Digitizing Luxembourg’s Legal Corpora:
Experience and Vision

John Dann\textsuperscript{1}, Mehrdad Sabetzadeh\textsuperscript{2}, Nicolas Sannier\textsuperscript{2}, and Lionel Briand\textsuperscript{2}

\textsuperscript{1} Central Legislation Department (SCL), State Ministry of the Government of Luxembourg, Luxembourg\texttt{john.dann@scl.etat.lu}

\textsuperscript{2} Interdisciplinary Centre for Security, Reliability and Trust (SnT), University of Luxembourg, Luxembourg\texttt{\{sabetzadeh,sannier,briand\}@svv.lu}

An important thrust of e-Government is increasing citizens’ access to legal texts such as laws and regulations through online portals. While formats such as HTML and PDF have traditionally been the norm for these portals, there has been a rapid shift in recent years toward markup representations that provide legal information alongside the texts.

In 2016, the Government of Luxembourg started a major paradigm change for its Official Journal. The aim was not only to provide better search facilities for citizens and professionals, but also structured data in line with linked open data standards. An important goal for the Government of Luxembourg is to provide metadata, both to facilitate the interpretation of the texts by humans and further to enable advanced computerized analyses.

With the release in 2017 of the digital legal portal legilux.lu (\url{http://www.legilux.public.lu}), Luxembourg went a further step to provide an official and in-force electronic version of the texts, the paper version not having legal value anymore.

Legilux now uses the Akoma Ntoso schema (\url{http://www.akomantoso.org}) for capturing the hierarchical representation of legal texts and the European Legislation Identifier (ELI, \url{http://eur-lex.europa.eu/eli}) for identifying and referencing all legal resources. ELI is a widely-adopted European Union initiative aimed at providing a unified legal referencing mechanism. In particular, ELI defines a labeling framework based on a customizable template to enable the definition of universal resource names that are independent of the countries and legal jurisdictions to which the texts belong.

During the digitization effort, the need was raised to convert large amounts of consolidated “Codes”, available online only as PDF documents. An immediate challenge in this respect was, given the sheer scale of the legal corpora that need to be enhanced with metadata, a fully manual process would be extremely laborious and prohibitively expensive. This called for techniques that can provide automated assistance in identifying and annotating the metadata items. In the past two years, SCL – the government agency in charge of legilux.lu – and the SnT Centre at the University of Luxembourg have been engaged in a collaborative project aimed at automating the extraction of metadata from legal texts [1–3]. The metadata items currently covered concern the hierarchical organization of legal texts and their cross references.

As part of the above joint initiative, we have been developing an automated assistance tool, named ARMLET (Automated Retrieval of Metadata from Legal Texts), which leverages artificial intelligence technologies, particularly Natural Language Processing (NLP), for automated demarcation of legal metadata and converting legal texts
into a markup representation that can be readily disseminated on legilux.lu. In our presentation, we will be describing ARMLET alongside our preliminary experience applying it to legal texts in Luxembourg, as outlined below.

**The ARMLET Toolbox.** ARMLET is a tool for automated detection of metadata in legal texts. ARMLET has been implemented as a configurable framework. In particular, the toolbox can be tailored to legal texts with different internal structures and written using different legal drafting practices. ARMLET is targeted at processing both long and complex texts such as legislative codes and smaller texts which typically have a simpler and more homogeneous organization. ARMLET’s current main features are:

1. Fine-grained segmentation: ARMLET automatically and precisely segments legal texts into their constituent parts, going from high-level divisions, such as books, chapters and sections, all the way down to articles and article subdivisions, such as paragraphs, alines, and lists.
2. Advanced cross reference handling: ARMLET, to the best of our knowledge, has the most complete catalog of cross reference patterns developed to date [1]. ARMLET comes equipped with rules for automatic detection and resolution (hyperlinking) of cross references.
3. Customization facilities: ARMLET provides easy-to-use facilities for customizing the toolbox to different legal corpora in different countries and jurisdictions.
4. Standardized markup format: ARMLET provides a generic algorithm for XML generation, which at the moment supports the Akoma Ntoso schema and ELI.

**Experience and Next Steps.** ARMLET has been successfully applied for converting seven of Luxembourg’s legislative codes into digital resources, most of which have been published on legilux.lu or are under review. Overall and in this context, we observe that ≈91% of the markup elements generated by ARMLET are fully correct, ≈8% of the elements are partially correct, and the remaining 1% are missed. We have further developed a configuration of ARMLET aimed at converting the the smaller texts that the ministries produce on a near-daily basis. A systematic evaluation of the accuracy of ARMLET in this context is now underway; the initial results are very promising.

At the moment, ARMLET is primarily geared towards structural and citation metadata. We are now conducting research on the detection of semantic elements in legal provisions including, among others, actors, modalities, conditions and exceptions. Our ultimate, long-term goal is to use the extracted metadata to ease the interpretation of legal provisions and to support legal experts in their tasks, including the specification of legal compliance rules.

**References**