Legal Information Retrieval meets Artificial Intelligence (LIRAI)

Ernesto William De Luca
deluca@gei
Otto von Guericke University
Magdeburg, Germany
Leibniz Institute for Educational Media | Georg Eckert Institute
Brunswick, Germany

Manuel Fiorelli
manuel.fiorelli@uniroma2.it
Tor Vergata University of Rome
Rome, Italy

Davide Picca
davide.picca@unil.ch
University of Lausanne
Lausanne, Switzerland

Armando Stellato
stellato@uniroma2.it
Tor Vergata University of Rome
Rome, Italy

Sabine Wehnert
sabine.wehnert@gei.de
Otto von Guericke University
Magdeburg, Germany
Leibniz Institute for Educational Media | Georg Eckert Institute
Brunswick, Germany

ABSTRACT

The Legal Information Retrieval meets Artificial Intelligence (LIRAI) workshop series aims to provide a venue hosting discussion of novel ideas, evaluations, and success stories concerning the application of Artificial Intelligence (AI) and Information Retrieval (IR) to the legal domain. All around the world, lawmakers, legal professionals, and citizens must cope with the sheer amount of legal knowledge present in legal documents. These documents can be norms, regulations, directives, legal cases, and other relevant material for legal practitioners, such as legal commentary. The continuous evolution of legal documents is a challenging setting, with implicit relationships playing an important role beyond explicit references. Recently, the adoption of shared machine-readable formats and FAIR principles, as well as methods and practices from the Semantic Web, have certainly improved the accessibility of legal knowledge and its interoperability. Still, retrieving legal knowledge and making sense of it are not solved problems. The legal community often has special requirements for retrieval systems (e.g., high recall, explainability). Artificial Intelligence (AI) is positioned as a lever to enhance our ability to find, understand, and correlate legal information, and to comprehend its relationship to reality, in terms of compliance evaluation and risk/benefit analysis. We call contributions on these topics in the form of papers, which will be collected in an open-access proceedings published on CEUR-WS.org and thus indexed by Scopus, DBLP, Google Scholar, and other citation databases.

CCS CONCEPTS

• **Applied computing** → **Law**; • **Information systems** → **Information retrieval**; • **Computing methodologies** → **Artificial intelligence**.

KEYWORDS

Legal Informatics, Legal Information Retrieval, Legal Knowledge Representation, Legal Text Mining, Legal Compliance, FAIRness, Semantic Web, Linguistic Legal Linked Open Data, Explainable AI, High-Recall Retrieval

ACM Reference Format:
Ernesto William De Luca, Manuel Fiorelli, Davide Picca, Armando Stellato, and Sabine Wehnert. 2023. Legal Information Retrieval meets Artificial Intelligence (LIRAI). In *34th ACM Conference on Hypertext and Social Media (HT '24), September 4–8, 2023, Rome, Italy.* ACM, New York, NY, USA, 4 pages. https://doi.org/10.1145/3603163.3610575

1 DESCRIPTION OF THE WORKSHOP

Legal systems all around the world have developed a large and intricate body of legal knowledge, which is continually evolving as new norms are enacted and existing ones are either abrogated or amended. While citation is at the heart of modern legal systems, the interconnectedness between different acts can be implicit, as happens when a later act implicitly abrogates an earlier one by regulating the same topic differently. It is thus very difficult to retrieve relevant legal knowledge, make sense of it, and foresee the complex relationships between legal documents and the reality they regulate.

The intersection of Legal AI and hypertext has a long history, with legal retrieval systems [2, 13] and hypertext databases for law [20]. These days, use cases of hypertext in the legal domain still stem from the characteristics of legal documents and the relations between them:

- Hierarchical relationships between legal documents (e.g., European legislation compared to legislation within member states of the European Union).

HT '24, September 4–8, 2023, Rome, Italy

ACM ISBN 979-8-4007-0232-7/23/09.
https://doi.org/10.1145/3603163.3610575

- Temporal relationships between legal documents (e.g., different versions of the same legal norm).
- Referential relationships (e.g., court cases citing relevant legal norms), thereof explicit references (i.e., citations) and implicit references (e.g., use of legal jargon defined in a norm which is not cited).
- Cross-lingual relationships (e.g., different language versions of the same document).

There can be more types of relationships between legal documents, depending on the specific use case and the notion of relevance adopted for it [17].

Compared to systems from three decades ago, modern approaches such as contextual language models are applied to overcome matching problems and different abstraction levels across documents. Paired with knowledge graphs or other forms of modeling hypertext, there are many new applications of hypertext concepts, which makes this field of research interesting for the hypertext research community.

The core topics of this workshop are:

- Hypertext-based legal systems
- Legal information retrieval
- Legal information extraction
- Legal knowledge graphs
- Relation extraction from legal documents
- Explainability in legal document retrieval
- High-recall settings in legal document retrieval
- Legal ontologies
- Legal document formats
- FAIR publication of legal documents

Well-motivated submissions related to (but not directly fitting) one of the core topics are also welcome. Some example use cases relevant for this workshop are:

- Publication of legislation and other legal documents on official portals (e.g., EUR-Lex, Legislation.gov.uk, govinfo.org) complying with FAIR [19] principles.
- Organization of legislation and other legal documents (e.g., Linguistic Legal Linked Open Data (LLLOD) Lynx [8]).
- Languages for the publication of legal documents in a machine-readable form (e.g., Akoma Ntoso [4, 11], CEN MetaLex [1], United Kingdom Legislation schema [3], NIR [6], Formex [10], LegalHTML [14, 15]).
- Identification schemes (e.g., Akoma Ntoso naming convention, ELI scheme [5]) for citing documents or portions thereof.
- Consolidation of changes brought to a legal document by amending documents and corrigenda, and construction of point-in-time views of documents.
- Legal Assistants accessing external knowledge through hyperlinks [9].
- Hypertext use in educational settings in the legal domain [16].
- Legal advisory systems linking commentary to their results [7], similar use cases in comparative law [12].

In some use cases, such as patent retrieval and compliance checking, legal practitioners often face challenges in finding all relevant documents [18]. In such challenging settings, legal systems need to fulfill the requirements of high recall and explainability to be of value for legal practitioners.

Legal informatics has emerged as an interdisciplinary field of study, in between computer science and law, that has attempted to solve these problems in a computational way. Information technology has also improved the legal process itself, as machine-readable formats and intelligent systems have been adopted from first drafting to dissemination. The adoption of FAIR practices for the publication of legal knowledge has certainly the potential to make legal knowledge more accessible at every level (from citizens to legal professionals), making it easier to uncover information and explore relations between documents and concepts.

2 RELEVANCE OF THE WORKSHOP TO THE HYPERTEXT COMMUNITIES

In most legal traditions, legal knowledge evolved as an intricate hypertext spread across a large body of diverse documents woven together by explicit references. Since countries are digitizing their legal systems with the adoption of machine-readable formats for the interchange of legal information, their body of legal knowledge is provided as hypertexts. Furthermore, the adoption of Web technologies, in particular those of the Semantic Web, means that legal knowledge is not only available on the Web, but it is also part of the Web, which is perhaps the most successful example of hypertext to date.

As a hypertext, a legal document base has its own characteristics that distinguish it from other hypertexts and make it a special case study of interest in its own right. These include the practice of amending previous documents in subsequent ones, and the challenging nature of implicit relationships between documents, which are not fully covered by explicit hyperlinks.

3 WORKSHOP ORGANISERS' BIOS

Ernesto William De Luca is head of the Human-Centred Technologies for Educational Media department at the Georg Eckert Institute for International Textbook Research - Member of the Leibniz Association (GEI) and from October 2019 has been appointed as a Full Professor in Human-Centred Artificial Intelligence at the Otto von Guericke University Magdeburg, Germany. In addition, in May 2015 he was appointed by the Guglielmo Marconi University of Rome as an associate professor in "computational engineering". He studied computational linguistics and then gained his doctorate in computer science. His research includes Artificial Intelligence, Machine Learning, Natural Language Processing, Digital Humanities as well as the Semantic Web and Information Retrieval. He has written over 150 papers for national and international conferences and journals, organized and chaired numerous workshops and conferences, and is regular reviewer and programme committee member of different high-profile journals and conferences.

Manuel Fiorelli, Ph.D. is a Research Fellow at Tor Vergata University of Rome, researching on knowledge engineering and semantic technologies in the field of the Semantic Web. His current research interests include metadata modeling, data cataloging, semantic interoperability, FAIRness, and legal information representation. He is author of about 30 publications in workshops, conferences, and journals. He has been in the program committee of the MTSR conference since 2022. He is also a member of the W3C Ontology-Lexica Community Group, focusing on the metadata module LIME. He is a core developer of the platform VocBench and ShowVoc, both funded by the DIGITAL program. In the field of Legal Informatics, he contributed to the realization of Legal-HTML, an HTML domain language for the representation of legal acts on the Web: funded by the Publications Office of the EU, it is being adopted as a new dissemination format for EUR-Lex. He participated in the EU-funded research projects SEMAGROW and KATY.

Dr. Davide Picca is a confirmed researcher in Digital Humanities, focusing on cultural heritage and digital technologies. His work explores computational semantics and ontology, while also examining how legal domain shapes societies and cultures. With a passion for preserving cultural heritage in the digital age, Dr. Picca's works contribute to pave the way for meaningful interdisciplinary research.

Armando Stellato, Ph.D. is Associate Professor at Tor Vergata University of Rome, where he researches and teaches in the fields of Knowledge Engineering and Knowledge-Based Systems. He is author of over 100 publications and has been member of the program committees of over 100 international scientific conferences and workshops in the Semantic Web and Natural Language Processing fields. Armando Stellato participated in several EU-funded projects and working groups, such as the W3C Ontology-Lexica Community Group. He has collaborated with several institutions (FAO, ESA, UN, USDA, various governments, etc.), providing knowledge transfer, training, and consultancy for the management and publication of data, metadata, and documental archives. Professor Stellato is currently leading – under two projects funded by the DIGITAL program – the development of an ecosystem for knowledge acquisition and management, which includes widely adopted platforms such as VocBench and ShowVoc. He is also active in the field of Legal Informatics, being involved in a collaboration with the Italian government for the semantic organization of Italian laws and in the realization of a semantic representation model for legal acts, currently funded and under adoption by the Publications Office of the EU.

Sabine Wehnert, M.Sc is a PhD candidate at Otto von Guericke University Magdeburg, Germany. Her research focuses are in the fields of legal retrieval, information extraction, explainable artificial intelligence, knowledge graphs and usability. Coordinating a Usability Lab at the Leibniz-Institute for Educational Media | Georg Eckert Institute, she views interactive systems from a human-centered perspective. Since 2019, she participates in the Competition on Legal Information Extraction and Entailment (COLIEE) and won the title in the Legal Norm Retrieval category (Task 3) with her team in 2021. Since 2021, she is also member of the COLIEE program committee. In her dissertation, she develops a bottom-up knowledge graph from scientific textbooks (called HONto) for the downstream tasks of legal retrieval and recommendation.

4 MOTIVATION

Treating legal documents as the hypertexts they are by now, this workshop sheds light on this topic as an interesting and challenging use case for the conference audience. Given the availability of legal documents online, their binding nature and relevance for businesses, institutions, and individuals in their daily lives, it is evident that developing systems that help gather and understand relevant information is crucial.

5 WORKSHOP AND SUBMISSION FORMATS

The workshop is intended as a half-day workshop with presentations by the authors of the accepted papers and a keynote address by a distinguished researcher. For each paper presentation, 20 minutes will be allotted divided in 10-15 for the talk and 5-10 for discussion, discussion times can be extended when the topic is of particular interest. Talks may either provide anchors for discussion or be clustered in sessions, in which the presenters form a panel to interact with the audience about questions raised by the talks. This concept gives some flexibility to allow focused discussion on specific topics as well as longer discussions on controversial topics.

All submitted papers will be reviewed by members of the PC and published in proceedings. The submission guidelines are the same as for the full papers of the ACM Hypertext 2023.

The workshop will be announced on mailing lists and social networks, a website will be set up, calls for papers will be sent out by email, the event will be advertised at related conferences and workshops, and personal emails will be sent to distinguished researchers working on related topics.

Approximate audience size: 30 - 40.

6 LENGTH

Half day. We can extend it to a full-day event depending on conference requirements and the number of accepted submissions.

7 MEMBERS OF THE PROGRAM COMMITTEE

- Tommaso Agnoloni, ITTIG-CNR, Italy
- Ilaria Angela Amantea, University of Turin, Italy
- Vito Walter Anelli, Politecnico di Bari, Italy
- Dennis Aumiller, Heidelberg University, Germany
- Valerio Basile, University of Turin, Italy
- Luigi Di Caro, University of Turin, Italy
- Harshvardhan J. Pandit, Dublin City University, Ireland
- Mi-Young Kim, University of Alberta, Canada
- Rūta Liepiņa, University of Bologna, Italy
- Carlo Marchetti, Senate of the Republic, Italy
- Patricia Martín-Chozas, Universidad Politécnica de Madrid, Spain
- Elena Montiel-Ponsoda, Universidad Politécnica de Madrid, Spain
- Jack Mumford, University of Liverpool, UK
- Monica Palmirani, University of Bologna, Italy
- Ginevra Peruginelli, ITTIG-CNR, Italy
- Ken Satoh, National Institute of Informatics and Sokendai, Japan
- Emilio Sulis, University of Turin, Italy
- Andrea Tagarelli, University of Calabria, Italy
- Marc van Opijnen, Publications Office of the Netherlands, The Netherlands
- Eugene Yang, Human Language Technology Center of Excellence, Johns Hopkins University, USA

- Masaharu Yoshioka, Hokkaido University, Japan
- Tomasz Zurek, T.M.C. Asser Institute, University of Amsterdam

REFERENCES

[1] [n. d.]. *CEN MetaLex: Open XML Interchange Format for Legal and Legislative Resources.* Retrieved July 24, 2023 from http://www.metalex.eu/

[2] Maristella Agosti, Roberto Colotti, and Girolamo Gradenigo. 1991. A Two-Level Hypertext Retrieval Model for Legal Data. In *Proceedings of the 14th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (Chicago, Illinois, USA) *(SIGIR '91)*. Association for Computing Machinery, New York, NY, USA, 316–325. https://doi.org/10.1145/122860.122892

[3] The National Archives. [n. d.]. *XML Format | Legislation.gov.uk.* Retrieved July 24, 2023 from https://www.legislation.gov.uk/developer/formats/xml

[4] Luca Cervone, Monica Palmirani, and Fabio Vitali. [n. d.]. *Akoma Ntoso | Akoma Ntoso Site.* Retrieved July 24, 2023 from http://www.akomantoso.org/

[5] Thomas Francart, John Dann, Roberto Pappalardo, Carmen Malagon, and Marco Pellegrino. 2018. The European Legislation Identifier. In *Knowledge of the Law in the Big Data Age, Conference 'Law via the Internet 2018', Florence, Italy, 11-12 October 2018 (Frontiers in Artificial Intelligence and Applications, Vol. 317)*, Ginevra Peruginelli and Sebastiano Faro (Eds.). IOS Press, 137–148. https://doi.org/10.3233/FAIA190016

[6] Enrico Francesconi. 2006. The "Norme in Rete" project: Standards and Tools for Italian Legislation. *International Journal of Legal Information* 34, 2 (2006), 358–376. https://doi.org/10.1017/S0731126500001517

[7] Graham Greenleaf, Andrew Mowbray, and Philip Chung. 2018. Building sustainable free legal advisory systems: Experiences from the history of AI & law. *Computer Law & Security Review* 34, 2 (2018), 314–326. https://doi.org/10.1016/j.clsr.2018.02.007

[8] Julián Moreno Schneider, Georg Rehm, Elena Montiel-Ponsoda, Víctor Rodríguez-Doncel, Patricia Martín-Chozas, María Navas-Loro, Martin Kaltenböck, Artem Revenko, Sotirios Karampatakis, Christian Sageder, Jorge Gracia, Filippo Maganza, Ilan Kernerman, Dorielle Lonke, Andis Lagzdins, Julia Bosque Gil, Pieter Verhoeven, Elsa Gomez Diaz, and Pascual Boil Ballesteros. 2022. Lynx: A Knowledge-Based AI Service Platform for Content Processing, Enrichment and Analysis for the Legal Domain. *Inf. Syst.* 106, C (may 2022), 18 pages. https://doi.org/10.1016/j.is.2021.101966

[9] Andrew Mowbray, Philip Chung, and Graham Greenleaf. 2020. Utilising AI in the legal assistance sector—Testing a role for legal information institutes. *Computer Law & Security Review* 38 (2020), 105407. https://doi.org/10.1016/j.clsr.2020.105407

[10] Publications Office of the European Union. 2004. *Formalized Exchange of Electronic Publications (FORMEX).* Retrieved July 24, 2023 from https://op.europa.eu/en/web/eu-vocabularies/formex

[11] Monica Palmirani and Fabio Vitali. 2011. Akoma-Ntoso for Legal Documents. In *Legislative XML for the Semantic Web: Principles, Models, Standards for Document Management*, Giovanni Sartor, Monica Palmirani, Enrico Francesconi, and Maria Angela Biasiotti (Eds.). Springer Netherlands, Dordrecht, 75–100. https://doi.org/10.1007/978-94-007-1887-6_6

[12] Marylin J Raisch. 2007. Codes and hypertext: The intertextuality of international and comparative law. *Syracuse J. Int'l L. & Com.* 35 (2007), 309.

[13] Jacques Savoy. 1993. Searching Information in Legal Hypertext Systems. *Artif. Intell. Law* 2, 3 (sep 1993), 205–232. https://doi.org/10.1007/BF00871890

[14] Armando Stellato and Manuel Fiorelli. 2023. *LegalHTML.* Retrieved July 24, 2023 from https://w3id.org/legalhtml/

[15] Armando Stellato and Manuel Fiorelli. 2023. LegalHTML: A Representation Language for Legal Acts. In *The Semantic Web: 20th International Conference, ESWC 2023, Hersonissos, Crete, Greece, May 28–June 1, 2023, Proceedings* (Hersonissos, Greece). Springer-Verlag, Berlin, Heidelberg, 520–537. https://doi.org/10.1007/978-3-031-33455-9_31

[16] Natalia Udina. 2022. Developing Students' Competencies of Reading Hypertext in Legal Language and Translation Studies. In *Proceedings of New Perspectives in Science Education 11th Edition 2022.* Filodiritto Proceedings, Bologna, Italy.

[17] Marc Van Opijnen and Cristiana Santos. 2017. On the Concept of Relevance in Legal Information Retrieval. *Artif. Intell. Law* 25, 1 (mar 2017), 65–87. https://doi.org/10.1007/s10506-017-9195-8

[18] Sabine Wehnert, Gabriel Campero Durand, and Gunter Saake. 2019. ERST: Leveraging Topic Features for Context-Aware Legal Reference Linking.. In *JURIX.* 113–122. https://doi.org/10.3233/FAIA190312

[19] Mark D. Wilkinson, Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, Jan-Willem Boiten, Luiz Bonino da Silva Santos, Philip E. Bourne, Jildau Bouwman, Anthony J. Brookes, Tim Clark, Mercè Crosas, Ingrid Dillo, Olivier Dumon, Scott Edmunds, Chris T. Evelo, Richard Finkers, Alejandra Gonzalez-Beltran, Alasdair J.G. Gray, Paul Groth, Carole Goble, Jeffrey S. Grethe, Jaap Heringa, Peter A.C 't Hoen, Rob Hooft, Tobias Kuhn, Ruben Kok, Joost Kok, Scott J. Lusher, Maryann E. Martone, Albert Mons, Abel L. Packer, Bengt Persson, Philippe Rocca-Serra, Marco Roos, Rene van Schaik, Susanna-Assunta Sansone, Erik Schultes, Thierry Sengstag, Ted Slater, George Strawn, Morris A. Swertz, Mark Thompson, Johan van der Lei, Erik van Mulligen, Jan Velterop, Andra Waagmeester, Peter Wittenburg, Katherine Wolstencroft, Jun Zhao, and Barend Mons. 2016. The FAIR Guiding Principles for scientific data management and stewardship. *Scientific Data* 3, 1 (15 Mar 2016), 160018. https://doi.org/10.1038/sdata.2016.18

[20] Eve Wilson. 1990. Links and Structures in Hypertext Databases for Law. In *Hypertext: Concepts, Systems and Applications, Proceedings of the European Conference on Hypertext, INRIA, France, November 1990 (The Cambridge Series on Electronic Publishing)*, Antoine Rizk, Norbert A. Streitz, and Jacques André (Eds.). Cambridge University Press, 194–211.