Automatic Transformation of Plain-text Legislation into Machine-readable Format

Andrija Cvejić*, Katarina-Glorija Grujić*, Aleksandar Cvejić*, Marko Marković*, Stevan Gostojić*

* University of Novi Sad, Faculty of Technical Sciences, Novi Sad, Serbia
{andrijacvejic, katarina.glorija, aleksandar.cvejic, markic, gostojic}@uns.ac.rs

Abstract — Legislative documents are an important class of legal documents that regulate almost every area of people's lives. The introduction of modern information technologies in the legal domain brings numerous benefits for the legal profession, but the lack of a machine-readable representation of legal documents brings some difficulties. This paper proposes a method for a regulation transformation from the plain text format into the machine-readable Akoma Ntoso format. The XML tagging process is implemented in three layers: metadata layer, structural layer and text layer. To apply structural and semantic markup, we use a combination of rule-based methods and methods based on neural networks. The output of our system is XML documents that comply with the Akoma Ntoso schema. These documents are technology-neutral and machine-readable.

Keywords: legislation; Akoma Ntoso; regular expressions; NER; TF-IDF

I. INTRODUCTION

Laws regulate many aspects of peoples' lives. Although legal documents are written in a natural language they are best interpreted by lawyers. With every enactment of a law, it increases pressure on the legal profession since lawyers are expected to work with a larger information set. In the last few decades, computer science has become mature enough to offer a variety of methods applied in the legal domain. Using software tools to support lawyers in their everyday work leaves more time for lawyers to deal with the legal profession while data processing is delegated to machines.

To process legal documents using computers, it is important that these documents are in a machine-readable format. When data in these documents are connected with their meaning then these data are machine-understandable enabling more advanced document processing. In this paper, we present a transformation method for legislative documents published in the Republic of Serbia into a machine-readable format for legal documents.

We choose the Akoma Ntoso (Architecture for Knowledge-Oriented Management of African Normative Texts using Open Standards and Ontologies) standard [1] for representing legislation in the machine-readable format. The Akoma Ntoso is adopted by OASIS Technical Committee [2] as a basis for the LegalDocML standard. The applicability of Akoma Ntoso for legal acts in Serbia is already explained in [3, 4, 5].

Akoma Ntoso standard is mainly oriented to the representation of legislation and court decisions. It supports the identification of legislative documents, facilitating their retrieval in different versions. Metadata within the Akoma Ntoso document can univocally identify the authorities producing them, procedural events, and references to other documents. Akoma Ntoso supports the representation of changes to the law so that textual modifications can be clearly identified.

We propose a method that uses legal documents in plain text or HTML format and transforms them to Akoma Ntoso 3.0 standard. The transformation involves markup at three separate layers: metadata, structural (hierarchical), and textual.

The rest of this paper is organized as follows. Section 2 explains the literature review on the transformation of legislation into machine-readable formats. The method proposed in this research is described in Section 3. Section 4 presents the main results. Finally, Section 5 provides conclusions and directions for future work.

II. RELATED WORK

This section reviews the published research related to the transformation of legal text into machine-readable formats.

In [6] the author emphasizes that possibility for the transformation of law into a machine-readable format is supported by the fact that regulations are highly formalized and algorithmically determined. The transformation of legal texts into a machine-readable and machine-executable format is suggested as an important step for law to become more effective.

The authors in [7] argue that most of the published legal documents are available in unstructured and presentation-oriented formats. It becomes an obstacle when the end-user needs to perform a search based on document semantics. The authors designed a rule-based method to transform Greek legal documents to Akoma Ntoso format. Using this method, the legal text of generated documents is supplemented with structural and semantic markup. The authors report high accuracy of the method in comparison to manually performed annotations. The main issue of the method is in the detection of annexes and tables.

In [8] the authors present the system for the detection of semantics in provisions found in legal text. The system extracts provisions from legal text and determine its type (e.g., definition, prohibition, obligation, permission, penalty) and extracts its arguments. The system uses algorithms based on machine learning and natural language processing. The results obtained processing Italian legislative documents shows that the system accuracy depends on the volume of a training sample, structural patterns in provisions, specific lexical and syntactic constructions etc.

Semantic Finlex is the data service for publishing Finnish legislation and case law as Linked Open Data [9]. It is performed by automatic annotation of legal texts using natural language processing techniques. Furthermore, personal data are automatically anonymized to protect the

privacy of parties in caselaw. The authors suggest the introduction of XML and RDF standards in the early phases of legislative and judicial processes to reduce costs and errors in further processing of legal documents.

The CLIEL (Commercial Law Information Extraction based on Layout) system enables automatic information extraction from commercial legal documents [10]. The system combines natural language processing with rule-based techniques to detect and annotate entities and structure in these documents. The system evaluation shows that the best results are obtained when the annotation process takes into account the layout information.

III. METHODOLOGY

The Akoma Ntoso standard defines a set of simple, technology-neutral electronic representations of legal documents. We chose this standard to represent legal acts in the machine-readable format. The standard uses separate layers to describe the legal text.

For using the Akoma Ntoso 3.0 schema [1] in this research, some customization is made to support Serbian laws. First, it is necessary to choose hierarchical elements from the standard that reflects the structure of Serbian regulations. Then it is necessary to represent relations between structural elements using the customized sub-scheme.

Table 1 proposes a mapping between structural elements of Serbian legal acts and Akoma Ntoso schema elements. An asterisk (*) next to a subpoint (*podtačka* in Serbian) indicates an exception because the subpoint is represented using an extensible `hcontainer` element with an obligatory attribute `name` containing the value `subpoint`.

TABLE I. HIERARCHICAL ELEMENTS

THE NAME OF THE HIERARCHY IN THE REGULATION (IN SERBIAN)	THE NAME OF THE ELEMENT IN THE AKOMA NTOSO STANDARD
DEO	*PART*
GLAVA	*CHAPTER*
ODELJAK	*SECTION*
PODODELJAK	*SUBSECTION*
ČLAN	*ARTICLE*
STAV	*PARAGRAPH*
TAČKA	*POINT*
PODTAČKA	*HCONTAINER**
ALINEA	*ALINEA*

Customization of the Akoma Ntoso schema is facilitated using the Akoma Ntoso schema generator [11]. The generator simplifies version updates and reduces the possibility of collision with the original schema. Besides hierarchical sub-elements provided by the Akoma Ntoso (e.g., `num`, `heading`, `subheading`, `content`), the schema we made allows additional hierarchical sub-elements shown in Table 2. Using the generated sub-schema, it is possible to automatically validate Serbian legal regulations in Akoma Ntoso format.

TABLE II. RESTRICTIONS ON SUBELEMENTS

Element name	Additional sub-elements
Part	*chapter*
Chapter	*section, article, intro*
Section	*subsection, article*
Subsection	*article*
Article	*paragraph, intro*
Paragraph	*point, alinea, intro*
Point	*hcontainer, intro*
Hcontainer	*alinea, intro*
ALINEA	*//*

To perform automatic conversion of legal regulations we use publicly available legislation as source data. Legal regulations are downloaded from the Legal Information System of the Republic of Serbia [12]. The documents are written in formal legal language, mostly following the Uniform Methodological Rules for Drafting Regulations of the Republic of Serbia [13]. The downloaded data consists of 1130 laws, 2667 decisions, 2933 ordinances and 1388 decrees. Meta information provided for each published document is also collected in this phase. The documents are preprocessed by removing HTML and CSS elements to extract the pure plain text.

Our method for the conversion of legislative documents into the machine-readable Akoma Ntoso format is based on the legal document layers. At the metadata layer, we combine a rule-based approach with machine learning (TF-IDF for detecting concepts and CRF for detecting dates, locations, organizations and persons). To annotate the hierarchical layer, we use token parsing and finite-state reasoner. Annotation of the references in the text layer is performed by a rule-based approach.

A. METADATA LAYER

The Akoma Ntoso standard clearly separates content and metadata, where metadata is described as a concept of adding "knowledge to knowledge". Although the metadata information is understandable to humans, it plays an important role in legislation search performed by a computer [14].

Akoma Ntoso uses the FRBR [15] (Functional Requirements for Bibliographic Records) standard to describe a legal document at four levels of abstraction making the document unambiguously recognizable. The first three layers are Work (an abstract concept of a document), Expression (a special version of an act that differs on some basis such as language or version), Manifestation (a specific data format in which the document is represented), and Item (a unique physical manifestation of the document). We decided not to use the Item level because of its too specific nature. Figure 1 shows an example of an Akoma Ntoso identification block.

```
<identification source="#somebody">
 <FRBRWork>
  <FRBRthis value="akn/rs/act/2009/36-3/main"/>
  <FRBRuri value="akn/rs/act/2009/36-3"/>
  <FRBRdate date="2009-01-01" name="Generation"/>
  <FRBRauthor as="#author" href="#somebody"/>
  <FRBRcountry value="rs"/>
 </FRBRWork>
 <FRBRExpression>
  <FRBRthis alue="akn/rs/act/2009/36-
3/srp@/main"/>
  <FRBRuri value="akn/rs/act/2009/36-3/srp@"/>
  <FRBRdate date="2009-01-01" name="Generation"/>
  <FRBRauthor as="#editor" href="#somebody"/>
  <FRBRlanguage language="srp"/>
 </FRBRExpression>
 <FRBRManifestation>
  <FRBRthis value="akn/rs/act/2009/36-
3/srp@/main.xml"/>
  <FRBRuri value="akn/rs/act/2009/36-3/srp@.akn"/>
  <FRBRdate date="2009-01-01" name="Generation"/>
  <FRBRauthor as="#editor" href="#somebody"/>
  <FRBRformat value="xml"/>
 </FRBRManifestation>
</identification>
```

Figure 1. Akoma Ntoso identification block

The information necessary for the formation of the identification block is the date of adoption, version and language of the act. The language code is written according to the ISO 639-2 standard for language marking [16], a "srp" is used regardless of Latin or Cyrillic alphabet [17].

Along with the legal documents themselves, which were downloaded from the Legal Information System website, metadata was downloaded for each document. The Akoma Ntoso in its schema defines elements that support the representation of these metadata. Table 3 shows how we paired Akoma Ntoso elements with the corresponding metadata.

TABLE III. METADATA OF LEGAL DOCUMENTS

Element	Mapping relevant information
publication	"government gazette" and "date of publication"
workflow	"Effective date of first legislation text", "Date of application" and "Date of adoption"
classification	"Type of regulation", "Area" and "Group"
notes	"Publisher notes" and "More information"

Besides identification and description of the legal document, the metadata layer provides a connection of entities found in the legal text with their meaning. To connect these entities with their semantic, concept recognition and named entity recognition are performed at the document level. As the result, we generate TLC (Top Level Class) references from Akoma Ntoso informal ontology. Table 4 shows TLC references that we use to describe entities mentioned in the text of a legal document.

TABLE IV. MEANING OF SELECTED TLC REFERENCES

Element	Meaning in the document
TLCConcept	It represents an abstract concept or idea.
TLCLocation	It represents a location, be it territorial, historical, geographical or geopolitical.
TLCOrganization	Represents the name of an organization or group of people who identify with a particular name.
TLCPerson	Represents the name of an individual.
TLCEvent	Includes time intervals and dates.

The tf-idf [18] (term frequency – inverse document frequency) algorithm is used to extract the most relevant terms in the document. The input for the statistical algorithm is the word vector. Serbian words that do not contribute to the meaning in sentences (stopwords) have been removed from the vector. The algorithm separately analyzes the articles in the document to determine the word frequency.

In this research, we choose to reject all terms that did not receive at least 0.1 tf-idf value. Also, we exclude the 50 most frequent concepts on a sample of 50 laws (concepts such as law, lawful, government, organization, process). The remaining concepts are included as the TLC references in the output document.

Another approach for detecting other TLC references is performed using the NER (Named Entity Recognition) algorithm, where the standard CRF (Conditional Random Field) model was used to predict label sequences. CRF is a representative of the method for supervised learning and therefore a set of labelled data in the Serbian language is needed. SETimes.SR [19] was selected as a dataset for the Serbian language. The dataset contains the annotated entities with the maximum number of tags around 11 thousand none-other tags (the IOB tagging system was used when annotating). In addition, we manually annotated dates in the dataset.

Input to the prediction algorithm requires POS (part-of-speech) tags for new words, which is why the ReLDI tokenizer and tagger are used [20].

The LGBFS algorithm (Gradient descent using the L-BFGS method) is used for implementation. Figure 2 shows the result of training on the selected data after validation (5-cross-validation was used).

	precision	recall	f1-score	support
B-date	0.93	0.93	0.93	681
B-deriv-per	0.84	0.89	0.86	75
B-loc	0.89	0.93	0.91	2678
B-misc	0.65	0.26	0.37	298
B-org	0.87	0.83	0.85	1953
B-per	0.90	0.92	0.91	1884
I-date	0.97	0.95	0.96	427
I-loc	0.82	0.76	0.79	459
I-misc	0.66	0.33	0.44	488
I-org	0.78	0.71	0.74	1416
I-per	0.91	0.96	0.94	1161
micro avg	0.87	0.84	0.86	11520
macro avg	0.84	0.77	0.79	11520
weighted avg	0.86	0.84	0.85	11520

Figure 2. Result of CRF training

Randomized CV Search is used to optimize the hyperparameters (c1, c2) of the CRF model.

B. HIERARCHICAL LAYER

Legislative documents have a hierarchical structure that should be automatically annotated by the conversion process.

A preamble, an introductory part and a title can be found at the beginning of legislation [13]. However, not all legislation shares this structure (e.g., laws do not have a preamble). For some of the parts of the legislation, there are specific elements defined by the Akoma Ntoso schema. The content of the preamble is placed in the `preamble` element, while the `longTitle` element contains the title of the regulation. All other information is placed in the sub-element `p` of the `preface` element as shown in Figure 3.

The main part of the regulation consists of elements in a hierarchical (parent-child) relationship. Available classification units arranged from the largest to the smallest are part, chapter, section, subsection, article, paragraph, point, subpoint and alinea [13]. The basic classification unit of the hierarchy of regulation is an article, where it is necessary for each article to have a unique number. Therefore, there are logical units that are larger than an article (part, head, section, subsection) that serve only for the meaningful separation of content. Similarly, there are the smaller classification units (point, subpoint and alinea) that serve as structural support for better decomposition of the article. For each classification unit, we choose a corresponding element in the Akoma Ntoso schema. For recognition of each classification unit, a REGEX template is created.

```
<preface>
  <longTitle>
    <p>Уредба о утврђивању изворника
Великог и Малог грба, изворника заставе и
нотног записа химне Републике Србије
„Службени гласник РС”, број <ref
wId="ref0"
href="akn/rs/act/2010/85/srp@">85 од 15.
новембра 2010</ref>.
    </p>
  </longTitle>
</preface>
<preamble>
  <p>Влада доноси на основу члана 9. став
1. Закона о изгледу и употреби грба,
заставе и химне Републике Србије<ref
wId="ref1"
href="akn/rs/act/2009/36/srp@">(„Службени
гласник РС”, број 36/09</ref> и члан 42.
став 1. Закона о Влади („Службени гласник
РС”, бр. 55/05,)<ref wId="ref2"
href="akn/rs/act/2005/71/srp@">71/05</ref
> <ref wId="ref3" href=
"akn/rs/act/2007/101/srp@">- исправка,
101/07</ref>и<ref wId="ref4" href=
"akn/rs/act/2008/65/srp@">65/08</ref>
  </p>
</preamble>
```

Figure 3. Preface and preamble of the regulation in the Akoma Ntoso format

C. TEXT LAYER

Recognition of references is performed using REGEX templates. Initially, the templates were written on the basis of Article 36 of the Unique Methodological Rules for Drafting Regulations. However, we noticed that there is a large number of cases where referencing does not follow these rules. To overcome this issue an approach using machine learning techniques is attempted, but the results were not satisfactory. This is caused by the variability of the word forms found in the text (gender, number, use of abbreviations, etc.). For this reason, REGEX templates are selected for the implementation of this task, but this time taking into account the variety of referencing styles found in the sample data. The examination of patterns is performed manually, analyzing different types of references in legal acts. Templates are divided into 4 groups:

1. Reference of an article, paragraph or point within the same legal act,
2. Reference of an article that is in another legal act,
3. Reference of the government Official Gazette,
4. Reference of a range of articles, paragraphs, or points.

References are formed according to the Akoma Ntoso referencing style [1]. The general form of references at the level of FRBR manifestation is given in Figure 4. The reference consists of the following elements: **akn** (indicates the Akoma Ntoso reference), **<CO>** (country according to the ISO 3166-2 standard), **<TY>** (type of the legal document), **<STY>** (subtype of the legal document),

<ACT> (actor i.e. the name of the institution that proclaimed the act), **<DATE>** (the date of publication in YYYY-MM-DD format), **<N-AC>** (ordinary number of the act in a year), **<LG>** (language of the regulation according to the ISO 639-2 standard), **<V>** (version of the document), **<PDOC>** (part of the document where "main" refers to the main part of the document), **<EXT>** (optional indication of the type of the document, e.g. .pdf, .doc, .rtf), **<RDOC>** (reference to an element within the document).

```
akn/<CO>/<TY>/<STY>/<ACT>/<DATE>/
<N-AC>/<LG>@<V>/<PDOC><EXT><RDOC>
```

Figure 4. The general form of reference

Figure 5 shows examples of different types of references in the Akoma Ntoso format. In the first example, a single point of a legal act is referenced. The second example is referencing a range of articles using the operator "->". The third and the fourth example are referencing enumerated articles. As can be seen, references are formed differently if the range is specified using the preposition "to" ("до" in Serbian), or by enumerating the elements. When forming references, it is unclear if two or more elements are successively arranged in the referenced legal act (for example, between Article 41 and Article 42, an Article 41a could exist). For this reason, references are formed separately for each element and only if the preposition "to" or the sign "-" is found the reference is treated as a range.

```
1)<ref wId="ref152"
href="akn/rs/act/2013-02-06/2013-104-
4477/srp@/!main~art_87__para_1__point_6">
члан 87. став 1. тачка 6</ref>
2)<ref wId="ref13" href="akn/rs/act/2013-
02-06/2013-104-nn/srp@/!main~art_7-
&gt;art_11">чл. 7 до 11.</ref>
3)<ref wId="ref8" href="akn/rs/act/2013-
02-06/2013-104-nn/srp@/!main~art_41">чл.
41</ref>
4) <ref wId="ref9" href="akn/rs/act/2013-
02-06/2013-104-nn/srp@/!main~art_42">и
42</ref>
```

Figure 5. Example of different types of references

IV. Results and Discussion

This section presents the results obtained using the proposed method to transform legal documents into the Akoma Ntoso format.

The accuracy of document conversion is calculated in comparison with manually annotated documents. For this purpose, 18 legal acts are randomly selected and manually transformed into the Akoma Ntoso format.

The validation is performed on every layer of the legal document. The similarity of textual layers in generated documents and in manually annotated documents is calculated as string similarity using Ratcliff's algorithm [21]. The accuracy of hierarchical structure in the generated documents is evaluated using DOM parser and expressed by the F1 score. The accuracy of the metadata is obtained by string comparison and expressed using the F1 score. The results are shown in Table 5.

TABLE V. RESULTS

Element	F1 Score	Layer
Hierarchy structure	0.95	Hierarchy
Metadata information	0.72	Metadata
References	0.76	Text
Whole document	0.82	Hierarchy, Text, Metadata

The highest score is achieved for the validity of the hierarchy structure. The textual layer and the metadata layer obtained a lower score. It might be an issue if the searchability of the legal corpus as a whole is the main goal.

Some improvements could be made in the application of Named Entity Recognition. The model is trained using the corpus created from sentences used in common speech, but for application on a legal text, the model should be trained using a legal regulations corpus.

After analysis of the results, some obstacles are identified that make the generated documents differ from the expected output.

On the hierarchical level, most issues arise from the inconsistency of legislative documents with the rules for regulations drafting. The legal acts published before the Unified Methodological Rules for Drafting Regulations were pronounced does not follow a uniform style. There are also variations in the labelling of the document parts (e.g., missing punctuation marks, ambiguities between paragraphs and titles, diverse numbering styles of sections and points). One more issue is the processing of the Agreement documents because these regulations are written in a foreign language.

The accuracy of the generated metadata layer depends on the characteristics of the legal text. Generally, the NER sometimes fails to differentiate between organization and location, but efficiently detects details of the entity. Date category is also error-prone when dealing with incomplete date formats found in legal text. Currently, only full dates are supported. This particular problem might be solved with a specialized date detection model.

Considering the references in the text layer, there are some overlaps between the TLC references. It occurs when an entity is not recognized as a whole, producing multiple partial references instead of a single reference.

V. Conclusion

In this paper, we propose the method for the automatic conversion of legislative documents in plain text into the Akoma Ntoso format. The system for automatic annotation of legal documents is developed and evaluated by comparing generated documents and manually transformed documents.

It is concluded that compliance between the legislation and the Unified Methodological Rules for Drafting Regulations plays a major role in the accuracy of the obtained results. Because some regulations do not comply with these rules, the system is not able to correctly annotate these documents. Better results could be obtained if the number of legal acts that follow these rules increases. The accuracy of the generated Akoma Ntoso documents shows that the initial phase of the conversion can be performed automatically. The current results do not indicate that the whole legal corpus can be transformed without any assistance from human experts.

The system could also be used as a tool in the legislation drafting phase to automatically produce machine-readable documents. This could improve the searchability of legislation, automatic processing of these documents and interoperability of legal information systems.

The extraction of features with the neural network can be improved by using recent transformer models, trained for the Serbian language, that can better handle sequence dependencies. In further research, different algorithms for text and sequence embedding should be tested and parser improvements could be made.

ACKNOWLEDGMENT

This research was supported by the Science Fund of the Republic of Serbia, Program DIASPORA, #6497472, AI4Law.

REFERENCES

[1] OASIS, "Akoma Ntoso version 3.0 XML schema", 2018, available at: http://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part2-specs/schemas/akomantoso30.xsd (accessed: April 2021)

[2] OASIS LegalDocumentML TC, 2020, available at: https://www.oasis-open.org/committees/legaldocml (accessed: April 2021)

[3] M. Marković, S. Gostojić, Z. Konjović, and M. Laanpere, "Machine-readable identification and representation of judgments in Serbian judiciary", *Novi Sad J. Math*, 44(1), 2014, pp. 165-182.

[4] M. Marković and S. Gostojić, "A knowledge-based document assembly method to support semantic interoperability of enterprise information systems", *Enterprise Information Systems*, 2020, pp. 1-20.

[5] M. Marković, S. Gostojić, and Z. Konjović, "Structural and semantic markup of complaints: Case study of Serbian Judiciary", In *2014 IEEE 12th International Symposium on Intelligent Systems and Informatics (SISY)*, pp. 15-20, IEEE, September 2014.

[6] I. Ponkin, "The concept of machine-readable and machine-executable law: relevance, purpose, place in RegTech, content, ontology and prospects", *International Journal of Open Information Technologies*, 8(9), 2020, pp. 59-69.

[7] M. Koniaris, G. Papastefanatos, and Y. Vassiliou, "Towards automatic structuring and semantic indexing of legal documents", In *Proceedings of the 20th Pan-Hellenic Conference on Informatics*, pp. 1-6, November 2016.

[8] C. Biagioli, E. Francesconi, A. Passerini, S. Montemagni, and C. Soria, "Automatic semantics extraction in law documents", In *Proceedings of the 10th international conference on Artificial intelligence and law*, pp. 133-140, June 2005.

[9] A. Oksanen, J. Tuominen, E. Mäkelä, M. Tamper, A. Hietanen, and E. Hyvönen, "Semantic Finlex: Transforming, publishing, and using Finnish legislation and case law as linked open data on the web", *Knowledge of the Law in the Big Data Age*, 317, 2019, pp. 212-228.

[10] M. García-Constantino, K. Atkinson, D. Bollegala, K. Chapman, F. Coenen, C. Roberts, and K. Robson, "CLIEL: Context-based information extraction from commercial law documents", In *Proceedings of the 16th edition of the International Conference on Artificial Intelligence and Law*, pp. 79-87, June 2017.

[11] University of Bologna, "AKN4UN Subschema Frontend", 2018, available at: http://akn.web.cs.unibo.it/akgenerator (accessed: April 2021)

[12] Official Gazette RS, "Information System of the Republic of Serbia", 2021, available at: http://www.pravno-informacioni-sistem.rs/ (accessed: April 2021)

[13] Legislative Committee of the National Assembly of the Republic of Serbia, "Unified Methodological Rules for Drafting Regulations", *Official Gazette of the RS*, no. 21/2010, March 30, 2010.

[14] G. Barabucci, L. Cervone, M. Palmirani, S. Peroni, and F. Vitali, "Multi-layer markup and ontological structures in Akoma Ntoso", In *International Workshop on AI Approaches to the Complexity of Legal Systems*, pp. 133-149, Springer, Berlin, Heidelberg, September 2009.

[15] T. B. Hickey, E. T. O'Neill, and J. Toves, "Experiments with the IFLA functional requirements for bibliographic records (FRBR)", *D-Lib magazine*, 8(9), 2002, pp. 1-13.

[16] J.D. Byrum, "ISO 639-1 and ISO 639-2: International Standards for Language Codes", ISO 15924: International Standard for Names of Scripts, 1999.

[17] F. Vitali, M. Palmirani, and V. Parisse, "Akoma Ntoso naming convention version 1.0. OASIS Committee specification draft 03 / public review draft 03", April, 2017, available at: http://docs.oasis-open.org/legaldocml/akn-nc/v1.0/csprd03/akn-nc-v1.0-csprd03.html (accessed: April 2021)

[18] J. Beel, B. Gipp, S. Langer, and C. Breitinger, "Research-paper recommender systems: a literature survey", *International Journal on Digital Libraries,* 17(4), 2015, pp. 305–338, doi: 10.1007/s00799-015-0156-0

[19] V. Batanović, N. Ljubešić, T. Samardžić, and T. Erjavec, "Training corpus SETimes.SR 1.0", *Slovenian language resource repository CLARIN.SI*, 2018, available at: http://hdl.handle.net/11356/1200 (accessed: April 2021)

[20] N. Ljubešić, F. Klubička, Ž. Agić, and I. P. Jazbec, "New inflectional lexicons and training corpora for improved morphosyntactic annotation of Croatian and Serbian", In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pp. 4264-4270, May 2016.

[21] J. W. Ratcliff, and D. E. Metzener, "Pattern-matching-the gestalt approach", *Dr Dobbs Journal*, *13*(7), 1988, pp. 46-51.