EU-WIDE LEGAL TEXT MINING USING BIG DATA PROCESSING INFRASTRUCTURES

Ongoing Research paper

Track New Directions for Digital Governance: Towards Government 3.0

Michalis-Avgerinos Loutsaris, University of the Aegean, GR, mloutsaris@aegean.gr

Aggeliki Androutsopoulou, University of the Aegean, GR, ag.andr@aegean.gr

Yannis Charalabidis, University of the Aegean, GR, yannisx@aegean.gr

Abstract

In their effort to keep pace with the rapid evolution of technologies, governments are entering the era of Government 3.0, through the provision of novel services to citizens, businesses, and administrations. On the other hand, society is overwhelmed by the large amount of legal provisions, they have to comply with, Access to constantly changing legal information across the countries of European Union remains fragmented. Although significant advancements in the 'legal informatics' research field are observed, there is no system capable of acquiring, storing and processing the large amounts of legal information, at an advanced level in various languages. To address this gap, the current paper presents a proposed novel framework and an ICT architecture for the introduction of a set of services for citizens, businesses, and administrations of the European Union, built upon text mining, advanced processing and semantic analysis of legal information. To identify the desired services two workshops have been organised in collaboration with experts from the Greek and Austrian Parliament respectively. A set of tools integrated into the proposed architecture will access to big legal data currently produced and published in multiple national or EU public data sources (e.g EUR-Lex, NOMOS), link it and transform it to structured open datasets. By combining data coming from a multiplicity of sources, our framework aims to achieve seamless and inclusive access to legal information across EU and improve the efficacy of decision making in legislative procedures operated by public bodies.

Keywords: legal text-mining, legal information system, gov 3.0, big open legal data.

1 Introduction

The rule of Law is a cornerstone, a fundamental foundation of every democratic state and thus legal framework of the EU is a vital tool for achieving the vision set in the way to the future. The operation of Europe as a well-functioning Digital Single Market, where Europeans are able to move among the EU member states, live, work and exploit new business opportunities, prerequisites that all EU citizens can be easily informed in the legal and policy framework that exists in each individual country and across EU. Although society is overwhelmed with an overload of legal information, only legal experts can follow the latest legislation and case law produced by parliaments and courts on a national and on a European level. A large amount of information about laws that apply in the EU countries currently remains fragmented across multiple national databases, inaccessible systems, mainly consisting of documents (legislation acts, bills, case laws, resolutions, decisions), published in each Member States' language. On the other hand, in the era of the World Wide Web and the immense data (and information) availability, we are witnessing a technical, social, and societal revolution.

Mass customization tools can help to filter and thereby reduce the flood of legal information and make it easier to be followed even for citizens without legal expertise. Despite the rapid evolution of

technology to date, there is no system capable of acquiring, storing, integrating and processing such large amounts of legal information, at an advanced level in various languages, using the power of text mining, information processing, and visual analytics. The basic reasons for this are the considerable complexity of the law, the variety of legal systems.

The large amount of legal information in the countries of the European Union is a challenge of research and study, when aiming at facilitating legal issues and providing novel services to citizens, businesses and administrations. Despite the rapid evolution of technology to date and the appearance of Government 3.0 (Pereira et al., 2018) which refers to the use of emerging technologies (such as big data, artificial intelligence technologies), there is no system capable of acquiring, storing and processing such large amounts of legal information, at an advanced level in various languages, using the power of text mining, information processing and visual analytics. Searching across the national legal corpora in different languages, interrelating the European Parliament communications and directives with the national legal frameworks, comparing national laws which target the same life events, are some of the services that cannot be provided today at European level, with the current state of the art.

In this paper, we attempt to tackle the above challenge by making the following two contributions: i) we propose a framework for legal text mining using big data and ii) we design an ICT architecture for facilitating and implementation of the proposed framework The proposed method aims to address the challenge of fragmented information in the legal domain, by delivering a set of key services that facilitate seamless and ubiquitous access to legal data to citizens, businesses and administrations built upon the application of the aforementioned ICT methods, the integration of automated translation services and the utilization of HPC resources.

In the following Section 2 the methodology underlying our research is presented. The next section describes the major concepts and existing infrastructures of legal text mining domain. Section 4 proposes a framework for legal text mining using big data, followed by an ICT architecture needed to instantiate the above framework. The last section of the paper outlines concluding remarks and future work directions.

2 Research Methodology

This chapter presents the methodological approach followed for the definition of the proposed framework consisted of the following three steps. As a first step, we conducted a literature review on the field of "Legal Informatics" in order to identify established methods and tools to reuse, as well as research directions and gaps to be addressed. In the next step of our methodology, we performed desk research in order to identify relevant existing projects and infrastructures that we can built upon for the definition of our approach. Each project has been analysed and services have been identified, while particular focus has been placed on the identification of the two countries involved in the research, i.e. Greece and Austria.

In the final step. we had close co-operation for addressing the abovementioned problem, and the creation of new novel services for interacting with citizens, with higher levels of information 'richness', enabling the citizens to facilitate their legal issues, problems and needs with higher expressiveness through full text in their everyday language, with two European Parliaments: the Hellenic Parliament and the Austrian Parliament. With each of these government agencies we organized a workshop, in which several (between 4 and 8) experienced staff in e-government services provision, of about 2 hours duration. In each of these workshops initially we explained to them the basic idea, and then we asked them for their opinions about its feasibility and usefulness, as well as to elaborate this idea for their specific context. Then we collaboratively developed with them specific application scenarios of this idea for their government agencies, leading to specific advanced legal information e-services that can be useful for citizens and businesses. In the second step, we conducted a literature review that enabled us to assemble the basic definitions and characteristics of legal informatics, legal text mining, graph-based

visualizations, and multilingualism. Guided by the research papers, the next step of our methodology consists the identification of existing infrastructures relevant to legal text mining. Based on the material we collected in these workshops we designed in detail the Framework for Legal Text Mining using Big Data presented in Section 4, as well as the ICT platform architecture presented in Section 5.

3 Theoretical Background

3.1 Literature Review

Legal Informatics refer to the application of Information Technology within the context of legal environment (Erdelez et al, 1997) and is defined by Sartor and Francesconi (2010) as the «theory and practice of computable law, i.e. the cooperation between humans and machines in legal problem-solving». The area focuses on the opportunities and challenges faced in the legal system and thus involves all related organisations and information users within the legal domain. One of these challenges lies in the supply of legal services, which are currently under-consumed by individuals and companies (Hornsby, 2009). Therefore, the latest advancements in the legal informatics are targeted towards making services more open and promoting access to legal resources. During the last three decades, a number of online consultation services and knowledge systems have been appeared (e.g. FindLaw, LegalMatch, LegalZoom, Shake). However, with the advent of semantic web, more advanced capabilities have emerged that can accommodate professional, personal, organisational and business needs. In particular semantic interoperability can achieve access to heterogeneous data sources and multilingual information that can be processed by machines and is understandable by humans, the two basic elements of the field definition as stated above. Thus, it offers a unique opportunity for building unified services on aggregated data for the whole EU legal industry. Another research stream in this interdisciplinary field focuses in designing systems for better legal practice that can empower all parties in the legal system. This is also related with the knowledge acquisition issue described above, but yet includes the aspect of the human effort required for the effective implementation of law. Towards this direction, various legal information systems are built for supporting legal activities, legal document management, legal case and process management (e.g. Informer, Case Tracker, Courtnet). The aforementioned systems focus more on the ex-post implementation of law rather than on the design process carried out by administrations and regulatory bodies.

The standard text analysis pipeline performs several levels of analysis: morphological, syntactic, semantic, and discourse (Lacity & Janson, 1994).The morphological and syntactic analysis is usually performed with a syntactic parser (Lagos et al., 2017), which recognizes the syntactic word classes such as nouns and verbs, and the syntactic dependency structure of the constituents of a sentence (main verb, subject, object, etc.). In general language processing, recognizing the basic semantic roles of a sentence's constituents, i.e., the "who", "does what", "where", "when", and "how" constituents, is a well-established task for English. Coreference resolution is identifying when two mentions of an entity or event refer to the same underlying person, place, thing or event in the real world.

Semantic role labeling and event extraction focus on the extraction of propositional meaning (Christensen et al.,2011), that is, making assertions about the world that can be true or false. Non-propositional meaning conveys aspects of meaning that do not have a truth-value or where we cannot infer its truth value (modality) or that change the propositional meaning (negation) or are a combination of both. Textual entailment in natural language processing is seen as a directional relation between text fragments. Textual entailment methods recognize, generate, or extract pairs of natural language expressions, such that a human who reads (and trusts) the first element of a pair would most likely infer that the other element is also true. We think that the entailment approaches that use vector space models of semantics have value in discovering discourse relations.

Humans are very good in inferring causal relations in a discourse (e.g., recognizing that the arrest of a murderer by the police is the consequence of a murder), but for a machine, which lacks the world

knowledge that humans use in this process, this is not so obvious. The text might explicitly mention a causal relation between two events mentioned (e.g., the use of the word "because"), but often such cues are missing. To acquire knowledge on event causality automatically, current work focuses on novel methods developed in distributional semantics (Blei et al., 2003; Tomas et al., 2013; Scott et al., 2012; Baroni et al., 2014).

Another interesting line of research links the neural network architectures to legal reasoning, in which neural networks are used as a parallel computational model for argumentation and allow to combine argumentation, quantitative reasoning and statistical learning (Franca et al., 2014). Finally, it needs to be investigated whether word patterns could be translated into latent variable concepts, which would support current interest in the use of factors in legal texts.

Achieving shared conceptualisations of law is difficult in any legal system. The problem is confounded in Europe, which is increasingly governed by multiple jurisdictions. The principle of subsidiarity means that Member States are obliged to achieve the objectives of EU Directives, but have some flexibility in how they do so. This inevitably leads to differences among EU norms and various national norms. The term “law visualisation” can be understood as the visual communication of legal standards and/or their interrelations with the overall goal to present complex legal issues in a comprehensible manner. Law visualisation intends to support experts as well as non-experts in law during their assessment and analysis of legal aspects. Scientific approaches addressing the field of law visualisation can be mainly found in law faculties (Brunschwig, 2001). In addition, the visual representation of legal standards and their interrelations can be increasingly found in law theory (Röhl & Ulbrich, 2007). Therefore, specific data bases are developed and applied. An approach that goes beyond the visual but static representation can be found in “Fallnavigator”. Undoubtedly, for the great majority of people living in Europe, legal enactment is an inscrutable, complex process. Existing approaches for the interactive visualisation of legal standards and legal issues mainly apply graph-based visualisation for the presentation of relations.

3.2 Review of existing infrastructures

OpenLaws.gr (Garofalakis et al., 2016) constitutes a semi-automatic system for the analysis of Greek law texts, the consolidation of each law version by automatically applying modifications and their publication in a revision control system. Although, manual steps are required in order to correct law texts errors (e.g. syntax errors), it is capable of parsing xml documents used for the markup of law texts by analyzing the text and processing natural language in order to perform structural analysis and identification of the structural elements. Pattern matching occurs at the level of paragraphs, so paragraphs are distinguished in amending and non-amending. According to the regular expression that returns a match in each case, it is capable to know in which place there is the necessary information to perform the modification: modification category (addition, substitution or deletion of a text portion), cited law, referenced element, text to add or replace or text to delete.

A similar but cross-country approach is Openlaws.eu (Lampoltshammer et al., 2015), a compliance and legal information system in terms of adherence to legal regulations, regulatory standards and the fulfillment of other standards and requirements, which a company voluntarily agree to. The system integrates legal databases from Austria, Germany and the EU and makes it possible to search multiple databases at once. Openlaws.eu is a network of legislation, case law, legal literature and legal expert which automatically collects data from different sources, specifically EUR-Lex, Legal Information of the Republic of Austria and all of the German federal laws which provided from the German Federal Ministry of Justice and Consumer Protection. Users can use highlighting to emphasize what’s actually relevant from unstructured text, personal folders to collect laws that interest them and automatic search to constantly monitor specific terms and topics.

Another initiative in Greece is the PeriNomou system, developed by University of the Aegean, constituting an automated legal information system for analyzing Greek legislation and providing an easy and quick access to Greek laws for various scientific disciplines. Moreover, it is capable of encoding automatically Greek legislation in order to provide as much information as possible about Greek law texts.

The main components of PeriNomou system are the ability to extract data from laws such as correlations and vocabulary and to modify structural elements of Greek Laws (such as articles, paragraphs, subparagraphs etc.). Specifically, including a trained OCR (optical character recognition) tool, in order to process scanned law texts, converts pdf to text documents and by using the power of text mining, legal texts can be automatically analyzed. By analyzing legal texts and processing Greek Language, the system can identify the main content of any legal text providing it to its users as tags. Users have the ability to search Greek legal texts by vocabulary (such as specific words) or by unique number.

NOMOS Legal Information Data Base constitutes an innovative and all-inclusive service of valid, reliable and prompt legal information, constantly enriched and updated. It is primarily addressed to the legal, technical and financial circles of Greece, the broader public sector, the private companies, legal entities of public and private law, freelancers and the general public. NOMOS' main features are: convenience in use, quality and adequacy of information and multiple ways of quick searching, tracing and retrieval of legal information.

4 A Framework for Legal Text Mining using Big Data

In this section, we propose a framework which builds the proper environment and vision of semantically annotated Big Open Legal Data (BOLD), easily searchable and exploitable with proper visualization techniques. The ultimate objective is to provide the technical foundation and the tools for making legal information available to everybody, in a customizable, structured and easy to handle way. The framework concept is structured over three axes (Data, Processing, Services), as following (Figure 1):

4.1 Data

The information to be acquired, through web sources only, and primarily through web services communication where available, contains all the legal artefacts published by the European Parliament, the European Commission, the EU Council (EURlex, EUDOR), local parliaments, at national databases, in English and /or other languages (e.g. NOMOS). Furthermore, the information also contains news published in EU member states, concerning legal events (e.g. EU directive publication), other administration-generated content (e.g. local communications, regulations) and other citizen-generated relevant content (e.g. blogs, newsletters, social media posts).

We estimate that the above database will contain more than 1 trillion words in 21 different languages, corresponding to about 10 million “volumes” of classical books, when another 5,000 such “volumes” are added for study, on a daily basis.

4.2 Processing

The information processing stage of the infrastructure make use of massively parallel computing tools, balancing the load between batch and real-time service modes and contains two stages are envisaged: i) The Pre-processing stage is responsible to prepare data for text mining, ii) In the second stage, the data are converted to structured data using text mining techniques in order to offer a variety of services.

4.3 Services

Services are to be provided towards citizens, businesses and administrations, based on the most common needs of each user type. A thorough analysis of such needs within the project will al-low the proper selection and targeting of services to be developed. Through a user interface sup-porting simplification or advanced usage, the following services are to be provided at real time:

• Parallel search in many EU member-state legal frameworks (through parallel translation of search terms), using simple keywords

• Assessment of the degree of transposition of an EU Directive in a National Legal Framework

• Analysis of references to the European Legislation by National Laws

• Comparative analysis of equivalent or relevant laws from different EU member states or from the same member state

• Timeline analysis for all legal elements, visualising the progress and current status of a spe-cific national or European legislation (after amendment/extensions) over time including pre-paratory acts and agreements

• Interrelation of laws and news or social media posts, including sentiment analysis

• Various geo-related visualisations (e.g. EU maps indicating different parameters), text-related visualisations (e.g. wordle, sentiment graphs) and other common visual aids (e.g. graphs, charts etc.)

• Visualizations of correlations, dependencies and conflicts between different laws

• Decision Support Services (e.g. Impact Assessment) within legal procedures

Based on the above services and sub-products, a variety of add-on services can be developed after capturing new requirements from citizens, businesses and administrations.

5 The proposed ICT Architecture

The proposed ICT architecture supports the integration of the pool of identified services while keeping the structure flexible allowing the inclusion of further services and data sources. A layered approach supporting the data flow, from source data to visualised outputs is to handle the large volumes of data. The following figure presents the high-level IT architecture along with the indicative components to be incorporated.

5.1 Application Layer

The application layer features the following two modes of operation:

a) Community Mode: The Community Mode incorporates a number of social features that can promote system user engagement aspects, such as analytics provision, integration of news feed and social media connectors.

b) User Mode: In order to gain insight into the processed results, a user-centered approach has followed to design and develop a web interface providing to the system users interactive access to the Visual Analytics services. In the User Mode, meaningful and attractive visualizations of search outcomes and their semantics are displaying in order to allow not only experts but also ordinary people to explore and interact with versatile legal information in an intuitive manner. The whole layer follows a responsive web design in order for users to have access to the results through a mobile interface, apart from the web one.

5.2 Trust and Security Layer

a) Trust and Security Infrastructure: The trust and security infrastructure provide the trust management and the efficient security provisioning (like crypto-primitives, crypto-protocols, public key infrastructure / PKI integration, digital rights management / DRM systems cover-age, dependability assurance, risk prediction algorithms) needed for every component of the platform.

b) User MGT Authentication and Authorization Services: User MGT Authentication and Authorization Services ensure that users can rely on the platform and that the platform is protected against unauthorized access and attacks. The rights for the users in the Community Mode, the User Mode, the Business Mode and Administration Mode for the different users are cleared and managed in the User Management Authentication and Authorization Services. After registration users are able to log-in to the service and adjust their personal settings.

5.3 Service Layer

a) User Generated Queries: The proposed system infrastructure stores search queries made by users for analysis and optimization purposes. The terms and structure that make up the user query are feeding into the Semantic search Engine to enhance the relevance of the results based on inferred concepts and semantic annotations. In addition, due to the resource intensiveness of semantic querying, this component applies optimization methods to prioritize queries, cache results, or even explore innovative methods for scaling such as Linked Data Fragments7.

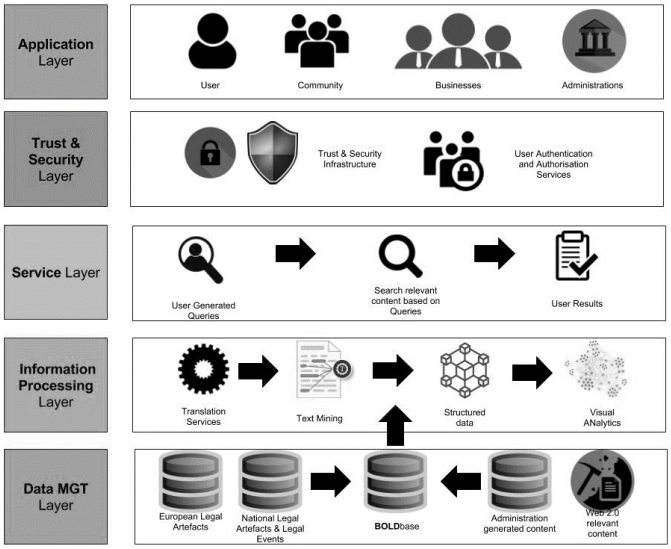

Figure 1: An ICT architecture for supporting the proposed legal text mining framework

b) Search relevant content based on queries: The Search Engine is used for searching through the system's triple store using a scalable Solr-based semantic search engine. Furthermore, the search engine is taking into account semantic relations between search terms and stored entities (e.g. synonyms). Best practices such as faceted search are also used to present the user with more search options, relevant to the search terms by semantic association.

c) Search results: This component is responsible for retrieving and presenting to the user the search results in an efficient and user-friendly manner.

5.4 Information Processing Layer

a) Data Preparation and Translation Services: In this stage, data is acquired and prepared for the text mining tools to follow. This stage includes data reading and initial cleansing, anonymization if needed, semantic annotation (so that to be indexed at the European Open Data Portal), and formulation for processing. Due to the diversified origin of the texts to be acquired, a large amount of effort and computational power is devoted to Optical Character Recognition and translation in English, if in a different member state official language. Translation is based on EUROVOC and automated translation services, both for the complete texts but also for the various indexes and n-grams to be created at the processing stage.

b) Text Mining: Various algorithms are being applied in different processing tasks, relying on a supercomputing infrastructure, in order to produce service – oriented intermediate results:

- Creation of reverse indexing, occurrence and frequency tables for millions of words
- Creation of various n-grams for the identification of important terms or phrases

- Semantic comparison of different law sets (e.g. EU Directive against national legal framework) performing full word-level, document-to-document comparison for billions of pages – needing the power of thousands of processors
- Interrelation of all the original and translated terms and texts
- Term extraction analysis of news, social media, blogs and other content

c) Structured Data: This component represents the information collected from the various sources harvested by the Data Sources Layer and adhering to a common model and format, that can then be used more effectively by the Visual Analytics Service. The data stored include any harvested and derived information that is necessary to realize the project's use cases.

d) Visual Analytics Service: The Visual Analytics Service provides the ability to access the entire data transformation pipeline from raw or semantic data to interactive visual representations. The main goal is to enable user-centered and comprehensible solutions for getting insights and knowledge about the entire domain. Visual Analytics as a service further enables solving the variety of knowledge related questions in the domain of law. The service characteristics enable to exploit the technology to other domains. The variety of visualizations and transformation technologies is adaptable to user's requirements and support her in the knowledge acquisition process.

5.5 Data Sources Layer

Data Repository: The data sources to be accessed primarily through web services communication where available.

6 Conclusions and further research directions

This study has conducted a literature review towards the identification of used technologies and existing infrastructures in legal text mining. According to our findings from workshops, we have proposed a framework analysis for legal text mining using big data that facilitates to solve legal issues and helps citizens and enterprises to follow the latest legislation. In particular, the establishment of the proposed framework foster transparency and promote social inclusion in lawmaking. Citizens find and understand relevant legislation, case law, and other legal information and so that they understand how EU-law is created and implemented. Furthermore, legal comparison functionalities of member states' implementations of directives will support business makers in their quest to market relevance and lead to a more harmonized legal EU body. Moreover, A key constraint on contemporary policymakers is removed, at both local and national level, as they operate in an information environment with unlimited data on the current policy context or the effectiveness of proposed solutions. Lastly, we proposed an ICT architecture to implement the above framework. We intend to implement our framework by developing a web platform which enables providing the above services. afterward, we plan to execute real-scenarios which confirms the usefulness of services and to confirm them by legal experts.

References

Erdelez, S, and S & O'Hare (1997). "Legal informatics: application of information technology in law", *Annual Review of Information Science and Technology*, 32: 367-402.

Christensen, J., Soderland, S., & Etzioni, O. (2011, June). An analysis of open information ex-traction based on semantic role labeling. In Proceedings of the sixth international conference on Knowledge capture (pp. 113-120). ACM.

Sartor G. and E. Francesconi (2010). "Legal informatics and legal concepts", *Eurovoc Conference* – 18-19 November 2010, Luxembourg.

Hornsby, W. (1999). Improving the Delivery of Affordable Legal Services Through the Internet: A Blueprint for the Shift to a Digital Paradigm.

Blei, D., Ng, A., and Jordan, M. (2003). "Latent Dirichlet allocation." *Journal of Machine Learning Research*, 3:993.

Mikolov, T., Yih, W. T., & Zweig, G. (2013). Linguistic regularities in continuous space word representations. In Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (pp. 746-751).

Yih, W. T., Zweig, G., & Platt, J. C. (2012, July). Polarity inducing latent semantic analysis. In Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (pp. 1212-1222). Association for Computational Linguistics.

Baroni, M., Bernardi, R., and Zamparelli, R. (2014). Frege in space: A program for compositional distributional semantics. *Linguistic Issues in Language Technology*, 9.

Garofalakis, J., Plessas, K., & Plessas, A. (2016, November). A semi-automatic system for the consolidation of Greek legislative texts. *In Proceedings of the 20th Pan-Hellenic Conference on Informatics* (p. 1). ACM.

Lagos, N., Gallé, M., & Chernov, A. (2017). U.S. Patent Application No. 14/850,060.

Lampoltshammer, T. J., Sageder, C., & Heistracher, T. (2015). The openlaws platform—An open architecture for big open legal data. *In Proceedings of the 18th International Legal Informatics Symposium IRIS* (Vol. 309, pp. 173-179).

Pereira, G., Charalabidis, Y., Alexopoulos, C., Mureddu, F., Paryced, P., Ronzhyn, A., Flak, L. and Wimmer, M. A. (2018). Scientific foundations training and entrepreneurship activities in the domain of ICT-enabled Governance. *In Proceedings of DG.O 2018*, Delft, Netherlands, May2018.

Lacity, M. C., & Janson, M. A. (1994). Understanding qualitative data: A framework of text analysis methods. Journal of Management Information Systems, 11(2), 137-155.

França, M. V., Zaverucha, G., & Garcez, A. S. D. A. (2014). Fast relational learning using bottom clause propositionalization with artificial neural networks. Machine learning, 94(1), 81-104.Brunschwig, Colette (2001): Visualisierung von Rechtsnormen, Legal Design, Zürcher Studien zur Rechtsgeschichte, hg. von M. T. Fögen [u.a.], Zürich: Schulthess Juristische Medien

Röhl, K. F., & Ulbrich, S. (2007). Recht anschaulich: Visualisierung in der Juristenausbildung (Vol. 3). Herbert von Halem Verlag.