

# An LLM-assisted ETL pipeline.pdf

English -> Polish

TransPDF

Date: **Thu Jun 19 12:52:20 2025**  
Owner: **miki29042002@wp.pl**  
Translator: **miki29042002@wp.pl**  
Complete: **100%** (0% XLIFF, 0% Pseudo, 100% MT)  
Pages: **25**

## Font Problems

**OpenSans-Regular -> Alegreya Sans Black:**

**IBMPlexSans-Bold -> Alegreya Sans Black:**

**IBMPlexSans-Italic -> Alegreya Sans Black:**

**IBMPlexSans -> Alegreya Sans Black:**

**Montserrat-Regular -> Noto Sans Regular:**

**FreeSerif -> Noto Serif Regular:**

Missing Characters: [1B82]

This list details all problems with and substitutions of fonts embedded in the document.

These problems are typically caused by characters not embedded in the original document but which are now required by the new translated text.

Throughout the PDF, any missing characters are shown in **red** using a default substitute font.

You can resolve these issues at [transpdf.iceni.com](https://transpdf.iceni.com) by visiting the flight-checking summary for this document. You will then be able to substitute more appropriate fonts containing all the character shapes required for this translation.

Once you change the substitutions, re-generate the preview to see if all problems are resolved.

**Produced by [transpdf.iceni.com](https://transpdf.iceni.com)**

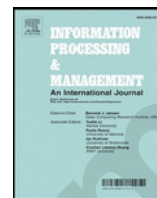
- the web-based service for translating PDF documents.

Make the task of PDF translation faster and more accurate while using your own CAT tools such as Trados and MemoQ .

Register for your free usage allowance by visiting: [transpdf.iceni.com](https://transpdf.iceni.com)



## Przetwarzanie i zarządzanie informacją

Strona domowa czasopisma: [www.elsevier.com/locate/ipm](http://www.elsevier.com/locate/ipm)

# Potok ETL wspomagany przez LLM w celu zbudowania wysokiej jakości grafu wiedzy na temat prawodawstwa

,Anna Bernasconi, Stefano Ceri

Dipartimento di Elettronica, Informazione e Bioingegneria, Politecnico di Milano, Via Giuseppe Ponzio, 34, Mediolan, 20133, Włochy

## Art.

## INFO

## ABSTRAKT

## Słowa kluczowe:

Wpływ wiedzy  
Wykresy właściwości  
Modele wielkojęzykowe  
Jakość danych

Rosnąca złożoność systemów prawnych, charakteryzująca się stale rosnącą liczbą aktów prawnych i ich współzależnościami, uwypukliła użyteczność grafów wiedzy (KG) jako skutecznego modelu danych do organizowania takich informacji, w porównaniu z tradycyjnymi metodami, często opartymi na modelach relacyjnych, które mają trudności z efektywnym reprezentowaniem wzajemnie powiązanych danych, takich jak odniesienia w przepisach, utrudniając skuteczne odkrywanie wiedzy.

Zmiana paradygmatu w modelowaniu danych legislacyjnych jest już w toku wraz z przyjęciem wspólnych międzynarodowych standardów, głównie opartych na XML, takich jak Akoma Ntoso (AKN) i Legal Knowledge Interchange Format, które mają na celu uchwycenie podstawowych aspektów prawa wspólnych dla różnych aktów prawnych i uproszczenie zadania tworzenia grafów wiedzy za pomocą znaczników i identyfikatorów XML. Jednak, aby umożliwić zaawansowaną analizę i odkrywanie danych w tych węzłach KG, konieczne jest dokładne sprawdzenie, uzupełnienie i wzbogacenie węzłów KG o właściwości, metadane lub dodatkową wiedzę pochodną, które poprawiają jakość i użyteczność modelu, na przykład poprzez wykorzystanie możliwości najnowocześniejszych dużych modeli językowych.

W niniejszym artykule przedstawiamy proces modelowania i badania włoskiego prawodawstwa w Grafie Wiedzy, przyjmując model grafu właściwości i standard AKN zaimplementowany we włoskim systemie. Model grafu właściwości oferuje dobry kompromis między reprezentacją wiedzy a możliwością wykonywania analizy grafów, co uważamy za niezbędne do umożliwienia zaawansowanego wykrywania wzorców. Następnie wzbogacamy KG o cenne właściwości, stosując starannie dopracowane LLM typu open source, tj. modele i Mistral-7B, które wzbogacają i zwiększają jakość KG, umożliwiając dogłębną analizę danych legislacyjnych.

## 1. Wprowadzenie

Przyjęcie powstających baz danych i technologii reprezentacji wiedzy, takich jak wykresy wiedzy, wzrosło ostatnio uwagę wielu społeczności poszukujących dostępnych i skutecznych podejść do przedstawiania złożonej wiedzy. Wśród nich, społeczność prawa komputerowego była bardzo aktywna w proponowaniu rozwiązań KnowledgeGraph dla reprezentowania złożonych dziedzin, takie jak ustawodawcze, z prawami powiązanymi ze sobą za pomocą cytatów (Anelli et al., 2023; Angelidis, Chalkidis, Nikolaou, Soursos, 2018; Rodríguez-Doncel, Navas-Loro, Montiel-Ponsoda, & Casanovas, 2018).

Jednym z głównych wyzwań związanych z danymi legislacyjnymi jest tekstowy charakter aktów prawnych, który w związku z tym zawiera nieustrukturyzowane informacje. Podczas gdy aplikacje LLM do bezpośredniej konstrukcji text-to-KG oferują obiecujące rozwiązanie tego problemu, ich dane wyjściowe są zbyt niedokładne, aby można było uzyskać wysokiej jakości reprezentację danych tekstowych, która zapewni poprawność schematu (Dong,

c

\*

<https://doi.org/10.1016/j.ipm.2025.104082>

Otrzymano 1 września 2024 r.; Otrzymano w dniu 27 grudnia 2024 r.; Zaakceptowane 27 stycznia 2025 Dostępne online 10 lutego 2025  
0306-4573/© 2025 Autorzy.

Published by Elsevier Ltd. Ten artykuł jest dostępny w otwartym dostępie pod adresem

(<http://creativecommons.org/licenses/by-nc-nd/4.0/>).

Licencja CC BY-NC-ND

w 2023 r.; [Mihindukulasooriya, Tiwari, Enguix, & Lata, 2023](#)). Aby temu zaradzić, społeczność prawa komputerowego włożyła wiele wysiłku w zaproponowanie odpowiednich standardów międzynarodowych, które reprezentowałyby, w ramach tego samego schematu, prawa uchwalone w różnych systemach prawnych. Większość wcześniejszych prac wykorzystywała format eXtensible Markup Language (XML), częściowo ustrukturyzowany model danych, który był naturalnie używany do modelowania w społeczności prawa komputerowego do reprezentowania danych tekstowych, takich jak prawa ([Lupo i in., 2007](#)).

Tagi XML można łatwo zmapować do wiarygodnych grafów wiedzy, które przechowują prawa i ich artykuły jako węzły, połączone krawędziami cytatów ([Sana & Suganthi, 2017](#)). Odpowiednie propozycje oparte na XML obejmują Legal Knowledge Interchange Format (LKIF) ([Hoekstra, Breuker, Marcello, & Boer, 2007](#)), LegalRuleML ([Athani i in., 2013](#)) oraz Akoma Ntoso ([Barabucci, Cervone, Palmirani, Peroni, & Vitali, 2009](#)). Ten ostatni został niedawno oficjalnie przyjęty przez wiele organów międzynarodowych i krajowych jako wspólny standard ([Vitali, Palmirani i in., 2019](#)); wśród nich został przyjęty przez włoskiego ustawodawcę ([Palmirani, 2021](#)), co umożliwiło realizację solidnych rurociągów w oparciu o ten standard, w tym uzyskanie wiarygodnego wykresu wiedzy włoskiego ustawodawstwa.

Podczas gdy dostępność standardu XML upraszcza zadanie tworzenia Grafu Wiedzy dla danych legislacyjnych, jego praktyczna użyteczność zależy od (i) wyboru, który model grafu i schemat zostanie przyjęty oraz (ii) jego bogactwa pod względem węzłów, krawędzi, a zwłaszcza właściwości, które pozwalają użytkownikom na przeprowadzanie zaawansowanych analiz nad KG. Jeśli chodzi o tę pierwszą kwestię, na podstawie naszej pracy z ekspertami politycznymi i badaczami z Istituto Einaudi per l'Economia e la Finanza ([EIEF, 2024](#)), zauważyliśmy potrzebę obliczalnej i elastycznej reprezentacji wiedzy legislacyjnej. W przypadku tych ostatnich uważamy, że LLM mogą odegrać kluczową rolę we wspomaganie potoku ETL poprzez uzupełnianie i wzbogacanie KG o dodatkowe obiekty grafów, zwłaszcza gdy są specjalnie dostrojone do zadań ekstrakcji informacji.

W tym celu skierowaliśmy naszą uwagę na postępy w grafowych bazach danych, a zwłaszcza na standaryzację Graph Query Language (GQL) ([Deutsch et al., 2022](#)) - języka zapytań Property Graph; Na tej podstawie proponujemy pierwszy schemat Property Graph do modelowania danych legislacyjnych. W naszym schemacie wykorzystujemy elastyczność i zalety GQL, aby wyrazić je w zwartej i intuicyjnej formie [Hogan i inni. \(2021\)](#) złożoność danych legislacyjnych, np. poprzez wykorzystanie zaawansowanych struktur danych jako właściwości węzłów i krawędzi, które pozwalają nam płynnie uchwycić wymiar czasowy (naturalną ewolucję praw w sposób naturalny w czasie), jedną z najbardziej krytycznych cech tej dziedziny.

Wdrożyliśmy schemat i opracowaliśmy Graf Wiedzy o włoskim ustawodawstwie, przechowywany w Neo4j (najpopularniejsza baza danych wykresów nieruchomości [Guia, Soares, & Bernardino, 2017](#); [Solidne doradztwo IT, 2024](#)). Aby to osiągnąć, stworzyliśmy kompleksowy potok ETL, który począwszy od praw opublikowanych w formacie XML Akoma Ntoso poprzez Normattiva ([Istituto Poligrafico e Zecca dello Stato, 2024](#)) – oficjalny punkt końcowy włoskiego ustawodawstwa – stosuje zestaw przekształceń, które mapują znaczniki XML na obiekty grafu, uzyskując spójną, wzajemnie powiązaną reprezentację domeny. Następnie zintegrowaliśmy dodatkowe informacje, do których można łatwo uzyskać dostęp z oficjalnych punktów końcowych prawodawstwa, a także zastosowaliśmy kroki wykrywania i korygowania błędów, wykorzystując strukturę wykresu. Na koniec zastosowaliśmy LLM w celu ulepszenia grafu poprzez uzupełnienie brakujących informacji lub wyprowadzenie dodatkowych właściwości, zgodnie z linią podejść do grafu wiedzy rozszerzonego przez LLM, które mają na celu wykorzystanie możliwości LLM do zadania uzupełniania i budowy grafów ([Pan i in., 2024](#)). W tym celu przyjęliśmy połączoną strategię kilku strażów i dostrajania, aby poprawić jakość wyników LLM i umożliwić stosowanie lepszych modeli, co korzystnie wpłynie na przyszły zrównoważony rozwój rurociągu. W szczególności użyliśmy modelu opartego na do klasyfikowania praw zgodnie z ich domeną oraz dwóch modeli Mistral-7B do uzupełnienia brakujących tytułów i przypisania tematów do ustaw, artykułów i załączników. Uczenie się w kilku ujęciach pozwala nam zwiększyć wydajność LLM w ekstrakcji ustrukturyzowanych informacji w postaci relacji lub właściwości, jak omówiono w [Wadhwa, Amir i Wallace \(2023\)](#) oraz [Xu, Zhu, Wang i Zhang \(2023\)](#), podczas gdy, przyjmując strategię dostrajania, możemy opracować wyspecjalizowane, ale lekkie LLM, które mogą skutecznie poradzić sobie z zadaniami ekstrakcji.

Nasz pipeline ETL gwarantuje regularną aktualizację wykresu nieruchomości we włoskim ustawodawstwie krajowym, poprzez codzienne uruchamianie dedykowanego zadania, które przetwarza nowo opublikowane przepisy w Dzienniku Urzędowym i integruje je z ekosystemem wykresów. Aby zademonstrować przydatność i potencjał naszego KG, badamy jego główne cechy za pomocą zapytań grafowych, które wykorzystują jego bogactwo i zaawansowane struktury danych, które są odblokowywane przez model grafu właściwości i funkcjonalne właściwości wywodzące się z LLM. W tym celu obliczamy wskaźniki i statystyki, które charakteryzują włoskie ustawodawstwo, w tym wzorce działań legislacyjnych i trendy w stanowieniu prawa, a także odkrywamy wzorce specyficzne dla rządu, wykorzystując domeny i tematy (np. w celu scharakteryzowania obszarów interwencji rządowej). Inspiracją dla nas są typowe, ręcznie obliczane wskaźniki wykorzystywane przez niezależne organy nadzorcze ds. prawodawstwa do celów sprawozdawczości rocznej ([Osservatorio sulla legislazione della Camera dei Deputati, 2023](#)). Następnie omawiamy jakość wyników końcowych, analizując KG w wielu wymiarach ([Wang i wsp., 2021](#); [Xue & powiedział: Zou, 2023](#)); w szczególności analizujemy dokładność KG – poprzez doraźne porównanie ze zaktualizowanymi prawami tekstowymi – spójność – podkreślając, w jaki sposób radzimy sobie ze sprzecznościami w danych – kompletność – ilustrując ilościowe ulepszenia na każdym etapie naszego procesu produkcyjnego – terminowość – analizując wydajność aktualizacji – i wreszcie wiarygodność (źródeł danych) i interoperacyjność – pod względem możliwości ponownego zastosowania z tego samego rurociągu do innych systemów prawnych.

Wkład tego artykułu można podsumować w następujący sposób:

- Proponujemy pierwszy schemat Property Graph (PG) do modelowania danych legislacyjnych, zdolny do uchwycenia głównych złożoności domeny poprzez wykorzystanie struktur danych zgodnych z GQL w praktycznym i zwartym schemacie.

- Wdrażamy kompleksowy pipeline ETL w celu stworzenia Grafu Wiedzy o włoskim prawodawstwie w oparciu o solidną technikę mapowania XML do grafu. W tym celu wykorzystujemy międzynarodowy standard Akoma Ntoso wdrożony przez włoskiego ustawodawcę i integrujemy KG z dodatkowymi danymi legislacyjnymi. Następnie używamy paradygmatu Graph Query Language do wykrywania błędów i identyfikowania niespójności we wzorcach wykresów.

- Tworzymy minimalny zestaw (lekkich) dopracowanych LLM, które pozwalają nam uzupełniać i wzbogacać KG poprzez używanie, jeśli to możliwe, samego wykresu do szkolenia i przestrzeganie zasad przewodnich oprogramowania Zrównoważony rozwój i Open Source

([Kukreja, Kumar, Purohit, Dasgupta, & Guha, 2024](#); [Raiaan i in., 2024](#); [Wu i wsp., 2022](#)).

\* Badamy i omawiamy jakość wynikowego grafu wiedzy, analizując jego główne cechy i oceniając go w wielu wymiarach.

## 2. Praca pokrewna

Wykorzystanie grafowych baz danych i technologii Grafów Wiedzy w społeczności prawa komputerowego przyciągnęło znaczną uwagę w ostatnich latach. W szczególności społeczność sieci semantycznej poczyniła ważne postępy, rozwijając ontologie i Paradygmat RDF (Resource Description Framework) do przedstawiania informacji prawnych na wykresach wiedzy, oparty na potrójnym RDF paradygmat (Anelli *in.*, 2023). Prace te mają zasadnicze znaczenie dla łączenia baz wiedzy poprzez oferowanie unikalnych identyfikatorów w wielu Domen. Są one jednak ograniczone do korzystania z modelu danych grafu z etykietami krawędziowymi, którego specyficznym typem są wykresy RDF (Kąty *in.*, 2017). RDF operuje na trójkach – składających się z podmiotu, orzeczenia i dopełnienia – które służą jako zdania opisujące relacje między podmiotem a przedmiotem. Te wykresy RDF mogą być odpytane za pomocą SPARQL (Pérez, Arenas, & Cutierrez, 2009), semantycznego język zapytań. Jednak niedawne pojawienie się międzynarodowego standardowego języka zapytań dla wykresów właściwości (ISO, 2024) rzuca szansę na innowacje w systemach legislacyjnych.

Dogłębna dyskusja na temat zalet i wad przyjęcia RDF lub wykresów właściwości (PG) wykracza poza zakres tej pracy. Tu Omawiając powiązane prace, przypominamy główne różnice między modelami, wskazując na specyficzne dla danej dziedziny zalety Przyjęcie wykresów właściwości podczas modelowania danych legislacyjnych. Wykresy właściwości modelują dane jako mieszane, tj. częściowo ukierunkowane multigrafia (Deutschetal., 2022). Oba węzły i krawędź są oznaczone i obecne – prawdopodobnie wiele – właściwości (to znaczy, że są skojarzone z parami właściwość/wartość). Ponieważ prawa mogą być naturalnie postrzegane jako węzły w KG, model grafu właściwości umożliwia przypisanie konkretne funkcje bezpośrednio do węzła (np. lista tematów regulowanych przez prawo). Ponadto, jak tradycyjnie przewidziano w każdym akcie prawnym Niektóre cechy, które czynią go unikalnym, to elastyczna reprezentacja atrybutów, która przypisuje taką specyfikę węzłowi (zobacz dyskusja w: Das, Srinivasan, Perry, Chong, & Banerjee, 2014). W RDF dołączanie dodatkowych informacji kontekstowych do poszczególnych Trójki są mniej trywialne, prawdopodobnie wymagają reifikacji, techniki, która pozwala na formułowanie twierdzeń na temat zdań, ale utrudnia wykonywanie zapytań o wydajność, efektywność pamięci masowej i użyteczność (Orlandi, Graux, & O'Sullivan, 2021). Model danych grafu właściwości również pozwala na bardziej naturalne wyrażanie ścieżek i wzorców grafów, ponieważ GQL ułatwia wyrażanie struktur ścieżek poprzez narzucanie przyjazne dla użytkownika ograniczenia składniowe (Francis *in.*, 2023). Zamiast tego w RDF tabelaryczny format wyników SPARQL ogranicza naturalną wyrażanie wzorców grafowych (Libkin, Martens, & Vrgoč, 2016; Seaborne, 2013), co utrudnia wyrażanie ścieżek złożonych struktur. Na przykład w dziedzinie prawodawstwa połączenie ścieżek i atrybutów ma kluczowe znaczenie dla wykrywania wzorców w danych, takie jak niespójności, które można łatwo zidentyfikować za pomocą zapytań dotyczących przechodzenia przez grafy. Wreszcie, ważną cechą dla coraz bardziej rozwijającej się dziedziny, jaką są systemy legislacyjne, jest również wydajność, przy czym PG jest bardzo odpowiedni do szybkiego przechodzenia przez relacje (Ciglan, Averbuch, & Hluchy, 2012), podczas gdy przechowywanie w RDF potraja się – zwłaszcza w połączeniu z reifikacją – szkodzi wydajności zapytań (Robinson, Webber, & Eifrem, 2015).

W Grecji wdrożono wstępne prototypy grafów do modelowania systemów legislacyjnych (Angelidis *in.*, 2018) oraz w Hiszpanii (Rodríguez-Doncel *in.*, 2018) oraz we Włoszech (Anelli *in.*, 2023). W takich modelach – wszystkie oparte na RDF – węzły prawne są połączone relacjami takimi jak "poprawia", "pochodzi od", "cytuje". Niemniej jednak każdy z takich prototypów ma swoje ograniczenia. Po pierwsze, Stopień szczegółowości jest na poziomie praw, a artykuły nie są traktowane jako węzły wykresu. Niektóre z nich, takie jak Hiszpania jeden (Rodríguez-Doncel *in.*, 2018), silnie opierają się na NLP i technikach rozpoznawania nazwanych jednostek w celu zbudowania KG, które może skutkować czynnikiem niskiej jakości, ponieważ zadanie prawidłowej identyfikacji prawa za pomocą zwykłej techniki opartej na sztucznej inteligencji jest skomplikowane zadanie, co może powodować pominięcia i niespójności (de Maat, Winkels, & van Engers, 2006; Sadeghian *in.*, 2018). Włoski prototyp (Anelli *in.*, 2023) poświęca swoje wysiłki na opracowanie narzędzi wspierających nawigację we włoskim systemie legislacyjnym, Podkreślają to również proponowane główne zastosowania zastosowań, które są w większości zorientowane na konkretne wyszukiwanie praw i powiązań oraz do tworzenia wizualizacji graficznych (Crotti Junior *in.*, 2020; Curtotti Curtotti & powiedział: McCreath, 2012; Curtotti, McCreath, & Sridharan, 2013; Oliveira & Oliveira, 2023). Należy również zauważyć, że żadne z wymienionych ćwiczeń nie wykorzystuje międzynarodowego standardu – takiego jak standard AKN oparty na XML – który pozwoliłby na wyższą jakość dzięki zastosowaniu znaczników XML wspierających je w odniesieniu do wykresu. wykorzystali pliki XML do zbudowania wykresu wiedzy legislacyjnej, na przykład przy użyciu języków mapowania, które jednak stają się specyficzne dla danego kraju (Crotti Junior, Orlandi, O'Sullivan, Dirschl, & Reul, 2019).

Rozwój rurociągów, które mają na celu wydobycie i uporządkowanie ustrukturyzowanych informacji z dokumentów ustawodawczych, był Trudne zadanie od czasu rozpowszechnienia się zdigitalizowanych wersji tekstowych dokumentów prawnych. Na przykład w Purpura i Hillard (2006), Autorzy opracowują klasyfikator ustawodawstwa Kongresu USA, który wykrywa temat o podstawowym znaczeniu dla projektu ustawy. Jednak Ich podejście wymaga, aby dostępna była już nieuseregowana lista tematów, a zatem ograniczona do aktów prawnych, które już takie przewidują. metadane. W Wulczyni i wsp. (2016), autorzy opracowują algorytm parsowania tabel w celu wyodrębnienia alokacji budżetowych za pomocą maszyny Klasyfikatory uczące się. Jednak, jak stwierdzili autorzy, wiele trudności wynikało z braku ustrukturyzowanego sposobu dostępu do Składniki dokumentu. Ostatnio zastosowania przetwarzania języka naturalnego i LLM, zwłaszcza do konstruowania wiedzy Wykres w domenie prawnej, rozprzestrzeniają się (Sansone & Sperli, 2022). Jednym z nich jest projekt Lynx (Moreno-Schneider *in.*, 2020), który łączy wiele technik NL w korpusie prawnym w celu skonstruowania KG. Jednak ostatnie prace przetestowały wydajność LLM w domenie bezpośredniej aplikacji text-to-KG, co pokazuje, że nadal brakuje im elastyczności w tworzeniu wysokiej jakości KG. Jednocześnie mogą być nadal wykorzystywane jako asystenci w celu zwiększenia dokładności faktograficznej KG specyficznych dla danej domeny (Zhu *in.*, 2024).

Zainspirowany rozwojem standardu GQL i międzynarodowym przyjęciem standardu XML do reprezentowania prawa, W niniejszej pracy proponujemy nowatorski paradygmat przedstawiania złożoności systemów legislacyjnych poprzez wykorzystanie najnowszych osiągnięć technologia grafowych baz danych i używanie LLM jako asystentów w celu wzbogacenia węzłów i krawędzi naszego KG. Nasz wybór oferuje dobry kompromis między intuicyjnością a elastycznością w reprezentacji danych (Angles, 2018), w połączeniu z bardziej kontrolowanym wykorzystaniem LLM.

Tabela 1

Podstawowe elementy składowe standardu Akoma Ntoso, które są wykorzystywane do reprezentowania prawa w wielu tradycjach legislacyjnych.

	Znacznik XML	Zawartość
Metadane	ERBRten	Niepowtarzalny identyfikator aktu, zgodnie z
	dokumentTytuł	przepisami prawa i tytułem prawnym.
	docTyp	Rodzaj prawa.
	dokumentData autorskaUwaga	Detal publikacji jest statystycznie informatywny zawierający istotne informacje na temat aspektów prawa.
Tekst ustawy	Przedmowa/ nagłówki	Informacje o tytule ustawy, (progresywnym) numerze identyfikującym ustawę, dacie jej wprowadzenia. Część tekstu, która określa podstawę prawną i wprowadza ustawę.
	preambuła	Zasadnicza treść prawa, obejmuje wszystkie podstawowe jednostki prawa. Podstawowa
	Treść artykułu/ sekcji/zasady	jednostka prawa, tj. główny podział ciała
	Wnioski	Znacznik zawierający oświadczenia końcowe i podpisy ministrów.
	Nagłówki załączników	Dokumenty tekstowe lub graficzne, które integrują informacje zawarte w treści.
Odwołania	Nagłówki	Nazwa podstawowej jednostki organu prawnego.
	cytaty activeMod	Cytaty z ustaw lub artykułów, które stanowią podstawę prawną tego, co jest uchwalane. Blok zawierający poprawki/uchylenia dokonane w innym dokumencie.
	textualMod	
	powiedzial	Wzrost i wzrost jest jednostką modyfikacji. Przedstawia on artykuł w tekście, w którym podana jest modyfikacja.
	Miejsce docelowe	Podobnie jak w przypadku tagu źródłowego, ale odnoszącego się do ustawy lub artykułu, który jest modyfikowany.
	href powiedzial	Identyfikator dokumentu docelowego lub części dokumentu dla cytatu.

3. Podstawy i schemat grafu

Opierając się na niedawnym przyjęciu wspólnych międzynarodowych standardów reprezentacji prawnej opartych na XML, koncentrujemy się na składnikach procesu ETL wykorzystanego do skonstruowania grafu wiedzy zgodnie z włoskim prawodawstwem. W pierwszej kolejności przypominamy jeden z najpopularniejszych standardów międzynarodowych, który jest przyjmowany w coraz większej liczbie krajów. Następnie przedstawiamy pierwszy schemat modelowania systemu w Grafach Wiedzy w oparciu o model danych wykresu właściwości. W kolejnych sekcjach zastosujemy te składniki do włoskiego ustawodawstwa i poprawimy jakość Grafu Wiedzy poprzez zastosowanie dużych modeli językowych.

3.1. Międzynarodowy standard XML Akoma Ntoso

Przyjmując i wykorzystując standard XML przyjęty na całym świecie, proces tworzenia, analizowania i porównywania systemy legislacyjne zostałyby znacznie przyspieszone. Wśród standardów XML Akoma Ntoso wyróżnia się jako jeden z najbardziej obiecujących ponieważ został oficjalnie przyjęty przez wiele krajów (Witalij i in., 2019). Jedną z jego kluczowych zalet jest zdolność do Uchwycić podstawowe cechy dokumentów zgodnych z prawem w różnych systemach, takie jak identyfikacja podstawowych jednostek prawa oraz pomocnicze znaczniki identyfikacyjne do modelowania odniesień do innych praw. Specyfikacje standardu AKN zostały również zatwierdzone przez organ OASIS (OASIS, 2018), co świadczy o jego wysokiej jakości i interoperacyjności między systemami legislacyjnymi. Do najważniejszych instytucji, które przyjęły AKN, możemy znaleźć Parlament Europejski (Urząd Publikacji Unii Europejskiej, 2023), który prawdopodobnie zachęci wiele państw członkowskich UE do dostosowania swoich systemów do tego formatu.

W USA Biblioteka Kongresu próbowała przekształcić Kodeks Stanów Zjednoczonych w standard AKN (Legix.Info, 2012). Akoma Ntoso została oficjalnie przyjęta i wdrożona we Włoszech, a wszystkie jej przepisy są publikowane w tym formacie na oficjalnym portalu Normattiva, Wielka Brytania, 2 Szwajcaria, a także przez instytucje międzynarodowe, takie jak Organizacja Narodów Zjednoczonych (Rada Dyrektorów Naczelnych Systemu ONZ ds. Współpracy lub Dynacjologii, 2017) i FAO (Palmirani, 2018). W związku z tym, chociaż niniejsza praca koncentruje się na ustawodawstwie włoskim, podejście do budowania Wiedzy Grafu będzie miało bezpośrednie zastosowanie do innych krajów, które wdrożyły i opublikowały przepisy w AKN.

Główne bloki konstrukcyjne AKN. Tabela 1 zawiera szczegółowe informacje na temat głównych elementów składowych AKN, które omówimy w kolejnych sekcjach naszego rurociągu ETL oraz do zbudowania wykresu Knowledge włoskiego ustawodawstwa (zob. sekcja 4). Znaczniki AKN są przeznaczone do przechwytywania Nieco inne aspekty tradycji legislacyjnych – demokratycznych. Na przykład preambuła zawsze ujmie formuły w celu określenia "podstawy prawnej", tj. innych aktów prawnych, które są niezbędne do zapewnienia podstaw prawnych nowej ustawy, oraz do opisanie "Uchwalanie zdań", czyli wyrażenia językowych, które są regularne dla danej tradycji i służą do wprowadzenia tekstu prawa. Standard AKN definiuje również wiele znaczników, które mogą być używane w częściach treści (rozdział, sekcja, artykuł, reguła itp.), oznaczając podstawowych jednostek systemu legislacyjnego. Taki znacznik zależy od konkretnej tradycji legislacyjnej, np. artykuł i reguła są tym samym przedmiotem dla różnych aktów prawnych. Ponieważ skupiamy się na ustawodawstwie włoskim, będziemy używać tagu article, aby odnieść się do jednostki prawa zasadniczego. Dla Na przykład, w przypadku tradycji amerykańskiej, jednostka prawna jest "sekcją" prawa. Szczególną uwagę należy zwrócić na przepisy prawa i Cytaty z artykułów. W rzeczywistości istnieje wiele typów cytatów, a każdy z nich jest przechwytywany w dedykowanym znaczniku XML, w zależności od cytatu

1 <https://www.normattiva.it/>, <https://www.fedlex.admin.ch/eli>.  
2 one w AKN (<https://www.fedlex.admin.ch/eli>).  
3



innych ustaw oraz cytaty w preambule, tj. odniesienia do innych ustaw lub artykułów, które stanowią podstawę prawną ustawy. Inne cytaty mogą pojawiać się w całym tekście wewnątrz ogólnego znacznika href, którego nie można sklasyfikować jako modyfikacji lub jako podstawy prawnej. Wreszcie, w wielu tradycjach legislacyjnych jesteśmy świadkami obecności załączników, tj. dodatkowych dokumentów w formie tekstowej lub graficznej, na przykład tabel, które nie pojawiają się w treści prawa, ale z jakichkolwiek względów praktycznych lub innych. Na przykład umowa międzynarodowa zatwierdzona przez prawo pokrewne jest zawsze dostarczana jako załącznik. Takie obiekty są przechwytywane w dedykowanych znacznikach XML, czyli dokumentach załączników. W następnej sekcji szczegółowo omówimy, w jaki sposób każdy tag jest używany do tworzenia obiektów wykresu.

### 3.2. Schemat wykresu właściwości

Aby wyrazić nasz schemat, bierzemy pod uwagę Cypher (Neo4j, 2024), deklaracyjny język zapytań dla grafów właściwości (Angles i in., 2017), który jest bardzo zbliżony do niedawno ustandaryzowanego języka Graph Query Language (GQL) (ISO, 2024). Cypher jest wspierany przez Neo4j, jeden z najpopularniejszych grafowych systemów zarządzania bazami danych (Francis i in., 2018), który przyjmujemy w tej pracy. Ten Model danych wykresu właściwości składa się z węzłów, które mogą mieć etykiety i wiele obiektów atrybutycznych (które będziemy nazywać właściwościami), jak a także ukierunkowane relacje, które mogą być również oznaczone etykietami i mieć swoje atrybuty. Na Fot. 1, przedstawiamy proponowany wykres właściwości schemat.

#### 3.2.1. Opis schematu grafowej bazy danych

W tej sekcji szczegółowo omówiono schemat grafu i motywację stojącą za wyborami modelowania.

**Węzły prawne.** Każde uchwalone prawo jest modelowane jako węzeł na wykresie. Węzły prawa są identyfikowane za pomocą klucza opartego na ciągach znaków przyjętego w każdym

ustawodawstwo. Na przykład w UE europejski identyfikator prawodawstwa (ELI) (Urząd Publikacji Unii Europejskiej, 2024) służy do określenia niepowtarzalnie akty ustawodawcze. Ponadto pozyskujemy wszystkie istotne metadane i przypisujemy je jako właściwości prawa; Należą do nich tytuł, rodzaj ustawy, datę publikacji i datę wejścia w życie. Aby poprawić użyteczność, dołączamy również do właściwości węzła liczbę artykułów i załączników; Chociaż zapytanie może łatwo uzyskać te informacje, my decydujemy się na jego Wstawienie jako właściwość w węzle, ponieważ może to być natychmiastowe zainteresowanie użytkowników. Na koniec dodajemy właściwość domeny (opisując ministerstwa lub departamenty zaangażowane w nowe prawo) oraz właściwość tematu (opisując konkretne tematy, których dotyczy ustawa). Te właściwości są modelowane w postaci list, które mogą mieć wiele wartości (funkcja obsługiwana przez modele danych wykresów właściwości).

**Węzły artykułów.** Każde prawo składa się z jednego lub więcej artykułów, modelowanych jako dodatkowy węzeł schematu, połączony HAS

związek. Artykuł ten jest w istocie podstawową jednostką prawa i zawsze można go zidentyfikować za pomocą liczby progresywnej, która jest następnie łączony z identyfikatorem prawa w celu utworzenia identyfikatora artykułu. Artykuły mają swoje właściwości: tytuł, numer i pełny tekst. Aby wyodrębnić pełny tekst prawa, wystarczy napisać zapytanie, które łączy się z tekstem zawartym w (stopniowo) ponumerowanym artykułach (zob. dodatek A.2). W węzłach prawnych tematy są dodatkową właściwością o praktycznej użyteczności; Poszczególne artykuły mogą jednak dotyczyć odrębnych przepisów prawa. Na przykład ustawa zatytułowana "Przepisy dotyczące reorganizacji uprawnień departamentów" może poświęcić swoje artykuły każdemu Departamentowi. W związku z tym każdy z nich byłby poświęcony odrębnemu tematowi.

**Węzły załączników.** Możliwe, że prawo może również obejmować załączniki lub załączniki. Te specjalne dokumenty określają aspekty relacji z nadrzędnym węzłem prawnym. Załączniki – z definicji – nie

HAS\_ATTACHMENT

może orzekać o zmianach w innych przepisach i wskazywać, że takie zmiany są szczegółowo opisane w załączniku. Z formalnego punktu widzenia Źródło modyfikacji pozostaje artykułem prawnym. W związku z tym bierzemy pod uwagę odrębny typ węzła w naszym schemacie. Niektóre załączniki Mogą to być tabele, które określają dodatkowe informacje w formie tabelarycznej. Na przykład wartości nowych taryf za prawa jazdy lub wykazy opisujące realokację zasobów ludzkich między działami. W związku z tym, oprócz tych samych właściwości węzłów artykułu, Dodajemy właściwość type, wskazującą charakter zawartości załącznika.

**Krawędzie odniesienia.** Modelujemy pięć typów możliwych krawędzi odniesienia, które ujmują wzajemne powiązania między prawami, artykułami i załącznikami, a mianowicie krawędzie IS\_LEGAL\_BASIS\_OF, AMENDMENTS, INTRODUCES, UCHYLES i CITES. IS\_LEGAL

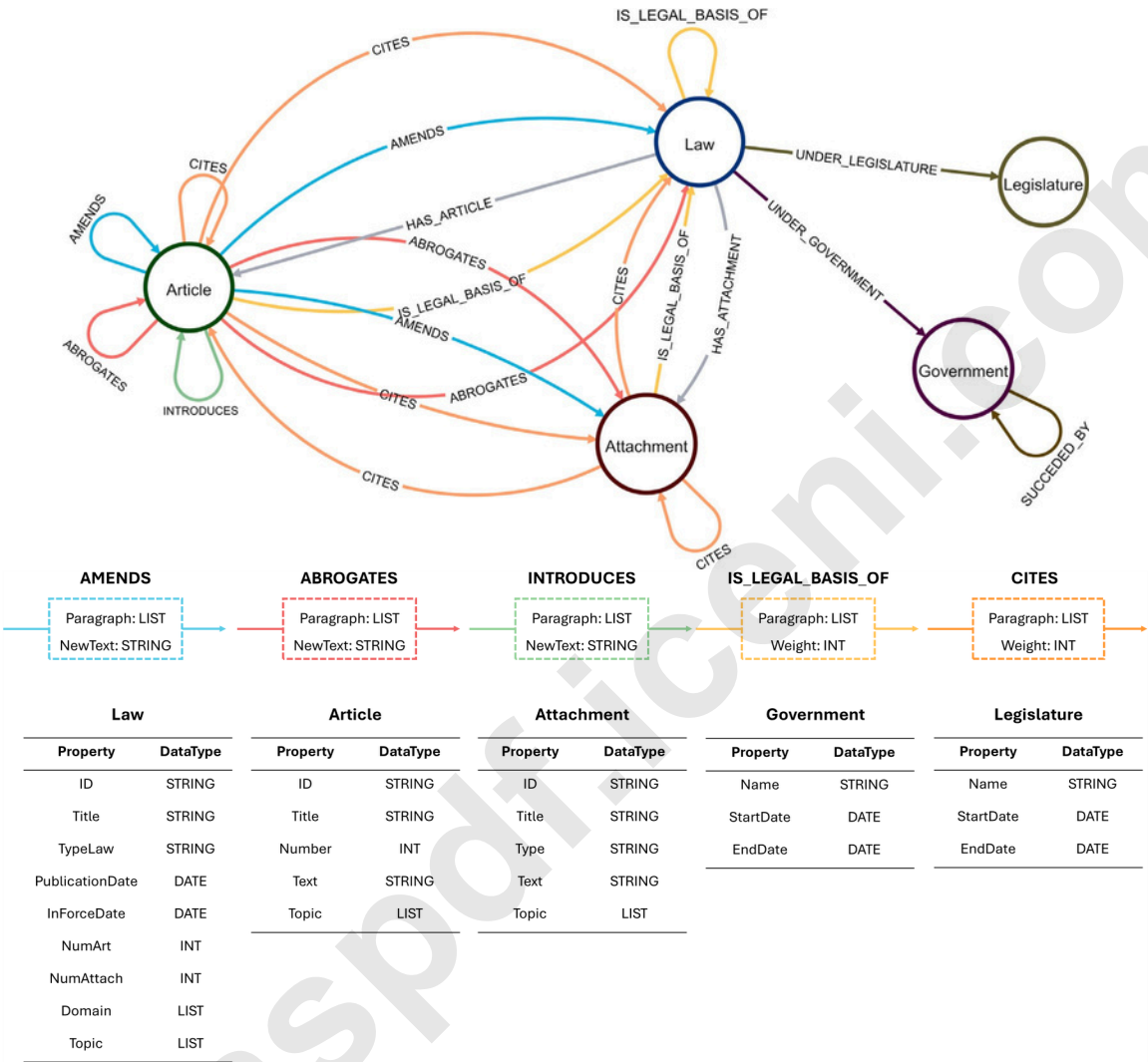
Krawędzie oznaczają odniesienia w części wprowadzającej dokumentu, które określają jego podstawę prawną. W związku z tym węzeł źródłowy dla takich krawędzi

mogą to być ustawy, artykuły lub załączniki (zawarte w preambule ustawy o miejscu przeznaczenia). POPRAWKI, WPROWADZENIE I AB

Krawędzie to odwołania, które odpowiednio zastępują, dodają lub usuwają część lub pełny tekst artykułów i załączników, które wcześniej występowały Opublikowany. Zgodnie z normatywnymi regułami redakcyjnymi (Karpen, 2008) załączniki muszą zawierać treść, która nie może być sformułowana w normatywnym sposób (z wyłączeniem zasad modyfikacji). W związku z tym te trzy rodzaje krawędzi zawsze prezentują węzeł artykułu jako źródło. Zamiast tego ich miejscem docelowym może być również prawo (np. gdy tytuł prawny zostanie zmieniony). Wreszcie, prawa, które występują w krawędzie oznaczają ogólne odniesienia do innych całym tekście, aby przypomnieć odpowiednie prawo, artykuł lub załącznik, który może być ważny do zacytowania w tekście podania

Informacje kontekstowe. Na przykład odniesienie jest używane (i wymagane) przy podawaniu definicji lub specyfikacji terminów i przedmiotów używanych w całym prawie, np. wykazu szkodliwych substancji chemicznych.

Do każdej z opisanych krawędzi przypisana jest właściwość paragraph, czyli lista wskazująca akapity docelowe, które są zainteresowane przez odniesienie. Jeśli odwołanie wskazuje cały artykuł lub załącznik, właściwość paragraph ma wartość null. ZMIENIA, INTRODU ABROGATESedgesprzedstawia również właściwość newText, która przechowuje tekst, a modifiedbythesourcenode.IS\_LEGAL\_BASIS\_i krawędziom CITES przypisywana jest właściwość weight, która zlicza, ile razy to samo odniesienie ma zastosowanie.



Rys.1. PropertyGraphschema wizualizacjaodes i directededges, modelowanie systemu autonomicznego, np. włoskiego; formalny schemat PG (zgodnie z definicjąkątów etal.(2023)) znajduje się w dodatku A.1. Na spodzie wyszczególniamy krawędzie i węzły wraz z ich właściwościami; Krawędziebez określonych właściwości są pomijane ze względu na zwięzłość.

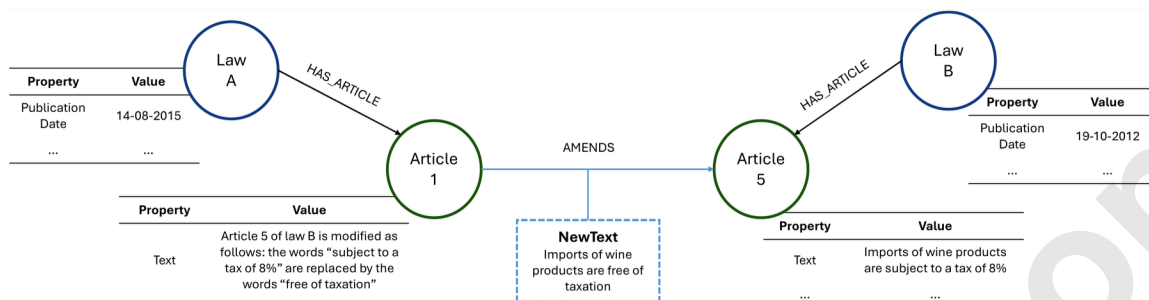
w preambule lub w tekście. Węzły rządowe i ustawodawcze. Każdy akt prawny jest uchwalany w odrębnych krajobrazach legislacyjnych,

które przechwytyjemy, dodając

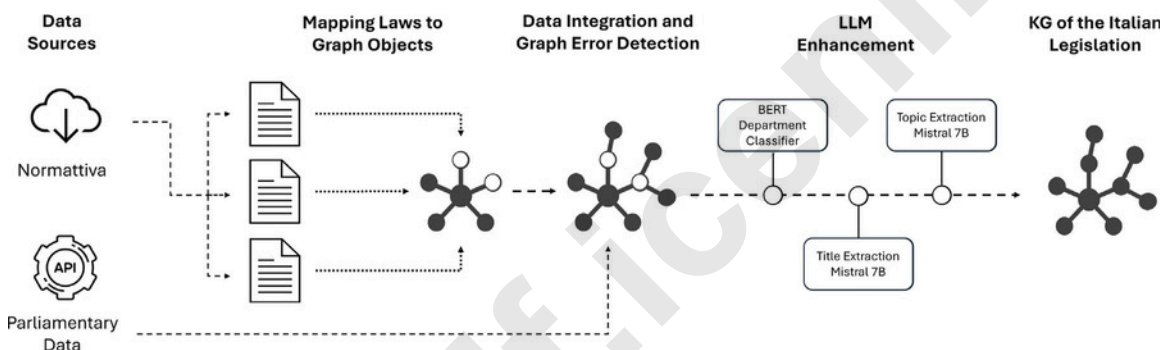
węzły rządowe i ustawodawcze. Takie węzły są naturalnie i jednoznacznie identyfikowane przez ich nazwę, która jest zwykle oznaczana za pomocą liczby progresywnej, takiej jakLegislatura I, Legislatura II lubBerlusconi-IiBerlusconi-II dla rządów, przy użyciu nazwiska odpowiedzialnego premiera. Oba obiekty grafu mają wspólne właściwości daty początkowej i końcowej, które charakteryzują czasową ewolucję władzy ustawodawczej i rządów. Należy zauważyć, że w większości krajów demokratycznych daty rozpoczęcia i zakończenia kadencji ustawodawczych i rządów nie pokrywają się, ponieważ rządy zwykle oficjalnie pozostają u władzy nawet po wyborze nowego parlamentu, tj. rozpoczęciu nowej kadencji; Powodem jest brak bezpośredniej przewagi między władzą ustawodawczą a rządami, podczas gdy oba są połączone z węzłami prawnymi za pośrednictwem UNDER\_LEGISLATUREandUNDER\_GOVERNMENTedges. Modelujemy je jako niezależne węzły w naszym schemacie.

W sekcji 3.1 ten wybór modelowania okazuje się pomocny w uzyskiwaniu informacji na temat krajobrazu legislacyjnego. Na przykład poprzez wykorzystanie SUCCEEDED\_BYrelationship zapytania o ścieżki cykliczne umożliwiają przechodzenie przez wzorce czasowe. W przyszłych iteracjach więcej

mogłybyuwzględnićdodatkowe informacje o krajobrazie legislacyjnym, takie jak na przykład skład parlamentu partii politycznych.



Rys.2. Przykład ilustrujący, w jaki sposób uchwyciliśmy czasową ewolucję praw. W tym przypadku art. 5 prawa B, opublikowany w 2012 r., został zmieniony przez art. 1 ustawy A, opublikowany w 2015 r. Oryginalny tekst art. 5 jest przechowywany jako własność węzła. Jego nowa wersja, zmodyfikowana przez art. 1 prawa A, jest przechowywana z właściwością NewText krawędzi poprawki. W związku z tym pełny tekst ustawy B w określonym czasie można uzyskać za pomocą zapytania graficznego (zob. [dodatek A.2](#)).



Rys.3. Proces integracji danych z KG. W tym przypadku dane z różnych źródeł (np. z API, z plików, z bazy danych) są przetwarzane i integrowane z KG. Następnie stosujemy kroki czyszczenia oparte na grafach w celu skorygowania błędów wykrytych w zapytaniach. Wreszcie, dostosowane LLM integrują KG z dodatkowymi funkcjami, które uzupełniają i zwiększają jakość i bogactwo bazy danych.

### 3.2.2. Uchwycenie wymiaru czasowego

Systemy legislacyjne w naturalny sposób ewoluują, z ciągłym napływem nowych praw, które modyfikują lub uchylają stare. Każda z takich zmian oznacza nowe wersje praw, które przechwytują nowy tekst za każdym razem, gdy następuje zmiana; prowadzi to do wykładniczego wzrostu liczby dokumentów tekstowych, ponieważ niewielkie zmiany w tekście wymagają przechowywania dodatkowego pliku. Wykorzystujemy nasz model danych grafowych, aby przezwyciężyć takie ograniczenia, przechowując zmodyfikowane artykuły jako właściwość krawędzi. Dzięki temu możemy uzyskać żadaną wersję prawa poprzez zapytanie o KG. Rzeczywiście, przechowujemy tylko oryginalną wersję prawa w węzłach; Wszelkie zmiany można odzyskać, poruszając się po wykresie w celu wyszukania informacji o prawie w określonym znaczniku czasu za pośrednictwem właściwości krawędzi. [Figa. Rysunek 2](#) ilustruje, w jaki sposób używamy funkcji wykresu właściwości do śledzenia wielu wersji tekstowych tego samego artykułu. W [dodatku A.2](#) przedstawiamy praktyczne zapytania, które ilustrują, w jaki sposób przepytujemy wykres właściwości pod kątem cech zależnych od czasu, takich jak wyprowadzanie wersji prawa lub wykrywanie liczby praw nadal obowiązujących w danym znaczniku czasu.

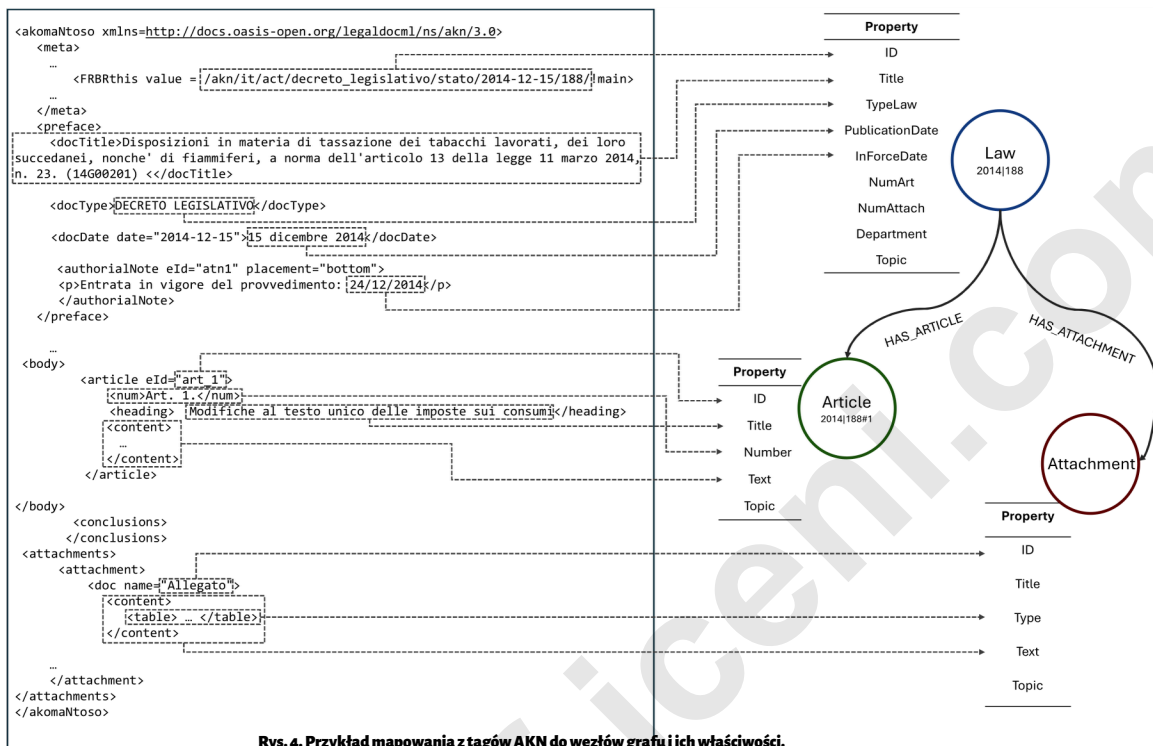
## 4. Tworzenie wykresu wiedzy na temat włoskiego prawodawstwa

W tej sekcji opieramy się na fundamentach przedstawionych w sekcji 3. Opracowujemy potok Extract-Transform-Load (ETL) dla Włoski system prawny. Wykorzystujemy postępy w oficjalnym modelowaniu danych (tj. standardzie AKN) i prezentujemy szereg technik przekształcania takich dokumentów i ich treści w obiekty danych grafowych ([Sana & Suganthi, 2017](#)), tj. przedstawionego schematu wykresu właściwości. W tym celu integrujemy również publicznie dostępne źródła danych i pokazujemy, jak łączymy dane wejściowe z zestawem starannie dostosowanych modeli językowych w celu uzyskania kompleksowej i wysokiej jakości reprezentacji danych legislacyjnych, co również przedstawimy na przykładzie. Przegląd rurociągu ETL przedstawiono na [rys. 1.3](#). Potok działa codziennie i aktualizuje naszą bazę danych wykresów nieruchomości na Neo4j, publicznie dostępną (w wersji zamrożonej) w repozytorium Zenodo ([Colombo, 2024b](#)).

Aktualizacja wykresu opiera się na oficjalnych źródłach danych (tj. danych rządowych i parlamentarnych), które zapewniają w ten sposób regularną i terminową publikację nowych danych, gdy tylko staną się one dostępne, tj. gdy zostanie opublikowana nowa ustawa.

Wymagania dotyczące zrównoważonego rozwoju i odtwarzalności. Zaprojektowaliśmy potok ETL, stawiając na pierwszym miejscu zrównoważony rozwój (w szczególności wydajność) i odtwarzalność. W tym celu przyjęliśmy: (i) lekkie LLM (które przyczyniają się do usprawnienia operacji, reducing powiedz: [4.5](#) omówiono potencjalne uogólnienie naszego rurociągu na inne systemy legislacyjne.





#### 4.1. Źródło danych prawa włoskiego

Współczesny włoski system legislacyjny wywodzi się z przyjęcia konstytucji republikańskiej w 1948 roku, która służy jako punkt graniczny, aby wykluczyć przestarzałe prawa z okresu Królestwa. Zrobiliśmy jednak wyjątki dla dwóch istotnych ustaw, Kodeksu cywilnego oraz kodeksy karne, które pozostają w mocy mimo daleko idących modyfikacji. Wszystkie prawa są publicznie dostępne za pośrednictwem Normattiva portalu, który korzysta ze standardu Akoma Ntoso. W celu zbudowania Grafu Wiedzy przyjęcie przedstawionego schematu w sekcji 3.2 zebraliśmy wszystkie akty prawne opublikowane po dacie granicznej w ich pierwotnej wersji, tj. w takiej postaci, w jakiej zostały opublikowane. Wtedy Pipeline automatycznie pobiera nową enuchwalone prawa na co dzień. Podczas gdy systemy legislacyjne zazwyczaj rozróżniają wersje a law, jak omówiono w sekcji 3.2.2, musimy jedynie pobrać nowe prawa w ich oryginalnej formie, ponieważ proponowany schemat obejmuje wszystkie późniejsze zmiany.

**Struktura prawa włoskiego.** Zgodnie z oficjalnymi regułami (Senato della Repubblica, 2001) prawo włoskie musi być zgodne z wcześniej określonym ogólnym strukturą. W pierwszej części, po tytule, preambuła wskazuje podstawy prawne ustawy, jeżeli są dostępne; Następnie, korpus Ustawa, zawierająca artykuły, które są podzielone na artykuły wprowadzające (z ogólnymi i głównymi przepisami prawa), artykuły główne (ze szczegółowymi zasadami tego, co jest regulowane) oraz artykuły końcowe (zawierające informacje o obowiązujących postanowieniach ustawy). Każdy Artykuł jest podzielony na akapity, z których każdy kończy się podziałem wiersza. Po oświadczeniach końcowych – zawierających podpisy odpowiedzialnych urzędników – tabele, prospekty emisyjne, wykazy itp. mogą zostać zamieszczone w załączniku do tekstu legislacyjnego.

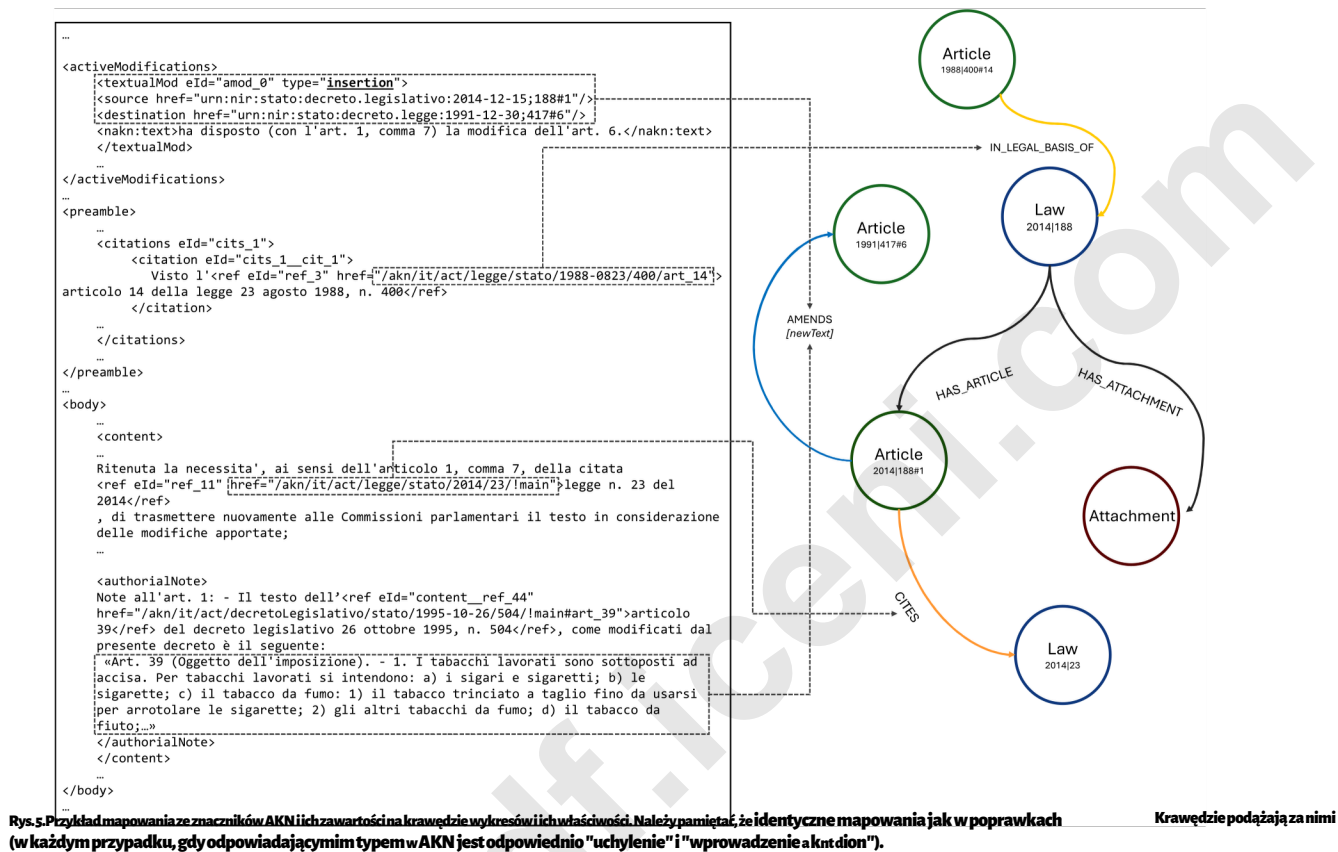
#### 4.2. Mapowanie dokumentów AKN i znaczników do obiektów wykresu właściwości

Ustawy w portalu Normattiva są dostępne w standardzie AKN (patrz punkt 3.1); prezentujemy, w jaki sposób wykorzystujemy tagi AKN tabeli 1 w celu odwzorowania jego obiektów na wykresie właściwości, tj. węzłów, krawędzi i właściwości.

**Schema Nodes.** Po pierwsze, wyprowadzamy węzły schematu, tj. węzły praw, artykułów i załączników, jak przedstawiono w sekcji 3.2. Fig. 4 wizualnie ilustruje mapowanie.

**1. Węzły prawne.** Każde prawo z Normattiva, tj. każdy dokument AKN, reprezentuje węzeł prawny na wykresie. Dla każdego z tych przepisów metadane przechwycone w określonych znacznikach XML są używane do wyprowadzania i wyodrębniania właściwości pierwszego węzła, jak na rys. 1.4. W szczególności:

Pobieramy tytuł, datę publikacji i rodzaj aktu. Zliczając obecność tagów article i attachment dostępnych w całym XML, otrzymujemy całkowitą liczbę artykułów i załączników. Następnie wyprowadzamy obowiązującą właściwość daty wyszukując w tagach authorialNote tag "Entrata in vigore del provvedimento" (tj. data wejścia w życie ustawy), który zawiera konkretną datę.



## UCHYLA WPROWADZENIE

**2. Węzły artykułów.** W ustawodawstwie włoskim każde prawo składa się z artykułów, podstawowej jednostki prawnej, która szczegółowo opisuje różne aspekty tego samego prawa. Artykuły mają swoje własne tytuły, zwane epigrafami, i są ujęte w określonym tagu nagłówka. W związku z tym każdy znacznik artykułu w pliku AKN definiuje węzeł artykułu w grafie, który jest naturalnie połączony z węzłem nadrzędnym, który mapuje dokument AKN w pliku wykres. Każdy artykuł ma znacznik nagłówka, który jest wykorzystywany do pobierania właściwości tych ten epigraf i znaczniknum, reprezentujący jego progresywny numer w prawie. Wywodzimy IDof Węzły artykułów poprzez połączenie identyfikatora węzła z numerem artykułu. Suchaconstructisexploitedinhreftags przez cały okres Tekst do cytowania inny Artykuły (zezwalające na umieszczanie odniesień do mapy, jak zobaczymy w poniższych sekcjach).

**3. Węzły mocowania.** Załączniki są modelowane jako dodatkowe obiekty wewnętrzne znaczników załączników. Foreachattachment, tagdokumentu zawiera jego nazwę, której używamy do ustawienia identyfikatora węzła, połączonego z identyfikatorem prawa nadrzędnego. Na przykład, jeśli załącznik jest tabelą, której identyfikator jest uzyskiwany przez konkatencję: AKN\_ID#Tabela1, gdzie Tabela to nazwa dokumentu. Ponadto czerpiemy Rodzaj załącznika poprzez sprawdzenie, czy znacznik tabeli jest używany w części dokumentu.

Krawędzie odniesienia do prawa. Odniesienia mogą zawierać cytaty z innych dokumentów znajdujących się w tekście. różne cele i charaktery, takie jak substytucja softext, dodatki nowych części lub słów lub bardziej ogólne odniesienia do przypomnij sobie pewne aspekty. Wykorzystujemy standard AKN, aby uchwycić takie rozróżnienie; Norma obejmuje wstępnie zdefiniowany zestaw możliwych cytaty, wraz z dedykowanymi znacznikami XML, których prawodawcy muszą przestrzegać, aby zachować zgodność z międzynarodowym standardem (tj. su bstitution, wstawianie, dzielenie, łączenie, zmiana numeracji, uchylenie). Poniżej opisujemy, w jaki sposób mapujemy AKN do obiektów KG (patrzrys. 1). Sdla wsparcie wizualne:

### 1. IS\_LEGAL\_BASIS\_OFedges, stanowiąc podstawę prawną prawa. Standard AKN przechwytuje takie odniesienia w ramach

cytatTagi. Były one obecne jako bezpośrednie krawędzie: ich przeznaczeniem jest zawsze węzeł prawny, którego preambuła jest analizowana; Właściwość wagi jest wielokrotnie taka sama

Para Źródło-miejsce docelowe danego typu jest wymieniona w całym tekście.

### 2. Zmieniamy krawędzie. Wyprowadzamy modyfikację krawędzi, wyszukując znaczniki tekstowe Modaktywnym bloku XML Modyfikacje

Zgodnie ze standardem, każdy aktywny tag modyfikacji reprezentuje modyfikację, tj. krawędź czterogracznego.

### 3. WPROWADZAMY krawędzie. Z poprawek izolujemy znaczniki tekstowe Mod, których typ jest tylko do wstawiania i wyprowadzamy

Krawędzie wprowadzenia, które dodają dodatkowy tekst bez modyfikowania poprzednich akapitów.

4. Krawędzie. Podobnie znaczniki, których trybem jest uchylenie, są zamiast tego modelowane jako krawędzie uchylone, które są
5. ~~uchylone~~ dla części tekstu. W przypadku, gdy artykuł zostanie całkowicie uchylony, krawędź nie ma właściwości akapitu, co wskazuje na pełne uchylenie. krawędzie, tj. inne, bardziej ogólne cytaty z innych aktów prawnych, które mogą być innym aktem prawnym, artykułem, załącznikiem, a nawet konkretnymi aktami. Takie cytaty uzupełniają tekst, przywołując inne przydatne informacje i są gromadzone przez W przypadku wszystkich typów cytatów używamy struktury danych listy, aby uchwycić przypadki, w których cytowanych jest wiele odrębnych akapitów tego samego artykułu. Na przykład, jeśli istnieją dwie krawędzie poprawek z tą samą parą źródło-miejsce docelowe -np. (Prawo A Art. 2)-[r:POPRAWKA]- (Prawo B Art. 1) – ale odwołujące się do odrębnych ustępów docelowych – np. odpowiednio ust. 1 i 4 art. 1 prawa B – reprezentujemy obie w obrębie tej samej krawędzi, ale z właściwością akapitu mającą dwa elementy (np. r. ust. [1,4]).

**Rola znaczników AuthorialNote.** Implementacja standardu AKN w systemie włoskim wykorzystuje znaczniki AuthorialNote do dodawania adnotacji w całym tekście. Jako przykład służy do dodawania inForceDate do metadanych, ponieważ dla takich informacji nie zdefiniowano żadnego konkretnego znacznika AKN (patrz mapowanie na rys. 1). 4). Służy również do wstawiania użytecznych informacji ciągłych w całym tekście ustawy, np. tekst docelowy jako modad zart. ustawy; na przykład na Fot. Art. 5 ust. 1

Ustawa 2014/188 zmienia art. 6 ustawy 1991/417 (zob. znacznik aktywna Modyfikacje na początku dokumentu XML). Ten Pełną nową wersję tekstową tego ostatniego można odszukać, analizując notatkę autorską na końcu art. 1, która określa nową wersję, zawierającą modyfikacje wprowadzone przez nowy artykuł. Innymi słowy, za każdym razem, gdy rzeczywisty tekst prawa orzeka Zastąpienie niektórych fragmentów innego artykułu notą autorską – która nie jest częścią właściwego tekstu ustawy – wskazuje nową wersję artykułu docelowego. W naszym potoku nowy tekst jest przypisywany jako właściwość do krawędzi modyfikacji.

#### 4.3. Integracja danych i wykrywanie błędów na podstawie wykresów

Aby poszerzyć zakres czterech wykresów własności, w pierwszej kolejności przystępujemy do integracji danych, które opisują kontekstowe ramy prawne krajobraz. Następnie, za pomocą zapytań grafowych, wykrywamy błędy i niespójności w danych, a gdy tylko jest to możliwe, bezpośrednio przyjąć strategię korekty mającą na celu poprawę spójności KG. Pokazujemy również, w jaki sposób wykorzystujemy model danych do sygnalizowania ustawodawcy niespójności systemowych, tj. błędów, które wynikają z nieprawidłowych działań legislacyjnych.

##### 4.3.1. Władza ustawodawcza i węzły rządowe

Zgodnie z naszym schematem integrujemy informacje o rządach i organach ustawodawczych, pod którymi prawo zostało uchwalone Opublikowany. Informacje te pozwalają nam, jak zobaczymy w sekcji 5, na analizę cech systemów legislacyjnych na wyższym poziomie. W przypadku ustawodawstwa włoskiego gromadzimy takie dane z punktu końcowego dostarczonego przez parlament włoski (Camera dei Deputati, 2024), który dostarcza aktualnych informacji o rządach i danych parlamentarnych. Krawędzie łączące prawa z Rządy i węzły ustawodawcze są wyprowadzane przez wykorzystanie wymiaru czasowego, aby zrozumieć, w ramach którego rządu i legislatury prawo zostało opublikowane.

##### 4.3.2. Wykrywanie błędów na podstawie wykresu

Mimo że źródła dokumentów AKN są wysokiej jakości (są one dostarczane bezpośrednio przez Dziennik Urzędowy), zestaw wykresów Wzorce można uruchamiać w celu sprawdzenia niespójności w źródle danych. Takie wzorce wykresów można łatwo zaimplementować za pomocą naszych Model PG w postaci zapytań Cypher. Należy pamiętać, że obecność błędów wpływa na wyniki zapytania, na przykład podczas obliczeń systemowych statystyki oparte na wzorcach wykresów; W związku z tym ich wykrywanie i zgłaszanie ma zasadnicze znaczenie dla osiągnięcia wysokiej jakości reprezentacji danych legislacyjnych.

1. Krawędzie automatycznego cytowania, tj. krawędzie odniesienia z tym samym węzłem co źródło i miejsce docelowe. Chociaż odwołania wewnętrzne są dozwolone, wykluczamy je z naszego KG, ponieważ ich charakter różni się od innych krawędzi cytowań. Identyfikujemy je za pomocą prostego szyfru (patrz poniżej), a następnie usuwamy je z wykresu.

MATCH p=(l:Prawo)-[:HAS\_ARTICLE]->(a:Artykuł)-[:CITES|POPRAWKI|ZNOSI SIĘ|PRZEDSTAWIA]->

```
(a2:Artykuł)-[:HAS_ARTICLE]-(l:Prawo) RETURN p
UNIA
PODAJ P=(l:Prawo)-[:IS_LEGAL_BASIS_OF]-(l:Prawo) RETURN p
UNION MATCH p=(l:Prawo)-[:IS_LEGAL_BASIS_OF]-(a:Artykuł)-
```

```
[:HAS_ARTICLE]-(l:Prawo) RETURN p
```

gdzie istnieją, który może być połączony ze sobą za pomocą trzech różnych wzorców grafów, a mianowicie odniesień między własnymi w celu usunięcia artykuły, bezpośrednie odwołania do samego siebie w preambule oraz odniesienia do jednego z jej artykułów w preambule. To zapytanie pozwoliło nam 90edges, głównie ogólnychOdniesienia do CITES (pierwszy typ wzorca zapytania) andIS\_LEGAL\_BASIS\_OFcitati (drugi typ wzorca).

2. Incorectedgessource, tj. znosi, zmienia lubwprowadza odnośniki, których artykuł źródłowy został nieprawidłowo wstawiony waknumencie (patrz znacznikactiveModificationnarys. 1). 5). Możemy wyprowadzić takie niespójności, uruchamiając następujące polecenie Zapytanie Cypher:

To zapytanie jest i może być uruchamiane tylko w fazie aktualizacji KG i unikatowo dla nowych węzłów; W związku z tym właściwośćNewtextnie ma na nią wpływu.



DOPASUJ p=(l1:Prawo)-[:HAS\_ARTICLE]->(a1:Artykuł)-[:ZRZEKA SIĘ | POPRAWKI | PRZEDSTAWIA] WH-E>R(aE2N:AOrTical1e.t)e<x-[t:HCOASN\_TAARINTISctLoES]t-r(iln2:gL(aaw2,).numer) //i.e.,numer artykułu

I NIE a1.tekst ZAWIERA split(l2.id,"|")[1] // tj. numer prawa  
I NIE a1.tekst ZAWIERA toString(l2.publicationDate.year)

RETURN p wykrywanie wzorca wykresu  gdzie rodzajnik , źródło

wszelkie odniesienia tekstowe do a2w jego tekście (tj. numer artykułu, numer ustawy i rok wydania ustawy). Oznacza to, że źródło w tagu activeModification było niepoprawne. Zidentyfikowaliśmy około 4 tys. krawędzi dotkniętych tym problemem i poprawiliśmy je, wyszukując ten sam wzór w innych artykułach.

**3. Ponowna klasyfikacja krawędzi CITES into IS\_LEGAL\_BASIS\_OF.** Wykryliśmy przypadki, w których preambuła mogła zostać niezgodnie z prawem umieszczona w pierwszym artykule ustawy. W związku z tym krawędzie cytowania w jego tekście są identyfikowane przez nasz pipeline as CITES krawędzie, podczas gdy zamiast tego należy je traktować jako as IS\_LEGAL\_BASIS\_OF edges. Aby to naprawić, możemy uruchomić następujące zapytanie C:

MATCH (n:Article)-[:CITES]->(s:Article)-[:HAS\_ARTICLE]-(l:Law) WHERE toLower(n.text) CONTAINS "presidente della repubblica" AND toLower(n.text) CONTAINS "decreta" AND split(n.text,"decreta")[0] CONTAINS toString(l.publicationDate.year) AND split(n.text,"decreta")[0] CONTAINS toString(s.number) RETURN r UNION MATCH (n:Article)-[:CITES]->(s:Prawo) WHERE toLower(n.text) CONTAINS "presidente della repubblica" AND toLower(n.text) CONTAINS "decreta" AND split(n.text,"decreta")[0] CONTAINS toString(s.publicationDate.year) AND split(n.text,"decreta")[0] CONTAINS split(s.id,"|")[1] RETURN r , który wykorzystuje formuły preambuły (tj. obecność słów kluczowych, które muszą być użyte do wprowadzenia prawa) do wykrywania krawędzi , które są CITES, i których artykuły źródłowe  zawierają preambułę w swoim tekście. W szczególności korzystamy z obecności rytuału: "presidente della repubblica" (tj. prezydent Republiki) i "decreta" (tj. uchwała) jako heurystyki do identyfikacji i analizy artykułów, których tekst zawiera preambułę. Podczas gdy ten pierwszy jest rytuałem wprowadzającym, który charakteryzuje preambułę. Spośród wszystkich ustaw Republiki Włoskiej to ostatnie jest słowem rytualnym, które zamyka preambułę i wprowadza tekst ustawy. Za pomocą tych heurystyk wyprowadzamy krawędzie, które zostały błędnie oznaczone jako CITES one, i przekształcamy je w IS\_LEGAL\_BASIS\_OF. Za pomocą tego zapytania zidentyfikowaliśmy 6273 krawędzie, które zostały niepoprawnie wyprowadzone jako CITES edges, i Przerobiliśmy je into IS\_LEGAL\_BASIS\_OF ones. W sumie 3255 różnych artykułów zostało dotkniętych taką niespójnością.

**4. Artykuły, które stanowiły podstawę prawną innego prawa, ale zostały już uchylone w tym czasie.** Śledząc oznaczone krawędzie na wykresie właściwości, identyfikujemy artykuły, które były cytowane po ich uchyleniu, reprezentujące błędy w systemie legislacyjnym.

MATCH p=(l:Prawo)-[:HAS\_ARTICLE]->(a:Artykuł)-[:ZCIĄGA]-(a2:Artykuł) <-[:HAS\_ARTICLE]-(l2:Prawo)  
MATCH (a)-[:IS\_LEGAL\_BASIS\_OF]-(l3:Law) WHERE r.paragraph IS NULL AND l3.publicationDate > l2.publicationDate RETURN l3.id as LawWithError, a.id as CitedAbrogatedArt  
Wykryliśmy 145 błędów w cytowaniu, stosunkowo równomiernie rozłożonych na przestrzeni lat. Charakter takich błędów różni się od wcześniejszych niespójności: wynikały one z nieprawidłowego redagowania prawa w trakcie całej działalności legislacyjnej. Dlatego tutaj nie wprowadzamy poprawek do danych, a jedynie obserwujemy, jak nasz model danych może wychwycić – i potencjalnie zgłosić – takie niespójności.

#### 4.4. Wzbogacenie grafu o duże modele językowe

Podczas gdy schemat wykresu właściwości jest na krawędzi i może być modyfikowany i korygowany poprzez implementację technik opartych na regułach, jak opisano w poprzednich sekcjach, niektóre istotne i użyteczne właściwości są trudne do odzyskania za pomocą zwykłej heurystyki lub danych techniki integracji. Jest tak w przypadku atrybutu prawdomenowego, który określa ministerstwa zaangażowane w opracowywanie ustawy,

Tabela 2

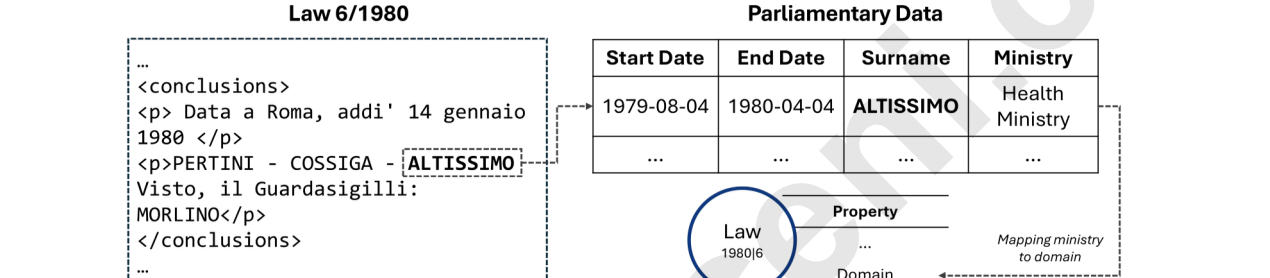
Przetwarzanie i zarządzanie informacją 62 (2025) 104082

Część par słowo kluczowe-domena, których używamy, aby wyprowadzić domenę od nazwy ministerstwa.

Zidentyfikowaliśmy listę 16 możliwych dziedzin: sprawy wewnętrzne, instytucje, rolnictwo, edukacja, gospodarka, komunikacja, prezydencja, transport, opieka zdrowotna, sprawy zagraniczne, wymiar sprawiedliwości, praca, obrona, administracja publiczna, sztuka i środowisko, sport i turystyka. Słowa kluczowe pozwalają nam jednoznacznie odwzorować Służba dla domeny.

Słowo kluczowe w nazwie służby

	Domena
lotnictwo	
zdrowie	transport
dplomacja	Sprawy
pensjonaty	zagraniczne
las	Służba zdrowia
...	Praca, rolnictwo
	...



Figa. 6. Przykład wyprowadzenia dziedziny prawa na podstawie podpisów złożonych we wnioskach dotyczących małego prawa. W tym celu integrujemy dane parlamentarne, aby uzyskać nazwę i rolę ministerstwa oraz skorzystać ze słownika (w prawym górnym rogu) w celu określenia domeny przypisanej do odpowiedniego węzła prawnego. Należy zauważyć, że pierwsze dwa podpisy są ignorowane, ponieważ zawsze należą one odpowiednio do Prezydenta Republiki i do Prezesa Rady Ministrów.

użyteczny sposób charakteryzowania praw do przeprowadzania analizy społecznej i ekonomicznej (Ciommoni, Morelli, & Paserman, 2022). Kolejną cechą o dużej użyteczności jest tematyka prawna, która umożliwia przeszukiwanie ustaw, artykułów i załączników odnoszących się do tej samej treści. Takie informacje są również istotne dla statystyk rocznych, a ich obecność wspiera automatyzację działań sprawozdawczych (Osservatorio sulla legislazione della Camera dei Deputati, 2023). Ponadto, mimo że przyjęty standard publikacji oparty na XML pozostaje niezwykle pomocny w gromadzeniu danych potrzebnych do budowy KG, doświadczyliśmy braków w zakresie prawidłowego wykorzystania standardu i publikowanych tekstów, zwłaszcza tytułów artykułów. Należy pamiętać, że tytuły są podstawowym elementem, który podsumowuje treść prawa lub artykułu, odgrywając kluczową rolę w opracowywaniu technik, które umożliwiają Retrieval Augmented Generation (RAG) wykresy lub potoki wyszukiwania informacji na szczycie legislacyjnego KG, takiego jak nasz. Aby zaradzić tym niedociągnięciom, wdrożyliśmy kilka kroków opartych na LLM, które pozwalają nam poprawić kompletność KG poprzez integrację i wzbogacenie węzłów o dodatkowe właściwości, przechwytywanie domen, tytułów i tematów.

4.4.1. Klasyfikator domen Ministerstwa

Każdy akt prawny może być klasyfikowany zgodnie z departamentami rządowymi lub ministerstwami, które są zaangażowane w prace nad projektem. Standard AKN nie określa żadnego znacznika, który przechwytywa takie informacje. Aby go uzyskać, wykorzystujemy fakt, że każde zatwierdzone prawo musi być podpisane przez jedno lub więcej ministerstw (reprezentujących swoje ministerstwo), informacja znajdująca się w wnioskach ustawy.

Pierwsze wyzwanie wynika z faktu, że mianowani ministrowie zmieniają się w czasie, głównie ze względu na zmianę rządów. W niektórych przypadkach podpisy pod konkluzjami są również ozdobione odpowiednią nazwą ministerstwa, np. Franco, Ministro dell'Economia e delle Finanze; Trybunał stwierdził jednak, że w większości aktów prawnych brakuje określenia nazwy ministerstwa, ponieważ tak nie jest.

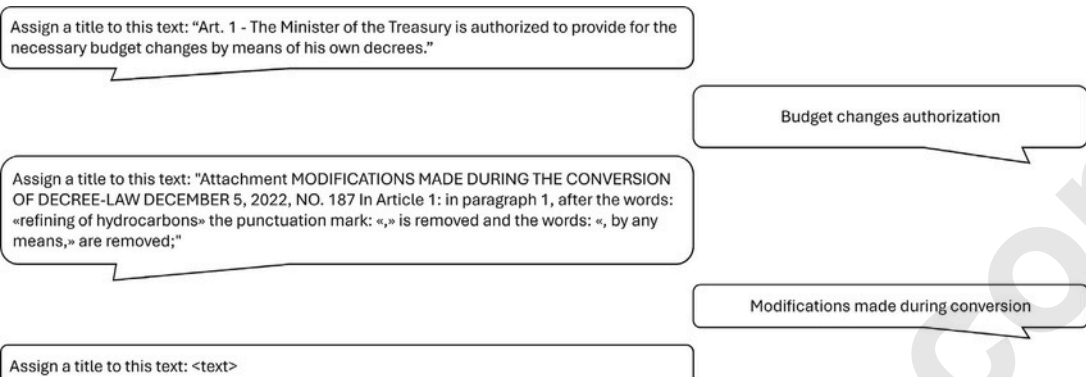
Pole obowiązkowe, ale tylko fakultatywne. W tym przypadku wykorzystujemy punkt końcowy danych parlamentarnych, aby uzyskać dane historyczne dotyczące ministrów i ich departamentów, co pozwala nam powiązać każde nazwisko z ministerstwem, do którego się odnosi (patrz rys. 1). 6). Spośród 74 tys. praw, które są w naszej bazie danych około 65 tys. ustaw wymagało tego działania związanego z powiązaniami.

Drugim wyzwaniem jest to, że nazwy służb również zmieniają się z czasem; W związku z tym muszą one być odpowiednio zgrupowane w oparciu o rzeczywiste domena. Na przykład ministerstwo skarbu zmieniło wiele nazw i kiedyś zostało podzielone na odrębne ministerstwa, a mianowicie Ministero del Tesoro, Ministero dell'Economia i Ministero delle Finanze, które muszą wywodzić się z tej samej dziedziny.

czyli gospodarka. Na przestrzeni dziejów Republiki Włoskiej udało nam się zidentyfikować 229 odrębnych nazw ministerstw (zob. Dodatek A.3), które powinny być pogrupowane według domeny, aby przeprowadzić analizę czasową. Na podstawie tego zestawu ręcznie stworzyliśmy zestaw słów kluczowych, które Połącz nazwę ministerstwa z domeną. Ogólnie rzecz biorąc, zidentyfikowaliśmy 16 domen i 107 słów kluczowych, które mogą łączyć służbę z domeną. Tabela 2 przedstawia kilka przykładów par słowo kluczowe-domena, których używamy do wykrywania domeny na podstawie słów kluczowych w jej obrębie. MiniStry NamE.

Klasyfikator oparty na. Biorąc pod uwagę, że połączenie integracji danych i podejścia opartego na słowach kluczowych pozwala na wzmocnienie ministerstwa





Rys. 7. Ręcznie wykonane przykłady ekstrakcji tytułów dostarczone do modelu LLM do nauki w kilku ujęciach, tj. instruowania modelu o zadaniu, które wykona za pomocą przykładów. Analizowany tekst jest następnie przekazywany w trzecim monicie.

Z biegiem czasu konsekwentnie buduje się rurociąg typu end-to-end, takie deterministyczne podejścia nie są idealne, ponieważ muszą być ręcznie ponownie przeanalizowane

za każdym razem, gdy pojawia się nowa nazwa ministerstwa, s. Totackletis, korzystamy z rozwiązania opartego na LLM, które zapewnia większą elastyczność niż słownictwo oparte na słowach kluczowych pokazane w Tabeli 2. W rzeczywistości nie będzie on wymagał konserwacji w przyszłości, a jednocześnie przewyższy problem niepoprawne formatowanie dokumentów opartych na AKN z powodu braku podpisów ministerstwa. Biorąc pod uwagę problem klasyfikacji wieloetykietowej, stworzyliśmy zestaw danych 45 tys. ustaw, łącząc tytuły prawne z ministerstwem/ministerstwami wywodzącymi się z omówionych wcześniej podejść deterministycznych. Następnie podzieliśmy zestaw danych na zestaw 90–10 pociągów i walidacji. Rozważyliśmy dwukierunkowy model reprezentacji kodera z transformatorów () (Devlin, Chang, Lee, Toutanova, 2018 &), który wykazał najnowocześniejsze wyniki klasyfikacji w różnych dziedzinach (Chen, Du, Allot, & Lu, 2022; Zahera, Elgendy, Jalota, Sherif, & Voorhees, 2019).

Dostosowaliśmy model za pomocą optymalizacji Adama (Loshchilov & Hutter, 2017) i użyliśmy funkcji aktywacji sigmoid, aby uwzględnić wieloklasowy charakter problemu, tj. więcej domen dla tego samego prawa. Wytrenowaliśmy model dla 5 epok na wielkości partii 32, ze współczynnikiem uczenia się 2e-5. Szkolenie trwało około 40 minut; Epoka 4 była najlepsza, jaką uzyskaliśmy, ze średnią stratą wyszkolenia wynoszącą 0,22 i dokładnością 90%, co uważamy za dość wysokie i akceptowalne, biorąc pod uwagę dodatkowe wyzwanie związane z językiem włoskim, w którym model (wyszkolony głównie na tekście angielskim Devlin i in., 2018) nie był ekspertem. Szczegółowe informacje na temat kroków dostrajania, z utratą trenowania i walidacji na każdym kroku, można znaleźć na rys. A.1 w aplikacji endix. Chociaż do tej pory dostępne są bardziej zaawansowane duże modele językowe, biorąc pod uwagę ogólnie dobre wyniki tuning, postanowiliśmy postawić na ten model. Wybór ten jest zgodny z naszymi wymogami w zakresie zrównoważonego rozwoju i repo ucją się, w przypadku gdy bierzemy również pod uwagę etyczne konsekwencje używania bardzo dużych modeli do prostych zadań (Gunasekar i in., 2023; Promień, 2023) jak klasyfikacja z wieloma etykietami. Nasz dopracowany model jest dostępny na Huggingface (Colombo, 2024a). W związku z tym model ten został wykorzystany do uzupełnienia KG poprzez wyprowadzenie domen dla wszystkich węzłów prawnych.

#### 4.4.2. Wyodrębnianie tytułów artykułów

W naszym schemacie przypisujemy tytuły do wszystkich węzłów związanych z prawem (tj. ustaw, artykułów i załączników). Informacje te są przydatne w przypadku budowania potoków wyszukiwania informacji (np. RAG), które muszą identyfikować odpowiedni tekst na podstawie danych wejściowych. Chociaż byłoby to możliwe do wykorzystania tekstu artykułu lub załącznika, potoki działałyby gorzej, gdyby zostały przedstawione z długimi tekstami (Wang et al., Huang, & Sheng, 2024).

Chociaż tytuł prawa jest zawsze dostępny i uchwycony przez metadane w tagu docTitle AKN, doświadczyliśmy znaczącego błędów w tagu nagłówek artykułów: Z 318 tys. artykułów tylko 108 tys. miało tag nagłówek. W przypadku załączników nie jest nawet używany znacznik tytułu. Do zajmię tym, wdrożyliśmy krok oparty na LLM, aby wyprowadzić tytuł z niestandardowego tekstu, tj. treści artykułu. Ponownie, tutaj możemy zastosować najnowocześniejsze, bardzo duże, wstępnie wytrenowane duże modele językowe, które są w stanie poradzić sobie zarówno z językiem włoskim, jak i języka i wracając bezpośrednio do tytułu prawnego. Zamiast tego, zgodnie z wymogami zrównoważonego rozwoju dla naszego rurociągu, zdecydowaliśmy się dla mniejszego modelu, który możemy dostroić, aby osiągnąć osiągi podobne do większych.

Mistral-7B. W swojej wersji opartej na czacie instruującym, Mistral 7B jest modelem językowym z 7 miliardami parametrów, który osiąga dobrą równowagę między dokładnością a wydajnością obliczeniową (Jiang i in., 2023). Jest znacznie mniejszy niż większe modele, takie jak GPT-4 czy Llama3-70B, przewyższając inne porównywalne duże modele językowe. Jest wydawany na licencji Apache 2.0, dzięki czemu użytkownicy mogą łatwy sposób na dostrojenie do konkretnych zadań. Jego mniejszy rozmiar, przy ogólnej dobrej wydajności i dostępności oprogramowania typu open source, sprawił, że Dobrej wybór dla naszego rurociągu.

Najpierw eksperymentowaliśmy bezpośrednio z dostępnym, wstępnie wytrenowanym modelem Mistral-7B do zadania ekstrakcji tytułów, dostarczając treść tekstową artykułów i podpowiadając przykładami, wykonując uczenie się w kilku ujęciach, podejście, w którym model się uczy uogólnianie na podstawie bardzo małej ilości danych dotyczących opadów deszczu, które są bezpośrednio dostarczane jako kontekstowe dane wejściowe do modelu (Parnami & Lee, 2022; Wang, Yao, Kwok, & Ni, 2020). Figa. Rozdział 7 ilustruje przykłady nauczania w kilku ujęciach. Mimo to model dał

Tabela 3

Różnice między wstępnie wyszkolonym modelem Mistral7B a naszą dostrojoną wersją w zadaniu wyodrębniania tytułów, gdy dostarczana jest ta sama zawartość systemu, tj. poprzez wykonywanie tego samego uczenia się w kilku ujęciach. Przykłady odnoszą się do wyodrębniania tytułów dla artykułów, których model nie widział podczas trenowania.

Tekst	Mistral-7B Wstępnie wyszkolony	Mistral-7B Dostrojony
	Artykuł 13	Postanowienia końcowe i czas trwania
Artykuł 13.1. Niniejszy Protokół wchodzi w życie w dniu uzgodnionym przez Strony w wyniku późniejszej wymiany not. 2. Protokół pozostaje w mocy przez 5 lat. Chyba że jedna ze Stron poinformuje z co najmniej sześciomiesięcznym wyprzedzeniem przed wygaśnięciem umowy o zamiarze nieprzedłużania umowy...	Zmiany w systemie koncesji na produkt oznaczony numerem 13 w tabeli A	Zmiany w systemie koncesji na paliwo w Słowenii
Art. 7.1. Ustrój ustanowiony ustawą z dnia 1 grudnia 1948 r., nr 1438, oraz Późniejsze zmiany, ograniczone do produktu oznaczonego numerem 13 w tabeli A, załączonym do ustawy z dnia 27 grudnia 1975 r., nr 700, zostały ponownie zdefiniowane zgodnie z przepisem dotyczącym jego artykułu, niezgodnie z artykułami 30 i 32 umowy o współpracy między Europejską Wspólnotą Gospodarczą a Republiką Słowenii...	Artykuł 33. organizacje wolontariackie	Finansowanie organizacji wolontariackich
Art. 33.1. Organizacje wolontariackie mogą zatrudniać pracowników lub korzystać z osób samozatrudnionych lub innych rodzajów pracy tylko w zakresie niezbędnym do ich regularnej działalności lub do Zakres wymagany do zakwalifikowania się...		

nierzadkowi. Wyniki Ponieważ po ręcznej kontroli tytuł często nie wyjaśniał treści artykułu lub był zbyt ogólny (zob. przykłady w tabeli 3). W rzadkich przypadkach (1%) spotkaliśmy się również z odpowiedziami w językach innych niż docelowy, tj. po włosku.

Dostrajanie zadania ekstrakcji Title. Aby poprawić wyniki uzyskane z LLM, dopracowaliśmy model Mistral-7B. Zbudowaliśmy duży zestaw danych treningowych, gromadząc artykuły, których nagłówki były dostępne (tj. odpowiedni tag AKN został wypełniony poprawnie); Pozwoliło nam to zgromadzić łącznie 108 tys. wysokiej jakości par tytuł-tekst, z pożądanym językiem (włoskim) i odnoszącymi się do zadania będące przedmiotem zainteresowania (ekstrakcja tytułów artykułów prawnych). Przyjęliśmy technikę Low-Rank Adaptation (LoRA), która pozwala na szybką adaptację LLM do określonych zadań poprzez zamrożenie oryginalnych, wstępnie wytrenowanych ciężarów i trenowanie tylko nowo wprowadzonych, które można trenować parametry (Hu i in., 2021). Wykazano, że Mistral 7B radzi sobie nieco lepiej w specjalizacji zadań (Zhao i in., 2024), dzięki czemu To idealne rozwiązanie dla naszego rurociągu.

Wyniki. Model został przeszkolony dla 5 epok o wielkości partii 4, 4-bitowej kwantyzacji przy użyciu bitsandbajtów i rangi LoRA z 64. Używamy strony Adamoptimizer, współczynnik uczenia się 0,004 i harmonogram tempa uczenia się cosinusa z 0,03 rozgrzewki ułamek. Użyliśmy procesora graficznego A100 z 40 GB pamięci, a najlepszy model zgłosił stratę oceny na poziomie 1,003 (dostępne na stronie PrzytulanieFace Colombo, 2024c). Szkolenie wymagane około 9 godzin. Szczegółowe informacje na temat ewolucji strat związanych z uczeniem i walidacją przedstawiono na rys. A.1 w dodatku IX.

Model finezyjny jest w naszym pipeline zamknięty jako dodatkowy komponent, który wzbogaca węzły artykułów o tytuły, gdy są niedostępne, tzn. zamknięte dokumenty AKN nie zgłaszają tytułów. Ponieważ charakter zadania i treść są bardzo podobne i Skorelowany, zastosowaliśmy również ten sam model do wyprowadzania tytułów załączników.

4.4.3. Wyodrębnianie tematów

Chociaż domeny są użyteczne jako ogólny typ klasyfikacji, ich zakres jest nadal zbyt szeroki w porównaniu z dużym zestawem Tematy które mogą regulować ustawy i artykuły. Zamiast tego tematy to słowa kluczowe, które w krótki sposób oddają treść tekstu (ustaw, artykułów lub załączników) i które mogą się stale zmieniać w czasie.

Chociaż już tytuł zawiera informacje o treści prawa, nie pomaga w przeprowadzaniu ustrukturyzowanych zapytań: Jego treść może się często różnić ze względu na niewielkie zmiany w tekście. Wyodrębnienie tematów wymaga (i) zidentyfikowania słów kluczowych charakteryzujących tekst oraz (ii) uogólnienia słów kluczowych na wykorzenione/częściej używane – i silnie powiązane – słowa (wykraczające poza specyfika rzeczywistego słowa kluczowego). Podczas gdy te pierwsze można osiągnąć za pomocą technik NLP lub systemów nienadzorowanych, te drugie mogą można osiągnąć jedynie poprzez zastosowanie najnowocześniejszych dużych modeli językowych zdolnych do uchwycenia pojęć, semantyki wykrywanie zwykłych słów kluczowych (Invernici, Bernasconi, & Ceri, 2024; Mu, Dong, Bontcheva, & Pieśń, 2024; Wu, Gong, Shou, Liang, & Jiang, 2023).

Co więcej, LLM mogą bezproblemowo dostosowywać i uwzględniać pojawiające się tematy (np. przepisy dotyczące sztucznej inteligencji), zapewniając, że Uchwyczone są nowe trendy. Weźmy na przykład pod uwagę dwa tytuły praw, które zawierają odpowiednio słowa szczepionki covid i wirus SARS-CoV-2, które można zidentyfikować jako słowa kluczowe. Optymalnym wspólnym tematem dla obu przypadków byłby covid-19, co pozwoliłoby nam na zapytanie o oba prawa za pomocą tego samego, bardziej ogólnego słowa kluczowego.

Dostrajanie do wyodrębniania tematów. Podobnie jak w przypadku zadania ekstrakcji tytułu, dopracowaliśmy kolejny Mistral-7B He, Huang, & Li, 2024), również w przypadku języka włoskiego. Wzięliśmy pod uwagę model Mixtral-8 22B i podaliśmy mu kilka przykładów do kilkustrzałowej nauki wraz z tytułem praw (patrz przykłady na rys. 1). 8). Utworzyliśmy zestaw danych przy użyciu tytułów prawnych, podpowiadając modelowi następujące instrukcje: Wyodrębnij tematy z tego tytułu: text i jako kontekst systemowy: Jesteś asystentem, który wyodrębnia tematy z tytułów. Każdy temat musi zawierać kilka słów. Zwraca tylko concise listę. < >



Figa. 8. Maniualnie opracwane przyklady ekstrakcji tematow dostarczone do LLM w celu nauki w kilku ujeiach. Proces jest podobny do tego, który jest używany w zadaniu ekstrakcji tytułu (patrzrys. 1). 7).

Tabela 4

Różnice między tematami wyodrębnionymi za pomocą wstępnie wyszkolonego Mixtral-8×22B i (dostrojonego) mniejszego Mistrala-7B. Podczas gdy niektóre tematy są powszechne, wstępnie wytrenowany model wykazuje mniejsze możliwości uogólniania, a niektóre tematy są nieistotne i powtarzają się, np. drugi wiersz.

Tytuł	Mixtral-8×22BPre-przeszkolony	Mistral-7BFuntrój wewnętrzny
Charakterystyka CapitalIncrease	Kapitał, wzrost, charakterystyka	Podwyższenie kapitału, finanse przedsiębiorstw
Wejście w życie	Data obowiązywania, aktywacja, aktywacja regulacyjna, aktywacja regularna, aktywacja	Data obowiązująca
Wzrost rocznego wkładu osobistego	Wkład osobisty, rok, wzrost	Składka osobista, emerytura
Uprawnienia kontrolne i poszukiwawcze przez siły policyjne	Policja, kontrola, uprawnienia	Policja, kontrola

Mimo że zwiększenie rozmiaru modelu złagodziło problemy w danych wyjściowych, doświadczyliśmy mieszanej jakości, często wyodrębniając nieistotne elementy takie tematy, jak rodzaj/data wydania aktu prawnego lub liczba aktów prawnych wymienionych w tytule. Poprzez losowe próbkowanie zestawu 4 tys. węzłów prawnych, zaobserwowano, że najczęstszymi tematami były "regulacja" (19 proc. przypadków), "prawo" (10 proc.), "ratyfikacja" (9 proc.) i "dekret" (7 proc.). Poza "ratyfikacją" zauważyliśmy, że takie tematy były niewystarczające do scharakteryzowania konkretnych aspektów ustawy. Aby temu zaradzić, przed przystąpieniem do dostrajania do naszego mniejszego modelu przeanalizowaliśmy zestaw danych par tytuł-temat i zastosowaliśmy heurystykę opartą na ciągach znaków do (i) zmniejszyć liczbę nieistotnych tematów w zbiorze danych i (ii) zharmonizować je, tj. wyprowadzić rdzeń w taki sposób, abyśmy mogli uwzględnić dla wyrażnej deklinacji tego samego słowa. W szczególności usunęliśmy ogólne najczęściej używane słowa kluczowe i użyliśmy typu/daty prawa node, aby usunąć tematy związane z tą funkcją. Następnie zdylematyzowaliśmy słowa, stosując spaCy (Honnibal, Montani, Landeghem, & Boyd, 2020), wielojęzyczne, najnowocześniejsze narzędzie do redukcji słowa do jego podstawowej lub podstawowej formy, znanej jako "lemma". Pozwoliło nam to poprawić jakość zestawu szkoleniowego, aby uzyskać bardziej znaczącą ekstrakcję tematów.

Szkolenie i wnioskowanie. Użyliśmy tej samej konfiguracji dostrajania modelu wyodrębniania tytułów i uzyskaliśmy ocenę Colombo, 2024d). Podobnie jak w przypadku modelu ekstrakcji tytułu, wymagane jest odbycie szkolenia około 9 godz. Szczegółowe informacje na temat ewolucji strat związanych z uczeniem i walidacją przedstawiono również na rys. 1. A.1 w dodatku.

Następnie użyliśmy dopracowanego modelu do wyprowadzenia tematów zarówno do artykułów, jak i załączników. Model ten jest regularnie wykorzystywany do wyprowadzania tematów również dla nowo publikowanych aktów prawnych. W tabeli 4 przedstawiono fragment wyników działań eksplenujących temat, które zostały za cznione za pomocą wstępnie wytrenowanego modelu Mixtral-8 22B oraz za pomocą Mistral-7B podopracowaniu go nad niektórymi tytułami artykułów, które znajdują się poza zbiorem treningowym.

#### 4.5. Uogólnienie na inne systemy legislacyjne

Chociaż wdrożyliśmy pełną linię produkcyjną dla włoskiego systemu legislacyjnego, elementy naszego procesu ETL mogą być powielane

w odniesieniu do innych aktów prawnych z odpowiednimi niewielkimi zmianami. Najważniejsze z nich zależą od podejścia do publikacji przyjętego przez każde z nich ustawodawstwo. Wiele krajów opracowuje obecnie interfejsy API lub inne interfejsy przyjazne dla maszyn, aby poprawić dostęp do danych. Stół 5 przedstawiono przegląd oficjalnych źródeł danych dotyczących prawodawstwa w sześciu głównych krajach, a także dostępność danych osobowych Standard publikacji nadający się do odczytu maszynowego, który można wykorzystać do zbudowania zasobu opartego na grafie, podobnego do tego opisanego tym artykule.

W chwili pisania tego tekstu Wielka Brytania znajduje się w fazie eksperymentalnej opracowywania własnego API, który przyjmuje międzynarodowy standard AKN (jak omówiono w sekcji 3.1); Mapowanie praw do obiektów grafów naszego rurociągu wymagałoby jedynie niewielkich adaptacji, tj. modyfikacji konwencji tagów AKNnaszczelbu krajowym tradycji legislacyjnej Wielkiej Brytanii. W USA, API dla przyjęcia ustawodawstwa (USLibraryofCongress, 2024) został niedawno opracowany; standard AKN nie został przyjęty, ale przepisy są opisane w formacie opartym na XML, co

Tabela 5

Przetwarzanie i zarządzanie informacją 62 (2025) 104082

Przegląd źródeł danych legislacyjnych i ich dostępności w wielu krajach.

Kraj	Źródło danych	Dostępność API	Format publikacji
Stany	Biblioteka Kongresu	Format eksperymentalna	Specyficzny dla danego kraju kod XML
Zjednoczone	legislation.gov.uk	Nie	Tylko krajowy XML AKN
Wielka Brytania	Bundesgesetzblatt	Tak	PDF Tylko krajowy XML AKN
Niemcy	Official del Estado Y		
Francja	Fedlex	Yeess (punkt końcowy SPARQL)	
Hiszpania			
Szwajcaria			

Są to węzły charakteryzujące węzły związane z prawem (Ustawy, Artykuły i Załączniki) z wzajemnymi powiązaniami odniesień i węzłów "częściowych". Rząd Węzły i legislatura są pomijane, ponieważ łączą one tylko węzły prawne.

Węzeł docelowy			
Węzeł wykresu			
Prawo	Prawo 61.989JestPodstawa prawna	Artykuł	Załącznik
Artykuł	z dnia 44.954JestPodstawa prawna	318.286Zawiera artykuł	126.674Ma załącznik.
Węzeł źródłowy	z	70.990 Poprawki	3.561 Zmiany
	7.009 Zmienia 532	3800223 Przedstawia	2.393 Uchyla
	uchylenia	95.214 Cytaty	4.922 Miasta
	78.256 Miasta	19.295 Miasta	1.173 Cytaty
Załącznik	1131sPodstawa prawna		
	29.357 Miasta		

Stanowi podstawę do mapowania praw i ich zawartości do naszych obiektów schematu grafu. To samo dotyczy Niemiec i Hiszpanii, które publikują przepisy w ich własnym formacie XML; W Niemczech możliwość przyjęcia standardu AKN jest obecnie badana (Flatt, Langner, & Leps, 2022). We Francji Légifrance oferuje dostęp do wszystkich aktów prawnych wydanych na szczeblu krajowym. Dostępne są jednak tylko pliki PDF, co wymaga bardziej wymagającego etapu identyfikacji struktury każdego prawa przed przekształceniem danych w wykres.

Po wyprowadzeniu wykresu, komponent wykrywania błędów oparty na grafie naszego potoku (omówiony w sekcji 4.3.2) nie wymaga żadnej adaptacji ze względu na obecność ujednoliconego abstrakcyjnego schematu Grafu wiedzy. Ostatnim etapem procesu produkcyjnego są etapy ulepszenia LLM KG, które mogą być powielane w innych systemach poprzez (i) szkolenie w zakresie dostępnych danych specyficznych dla danego kraju

w celu zintegrowania brakujących informacji lub, w najgorszym przypadku, (ii) przyjęcia większych LLM, które z uczeniem się w niewielu ujęciach mogą nadal wykonywać zadania ekstrakcji informacji z tekstu (Wadhwa i in., 2023; Xu i in., 2023), biorąc również pod uwagę wiele różnych języków (OpenAI, 2024), bez konieczności dostrajania, choć kosztem zaszkodzenia zrównoważonemu rurociagowi.

Na koniec warto wspomnieć o systemie szwajcarskim, ponieważ (i) niedawno przyjęto w nim standard AKN oraz (ii) publikuje ustawy, między innymi, w języku włoskim, ze względu na wielojęzyczny charakter kraju. W związku z tym, jeśli API zostanie udostępnione, potok będzie miał zastosowanie bez niewielkich lub żadnych dostosowań, w tym LLM, które dostosowaliśmy do języka włoskiego.

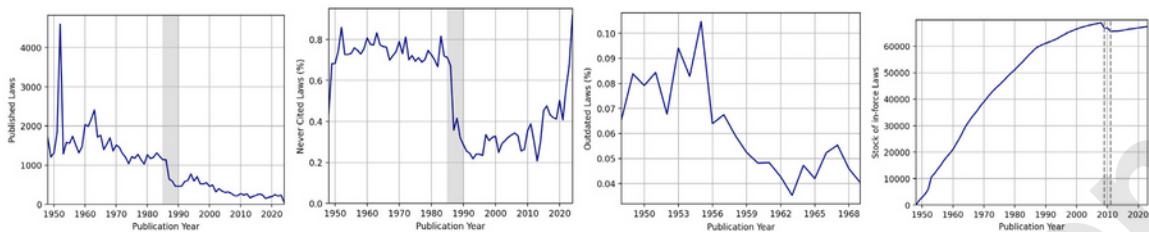
5. Wiedza eksploracja za pomocą zapytań graficznych

W niniejszej sekcji omówiono KG wynikającą z naszego toku ETL, ilustrując główne cechy włoskiego prawodawstwa. W tym celu proponujemy również wprowadzenie odrębnych typów zapytań graficznych, które są ułatwione przez nasz model danych. Odpowiadają one typowym statystykom, które są (ręcznie) obliczane przez urzędy statystyczne do celów sprawozdawczości rocznej (np. Osservatorio sull' legislazione della Camera dei Deputati, 2023) lub do opracowania interaktywnych aplikacji do monitorowania systemu legislacyjnego (Colombo, 2024). Wykorzystując Pokazaliśmy, jak takie działania mogą być wspierane przez model danych oraz przez wdrożony przez nas pipeline ETL. W tabeli 6 przedstawiono główne wymiary wykresu właściwości. Zamodelowaliśmy ponad 500 tys. węzłów i ponad 1 milion krawędzi, w tym piasek referencyjny Krawędzie Expressing Parthood.

Cechy czasowe ustawodawstwa włoskiego. W latach 80. i 90. możemy zaobserwować radykalną zmianę w sposobie tworzenia prawa: Roczna liczba aktów prawnych znacznie się zmniejszyła, długość każdego z nich wzrosła, a na jeden przypada więcej artykułów i załączników prawo. Chociaż analiza przyczyn tej zmiany wykracza poza zakres opisany w artykule, konieczne jest uwzględnienie tej tendencji podczas przeprowadzania zapytań, ponieważ może to znacząco wpłynąć na wyniki. Na przykład proste zapytanie mające na celu zidentyfikowanie instytucji rządowych To spowodowało, że te prawa mogą być wypaczone przez ten trend. Wreszcie, na szczególną uwagę zasługuje tzw. Decretia Semplificazione. Dekrety te zawierają liczne przepisy uchylające, które mają na celu usunięcie i oczyszczenie krajobrazu legislacyjnego z przestarzałych aktów. Podczas wykonywania zapytań dotyczących uchyleń krawędzi użytkownicy mogą chcieć odfiltrować takie prawa, w tym prawa z lat 2008/112, 2010/66 oraz 2010/212. Ponadto można opracować zestaw zapytań, aby uzyskać ogólne informacje na temat ewolucji w czasie włoskiego prawodawstwa system poprzez filtrowanie i agregowanie atrybutów opartych na kryteriach, takich jak rok, legislature, orgovernment.

Na przykład, rozważ następujące zapytania, których określenia szczyfrow znajdują się w dodatku A.7, a wyniki są przedstawione na rys. 9:

- Q1 Prawa opublikowane w roku publikacji, co oznacza liczbę praw w oparciu o rok publikacji spadkobiercy.
- Q2 Akty prawne, które nigdy nie były cytowane po publikacji, odnoszące się do ustaw, do których nie odniesiono się w żadnej preambule, ani nie otrzymały żadnych poprawek, uchyleń, wstępów lub innych cytatów po dacie ich publikacji.



(a) Published Laws (Q1)

(b) Laws Never Cited (Q2)

(c) Outdated Laws (Q3)

(d) Stock of in-force Laws (Q4)

Rys. 9. Panel (a) przedstawia wyniki **pierwszego kwartału**, w którym wykrywana jest liczba opublikowanych przepisów w ciągu roku. Początkowy szczyt, jaki można zaobserwować, spowodowany jest przejściem od monarchii do republiki, co wymagało wielu zmian w ustawodawstwie. Panel (b) przedstawia wynik **Q2** (część ustaw, które nie były cytowane po publikacji), a także podkreśla, że wiele ostatnio opublikowanych ustaw nie jest jeszcze cytowanych. W panelu (c) wykreśliśmy **Q3**; tutaj rozważaliśmy datę graniczną 1970. Następnie użyliśmy  $D = 1992$  – początek "drugiej republiki" we włoskiej polityce, aby obliczyć ułamek outowanych praw. Panel (d) przedstawia wyniki **IV kwartału**, w którym wyróżniono spadki odpowiadające dekretem upraszczającym (odpowiednio w 2008 i 2010 r.).

	Conte I (461 Days)	Conte II (527 Days)	Draghi I (616 Days)	Meloni I (608 Days)
Agriculture	0.03	0.06	0.04	0.09
Arts and Environment	0.15	0.1	0.15	0.08
Defense	0.04	0.07	0.09	0.15
Economy	0.43	0.49	0.44	0.5
Foreign affairs	0.26	0.31	0.27	0.21
Justice	0.26	0.28	0.24	0.25
Domestic Affairs	0.14	0.12	0.13	0.27
Institutions	0.12	0.18	0	0.18
Education	0.08	0.07	0.06	0.11
Labor	0.11	0.05	0.1	0.14
Presidency	0.18	0.11	0.16	0.13
Public Administration	0.13	0.08	0.11	0.12
Healthcare	0.07	0.16	0.15	0.15
Sport and Tourism	0	0.03	0.02	0.08
Transportation	0.11	0.1	0.11	0.15

Rys.10. Mapa cieplna z latamiumber ustaw podpisanych przez ministerstwa czterech ostatnich rządów od 2018 r. do 2024 r. (w nawiasach, ich czas trwania w dni), zgodnie z obliczeniami uzyskanymi za pomocą zapytania graficznego **Q5**, ilustrującym zakres zainteresowania każdego z sektorów rządowych. Ponieważ prawa mogą być wielodomenowe, tj. podpisane przez więcej niż jedno ministerstwo, jest wyższa niż jednostka. Należy zauważyć, że niektóre wartości mogą być zerowe, gdy rząd nie wyznacza stanowiska ministra dla określonej domeny.

**Q3** Przestarzałe prawa, zdefiniowane prawa, które przestały być cytowane po pewnym czasie. Aby je zidentyfikować, najpierw wybieramy datę graniczną. Następnie, wybierając kolejną datę  $D$ , na przykład datę oznaczającą ważne wydarzenie polityczne, wyodrębniamy zestaw praw cytowanych w każdym akcie prawnym opublikowanym po  $D$ . Pomaga nam to zidentyfikować akty prawne, które nie zostały przytoczone po  $D$ . **Q4** Stock of in-force laws, odnoszące się do całkowitej liczby ustaw obowiązujących w danym dniu. Wiąże się to z określeniem, które przepisy do tego czasu nie zostały uchylone. W kontekście prawa włoskiego oficjalne źródło danych Normattiva umożliwia użytkownikom przeglądanie czy ustawa weszła w życie, czy została uchylona w określonym momencie. Wymaga to jednak odzyskania wszystkich praw z pożądanym Data wybrana. Alternatywnie, wykorzystując krawędzie uchylecia na wykresie wiedzy, możemy określić, które prawa zostały uchylone — kiedy wszystkie ich artykuły zostały uchylone lub gdy całe prawo zostało bezpośrednio uchylone.

Odkrywanie prawa Domeny i Tematy. Wykorzystując właściwości dodatkowych węzłów domen i tematów, możliwe jest również **Osservatorio sulla legislazione della Camera dei Deputati (2023)**. Sprawozdania takie przedstawiają ogółowi społeczeństwa dane statystyczne, podsumowując tendencje i cechy charakteryzujące niektóre organy ustawodawcze. Podobny nbeOsiągnięteByqueringourgraph. Na przykład rozważmy proponowany **dodatku A.7**:





Figa. 11. Wizualizacja chmury słów wyników zapytania Q7, wykonanego dla ostatnich trzech rządów. Działalność dwóch pierwszych rządów była charakteryzowała się głównie prawodawstwem związanym z COVID-19, podczas gdy ostatni rząd skupiał się na rzecz pracy i tematów związanych z gospodarką (krajowe plany odbudowy i zwiększania odporności odnoszą się do krajowego planu naprawy gospodarczej finansowanego przez UE).

Top40 odpowiedzi z odpowiednią kardynalnością wyników zapytania Q7, wykonanego dla ostatnich trzech rządów odpowiedzialnych.

a) Rząd Hrabiego II		b) Rząd Draghiego I		c) Rząd Meloni I	
Temat	Hrabia	Temat	Hrabia	Temat	Hrabia
COVID-19	7 8 3	COVID-19 Praca	8 4 3	bezpieczeństwo	3 1 5
Przedsiębiorstwa	5 1 0	Zdrowie Przed-	6 0 9	pracy NRRP	2 7 7
Zdrowie	4 9 7	siębiorstwa ...	5 6 1	gospodarka ...	2 2 1
Pracy	4 4 0		3 9 4		1 8 3
...	...		...		...

Q5 Zaangażowanie ministerstw w produkcję legislacyjną. Korzystając z właściwości domeny , możemy uzyskać statystyki dotyczące charakteru, aktualności, a nawet cech językowych praw wydanych przez to samo ministerstwo pod różnymi rządami. Natyle możemy obliczyć częstotliwość podpisywania ustaw przez ministerstwa dla ostatnich czterech włoskich rządów. Narzys. 10 proponujemy wizualną ilustrację wyników zapytania. Q6Prawodawstwo UE wdrożone w systemie włoskim. Istotnym elementem sprawozdań rocznych jestto, że scharakteryzować źródła każdego aktu prawnego, pod względem tego, czy jest on ustawodawstwem krajowym, czy też wynika z wdrożenia Unia Europejska ustawodawstwo. W naszym schemacie możemy bezpośrednio sprawdzić takie wyniki, wykorzystując tematy przypisane do węzłów prawnych, wyszukując "rozporządzenie UE" lub "dyrektywa UE", dwa rodzaje prawodawstwa UE. Na przykład w sprawozdaniu rocznym za lata 2022–2023 (OsservatoriosullalegislazionedellaCameradeiDeputati,2023)policzone przez pierwsze siedem miesięcy XIX ustawodawca 13 praw, które wdrożonego prawodawstwa UE. Otrzymujemy 12 aktów prawnych, w których brakuje tylko jednego wpisu, który zidentyfikowaliśmy jako akt prawny zawierający modyfikacje poprzednich przepisów wykonawczych UE (tj. ustawę nr 54/2023), a zatem nie jest to akt wykonawczy bezpośrednio.

P7Tematy interwencji rządowych, tj. wyznaczenie tematów charakteryzujących podejście rządu do zmiany lub usunięcia poprzedniego prawodawstwa. Możemy prześledzić tematy wszystkich ustaw, które są zmieniane/uchylane przez specjalny wernment stworzyć prostą wizualizację obszarów, w których rząd był najbardziej aktywny (patrzrys.11itabela 7).

6. Dyskusja na temat jakości KG

W tej sekcji analizujemy jakość Wykresu Wiedzy zbudowanego poprzez rozwój naszego potoku. Biorąc pod uwagę dziedzinę zainteresowania, przy ocenie jakości czterech KG (zgodnie z badaniamiWangetal, 2021 r.;Xue&Zou, 2023 r.):

1. Dokładność, mierzenie, czy KG prawidłowo odzwierciedla przedstawione fakty. Z uwagi na to, że nie istnieje publicznie dostępny dokument KG włoskiego prawodawstwa o takim samym stopniu szczegółowości i terminowości, dokładność można zmierzyć jedynie poprzez porównanie eksperymentuje bezpośrednio z prawami w ich nieustrukturyzowanej wersji. Jako reprezentatywny scenariusz pomiaru dokładności przetestowaliśmy, jak skutecznie nasz KG radzi sobie z wymiarem czasowym. W tym celu skupiliśmy się na konkretnym mimestamp (2023-12-31) i zebrał wszystkie włoskie prawa w zaktualizowanej wersji, tj. z tekstem obowiązującym w tym czasie. Przeanalizowaliśmy tekst tych praw i policzyliśmy te uchylone, które charakteryzują się tekstem u abrogatina wzór w ich tekście. Następnie porównaliśmy takie wskaźniki z tym, co można wywnioskować z naszej reprezentacji KG (patrz Q4), która rozpatruje prawa tylko w ich pierwotnej wersji i pozwala nam wnioskować o cechach czasowych, takich jak uchylenie praw. To dało nam

Ponieważ nie jest dostępny żaden interfejs API, skrobienie zestawu danych trwało około trzech dni.

Tabela 8  
Wzbogacenia i stan węzłów KG oraz ich właściwości na różnych etapach potoku ETL.

	Mapowanie AKN	Integracja danych i wykrywanie błędów	Ulepszenie LLM
Węzeł	74 tys. utworzonych węzłów, nie	97% domen pobranych za pomocą danych	
Prawo	wyodrębniono domenę, nie	parlamentarnych	Ekstrakcja domen (3%)
Artykuł	wyodrębniono temat	Poprawiono 3 tys. węzłów artykułów z	Ekstrakcja tematów Mistral (100%)
Zajęcie	Utworzono 310 tys. węzłów	błędami podstawy prawnej	Ekstrakcja tytułu Mistral (66%)
Zajęcie	34% Wyodrębniono tytuły, nie		Ekstrakcja tematu Mistral (100%)
rzędu	wyodrębniono żadnego tematu		Ekstrakcja tytułów Mistral (100%)
Ustawodawca	216 tys. Utworzono węzły, Nie		Ekstrakcja tematów Mistral (100%)
	wyodrębniono żadnych tytułów,		
	nie wyodrębniono tematu.		
Utworzono 68 węzłów			
Utworzono 20 węzłów			

Tabela 9  
Wzbogacanie krawędzi KG na różnych etapach procesu ETL.

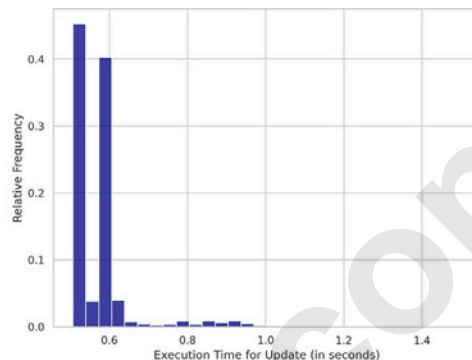
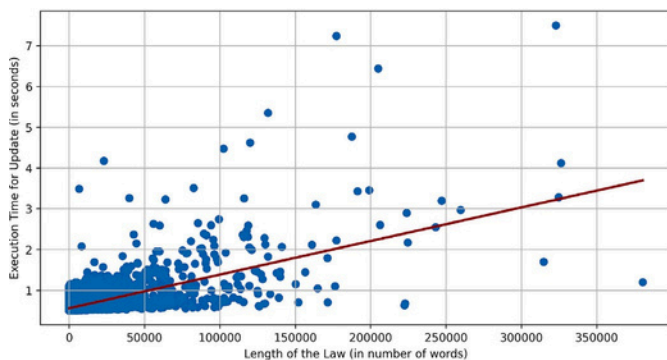
Brzeg	AKN mapping (mapowanie AKN)	Integracja danych i wykrywanie błędów
Ma artykuł	310 tys. krawędzi utworzono	
ma zajęcie	Krawędzie 216 tys. utworzono	
jest Podstawa	Krawędzie 100 tys. utworzono	6,3 tys. cytuje poprawki 145 zgłoszonych
prawna zmian	Krawędzie 82 tys. utworzono	nieśpójności
uchyla wprowadza	65 tys. krawędzi utworzono	Wykryto 3 K Nieprawidłowe źródło
Cites	Krawędzie 5 k utworzono	Wykryto 1 K Wykryto nieprawidłowe źródło
pod rządami	Utworzono 235 tys. krawędzi	
władzy		
ustawodawczej, po		74 K utworzonych krawędzi
której następuje		74 K utworzonych krawędzi
		67 utworzonych krawędzi

stosunek liczby prawdziwych uchyleń, które przechwytuje nasze zapytanie do KG, równy 0,98. W szczególności, 109 z 6283 "prawdziwych uchylonych praw" nie jest wnioskowanych przez nasz KG. Ręczna inspekcja takich praw sugerowała, że większość krawędzi w znacznikach activeModification była całkowicie nieobecna. W przyszłych pracach zastanowimy się, jak radzić sobie z takimi brakującymi krawędziami, aby osiągnąć idealną dokładność. Niemniej jednak dokładność pozostaje wysoka, co pokazuje, jak możemy bezproblemowo uchwycić wymiar czasowy za pomocą zapytań grafowych.

2. Completeness, coodnosi się do stopnia, w jakim wszystkie wymagane informacje są obecne w danych wyjściowych potoku ETL. Tabele 8 i 9 podsumowują, w jaki sposób nasz potok wzbogaca i ulepsza Graf wiedzy na każdym kroku. Na końcu naszego pipeline'u wszystkie węzły są ponownie wzbogacane o wszystkie właściwości, które zdefiniowaliśmy w schemacie, albo poprzez integrację danych (dla domen), albo poprzez LLMs. Jeśli chodzi o krawędzie, to chociaż nie możemy zmierzyć rzeczywistej ilości brakujących – poza przeprowadzaniem eksperymentów takich jak ten określający dokładność krawędzi znoszących – udaje nam się korygować i wykrywać różnego rodzaju błędy, co prowadzi do bardziej kompletnego wykresu.

3. Spójność, definiowana jako stopień, w jakim wiedza KG nie jest sprzeczna sama ze sobą, tj. nie określa sprzeczności w danych dotyczących konkretnej reprezentacji wiedzy. W sekcji 4.3.2 pokazaliśmy, w jaki sposób wykorzystujemy zapytania grafowe, które (i) wykrywają błędy w węzłach i krawędziach za pomocą heurystyki, co pozwala nam na podjęcie działań korygujących, oraz (ii) zgłaszają istotne niespójności w działalności legislacyjnej, takie jak w przypadku artykułów, które były cytowane, ale zostały uchylone. W ujęciu względnym te pierwsze stanowiły tylko 2% wszystkich krawędzi odniesienia (w KG o ogólnej wysokiej spójności, również biorąc pod uwagę nasze mechanizmy korekcyjne).

4. Aktualność, tj. stopień, w jakim wiedza jest aktualna. Rurociąg ETL może być aktualizowany codziennie, stale aktualizując KG o nowe informacje. W związku z tym uważamy, że nasze plany produkcyjne charakteryzują się wysoką terminowością w dziedzinie systemów legislacyjnych. Proponowany schemat PG pozwala nam skupić się tylko na nowych przepisach i pobrać tylko ich oryginalną wersję, ponieważ wyczyniłyśmy tempo naszej rejestracji za pomocą zapytań grafowych, unikając w ten sposób używania wersji (tj. nowego węzła) dla każdej "aktualizacji" legislacyjnej. Na Fot. 12, przetestowaliśmy czas aktualizacji wymagany do dodania każdego prawa do KG i porównaliśmy go z długością prawa, którą zapewniono w liczba słów. Uruchamiamy nasz pipeline na naszej dedykowanej maszynie serwerowej z 56-rdzeniowym procesorem Intel E5-2660 v4 i 384 GB pamięci RAM. Należy pamiętać, że wydajność może się różnić w zależności od wielu parametrów, takich jak liczba cytatów, artykułów i/lub załączników oraz możliwość aktywacji LLM, jeśli jest to wymagane do uzupełnienia właściwości. Wybraliśmy długość ustawy jako ogólny wskaźnik reprezentatywny. We wszystkich przypadkach czas realizacji jest bardzo krótki; codzienna aktualizacja KG zazwyczaj obejmuje co najwyżej dwie lub trzy ustawy. Niemniej jednak obliczyliśmy również czas wykonania wymagany do odtworzenia od podstaw całego wykresu włoskiego ustawodawstwa od 1948 r. i ogólnie rzecz biorąc, wymaga to około 12 godzin, wliczając w to połączenia LLM.



W tym artykule przedstawiliśmy kompleksowy pipeline, zaczynając od niedawno przyjętego międzynarodowego standardu Akoma Ntoso, a kończąc na kontrolowanym wykorzystaniu LLM, konstruuje wysokiej jakości wykres wiedzy na temat włoskiego ustawodawstwa. Zaproponowaliśmy schemat grafu oparty na schemacie na paradygmacie grafu właściwości i niedawno ustandaryzowanym języku Graph Query Language, który został zaprojektowany tak, aby efektywnie reprezentować systemów legislacyjnych oraz w celu uchwycenia złożonych aspektów związanych z prawem, takich jak wymiar czasowy. Zgodnie z naszą najlepszą wiedzą, ETL pipeline jest pierwszym, który łączy w sobie niedawno przyjęty standard XML do odczytu maszynowego do tworzenia grafu, Akoma Ntoso i wykres właściwości, podejście zgodne z CQL. Biorąc pod uwagę międzynarodowe przyjęcie tego standardu, uważamy, że ten sam rurociąg może być łatwo przystosowany do użytku w innych systemach z minimalnymi dostosowaniami atrybutów specyficznych dla danego kraju. W związku z tym uzyskane informacje z różnych wykresów legislacyjnych mogą być potencjalnie porównywalne, co ułatwi uzyskanie dalszych informacji. Rozszerzyliśmy kompletność wykresu, wykorzystując LLM do wyprowadzania lub uzupełniania właściwości węzłów. W tym celu skupiliśmy się również na dostosowaniu wystarczająco lekkie modele, które pozwalają nam zmniejszyć wymagania obliczeniowe i zrealizować zadania ekstrakcji informacji porównywalnie dobrze z najnowocześniejszymi modelami językowymi. Zbadaliśmy również, w jaki sposób ten model i jego ulepszenie pozwoliły nam uzyskać wgląd w system legislacyjny, pozwalający na automatyzację ręcznie wyliczanych statystyk i rozbudowę ich do nowych, wartościowych Metryki. Na koniec omówiliśmy ogólną wysoką jakość KG w wielu wymiarach. W szczególności wykazaliśmy, że dokładność i wydajność w uchwyceniu wymiaru czasowego oraz wykazała się ogólną wysoką spójnością i kompletnością, również dzięki integracji komponentów LLM. Nasz pipeline rozwiązuje jeden wystarczająco złożony przykład (przypadek włoski) w prawodawstwie pole i pokazuje pomyślne wyniki; W ten sposób dążymy do uutorowania drogi do łatwego w obsłudze zarządzania wiedzą na temat systemów legislacyjnych, być może umożliwiając również porównania międzysystemowe.

**5. Wiarygodność**, tj. stopień, w jakim informacje są akceptowane jako poprawne i wiarygodne. W naszej konstrukcji KG korzystamy z oficjalnych źródeł danych, takich jak Dziennik Urzędowy i Parlament Włoch. Omówiliśmy już, w jaki sposób nasz model reprezentacji KG może przyczynić się do poprawy jakości oryginalnych danych poprzez zgłaszanie niespójności do źródła danych. Ponadto, nasze zastosowanie LLM jest specyficzne dla zadań, a dzięki zastosowaniu tylko tych typu open source, ich wyniki i wydajność mogą być publicznie kontrolowane.

**6. Interoperacyjność**, tj. stopień, w jakim format i struktura informacji są zgodne z danymi pochodzącymi z innych źródeł. Konstrukcja KG dla włoskiego ustawodawstwa opiera się na wdrożeniu międzynarodowego standardu (AKN) w systemie krajowym, co oznacza, że w krajach, które przyjęły ten sam standard, rurociąg ETL jest w pełni możliwy do powielenia. z dostosowaniami specyficznymi dla danego kraju dla różnych krajowych źródeł danych.

## 7. Wnioski

W tym artykule przedstawiliśmy kompleksowy pipeline, zaczynając od niedawno przyjętego międzynarodowego standardu Akoma Ntoso, a kończąc na kontrolowanym wykorzystaniu LLM, konstruuje wysokiej jakości wykres wiedzy na temat włoskiego ustawodawstwa. Zaproponowaliśmy schemat grafu oparty na schemacie na paradygmacie grafu właściwości i niedawno ustandaryzowanym języku Graph Query Language, który został zaprojektowany tak, aby efektywnie reprezentować systemów legislacyjnych oraz w celu uchwycenia złożonych aspektów związanych z prawem, takich jak wymiar czasowy. Zgodnie z naszą najlepszą wiedzą, ETL pipeline jest pierwszym, który łączy w sobie niedawno przyjęty standard XML do odczytu maszynowego do tworzenia grafu, Akoma Ntoso i wykres właściwości, podejście zgodne z CQL. Biorąc pod uwagę międzynarodowe przyjęcie tego standardu, uważamy, że ten sam rurociąg może być łatwo przystosowany do użytku w innych systemach z minimalnymi dostosowaniami atrybutów specyficznych dla danego kraju. W związku z tym uzyskane informacje z różnych wykresów legislacyjnych mogą być potencjalnie porównywalne, co ułatwi uzyskanie dalszych informacji. Rozszerzyliśmy kompletność wykresu, wykorzystując LLM do wyprowadzania lub uzupełniania właściwości węzłów. W tym celu skupiliśmy się również na dostosowaniu wystarczająco lekkie modele, które pozwalają nam zmniejszyć wymagania obliczeniowe i zrealizować zadania ekstrakcji informacji porównywalnie dobrze z najnowocześniejszymi modelami językowymi. Zbadaliśmy również, w jaki sposób ten model i jego ulepszenie pozwoliły nam uzyskać wgląd w system legislacyjny, pozwalający na automatyzację ręcznie wyliczanych statystyk i rozbudowę ich do nowych, wartościowych Metryki. Na koniec omówiliśmy ogólną wysoką jakość KG w wielu wymiarach. W szczególności wykazaliśmy, że dokładność i wydajność w uchwyceniu wymiaru czasowego oraz wykazała się ogólną wysoką spójnością i kompletnością, również dzięki integracji komponentów LLM. Nasz pipeline rozwiązuje jeden wystarczająco złożony przykład (przypadek włoski) w prawodawstwie pole i pokazuje pomyślne wyniki; W ten sposób dążymy do uutorowania drogi do łatwego w obsłudze zarządzania wiedzą na temat systemów legislacyjnych, być może umożliwiając również porównania międzysystemowe.

## Oświadczenie o wkładzie autorskim CRediT

Andrea Colombo: Pisanie – redagowanie recenzji &, Pisanie – oryginalny szkic, Wizualizacja, Walidacja, Metodologia, Selekcja danych, Konceptualizacja. Anna Bernasconi: Pisanie – redakcja recenzji &, superwizja, konceptualizacja. Stefano Ceri: Pisanie – recenzja & montaż, nadzór.

## Oświadczenie o sprzecznych interesach

Autorzy oświadczają, że nie znają konkurujących ze sobą interesów finansowych lub powiązań osobistych, które mogłyby się pojawić w celu wywarcia wpływu na prace przedstawione w niniejszym dokumencie.

### A.1. Schemat grafu właściwości formalnych

**Formalna definicja schematu wykresu właściwości (Angles i in., 2023) zaproponowana dla KG to:**

```
CREATE GRAPH TYPE lawsGraphType STRICT{
(lawType: Law {id STRING, title STRING, typeLaw STRING, publicationDate DATE, inForceDate
    DATA, numArt INT,numAttach INT, domena LISTA, temat LIST}),
(articleType: Artykuł {id STRING, tytuł STRING, liczba INT, tekst STRING, temat LIST}),
(attachmentType: Załącznik {id STRING, tytuł STRING, typ STRING, tekst STRING, temat LIST}),
(legislatureType: Legislature {nazwa STRING, startDate DATE, endDate DATE}), (governmentType:
Government {nazwa STRING, startDate DATE, endDate DATE}), (:lawType)-[hasArticleType:
has_article]->(:articleType),
(:lawType)-[hasAttachmentType: has_attachment]->(:attachmentType), (:lawType)-
[underGovernmentType: under_government]->(:governmentType), (:lawType)-
[underLegislatureType: under_legislature]->(:legislatureType), (:governmentType)-
[succeededByType: succeeded_by]->(:governmentType), (:lawType)-[referenceType: is_legal_basis_of
(paragraph LIST, weight INT)]->(:lawType), (:articleType)-[referenceType: is_legal_basis_of| cites
{paragraph LIST,
waga INT}]->(:typprawa),
(:articleType)-[referenceType: zmienia |wprowadza | uchyla {akapit LIST, nowyTekst CİAG}]
->(:typ_prawa),
(:articleType)-[referenceType: zmienia |wprowadza | uchyla {akapit LIST, nowyTekst STRING}]
->(:typ_artykułu),
(:articleType)-[referenceType: zmienia |wprowadza | uchyla {akapit LIST, nowyTekst STRING}]
->(:typ_nretTeyrpeen)c,eType: cites{paragraphLIST,weightINT}]->(:articleType), (:articleType)-[referenceType:
cites{paragraphLIST,weightINT}]->(:attachmentType), (:attachmentType)-[referenceType: is_legal_basis_of| cites
{paragraph LIST, weight INT}]->(:lawType), (:attachmentType)-[referenceType: cites {paragraph LIST, weight
INT}]->(:articleType), (:attachmentType)-[referenceType: cites {paragraph LIST, weight INT}]->(:attachmentType)}
```

### A.2. Zapytania dotyczące wymiaru czasowego

W tej sekcji przedstawiamy dwa istotne, zależne od czasu zapytania szyfrów, które ilustrują zdolność naszego schematu do uchwycenia ewolucji korpusu legislacyjnego. Pierwsze zapytanie wyprowadza tekst w określonym momencie, zmodyfikowany przez prawodawstwa. Na przykład, `toderivelaw14/2010 asit wasinforceattimestamp2023-02-01:`

WYWOŁAĆ{

```

DOPASUJ (l:Pravo)-[:HAS_ARTICLE]->(a:Artykuł)
OPCJONALNE DOPASOWANIE (a)<-[r:ZNIEKSZTAŁCENIA|POPRAWKI|INTRODUCES]-(a2)<-
[:HAS_ARTICLE]-(l2:Law) WHERE l2.publicationDate < datetime("2010|14")
Z l.id JAKO IDLAW, a.id JAKO IDART, a.number AS NUMART, MAX(l2.publicationDate)
    JAKO LASTMOD
GDZIE IDLAW = "2010|14"
Z IDLAW, IDART, NUMART, LASTMOD
DOPASUJ (l:Pravo)-[:HAS_ARTICLE]->(a:Artykuł)<-[r:UCHYLONE|POPRAWKI|WPROWADZENIE]-(a2)
    <-[:HAS_ARTICLE]-(l2:Pravo)
GDZIE l.id = IDLAW I a.id = IDART I LASTMOD = l2.publicationDate
RETURN IDLAW, IDART, r.newtext AS TEXT, NUMART
UNION MATCH(l:Pravo)-[:HAS_ARTICLE]->(a:Artykuł)<-[r:ZNIEKSZTAŁCA|POPRAWKI|PRZEDSTAWIAJ-
    (a2)<-[:HAS_ARTICLE]-(l2:Pravo)
Z l.id JAKO IDLAW, a.id JAKO IDART, COUNT(r) JAKO NCHANGES

```

```

GDZIE IDLAW = "2010|14"
Z IDLAW, IDART, NCHANGES
DOPASUJ (l:Prawo)-[:HAS_ARTICLE]->(a:Artykuł)-[:r:ZCIĄGA|POPRAWKI|WPROWADZENIE]-(a2)
  <-[:HAS_ARTICLE]-(l2:Prawo)
WHERE a.id = IDART AND l2.publicationDate >= datetime("2010|14")
Z IDLAW, IDART, a.number AS NUMART, a.text AS TEXT, NCHANGES,
LICZENIE(*) JAKO PRZYSZŁE ZMIANY
GDZIE NCHANGES = PRZYSZŁE ZMIANY
ZWRACANIE IDLAW, IDART, TEXT, NUMART
UNIA
MECZ (l:Prawo)-[:HAS_ARTICLE]->(a:Artykuł)
GDZIE NIE (a)-[:ZNOSI SIĘ|POPRAWKI|INTRODUCES]-(l) I l.id = "2010|14" ZWRACA l.id JAKO
IDLAW, a.id JAKO IDART, a.text JAKO TEKST, a.number JAKO NUMART
}
Z TEKSTEM, NUMART
GDZIE TEKST NIE MA WARTOŚCI NULL
Z TEKSTEM, NUMART
ZAMÓWIENIE PRZEZ NUMART ASC
RETURN COLLECT TEKST)

```

**Drugie zapytanie Cypher wnioskuje o zaktualizowanej liście praw, które zostały uchylone:**

```

MATCH p=(l:Prawo)-[:HAS_ARTICLE]->(a:Artykuł)-[:r:ZNOSI ARTYKUŁ]-
(a2:Artykuł)-[:HAS_ARTICLE]-(l2:Prawo) GDZIE r.paragraf JEST NIEWAŻNY
Z l.id AS uchylonyPrawo, l.numArt AS N_Arts, COUNT(DISTINCT a)
  JAK N_Repeals GDZIE N_Repeals >= N_Arts
Z COLLECT(uchylonym) JAKO list_abrogations
MECZ (l:PRAWO) GDZIE l.id W list_abrogations RETURN l.id
UNIA
DOPASUJ (l:Prawo)-[:r:ZREKA]-(l2:Prawo) GDZIE r.paragraf IS NULL RETURN l.id
UNIA

```

### A.3. Integracja danych parlamentarnych

Z punktu końcowego *Camera dei Deputati* (*Camera dei Deputati, 2024*) możemy uzyskać historyczne nazwy wszystkich departamentów w całej Republice Włoskiej za pomocą następującego zapytania SPARQL:

```

WYBIERZ DISTINCT ?titolo
GDZIE: {
  ?governo rdf:type ocd:governo .
  ?governo dc:tytuł ? Nazwa.
  ?governo ocd:startDate ? Początek.
  OPCJONALNIE { ?governo ocd:endDate ? Koniec. }
  governo ocd:rif_membroGoverno ?membro . ?
  membro foaf:nazwisko ?cognome; dc:tytuł ?titolo .
}

```

### A.4. Lista domen włoskich i słownik

Podajemy listę 16 domen w języku włoskim, które bierzemy pod uwagę w całym artykule: *interno, istituzioni, agricoltura, istruzione, economia, comunicazioni, presidenza, trasporti, sanità, esteri, giustizia, lavoro, difesa, pubblica amministrazione, cultura e ambiente, sport e turismo*.

Przykładowy odwzorowanie słownika dziedzinowego jest dostępny w [tabeli A.1](#).

### A.5. Szczegóły dotyczące dostrajania LLM

Straty treningowe i parametry używane do dostrajania LLM są dostępne na [rys. 1. A.1](#) i [tabeli A.2](#).

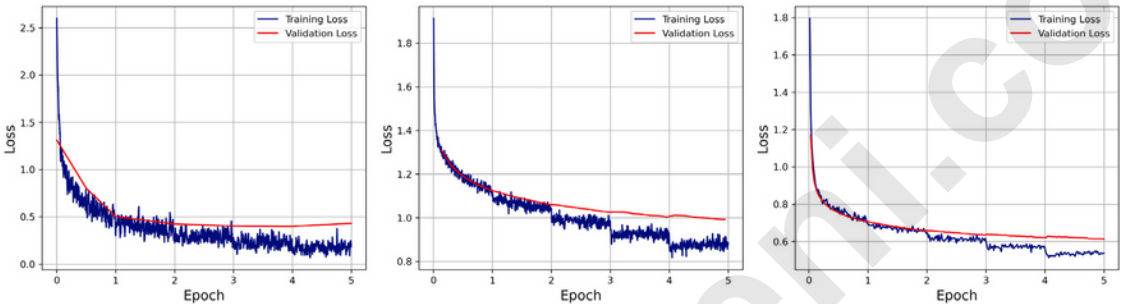
### A.6. Wyodrębnianie tytułów i tematów za pomocą LLM - wersje włoskie

Ze względu na to, że LLM są używane w języku włoskim, przedstawiamy oryginalne wersje dla [rys. 7, 8](#) i [tabeli 3, 4](#) (patrz [rys. D](#)). [A.2](#) oraz [A.3](#) i [Tabele A.3](#) i [A.4](#).



Tabela A.1  
Słownik domen włoskich mapujący słowa kluczowe do domeny.

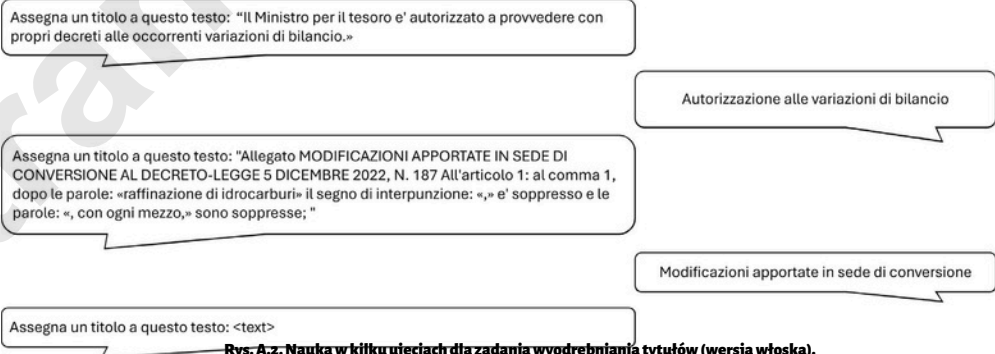
Słowo kluczowe w ministerstwie nameaviazione	Domena
Trasporti	
istituzioni	
esteri	
lavoro	
agricoltura	
...	



Rys. A.1. Szkolenie i walidacja kreślone dostrajania dla modelu klasyfikatora domen, ekstraktora tytułów Mistral-7B i ekstraktora tematów Mistral-7B. W przypadku wszystkich tych modeli straty związane z walidacją zaczynają być zbliżone wokół 4 modeli.

Tabela A.2  
Parametry używane do dostrajania modelu klasyfikatora domen, ekstraktora tytułów Mistral-7B i Mistral-7B Extractor.

orazBERTdomainclassifier		orazMistral-7Btitleextractor (Mistral-7B)		(c) Ekstraktor tematyczny Mistral-7B	
Parametr	Wartość	Parametr	Wartość	Parametr	Wartość
Wielkość partii	32	Wielkość partii	4	Wielkość partii	4
Szybkość uczenia się	0.45	Czas uczenia się	1.0	Szybkość uczenia się	0.01
Czas szkolenia	40 min	Czas szkolenia	1.0	Czas szkolenia	0.01
Bestwalidacjastrata	0.90	Optimalizacja best-walidacja	Adam	Optimalizacja walidacji	Adam
Dokładność	AdamW	Optimalizacja best-walidacja	108 tys.	Optimalizacja Trening	74 tys.
Optimalizacja Rozmiar	45 tys.	zestawu treningowego	0, 0 3	Wielkość zestawu	0,03
zestawu treningowego		Frakcja rozgrzewkowa	LoRa	Frakcja rozgrzewkowa	LoRa
		Technika dostrajania		Technika dostrajania	



Rys. A.2. Nauka w kilku ujęciach dla zadania wyodrębniania tytułów (wersja włoska).

A.7. Eksploracja grafów wiedzy - Zapytania Cypher

W tej sekcji przedstawiamy zapytania Cypher użyte do uzyskania wglądu w Sekcję 5:

1. Prawo uchwalone rocznie (Q1):

Tabela A.3.50. Tytuł ekstrakcja przed i po dostrajeniu (wersja włoska).

Tekst	Mistral-7B Wstępnie wyszkolony	Mistral-7B Dostrójony
Articolo 13. 1. Il presente Protocollo entra in vigore alla data concordata tra le Parti successive al cambiamento. 2. Il Protocollo resta in vigore per 5 anni. Salvo che una delle due Parti comunichi, con preavviso di almeno sei mesi, il suo dissenso alla scadenza, la propria intenzione di non rinnovare...	Articolo 13. Modifiche al regime di agevolazioni per il prodotto contraddistinto con il n. 13 della tabella A	Disposizioni finali e termini di vigore del Protocollo
Art. 7. 1. Il regime previsto dalla legge di dicembre 1948, n. 1438, e successive modificazioni, limitatamente al prodotto contraddistinto con il n. 13 della tabella A, allegata alla legge 27 dicembre 1975, n. 700, è determinato secondo quanto previsto nel presente articolo, in conformità agli articoli 130 e 132 dell'accordo di cooperazione tra la Comunità economica europea e la Repubblica di Slovenia...	Articolo 33. Organizzazioni di volontariato	Finanziamento delle organizzazioni di volontariato
Art. 33. 1. Le organizzazioni di volontariato possono assumere lavoratori dipendenti o avvalersi di prestazioni di lavoro autonomo o di altra natura esclusivamente nell'ambito necessario al loro regolare funzionamento oppure nei limiti occorrenti a qualificare...		

Estrai gli argomenti da questo titolo: «Restituzione dell'imposta generale sui prodotti esportati.»

imposta, esportazioni, tassazione

Estrai gli argomenti da questo titolo: «Disposizioni di attuazione della normativa comunitaria concernente l'Organizzazione comune di mercato (OCM) del vino.»

normativa EU, organizzazione comune di mercato, OCM, vino

Estrai gli argomenti da questo titolo: <text>

Rys. A.3. Nauka w kilku ujęciach dla zadania wyodrębniania tematów (wersja włoska).

Tabela A.4

Przykłady ekstrakcji tematów przed i po dostrajaniu (wersja włoska).

Tytuł	Mistral-8×22B Wstępnie przeszkolony	Mistral-7B Dostrójony
Caratteristiche dell'aumento di capitale	capitale, aumento, caratteristiche	Aumento di capitale, finanza aziendale
Entrata in vigore	entrata in vigore, attivazione, attivazione normativa, attivazione regolare, attivazione	Entrata in vigore
Aumento del contributo personale annuo	Contributo personale, anno, aumento	contributo personale, pensioni
Poteri di controllo e perquisizione delle forze di polizia	polizia, controllo, perquisizione, poteri	polizia, controllo, perquisizione

MECZ (l:Prawo)

RETURN l.publicationDate.year AS Data, count(l) AS Num

## 2. Ustawy, które nigdy nie były cytowane po publikacji (Q2):

```
MECZ (l:Prawo)-[:HAS_ART|HAS_ATTACHMENT]->(a)
GDZIE NIE (l)-[:IS_LEGAL_BASIS_OF]->(Prawo)
A NIE (a)-[:IS_LEGAL_BASIS_OF]->(Prawo)
I NIE (a)-[:POPRAWKI]-() I NIE (a)-[:UCHYLA]-()
I NIE (a)-[:CYTUJE]-() I NIE (a)-[:PRZEDSTAWIA]-()
AND NOT (l)-[:AMENDS]-() AND NOT (l)-[:ABROGATES]-() AND NOT (l)-[:CITES]-()
RETURN l.publicationDate.year jako data, COUNT(DISTINCT l)
```

## 3. Przeszarżone przepisy (Q3):

```
MATCH (l:Prawo)-[:IS_LEGAL_BASIS_OF]->(l2:Prawo)
GDZIE: l2.publicationDataData < /godzina("1960")
Z COLLECT(l.id) AS CitedBefore60s MATCH (l:Prawo)-[:IS_LEGAL_BASIS_OF]->(l2:Prawo)
```

```

GDZIE l2.publicationData, > data/godzina("1990")
I l.id W Cytowane Przed60s
Z COLLECT(l.id) AS StillCited, CitedBefore60s
UNWIND[x IN Cytowane przed60s GDZIE NIE
  ANY(z IN StillCited WHERE z CONTAINS x)] AS OutdatedLaws
RETURN OutdatedLaws

```

#### 4. Zapas obowiązujących przepisów (Q4):

```

MATCH p=(l:Prawo)-[:HAS_ARTICLE]->(a:Artykuł)-[:r:ZNOSI ARTYKUŁ]-
  (a2:Artykuł)-[:HAS_ARTICLE]-(l2:Prawo)
WHERE r.paragraph IS NULL AND l2.publicationDate <= datetime('2020') WITH
  l.id AS abuxedLaw, l.numArt AS N_Arts, COUNT(DISTINCT a)
  JAK N_Repeals
GDZIE N_Repeals >= N_Arts
Z COLLECT(uchylonym) JAKO list_abrogations
MECZ (l:Prawo)
WHERE l.publicationDate <= datetime('2020') I NIE l.id W list_abrogations I NIE
  ()-[:ZNOSI]-(L>:Prawo)
ZWRÓĆ COUNT(l.id) AS CountInForceLaws

```

#### 5. Udział ministerstw w tworzeniu aktów prawnych (Q5):

```

DOPASUJ (g:Rząd)-[:SUCCEEDED_BY*4]->(g2:Rząd) GDZIE NIE
  ISTNIEJE ((g2)-[:SUCCEEDED_BY]->(:Rząd)) Z g.name JAKO
  FIRSTGOV
PODAJ.MECZ (g:Rząd)-[:SUCCEEDED_BY*1..4]->(g2:Rząd)
DOPASUJ (l:Prawo)-[:UNDER_GOVERNMENT]-(g2:Rząd) GDZIE
  g.name = FIRSTGOV
Z g2.name JAKO GOVNAME, COUNT (l) JAKO NLAWS
DOPASUJ (l:Prawo)-[:UNDER_GOVERNMENT]-(g2:Rząd) GDZIE:
  g2.name = GOVNAME
Z g2.name JAK GOVNAME, NLAWS, l.domain jako allDomains
UNWIND allDomains as DOMAIN
ZWRÓĆ GOVNAME, NLAWS, DOMAIN, COUNT(DOMAIN) JAKO N

```

#### 6. EU Akty prawne wdrożone w systemie włoskim (Q6):

```

DOPASUJ (l:Prawo)-[:UNDER_LEGISLATURE]->(e:Legislatura)
WHERE ANY(x IN l.topic WHERE x IN ["direttiva ue", "regolamento ue"]) AND e.name =
  "Legislatura XIX" AND l.publicationDate <= e.startDate + Duration({months: 7}) RETURN
  COUNT(l) jako EUConversions

```

#### 7. Tematy interwencji rządowych (Q7)

```

DOPASUJ (l1:Prawo)-[:HAS_ARTICLE]->(a1)-[:ZNOSI PRAWA | POPRAWKI |
  PRZEDSTAWIA]-(a2)-[:WCELEWUJE | REPEALS | ABOLISHES]-(a3)-[:UNDER_GOVERNMENT]->(g:Rząd)

```

```

UNWIND l1.topic jako tematy
RETURN g.name AS GOVNAME, topics AS TOPIC, COUNT(topics) as N

```

#### Dostępność danych

Dane są udostępniane w repozytorium publicznym. Używane modele AI są również udostępniane w HuggingFace.