IfGPT, a Large Dataset Representing Bulgarian, with the Bulgarian National Corpus as its Core

12th Workshop on the Challenges in the Management of Large Corpora @ LREC 2026

Svetla Koeva, Ivelina Stoyanova
Department of Computational Linguistics
Institute for Bulgarian Language
Bulgarian Academy of Sciences

svetla@dcl.bas.bg | iva@dcl.bas.bg

Motivation

The development of large-scale corpora presents interconnected technical, linguistic and management challenges:

- **Technical** – scalable storage and efficient retrieval of text, metadata, and annotation layers.
- **Linguistic** – corpora must be diverse, representing a wide range of language use including low-resource languages, underrepresented phenomena, and historical texts.
- **Management** – coordinating data collection, quality control, licensing, and versioning across multiple sources.

Motivation

BulNC is a **standard reference corpus** designed to reflect the natural distribution of Bulgarian across text types, genres, styles, and time periods.

- **Annotation layers** – tokenisation, POS tagging, lemmatisation, dependency parsing, word sense, NER, noun phrase identification.
- **Linguistic integrity** – deduplication, removal of typographical errors, incomplete sentences, malformed words.

The main purpose on BulNC is to serve for linguistic studies on lexical and grammatical features of Bulgarian, dictionary creation, language change exploration.

Related work

There are many large and widely used text databases:

- **CC100** – ~2 TB of filtered monolingual text in 100 languages via the CCNet pipeline.
- **Pile** – 825 GiB of English text from 22 curated subsets; domain diversity improves generalisation.
- **Dolma** – 3 trillion tokens across six source types with full processing toolkit.
- **CulturaX** – 6.3 trillion tokens in 167 languages by merging mC4 and OSCAR.
- **Aya Collection** – 513 million instances across 114 languages via open participatory science.

Related work

Main directions:

- **Collecting, cleaning, and enriching large datasets.**
- **Efficient dataset management and data organisation.**
- **A shift from passive data collection to active dataset design** oriented towards particular tasks, ensuring prompt diversity and quality control.
- **Multilingual coverage and distribution.**

Bulgarian in multilingual corpora – 130 parallel corpora available via ELG and CLARIN; BG present in 159 multilingual corpora.

IfGPT Dataset

IfGPT integrates several Bulgarian text collections, including BulNC, applying cleaning, deduplication, and LLM-oriented metadata.

- **Primary language** – Bulgarian, with English materials included.
- **Authentic language data** – cleaned and deduplicated, like BulNC.
- **LLM-oriented metadata** – adds PII scores, bias scores, task categories.
- **Annotation** – sentence markup (compared to full BulNC annotation).
- **Project URL:** https://ifgpt.dcl.bas.bg/en/

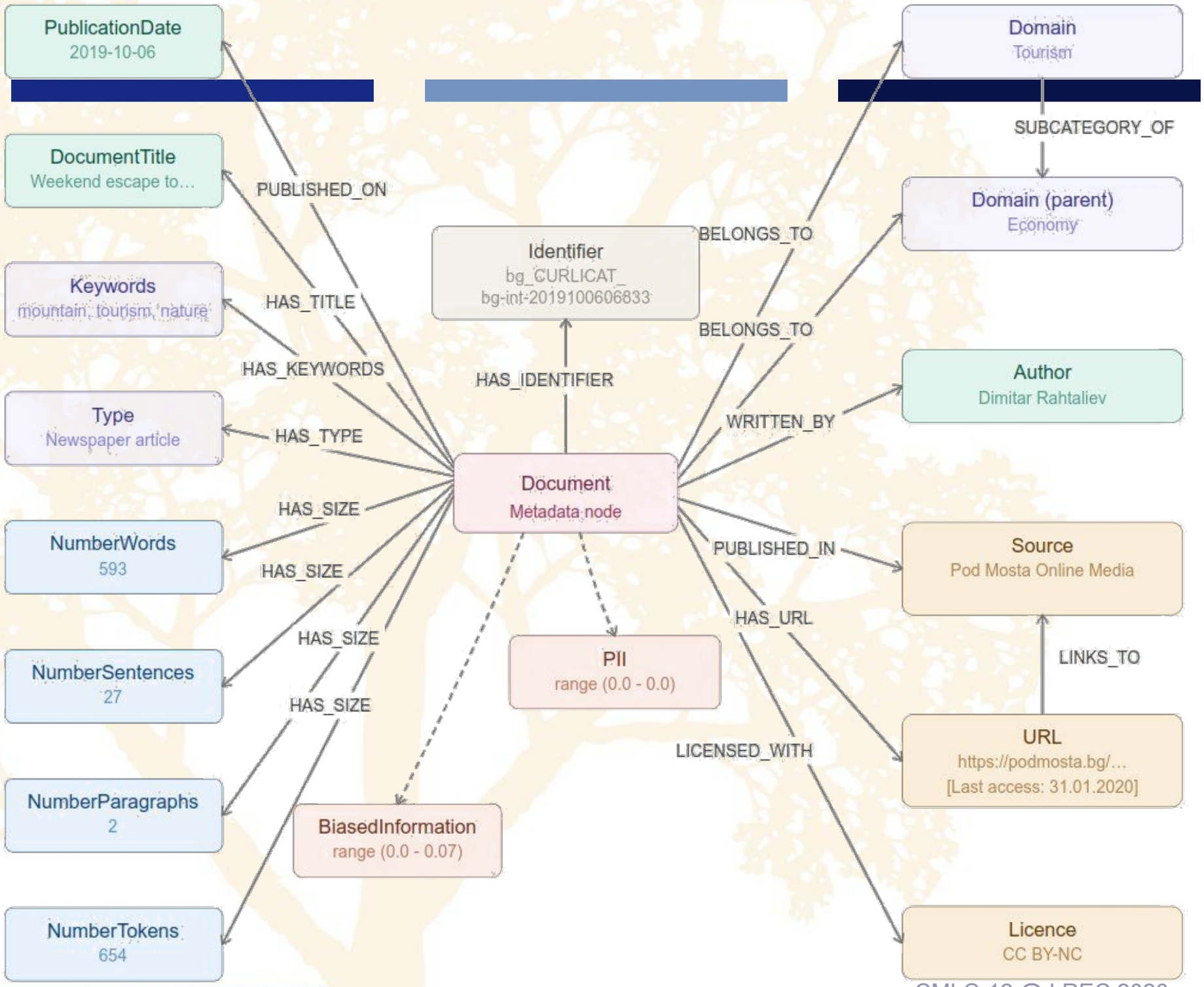
IfGPT Dataset

Dataset & Language(s)	Domains	Size	Format & annot.	Source & Licence
Bulgarian MARCELL BG 1946–2023	Legal (11 types: admin. court, agreements, amendments, conventions, etc.)	25K texts; 3.28M sents; 45M tokens	CoNLL-U+; morph., dep., NER, EuroVoc/IATE annotation	Bulgarian State Gazette Public Domain
Bulgarian CURLICAT BG	7 domains: Culture, Education, EU, Finance, Politics, Economics, Science	6K texts; 22.8M tokens	CoNLL-U+; JSON; full ling. annotation	BulNC; science sources: books, PhD theses; web CC-BY CC-BY-SA CC-BY-NC
Aligned and Normalised Parallel Data BG–EN	16 domains: General News, BG Presidency, Economics, Culture, Military, Politics, etc.	1.1M sent. pairs; 19.0M words (BG); 19.2M words (EN)	Sentence-aligned pairs; partly manual selection & correction	Web media; institutional websites Public Domain Various
General News in Bulgarian BG	185 domains	2.1M texts; 33.4M sents; 601M words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (11.8K domains, 2.1M pages) Various
General News in English EN	185 domains	5.9M texts; 166.7M sents; 3.3B words	JSON; metadata; automatic categorisation; normalisation & cleaning	Web crawling (324.5K domains, 5.9M pages) Various

IfGPT Metadata

Mandatory fields	Optional fields
Identifier	Author
Licence	Style
PublicationDate	Type
DocumentTitle	Subdomain
Source, Url	TranslatedDocument
Medium	CollectionDate
Domain	LicenseLink
Keywords	TaskCategories
NumberWords	
NumberSentences	
NumberParagraphs	
PersonallyIdentifiableInformation	
BiasedInformation	

IfGPT Metadata

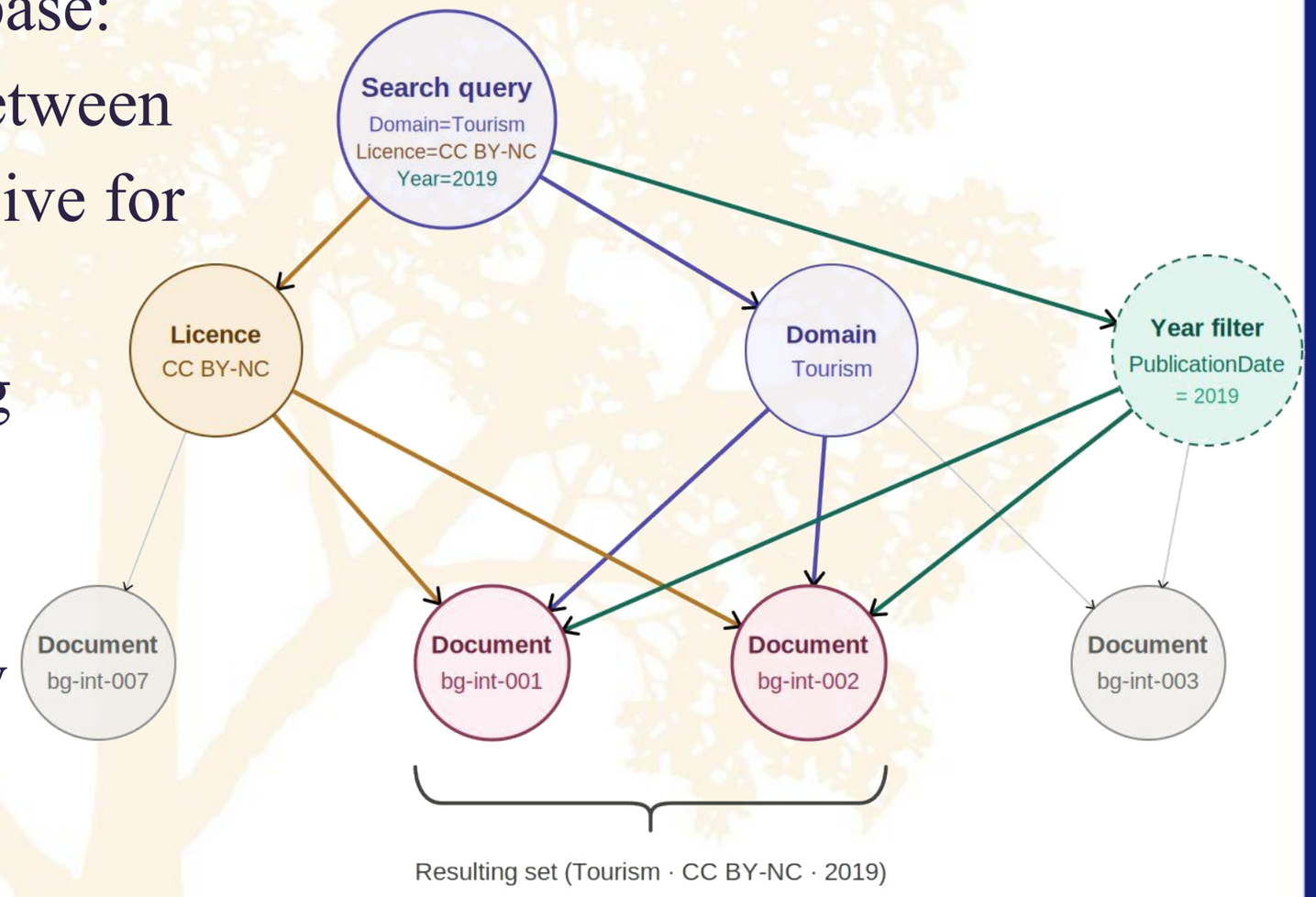


IfGPT Metadata

IfGPT metadata is managed in a Neo4J graph database and accessed through a web search interface allowing the selection of subdatasets.

Advantages of the graph database:

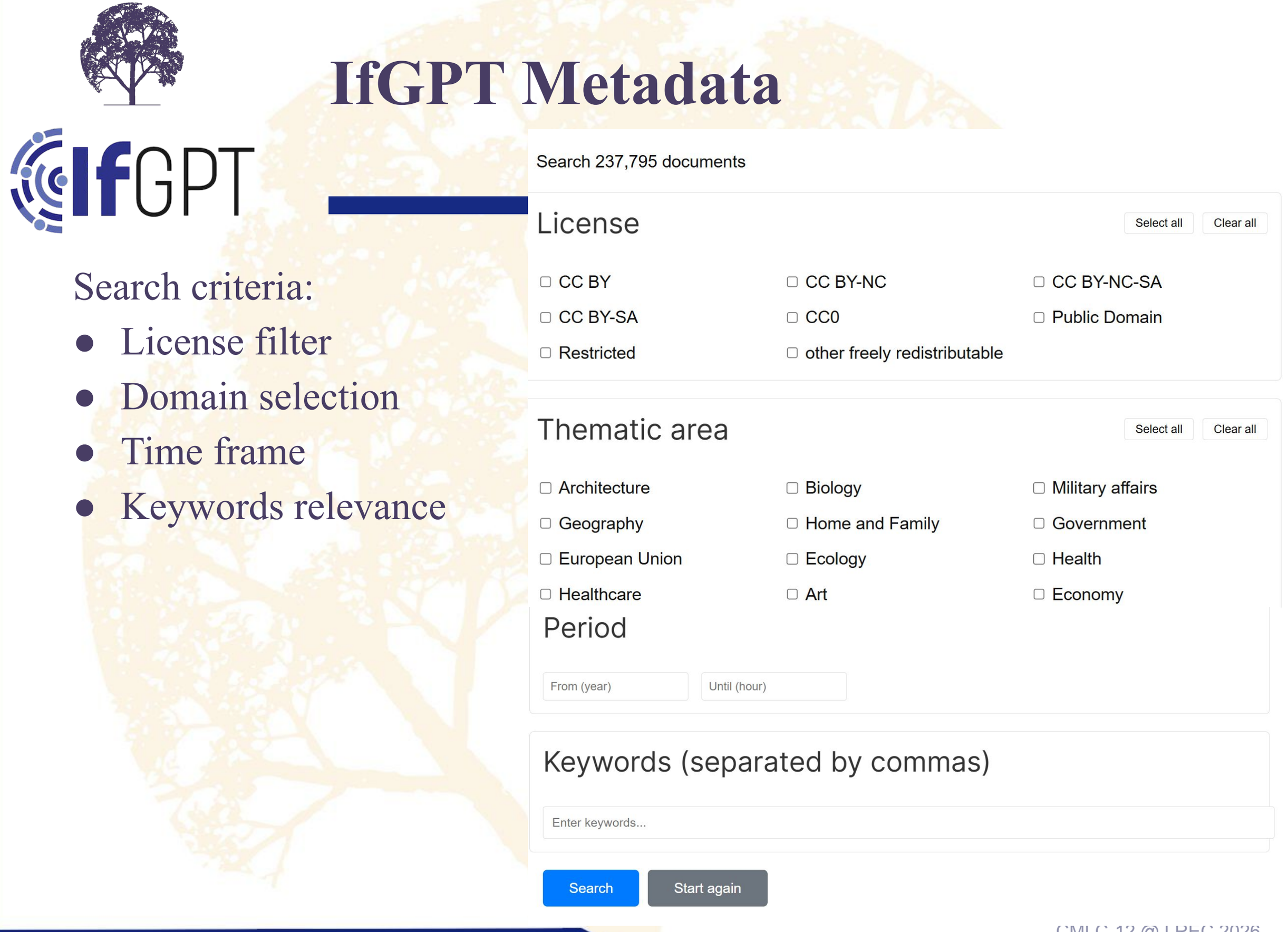
- models the rich relations between nodes, thus is more expressive for metadata exploration;
- complex queries combining multiple criteria through efficient graph traversals;
- scales efficiently when new relation types are added.



IfGPT Metadata

Search criteria:

- License filter
- Domain selection
- Time frame
- Keywords relevance



IfGPT Metadata

Search criteria:

- License filter
- Domain selection
- Time frame
- Keywords relevance

33,118
Total documents

103,422,498
Total words

1
Current page

METADATA (json) | LINKS (tsv) | DATA (zip) | 33,118 documents

← Previous. | 1 | 2 | 3 | … | 1656 | Next →

WHITE PAPER Agenda for Adequate, Secure and Sustainable Pensions

European Commission | Economy | CC0 | Administrative | indefinite | 2012-01-01

ID: bg_bnc_00230661nADZ **URL:** See **Paragraph:** 6159
Sentences: 6412 **Words:** 16680 **Media:** text

COMMISSION REPORT Second evaluation report on the EU Pilot project

European Commission | Economy | CC0 | Administrative | correspondence | 2011-01-01

ID: bg_bnc_00230099nADG **URL:** See **Paragraph:** 61
Sentences: 117 **Words:** 2631 **Media:** text

REPORT FROM THE COMMISSION TO THE EUROPEAN PARLIAMENT AND THE COUNCIL Mid-term evaluation of the European Metrology Research Programme - EMRP

European Commission | Economy | CC0 | Administrative | report | 2012-01-01

ID: bg_bnc_00230698nADC **URL:** See **Paragraph:** 86
Sentences: 154 **Words:** 4277 **Media:** text

IfGPT Metadata

JSON and TSV formats of data export

Search Results JSON

Generated on 9/11/2025, 8:36:15 AM


```
{
  "searchResults": {
    "statistics": {
      "numberOfTextSamples": 4570,
      "numberOfSentences": 272195,
      "numberOfTokens": 6560464
    },
    "selectedDocuments": [
      {
        "year": 2016,
        "PublicationDate": "2016-01-12",
        "Keywords": "",
        "NumberTokens": 118,
        "CollectionDate": "",
        "TranslatedDocument": "",
        "LicenceLink": "https://elrc-share.eu/static/metashare/licences/CC0.pdf",
        "TaskCategories": "",
        "Source": "",
        "Url": "http://dv.parliament.bg/DVWeb/showMaterialDV.jsp?idMat=100032",
        "BiasedInformation": "",
        "NumberWords": 107,
        "Type": "Решения на ЦИК",
        "Identifier": "bg_MARCELL_bg-100032",
        "Subdomain": "",
        "Medium": "text",
        "PersonallyIdentifiableInformation": "",
        "NumberParagraphs": 8,
        "Licence": "CC0",
        "Author": "Ministries and other institutions",
        "Style": "Legal",
        "DocumentTitle": "Решение № 3008-ПВР от 12 януари 2016 г. относно утвърждаване образци на изборните книжа
за произвеждане на избори за президент и вецепрезидент на Република България",
        "Domain": "Government",
```

IfGPT and BulNC: Future work

Future developments include:

- adding new and diverse text data to both resources;
- expanding metadata descriptions;
- validating and improving text data quality in both resources;
- enhancing accessibility and providing easy access to the IfGPT data for various purposes.

Acknowledgments

- The work is part of the project *Infrastructure for Fine-tuning Pre-trained Large Language Models*, Grant Agreement No. ПВУ – 55 from 12.12.2024 /BG-RRP-2.017-0030-C01/.

Funded by the European Union
NextGenerationEU

National Recovery and Resilience Plan of the Republic of Bulgaria

- https://ifgpt.dcl.bas.bg