b) *formal administrative and legal text register*. The aim here is not to describe these registers but merely identify them as making such a distinction may be of significant relevance in the data analysis phase.

Datasets pertaining to the journalistic(reportage) text register are

- Policy Watch Database
- EU action timeline
- Ireland action timeline

In the formal administrative and legal text register fall the following datasets

- EU Cellar COVID19

In the next section we proceed with describing the structure for each dataset and some of its peculiarities.

[11]Register (in Sociolinguistics) is a variety of language used for a particular purpose or in a particular communicative situation. This term is remotely synonym to text type, style and genre.

8 Dataset description: what is inside?

This section describes the datasets in terms of their structure, scope and manner in which the data have been collected.

8.1 Policy watch database (ds_pwdb)

Eurofounds COVID-19 EU PolicyWatch (PWDB)[12] collates information on the responses of government and social partners to the crisis, as well as gathering examples of company practices aimed at mitigating the social and economic impacts. Data has been mainly provided by the Network of Eurofound Correspondents, with quality control carried out by Eurofound staff.

PWDB includes large-scale government measures and wider collective agreements, as well as regional and local initiatives and support measures for smaller groups of workers. As the situation is evolving, measures are newly implemented, changed or cancelled and replaced at rapid speed. It is planned to update the cases with information on the actual uptake of the main measures. Outside the scope of this database are public health measures, travel and movement restrictions and company-specific job losses.

The original PWDB contains a rich set of attributes. Only a subset is considered relevant from the business perspective to the current project and is listed in the table below. The original structure is transformed into a simplified form for harmonisation with other datasets and easier usage. The data structure of the transformed core PWDB dataset (ds_pwdb_core) is described in Table 3.

Data attribute	**Description**	**BF**

continued on next page

[12] PWDB download page

continued from previous page

Category	Nine high-level categories for grouping the COVID19 measures (proposed by the EuroFond team[13]).	BF2
Subcategory	Further categorisation into fine grained categories, under a parent category. The two level taxonomy is not documented here but can be recreated from the dataset.	BF2
Target group (L1)	The database provides target groups for each measure. Target groups are organised on two levels: L1 & L2. The L1 level broadly differentiates between *workers, businesses* and *citizens*.	BF8
Target group (L2)	The L2 level contains more fine grained distinctions containing 42 specific target groups distinctions in total, including for example: • Workers: Self-employed, Seasonal workers, Platform workers, etc. • Businesses: SMEs, Start-ups, Larger corporations, etc. • Citizens: Parents, Older citizens, Migrants, etc.	BF8
Country	The country where the COVID19 measure is adopted by the government and social partners.	BF5, BF1

continued on next page

[13] About EuroFound

continued from previous page

Involved actors	Eurofound identified 757 legislations and other statutory regulations, 452 of which have been created entirely in the context of COVID19. This section gives an overview of social partner's involvement in designing and implementing these measures.	∼ (BF1 + BF8)
Funding	The sources of financing the measure, if the measure involves financial expenditures. Some measures do not have fundings.	∼ BF6
Type	Classification of the document types from where the description of the measures originates. Six types of source documents are distinguished, including legislations, collective agreements, recommendations and company practices.	
Start & End dates	The time period when the measure is applied.	BF4
Creation date	When the measure entry was created in the database.	∼ BF3
Update date	When last changes were made to the measure entry description.	∼ BF3
Background information	A short text providing the context and background information useful to understand the measure description.	BF10
Content of measure	A short text representing the abstract or a concise description of the measure.	BF10
Content updates	Short updates to the content of the measure.	BF10

continued on next page

continued from previous page

Use of measure	Information about the results and outcomes of executing/enacting the measure.	BF10
Title	A short text used to identify the measure, place it in context, and convey a minimal summary of its contents.	BF10

Table 3: The attribute structure for core PWDB dataset (ds_pwdb_core)

In the core dataset, a list of source references is provided. They represent links to the original documents elaborating on the contents of the measure. We proceeded with downloading, cleaning and injecting the content of the sources into the PWD dataset, this way extending it. An additional set of data attributes is provided in the extended version of the PWDB dataset, as is listed in Table 4.

Data attribute	**Description**	**BF**
Source URL	The measure descriptions are based on external sources referenced by URLs	
Source content	The content (in simple text) downloaded by accessing the document at the source URL.	BF10
Source title	The title of the source document, if possible to retrieve	BF10
Source language	The language of the source content is determined by a language identification system.	

Table 4: The additional set of attributes constituting the PWBD extension (ds_pwdb_ext)

Finally, the core and the extended dataset variants are merged and provided as a unified PWDB dataset.

8.2 EU Cellar COVID19 dataset (ds_eu_cellar)

The Cellar is the semantic repository of the Publications Office. It stores essential legal documents, general publications and other vital EU level documents. We query this repository to construct the EU level COVID19 datasets containing the document content and the associated metadata.

In the context of the current exercise, we distinguish the core and the extended datasets variants, which are results of querying Cellar with two different SPARQL queries.

The core dataset is the result of querying for documents (called works) that are annotated with a special "COVID19" tag in the Cellar repository. The tagging is performed manually by the EurLex team and its contractors. This tag marks documents that have been identified as dealing directly with issues of the COVID19 pandemic.

The extended dataset is also the result of querying for documents (called works) in Cellar, which are annotated with any of the EuroVoc concepts that have been identified as relevant to COVID19 pandemics. The list of selected EuroVoc concepts is provided in Table 5.

Concept URI	**Concept preferred label**
http://eurovoc.europa.eu/1005	EU financing
http://eurovoc.europa.eu/1439	innovation
http://eurovoc.europa.eu/1633	free movement of persons
http://eurovoc.europa.eu/1754	illness
http://eurovoc.europa.eu/1756	respiratory disease
http://eurovoc.europa.eu/1759	infectious disease
http://eurovoc.europa.eu/1802	labour market
http://eurovoc.europa.eu/1854	disease prevention
http://eurovoc.europa.eu/192	health control

continued on next page

continued from previous page

http://eurovoc.europa.eu/2916	applied research
http://eurovoc.europa.eu/2923	medical research
http://eurovoc.europa.eu/3730	health risk
http://eurovoc.europa.eu/3885	public health
http://eurovoc.europa.eu/4470	tourism
http://eurovoc.europa.eu/4505	air transport
http://eurovoc.europa.eu/5237	research and development
http://eurovoc.europa.eu/835	aid to undertakings
http://eurovoc.europa.eu/1280	occupational health
http://eurovoc.europa.eu/1634	free movement of workers
http://eurovoc.europa.eu/2062	standard of living
http://eurovoc.europa.eu/2479	health policy
http://eurovoc.europa.eu/5891	public awareness campaign
http://eurovoc.europa.eu/82	working conditions
http://eurovoc.europa.eu/2473	communications policy
http://eurovoc.europa.eu/3086	economic consequence
http://eurovoc.europa.eu/4636	vaccination
http://eurovoc.europa.eu/5992	economic activity
http://eurovoc.europa.eu/712	economic support
http://eurovoc.europa.eu/826	aid to disadvantaged groups
http://eurovoc.europa.eu/1596	health legislation
http://eurovoc.europa.eu/2870	quality of life
http://eurovoc.europa.eu/3956	social sciences
http://eurovoc.europa.eu/899	economic aid

continued on next page

continued from previous page

http://eurovoc.europa.eu/7983	European Centre for Disease Prevention and Control
http://eurovoc.europa.eu/83	living conditions
http://eurovoc.europa.eu/85	social situation
http://eurovoc.europa.eu/5764	organisation of health care
http://eurovoc.europa.eu/3552	teleworking
http://eurovoc.europa.eu/1742	job preservation
http://eurovoc.europa.eu/886	state of emergency
http://eurovoc.europa.eu/1926	working environment
http://eurovoc.europa.eu/4116	health service
http://eurovoc.europa.eu/5612	protective equipment
http://eurovoc.europa.eu/837	epidemic
http://eurovoc.europa.eu/2270	social participation
http://eurovoc.europa.eu/838	epidemiology
http://eurovoc.europa.eu/2793	aid programme
http://eurovoc.europa.eu/3588	restriction of liberty
http://eurovoc.europa.eu/6781	basic needs
http://eurovoc.europa.eu/3371	public hygiene
http://eurovoc.europa.eu/2013	mass media
http://eurovoc.europa.eu/7131	social impact
http://eurovoc.europa.eu/3906	freedom of movement
http://eurovoc.europa.eu/3370	patient rights
http://eurovoc.europa.eu/4881	social well-being
http://eurovoc.europa.eu/86	socioeconomic conditions

continued on next page

continued from previous page

http://eurovoc.europa.eu/1758	endemic disease
http://eurovoc.europa.eu/779	distance learning
http://eurovoc.europa.eu/6609	self-regulation
http://eurovoc.europa.eu/6770	disinformation
http://eurovoc.europa.eu/c_324b44f1	social media
http://eurovoc.europa.eu/c_5b447e3a	crisis management
http://eurovoc.europa.eu/c_31da5694	e-Health
http://eurovoc.europa.eu/c_60d3928d	patient safety
http://eurovoc.europa.eu/c_9b88f778	hospital infection
http://eurovoc.europa.eu/c_ece0a719	viral disease
http://eurovoc.europa.eu/c_814bb9e4	coronavirus disease
http://eurovoc.europa.eu/c_abfaf2ea	disease surveillance

Table 5: EuroVoc concepts considered highly relevant for COVID19 document search

The result of querying Cellar in both cases (core and extended datasets) contains the same data attributes. The difference is in the retrieved documents. The structure of the eu_cellar dataset is provided in Table 6.

Data attribute	**Description**	**BF**
Work URI	The URI, which is uniquely identifying the work.	
CDM type	The work type according to the Common Data Model ontology.	

continued on next page

continued from previous page

Resource type	The work type according to the OP classification used in the European inter-institutional exchange of legal documents. The resource types are organised in a classification scheme called Resource Type[14].	
EuroVoc concept	The EuroVoc concept[15] used as a topic and classifier of the document content. The EuroVoc thesaurus is developed by the OP and used in the inter-institutional context. The concepts are organised as taxonomies from the broad to more narrow concepts.	BF2
Subject matter	The subject matter concept used as topic and classifier of the document content. The Subject Matter controlled vocabulary[16] is developed by the OP and used in the inter-institutional context. The concepts are organised in a classification scheme called subject-matter, or FD_070[17].	BF2
Directory code	The directory code concept used to organise the legal documents in the EurLEx website. The concepts are organised in a classification scheme called FD_555[18].	$\sim$ BF2
Author	The authors of the legal document. The identifier of the authors provided here is from the controlled list called Corporate Body[19].	BF1

continued on next page

[14] Resource Type authority table
[15] The EuroVoc thesaurus is represented as Concepts organised in ConceptSchemes following the Simple Knowledge Organisation System (SKOS).
[16] Subject Matter authority table
[17] FD 070 ATTO table
[18] FD 555 ATTO table
[19] Corporate body authority table

continued from previous page

Date document	The date document was issued, entered into force, signed or other date relevant for the document legality.	BF3
Content	The actual content of the legal document, reduced to simple unstructured text.	BF10
Title	The document tile.	BF10

Table 6: The attribute structure for eu_cellar dataset

Finally, the core and the extended dataset variants are merged and provided as a unified EU Cellar COVID19 dataset.

8.3 EU action timeline dataset (ds_eu_timeline)

The European Commission (EC) is coordinating a common European response to the coronavirus outbreak. EC is taking resolute action to reinforce our public health sectors and mitigate the socio-economic impact in the European Union. EC is mobilising means to help the Member States coordinate their national responses and provide objective information about the spread of the virus and practical efforts to contain it.

The timeline of EU actions[20] during the COVID19 pandemics is a website published by the EC with the most important COVID19 responses. From 1104 actions (counted on 19/04/2021) published in the press corner only 171 were mentioned on the EU action timeline. So the timeline constitutes a refinement of the noteworthy and an aggregation of the EU actions.

We crawl this website, and its content is automatically organised using the attributes listed in Table 7. Unfortunately, only a very limited set of metadata can be identified in the crawled content: title, abstract, date of publication, and the actual content. And we know that the more metadata about a document, the better it is to answer some of the questions mentioned in the project goals section.

[20]EU actions timeline website

Following a discussion with the representatives of Directorate-General for Communication (DG COMM)[21], we identified a way to recover the possible topics using the name(s) of the authors for each article. This is possible because each author (spokesperson or press officer) is responsible for one or a few topics. These topics are assigned to each press contact who is on the spokespersons service page[22].

What we did was to first extract and structure the information about each person and the topics he/she covers. Then we extended the original crawler to take into consideration this mapping between the person name and the topics. A new data attribute is created for each article containing the array of possible topics that characterise the article.

Note that the list shall not be read as conjunction, that is: the article is about all of the topics provided; but as a disjunction, which means: the article is about one of the provided topics.

Data attribute	**Description**	**BF**
Abstract	A short summary of the article, an abstract.	$\sim$ BF3
Content	An extended description of the action as a press release article.	BF10
Date	The date when the press release was published	BF10
Title	The title of the press release article.	$\sim$ BF10
Topic	A list of possible topics that the article may be about, derived from the author's thematic responsibility.	BF2

Table 7: The attribute structure for ds_eu_timeline

[21]The Directorate-General for Communication is the Commission department responsible for explaining EU policies to outside audiences.

[22]Press contacts Spokesperson's Service - list of possible article authors each covering one or few topics.

8.4 Ireland COVID19 timeline dataset

Ireland was selected as a tryout member state country for which a COVID19 timeline shall be created similar to the EU action timeline. It was selected because it is the only member state country (UK having just left the EU) that publishes official documents and press releases in English.

An investigation was conducted searching for a comparable timeline summary of Ireland actions on COVID19, and none was found. However, the official government website press corner, www.gov.ie, was identified as a good source of information.

"gov.ie" website is a central portal for government services and information. It combines the websites of Irish government departments and is a trusted source that makes interactions with the government more user-focused.

We decided to use the search service of this website and search for the same set of EuroVoc concepts that were used to retrieve COVID19 relevant documents from Cellar. The search results are crawled and structured in a dataset using the set of data attributes listed in Table 8.

The preferred label of each EuroVoc concept is used as a search term. For each EuroVoc concept, a new search is launched, and only articles that are more recent than the 1st of January 2020 are considered.

Data attribute	**Description**	**BF**
Title	The title of the press release article.	BF10
Content	The press release article in simple, clean unstructured text.	BF10
Published date	The date when the press release was published	BF3
Update date	The date when the article was updated.	$\sim$ BF3
Content links	A list of links available in the text.	
Campaigns links	A list of links to the organised Campaigns	

continued on next page

continued from previous page

Department	The government department that was authoring the article.	∼ BF1
Policies links	The list of links to the broad policy category under which the article is placed.	∼ BF2
Keywords	A list of keywords assigned by the article authors.	∼ BF2
Page type	The type of article is similar to the classification from the Resource Type authority table used for the ds_eu_cellar dataset. The possible types are the following: press release, speech, news, policy information, reports, etc.	

Table 8: The attribute structure for ds_ireland_timeline

9 Architectural overview

This section presents the technology stack we employ and the design of the workflows for creating the datasets. An overview is depicted in the Figure 3 diagram.

The diagram in Figure 3 is split into three layers: *Platform as a Service (PaaS) layer, Software as a Service (SaaS) layer, and Application layer.*

At the bottom of the diagram is depicted the infrastructure layer. We have decided to operate based on a platform infrastructure because it abstracts away from the traditional physical servers and allows a deployment virtually in any environment: physical server, virtual machine, cloud. The chosen PaaS technology is Docker[23] for its popularity and relative simplicity over alternatives such as Kubernetes[24].

In the middle of the diagram are depicted five larger blue round-cornered rectangles five. They represent classes of services were employing in this project. These

[23]Docker is a set of platform as a service (PaaS) products that use OS-level virtualization to deliver software in packages called containers.

[24]Kubernetes is an open-source system for automating deployment, scaling, and management of containerized applications.

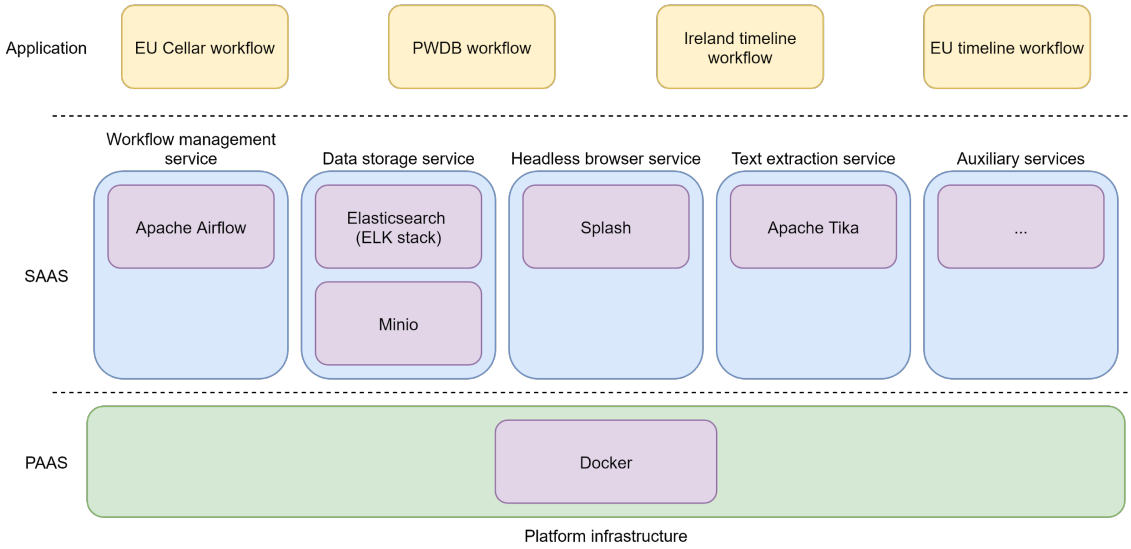

Figure 3: Technology stack overview

services are: *workflow management system, data storage service, headless browser service, text processing service,* and auxiliary services. We will explain below how each is used after we address the application layer workflow structure.

In the top part of the diagram is depicted the application layer which contains four workflows, one for each dataset: *EU Cellar workflow*, *EU timeline workflow*, *PWDB workflow*, and *Ireland timeline workflow*. These workflows have a prototypical extract, transform, toad (ETL) structure[25]. The generic ETL process implemented in our workflows is addressed in the next section.

10 Workflow structure: how it works?

This section addresses the general structure of dataset creation workflow. Because it is a simple sequence of tasks, without any major bifurcations we call them pipelines

[25]In computing, extract, transform, load (ETL) is the general procedure of copying data from one or more sources into a destination system which represents the data differently from the source(s) or in a different context than the source(s). The ETL process became a popular concept in the 1970s and is often used in data warehousing.

as well. The prototypical pipeline we implement is depicted in Figure 4 and can be conceptualised as a sequence of four steps:

- data *extraction* from the source and storage in the temporary object storage
- data *structure transformation* in the temporary storage
- data *content transformation* in the temporary storage
- final data *loading* into the document repository

The sources of data we employ in this project are:

- Cellar SPARQL endpoint[26] where the EU legal documents are stored and disseminated.
- EuroFound website[27] where the PWDB is published.
- EU action timeline page
- Ireland government press corner page

10.1 Extraction

The process starts with *extracting* the necessary data from the source and storing it in temporary object storage at our premises, making it available for further processing. Doing so enables us to process arbitrarily large amounts of data compared to the alternative of keeping the extracted data in memory, which is relatively limited on the traditional systems.

Besides simple HTTP requests[28], the extraction operations important to mention are the *SPARQL querying* and *Website crawling*.

SPARQL[29] is an RDF query language, a semantic query language for databases able to retrieve and manipulate data stored in [30] format. We query Cellar, the European semantic repository, for legal documents annotated with COVID19 tag or related EuroVoc concepts.

[26]Cellar SPARQL endpoint

[27]EuroFound PWDB download page

[28]HTTP is an application layer protocol for distributed, collaborative, hypermedia information systems. HTTP is the foundation of data communication for the World Wide Web.

[29]SPARQL Protocol and RDF Query Language

[30]Resource Description Framework (RDF)

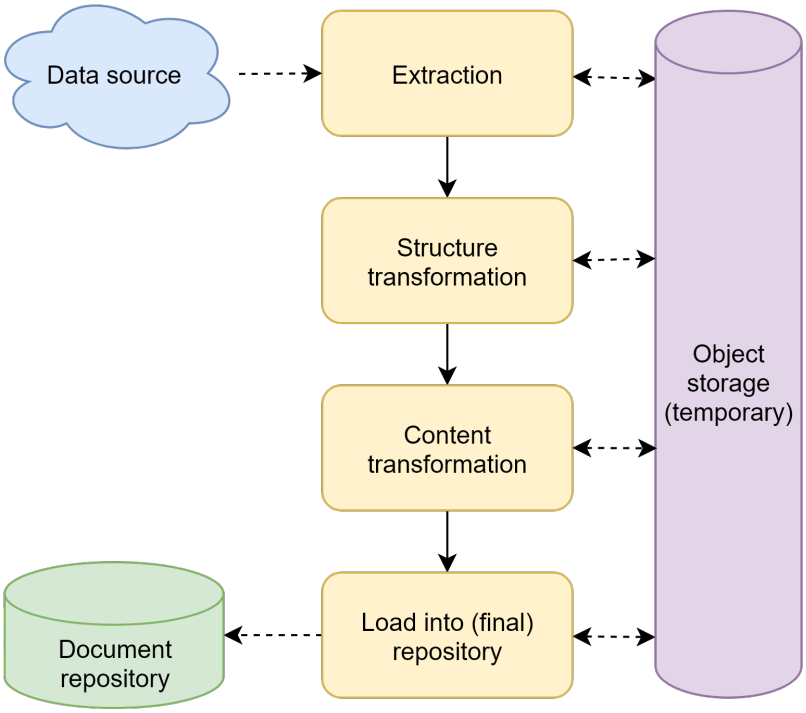

Figure 4: Generic ETL process for dataset creation

A *Web crawler*[31] sometimes called a *spider* and often shortened to *crawler*, is an Internet bot that systematically browses the World Wide Web, typically operated by search engines for the purpose of Web indexing. We developed two crawlers, using Scrapy library[32], first to fetch information from the EU action timeline website, and the second one from the Ireland government press corner. Scrapy is a free and open-source web-crawling framework written in Python. Originally designed for web scraping, it can also be used to extract data using APIs or as a general-purpose web crawler.

In the scraping process, often simple HTTP requests are not sufficient for performing full content extraction. Execution of custom JavaScript code is necessary to complete the page loading. For this purpose, headless browser services are typically

[31] Web crawler
[32] Scrapy, a fast high-level web crawling & scraping framework for Python.

implemented. They simulate the behaviour of a web browser but without having the interface of one. We employ Splash[33] to act as a headless browser. Splash is a javascript rendering service with an HTTP API. It's a lightweight browser with an HTTP API, implemented in Python 3. In combination with Scrapy, Splash allows us to crawl the two web sources successfully.

One other aspect essential to mention here is that we employ an object storage service called Minio[34] for the temporary persistence of the extracted data. MinIO is an Amazon S3 compatible server-side software storage stack. It can handle unstructured data such as photos, videos, log files, backups, and container images with the maximum supported object size of 5TB.

One may argue that such an object storage system is unnecessary as the created datasets, so far, are relatively small in size and can easily fit into the memory of most systems. However, this infrastructure, we intend to extend and use for the processing of much larger datasets exceeding the memory limits. That situation will inevitably invite the usage of such a persistence system, which we foresee and introduce upfront.

10.2 Structure transformation

Next, we proceed with *structural transformation* to *normalise* the data representation, simplify it, and increase its usability yet maintaining the maximally helpful structure. For convenience, we aim to represent the data in JSON format following the following system: *an array of objects with a unique identifier and an arbitrary number of atomically typed attributes.* The structure is depicted using an example of prototypical JSON objects in Figure 5.

The attributes may be of any atomic type, such as numbers, strings, dates, etc. or arrays of atomic types (as depicted on the left side of Figure 5), but not objects or arrays of objects (as depicted on the right side of Figure 5). We discourage, with some exceptions, using the embedded object structures. The reason for it is that the

[33]Splash javascript rendering service
[34]MinIO is an open source implementation of the Amazon S3 storage system.

embedded objects are no longer easily accessible in the data frame structures (using the numpy[35] or Pandas[36] libraries). The data frames or other tabular representations are the de facto representation for machine learning and data science exercises, which we plan to undertake in the current project.

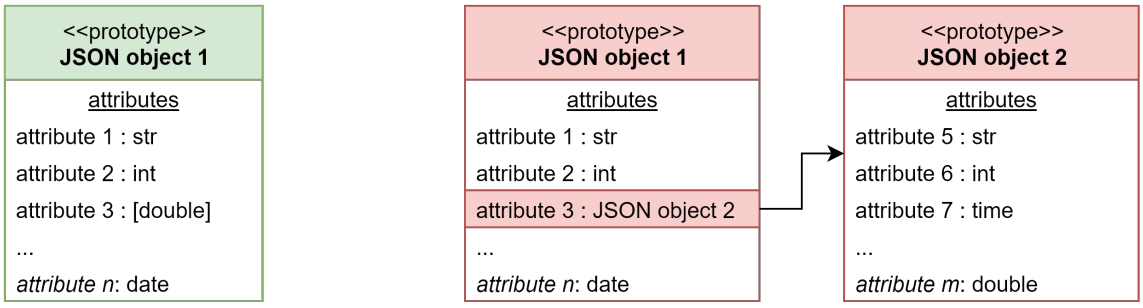

Figure 5: Example of prototypical JSON objects, on the left valid and on the right an invalid structure

For consistency and programming language neutrality, we chose to employ JQ JSON processor[37]. The transformation rules are written in JQ language and tailor-made for each data source. As our implementation is entirely written in Python, we use the python library to execute these JQ transformation rules.

10.3 Content transformation

fter the data structure is *normalised,* the content transformation generally consists of *text extraction*, *enrichment*, *restructuring, aggregation,* and other operations.

One peculiarity of this stage is that the initially extracted data contains links and references to other externally available resources. We have carefully analysed and decided to fetch the resources behind selected sets of links. Each dataset includes

[35]Numpy is a library for the Python programming language, adding support for large, multi-dimensional arrays and matrices, along with a large collection of high-level mathematical functions to operate on these arrays.

[36]Pandas is a software library written for the Python programming language for data manipulation and analysis.

[37]./jq is a lightweight and flexible command-line JSON processor

attributes with such links, which we access and inject the content as an additional attribute. This is what we call content enrichment: injecting extracted content into the dataset.

The fetched content is in the majority of cases of unpredictable serialisation format, structure and language. Therefore, before it is injected into the dataset, we first *reduce* it to simple texts, which is the foremost valuable data representation for the NLP tasks.

For the *text extraction task,* we decided to use the Apache Tika[38] toolkit. Apache Tika is a library that is used for document type detection and content extraction from various file formats. Internally, Tika uses existing different document parsers and document type detection techniques to detect and extract data. Tika is widely used while developing search engines to index the text contents of digital documents.

Some categorical data attributes are provided as items from a flat fine-grained classification. For practical reasons, it is helpful to reduce the classification scheme. This *restructuring is possible to undergo* by *aggregating* fine-grained categories in terms of more coarse-grained ones. One such example is the target group attribute from PWDB. There are 42 distinct values, which can be roughly categorised as belonging to three more prominent categories. We proceed to inject such aggregations as additional attributes to leverage exploratory data analysis in future stages of the project.

10.4 Loading into the repository

After the structure and content of the data have been restructured, the data is loaded into a repository where it is indexed and made available for querying and full-text search. Because the current datasets are designed for NLP tasks, the full-text search capability is critical. It plays an essential role in the exploratory data analysis and possible data segmentation or partitioning for machine learning experiments.

[38]Apache Tika is a toolkit that detects and extracts metadata and text from over a thousand different file types (such as HTML, PPT, XLS, and PDF).

For this purpose, we decided to use the Elasticsearch search engine[39] to act as the document repository. It is a real-time distributed and analytic engine that helps in performing various kinds of search mechanisms. It can achieve fast search responses because, instead of searching the text directly, it searches an index instead. Additionally, it supports full-text search, which is completely based on documents instead of tables or schemas that are easier to write queries and manipulate with this textual data. Some of the strongest points of elastic search are:

- Performing and combining various kinds of searches irrespective of their data type.
- Querying can retrieve data in any form required.
- Analyzing billions of records in a few seconds.
- Aggregating data enables us to explore trends and patterns.

When the datasets are loaded into the Elasticsearch, they are easily explorable using Kibana[40] discovery and dashboard functionality. Kibana offers histograms, line graphs, pie charts, sunbursts, geospatial map displays, and other standard visualisation options and the opportunity to create unique visualisations. It also makes it possible for users to spot and analyse relationships in Elasticsearch data and detect and explore anomalies in Elasticsearch data with machine learning features. Furthermore, the datasets are available for use in machine learning experiments and exploratory data analysis.

11 Workflow management system

The workflows mentioned above are deployed and executed in a workflow management system. Doing so drives automation and leads to increased control, transparency and trust in the execution results. It allows for close monitoring of each

[39]Elasticsearch is a search engine based on the Lucene library. It provides a distributed, multitenant-capable full-text search engine with an HTTP web interface and schema-free JSON documents.

[40]Kibana is a data visualization dashboard software for Elasticsearch. It provides visualization capabilities on top of the content indexed on an Elasticsearch cluster.

step, schedule executions, increased connectivity, eliminates manual tasks, reduces errors, retries on failure, and investigates causes and visualises the workflow, control panel, and other benefits.

We have chosen Apache Airflow system[41] due to its architectural choices, maturity, rich set of features, strong community, rich set of integrations and plugins, and because a system implemented in Python fits well our Python predominant technical stack.

Airflow is a platform to programmatically author, schedule and monitor workflows. It is written in Python, and workflows are created via Python scripts, called DAGs (Directed Acyclic Graphs). An example DAG structure is depicted in Figure 6.

Airflow is designed under the principle of "*configuration as code*" . While other "*configuration as code*" workflow platforms exist using markup languages like XML, using Python allows developers to import libraries and classes to help them create their workflows.

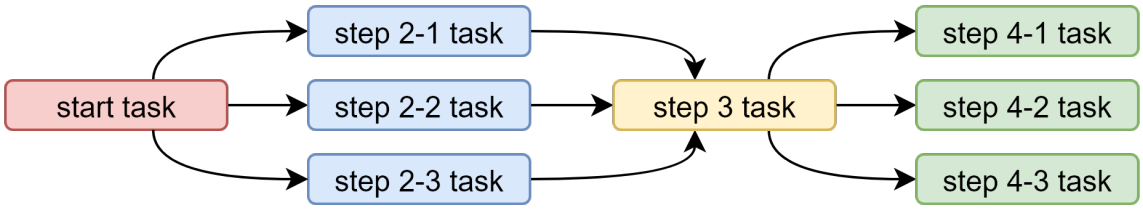

Figure 6: Example of a DAG structure

Airflow uses directed acyclic graphs (DAGs[42]) to manage workflow orchestration. Tasks and dependencies are defined in Python and then Airflow manages the scheduling and execution. DAGs can be run either on a defined schedule (e.g. hourly or daily) or based on external event triggers.

[41] Apache Airflow is an open-source workflow management platform. It started at Airbnb in October 2014 as a solution to manage the company's increasingly complex workflows.

[42] A DAG is a collection of all the tasks you want to run, organized in a way that reflects their relationships and dependencies.

DAG, or directed acyclic graphs, are a collection of all of the tasks, units of work in the pipeline. The tasks are organised by their relationships and dependencies between each other. A directed acyclic graph implies that your pipeline can only move forwards, not backwards. A task can retry, but a task cant be rerun after it has completed and another task downstream has begun.

Using Airflow enables us to organise our pipelines as DAGs, develop project-specific functionality and incorporate it seamlessly into the workflow architecture to deploy and execute in a robust environment easily.

12 Limitations and future work

In the process of developing the current dataset, a number of limitations were set to the scope of this project. This section covers a non-exhaustive list of these limitations and how they may be addressed in future work.

Currently, the dataset is limited to the English language. In future work, the dataset can be expanded to cover member state measures individually. Doing so will inevitably lead to creating a multilingual dataset containing texts in 24 official European languages. This is mainly because the authorities publish the measures in that countrys official language(s).

Using different text registers implies a disparate treatment of texts in the analysis and processing pipelines. Developing a mechanism to summarise a document to extract measure description from it, as, for example, would be necessary for large legal texts, would be of tremendous benefit for harmonisation of the dataset content across various sources. Addressing this in future work will yield an advantage for the exploratory data analysis and other machine learning exercises.

Another axis of homogenisation would be to bring all the texts to a common language, such as English, for example. To do so would necessitate translation services, which at the moment were not used. Having done so will decrease the linguistic fragmentation of the dataset, leading to a larger homogeneous corpus, which increases the statistical significance of the results produced based on this dataset.