Beyond Named Citations: A Corpus of Resolved Cross-References in EU Legislation

Jakob Vokač
jakob.vokac@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Marko Grobelnik
marko.grobelnik@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Dunja Mladenić
dunja.mladenic@ijs.si
Jožef Stefan Institute
Ljubljana, Slovenia

Abstract

EUR-Lex and its repository CELLAR make EU legislation freely available and machine-retrievable, but the cross-reference metadata they expose for legislation is essentially document-level and incomplete. Many legal automation tasks instead need each reference resolved to the provision it points to. We introduce an annotation scheme that pairs each reference, as a span in an act's text, with its target provision, and a ground-truth corpus of about 4,500 annotations over four EU acts, built by manually correcting the output of a mostly deterministic extraction pipeline. Measured against this corpus, two maintained EU linkers, the Commission's Ref2Link and LegalViz.EU, resolve the references they find well but miss most mentions, in particular indirect references that do not name their target. We also fine tuned and test a RoBERTa-large model, on our corpus using a cross-validation approach. It reached a high F1 score for detecting references on the validation sets. The dataset and code are available at https://github.com/JakobVokac/EURLex_CrossReferenceDataset_SiKDD2026.

Keywords

legal informatics, EU legislation, named entity recognition, entity linking, annotated corpus

1 Introduction

This work is one part of a larger effort to convert EU legislation into formal representations, in three stages: a structural knowledge graph of the legal text, whose nodes are the act's structural units and whose edges are the references between them; an ontology derived from the text; and a first-order logic world model built from the graph as source and the ontology as grounding. The motivation is to explore how different representations of legal text help an LLM agent acting as a legal assistant produce trustworthy and traceable reasoning paths, following previous work on building world models with LLMs in the educational domain [20]. The first stage is the problem of resolving cross-references: a reference reading *point (a) of Article 6(1)* is not an edge in any graph, and cannot be formalized or retrieved on its own, until that phrase is replaced by an unambiguous address of the unit it designates.

EU legislation is a favorable setting. CELLAR, the repository behind the EUR-Lex database of European Union law, serves every act as Formex, the XML dialect EUR-Lex uses to structure legal texts, whose markup fixes the units a reference can address — article, paragraph, subparagraph, point, annex — so the target space is closed and a resolution can be scored exactly rather than judged. Formex does not, however, resolve the references themselves. Prior work resolves legislative references in other jurisdictions, and two maintained linkers do so for EU law; none produces what this effort needs (see Related Work). Near-total reliability is nonetheless reachable, because both halves of the problem are constrained: Formex gives a well-defined structure to address, and EU drafting follows codified citation conventions [11, 24]. What remains is a residual of forms those conventions do not cover.

While one might be tempted to simply use an LLM agent with the proper tooling for finding references and retrieving the appropriate pieces of text, however, that approach is arguably unsuitable for the scale of this task. Querying CELLAR shows EUR-Lex holds roughly 1 million legislation documents documents and further documents covering individual legal disputes that might be applicable to individual interpretations of a legal problems, and the four documents covered in this work span already roughly ~4.5 thousand references. As such one can reasonably assume the total number of references spanning EUR-Lex would be in the hundreds of millions. Finding, disambiguating and retrieving the span of text for each such reference through an LLM approach would likely be prohibitively expensive.

As such, a more scalable approach must be devised. The work described in this paper makes the following contributions towards this:

(1) **An annotation scheme** pairing each mention, as a character span in the act's parsed source text, with a canonical target — a CELEX key, the identifier for most documents on EUR-Lex, and a subdivision path in the style of the European Legislation Identifier (ELI) — and the conventions the hard constructions force: conjunction under a shared qualifier, expanded ranges, several targets on one span, unnamed subparagraph steps.
(2) **A ground-truth corpus** of ~4,5k annotations carrying ~5,5k targets over four acts (GDPR, DSA, AI Act, SSM Framework Regulation).
(3) **A construction method built for expansion**: a staged extraction pipeline, mostly deterministic with LLM assistance for edge cases covering 0.35% of references, whose output a human reviews and corrects until a ground truth is reached.

2 Related Work

No freely available dataset publishes references and targets as defined in Section 3: mentions anchored as spans in the act's own text, every occurrence annotated, each resolved to a subdocument address. We are aware of no published evaluation of any system that extracts them. Two rule-based linkers come closest, and both compute what this dataset annotates without releasing it. Measured against the ground truth (Section 3) introduced here, they resolve direct mentions well and largely miss

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).
Information Society 2026, Ljubljana, Slovenia
© 2026 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2026.sikdd.107

Table 1: Comparison of existing datasets and methods for annotations of references in legal text to other documents or parts thereof (cross-references). Corpus - what set of documents the work covers. Reference classes covered - types of references covered (internal, external, document vs. sub-document, etc..). Annotations released - resulting dataset size.

Resource	Corpus	Reference classes covered	Annotations released	Ref.
Annotated resources				
CDM / ELI metadata	EU, all of EUR-Lex	act-to-act citation edges	targets only, no mentions	[16, 23]
Legal NER corpora	DE, BR, IN, US, UK	reference strings, unresolved	8.8k–93k spans, no targets	[3, 13, 14, 17, 21, 27]
LegalNERo	RO national	named citations	3,387 spans, document-level ELI	[22]
Legal Linking	US constitutional	named and short-form cites	1,241 paragraphs, provision targets	[29]
Darji et al.	DE court decisions	statutory citations in judicial prose	2,944 records, subdivision targets	[7]
LegalPincite	EU case law (CJEU)	case- and paragraph-level citations	paragraph-level links	[25]
Extraction methods				
National resolvers	NL, LU, JP, GR, DE	named citations, internal structural	none	[1, 2, 4, 6, 8, 26, 30]
Ref2Link	EU, 24 languages	direct mentions; orphan citations listed as future work	none	[9, 10]
LegalViz.EU	EU, EUR-Lex	direct and bare internal; no anaphora, pro-forms or insertion context	none	[15]
This work				
Proposed corpus	EU legislation, 4 acts	direct **and** indirect: anaphora, pro-forms, defined short names, positional, insertion context	**4,402 spans, 5,452 sub-document targets**	

indirect ones (mentions that do not name their target in their own wording, but derive it from the surrounding context), which account for roughly 75% of our annotations. Table 1 shows the succinct picture.

EU citation networks draw their edges from cataloged metadata rather than from text [16, 23]. The Common Data Model (CDM) behind CELLAR does carry sub-document detail, but it is a case-law product: of 1.34M such annotations only 898 have a legislative citing work, and none of the 162 citation edges of the four acts studied here carries one. It is incomplete at document level too: its 64 outgoing citations for the AI Act reach 47 of the 187 distinct external acts the text cites.

Sub-document resolution from legal text with target resolution does exist, but only for national law [1, 2, 4, 6, 8, 26, 30]. All of it resolves by deterministic construction from the citation's surface form (instrument type, year, number), which by design never meets a reference whose target the wording omits.

Released datasets resolve references, or cover EU legislative text, but never both. The closest comparable work decomposes German references into 21 resolved fields, but is drawn from court decisions [7]; the larger legal named-entity-recognition (NER) corpora make a law or statute type their largest class [3, 13, 14, 17, 21, 27]. EU sub-document work exists on case law [25, 28], both excluding references to legislation, or hand-built [12]. The nearest adjacent work covers the same three regulations one layer up, extracting deontic obligations into graphs that presuppose resolved targets [5].

Two maintained linkers compute what this dataset annotates, and publish neither the spans nor an evaluation. Ref2Link (Commission Legal Service, EUPL-1.2) emits a CELEX, an ELI subdivision URL and a character offset [9, 10]. LegalViz.EU (Maastricht Law & Tech Lab, GPL-3.0) parses the same Formex source [15]. Tested against our dataset, both fall well short on detection, as Section 4 shows.

3 Annotation Scheme and Corpus

For our dataset annotation scheme we require clear definitions of what constitutes a reference and what its target is; together they fix exactly what the corpus annotates and what it leaves out.

Reference. A reference is a mention that explicitly designates one or more units of legislative text (a whole act, or a part, title, chapter, section, article, paragraph, subparagraph, point or annex within one) whether or not the wording identifies which unit; indirect mentions such as `that Directive` or `Article 4(2) thereof` are references too. We annotate the mention as a character span in the act's own parsed source text, and we annotate every occurrence, not only each distinct unit referred to. A mention extends across conjoined units only where a shared qualifier governs all of them: `Article 5 and Article 6 of that Directive` is one mention, since `of that Directive` distributes over both, whereas `point (a) of Article 6(1) or point (a) of Article 9(2)` is two, since each head carries its own qualifier. A mention begins at the unit designation, not at the words describing its content: in `the requirements laid down in Article 35(2) of Directive 2013/36/EU`, the mention is `Article 35(2) of Directive 2013/36/EU`.

Not references. A mention of content that designates no unit (e.g. `..the rules on data portability..`) is not a reference, nor is a mention of a body rather than a text, such as `the European Council`. We also leave out structural text such as labels, headings and tables of contents.

Target. The target is the canonical address of the unit designated: a CELEX key with an ELI-style subdivision path, as in

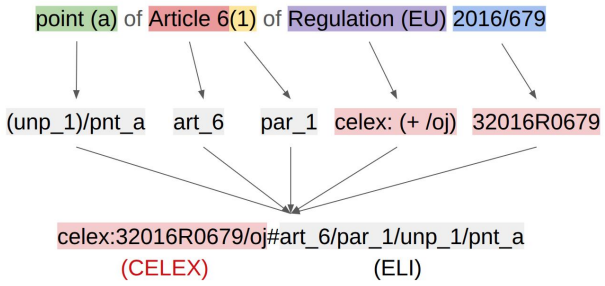

Figure 1: Illustrative example of a reference converted to a target.

Figure 1. The key identifies the act; `/oj` selects the version as published in the Official Journal rather than a later consolidated one; and the path names one structural unit per step, including steps the wording leaves implicit. A mention naming several units carries several targets on one span, and ranges are expanded. Instruments outside EUR-Lex are keyed to another authority, as `ext:std:iso-iec-17065`.

Corpus. The corpus covers four regulations — the GDPR, the Digital Services Act, the AI Act and the SSM Framework Regulation — carrying 4,509 annotated mentions and 5,452 targets (942, 1,167, 1,888 and 512 mentions per act, in that order); 451 mentions carry more than one target, and 5 targets are left explicitly unresolved. 4,402 mentions are anchored by character offsets; the remaining 107 were recorded by the reviewer without a span and are excluded from the span-level label set.

Construction. Annotation follows a recommend–revise scheme. A staged extraction pipeline proposes every candidate mention together with its target, and one human reviewer accepts, re-targets, rejects or adds, working in an interface that renders the act as published. The ground-truth is final pipeline independent file that correctly annotates all references in the review document.

Pipeline. The pipeline uses Formex as its input. Stage A parses Formex into a typed tree of citable units with ELI-style ids, and Stage B is a deterministic grammar over that immutable text; Stage C binds anaphora and pro-forms, and Stage D resolves footnote citations to a CELEX key. Every target any stage emits is validated against the registry built from the act's own parse, so a proposal addressing no unit that exists is never shown. Model calls are confined to two narrow jobs: binding bare anaphors, and one tier of footnote resolution.

Limits. There was only one annotator, who is not a legal expert, for the entire dataset. And because the reviewer works from proposals, a reference no rule proposed enters the corpus only if the reviewer spots it unaided; the 107 span-less annotations, 2.4% of the total, fall under that category.

4 Evaluation

Protocol. Every system is scored by one procedure against one gold set: the 4,402 mentions carrying character offsets, across all four acts. A gold mention counts as found when a system predicts at least one non-whitespace character inside it (boundaries vary by system and whether the mention span is fully identified is relevant, and expressed, in the target resolution accuracy). Identical prediction spans count once, and target accuracy is computed only over the mentions a system found, so detection and resolution do not confound. Neither linker has a notation for the unnumbered subparagraph step, so the subpath comparison tolerates it. One class of references, descriptive mentions, is held out of the scored export, and no figure here covers it.

Table 2: Mention (reference) detection (P - precision, R - recall, F1 - score) over the ground-truth references & target accuracy over the references each system found (where applicable).

System	P	R	F1	Target
Rule-based linkers				
Ref2Link	0.890	0.206	0.335	0.791
LegalViz.EU	0.765	0.406	0.531	**0.881**
Lexemo RoBERTa-large				
zero-shot	0.859	0.758	0.805	*N/A*
fine-tuned	0.946	0.993	**0.969**	*N/A*

Detector baseline. To test whether the corpus supports a learned fallback for acts no grammar has been written against, we fine-tune a token classifier on it. Lexemo's `roberta_large_legal_act_extraction` [18] is the only public checkpoint we found that extracts legal acts and articles from English EU legislative text; it is RoBERTa-large [19] and already task-adapted, so the zero-shot row is a strong baseline rather than a naive one. The original model had a 23-label head, which we collapsed to a single *detected* class. The fine-tuned model was trained and evaluation with cross-validation (3 training, 1 validation document split).

Reference detection. Fine-tuning on the corpus significantly improves the RoBERTa model performance. Mention F1 rises from 0.805 to 0.969. Additionally, fine-tuning the model raises the number of detected references that are correctly spanned from 13.5% to 94.5%. Ref2Link [9, 10] is an external-citation linker by design, and on external citations its recall (0.732) is ahead of LegalViz's (0.677). When accounting for all (internal and external) references in the proposed corpus, its recall drops significantly as shown in Table 2. This is because 72.5% of references in the corpus are internal. LegalViz [15] performs better on internal citations, yet still detects only 40.6% references overall.

Target resolution. On target resolution accuracy, both linkers are strong, and on external citations nearly identical: LegalViz 0.786, Ref2Link 0.789. LegalViz's higher performance in Table 2 is primarily due to its ability to detect internal references. The largest gap this corpus exposes for these two methods is on detection, not resolution.

Pipeline The dataset construction pipeline cannot be directly compared to any of the previous methods as it was built on the dataset documents themselves. The pipeline, when run on the corrected ground-truth gives: without the LLM feature — an F1 score of 0.985, and 0.999 target resolution; with the LLM feature — an F1 score of 0.986, and 0.998 target resolution.

5 Conclusion

We present an annotation scheme for references in EU legal documents, pairing each mention with a canonical target at sub-document granularity. Its purpose is to create a structural knowledge graph that keeps the text entire, connected along the references running within and between acts. We know of no existing resource that releases such annotations, or an evaluation over them.

We built a ground truth to that scheme by recommend–revise: a staged, mostly deterministic pipeline proposes every mention and its target, and a human reviewer accepts, corrects or adds. It spans four acts — GDPR, DSA, AI Act, SSM Framework Regulation — and holds 4,509 annotated mentions carrying 5,452 targets, available on [https://github.com/JakobVokac/EURLex_CrossReferenceDataset_SiKDD2026].

Measured against the corpus, the two maintained linkers find 20.6% and 40.6% of the annotated mentions, because indirect and act-internal references lie outside what they attempt; on the target resolution they both perform well. Fine-tuned on the corpus, a RoBERTa-large classifier finds 99.3% of the mentions in acts held out of its training. As such, given a large enough dataset, it can provide a better alternative for linking EUR-Lex legislation. That being said, a separate model would have to be trained specifically for target resolution, whose performance would have to at least match the existing linkers to make the entire approach a suitable alternative. Another caveat worth mentioning is that the documents selected are all relatively recent additions and mostly in the same area (apart from the SSM Framework Regulation). As such, the referencing styles are likely to coincide heavily, making overfitting a possible (at least partial) explanation for such good performance from the trained model.

Four acts are too few to predict how a method behaves across EUR-Lex, and the annotations carry one reviewer's judgment. Future work is to expand the corpus to a size on which a general method can be trained to an accuracy reliable enough for real users.

Acknowledgements

This research was financially supported by the Slovenian Research and Innovation Agency through the research program Knowledge Technologies (P2-0103) and the European Commission under the Horizon Europe project DataPACT, Grant Agreement No. 101189771. We would like to express our gratitude to all project partners for their contributions and collaboration.

References

[1] Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel C. Briand. 2014. Automated detection and resolution of legal cross references: approach and a study of Luxembourg's legislation. In *Proc. 22nd IEEE Int. Requirements Engineering Conf. (RE)*, 63–72.

[2] Iosif Angelidis, Ilias Chalkidis, and Manolis Koubarakis. 2018. Named entity recognition, linking and generation for Greek legislation. In *Legal Knowledge and Information Systems (JURIX 2018)*. IOS Press, Amsterdam, The Netherlands.

[3] Ting Wai Terence Au, Vasileios Lampos, and Ingemar Cox. 2022. E-NER — an annotated named entity recognition corpus of legal text. In *Proc. Natural Legal Language Processing Workshop (NLLP) at EMNLP*. Abu Dhabi, UAE, 246–255.

[4] Ilias Chalkidis, Charalampos Nikolaou, Panagiotis Soursos, and Manolis Koubarakis. 2017. Modeling and querying Greek legislation using semantic web technologies. In *The Semantic Web (ESWC 2017)* (Lecture Notes in Computer Science). Vol. 10249. Springer, Cham, Switzerland, 591–606.

[5] Thiago Raulino Dal Pont, Federico Galli, Galileo Sartor, and Giuseppe Contissa. 2025. "lost in EU regulation? Don't worry, AI found the obligation" — extracting and representing legal obligations in the GDPR, the DSA, and the AI Act. In *Proc. 20th Int. Conf. Artificial Intelligence and Law (ICAIL)*, 129–138. DOI: 10.1145/3769126.3769260.

[6] Harshil Darji, Martin Heckelmann, Christina Kratsch, and Gerard de Melo. 2026. Bundesrecht: an open library and corpus for German statutory reference processing. arXiv:2605.31338. Preprint. (2026).

[7] Harshil Darji, Jelena Mitrović, and Michael Granitzer. 2023. A dataset of German legal reference annotations. In *Proc. 19th Int. Conf. Artificial Intelligence and Law (ICAIL)*, 392–396. DOI: 10.1145/3594536.3595173.

[8] Emile de Maat, Radboud Winkels, and Tom van Engers. 2006. Automated detection of reference structures in law. In *Legal Knowledge and Information Systems (JURIX 2006)* (Frontiers in Artificial Intelligence and Applications). Vol. 152. IOS Press, Amsterdam, The Netherlands, 41–50.

[9] European Commission. 2021. Ref2Link. ISA^2 LegIT action 2016.38 presentation. [Online]. Available: https://interoperable-europe.ec.europa.eu/sites/default/files/document/2021-06/Ref2link_pres_1_7_short.pdf. Accessed: Aug. 27, 2026. (June 2021).

[10] European Commission Legal Service. 2025. Ref2Link. Kit v1.3.32, EUPL-1.2. [Online]. Available: https://interoperable-europe.ec.europa.eu/collection/justice-law-and-security/solution/ref2link. Accessed: Aug. 27, 2026. (Oct. 2025).

[11] European Commission, Legal Service. 2015. *Joint Practical Guide of the European Parliament, the Council and the Commission for Persons Involved in the Drafting of European Union Legislation*. Catalogue no. KB-02-13-228-EN-N. Print edition: ISBN 978-92-79-49121-4, doi 10.2880/89965. Publications Office of the European Union, Luxembourg. ISBN: 978-92-79-49084-2. DOI: 10.2880/5575.

[12] Nicola Fabiano. 2026. Qualified cross-references as a verification method: the normative environment of the EU AI Act. arXiv:2608.19194. Preprint. (Aug. 2026).

[13] Prathamesh Kalamkar, Astha Agarwal, Aman Tiwari, Smita Gupta, Saurabh Karn, and Vivek Raghavan. 2022. Named entity recognition in Indian court judgments. In *Proc. Natural Legal Language Processing Workshop (NLLP) at EMNLP*. Abu Dhabi, UAE, 184–193.

[14] Ioannis Karamitsos, Stefanos Roufas, Khalil Al-Hussaeni, and Andreas Kanavos. 2025. LegNER: a domain-adapted transformer for legal named entity recognition and text anonymization. *Frontiers in Artificial Intelligence*, 8, 1638971. DOI: 10.3389/frai.2025.1638971.

[15] Konrad Kollnig. 2026. LegalViz.EU. Law & Tech Lab, Maastricht University, GPL-3.0, `main` branch. [Online]. Available: https://github.com/maastrichtlawtech/legalviz.eu. Accessed: Aug. 28, 2026. (Aug. 2026).

[16] Marios Koniaris, Ioannis Anagnostopoulos, and Yannis Vassiliou. 2018. Network analysis in the legal domain: a complex model for European Union legal sources. *Journal of Complex Networks*, 6, 2, (Apr. 2018), 243–268.

[17] Elena Leitner, Georg Rehm, and Julián Moreno-Schneider. 2020. A dataset of German legal documents for named entity recognition. In *Proc. 12th Language Resources and Evaluation Conf. (LREC)*. Marseille, France, 4478–4485.

[18] Lexemo. 2024. `Roberta_large_legal_act_extraction`. RoBERTa-large token classifier, MIT licence. [Online]. Available: https://huggingface.co/Lexemo/roberta_large_legal_act_extraction. Accessed: Aug. 27, 2026. (2024).

[19] Yinhan Liu et al. 2019. RoBERTa: a robustly optimized BERT pretraining approach. *arXiv:1907.11692*.

[20] Mark David Longar, Erik Novak, and Marko Grobelnik. 2024. Creating local world models using LLMs. In *Data Mining and Data Warehouses (SiKDD), Proc. 27th Int. Multiconference Information Society (IS 2024), Volume C*. Ljubljana, Slovenia, (Oct. 2024). DOI: 10.70314/is.2024.sikdd.22.

[21] Pedro H. Luz de Araujo, Teófilo E. de Campos, Renato R. R. de Oliveira, Matheus Stauffer, Samuel Couto, and Paulo Bermejo. 2018. LeNER-Br: a dataset for named entity recognition in Brazilian legal text. In *Computational Processing of the Portuguese Language (PROPOR 2018)* (Lecture Notes in Computer Science). Vol. 11122. Springer, Cham, Switzerland, 313–323. DOI: 10.1007/978-3-319-99722-3_32.

[22] Vasile Păiș, Maria Mitrofan, Carol Luca Gasan, Vlad Ianov, Corvin Ghiță, Sergiu Coneschi, and Andrei Onuț. 2024. LegalNERo: a linked corpus for named entity recognition in the Romanian legal domain. *Semantic Web*, 15, 3, 831–844.

[23] Publications Office of the European Union. [n. d.] ELI for subdivisions: specifications for the identification of subdivisions in EU legislation using ELI. v2. [Online]. Available: https://eur-lex.europa.eu/content/eli-register/ELI-subdivisions-specifications-v2.pdf. Accessed: Aug. 27, 2026. ().

[24] Publications Office of the European Union. 2026. Interinstitutional style guide. [Online]. Available: https://style-guide.europa.eu/en/. Accessed: Aug. 31, 2026. Continuously revised; the site carries no version or revision date, so the access date is the only pin. (2026).

[25] Theresia Veronika Rampisela, Henrik Palmer Olsen, and Giovanni Colavizza. 2026. LegalPincite: multi-level legal information retrieval dataset. arXiv:2608.03756. Preprint. (Aug. 2026). DOI: 10.48550/arXiv.2608.03756.

[26] Nicolas Sannier, Morayo Adedjouma, Mehrdad Sabetzadeh, and Lionel C. Briand. 2017. An automated framework for detection and resolution of cross references in legal texts. *Requirements Engineering*, 22, 2, 215–237. Online-first Nov. 2015; issue 2017. DOI: 10.1007/s00766-015-0241-3.

[27] Holli Sargeant, Robert Östling, and Måns Magnusson. 2025. Detecting legal citations in United Kingdom court judgments. In *Proc. 2025 Conf. Empirical Methods in Natural Language Processing (EMNLP)*, 26810–26836.

[28] Giovanni Sartor, Piera Santin, Davide Audrito, Emilio Sulis, and Luigi Di Caro. 2022. Automated extraction and representation of citation network: a CJEU case-study. In *Advances in Conceptual Modeling (ER 2022 Workshops)* (Lecture Notes in Computer Science). Vol. 13650. Springer, Cham, Switzerland, 102–111. DOI: 10.1007/978-3-031-22036-4_10.

[29] Robert Shaffer and Stephen Mayhew. 2019. Legal linking: citation resolution and suggestion in constitutional law. In *Proc. Natural Legal Language Processing Workshop (NLLP) at NAACL*. Minneapolis, MN, USA, 39–44.

[30] Oanh Thi Tran, Bach Xuan Ngo, Minh Le Nguyen, and Akira Shimazu. 2014. Automated reference resolution in legal texts. *Artificial Intelligence and Law*, 22, 1, (Mar. 2014), 29–60. DOI: 10.1007/s10506-013-9149-8.