*REPORT*

**First webinar on the use of AI in the justice field:**

**Anonymisation and pseudonymisation of judicial decisions**

26 and 29 March 2021

**Day 1**

The webinar begun with an opening speech by Mr. Richard Sonnenschein, Acting Director of DG JUST, Directorate B (criminal justice), who talked about the growing role of AI and the need to have an ethical and human centred approach to the development of AI-based tools. He stressed the importance of making sure the justice sector is not left behind when it comes to innovative technologies and noted that this is a first webinar of a series, which will be organised following the Communication on digitalisation of justice of 2 December 2020.

The morning session stated with four presentations by **representatives of the academia**. a professor from the Aalto University and the Helsinki Centre for Digital Humanities at the University of Helsinki, presented Finland’s vision of publishing and using legal documents as Linked Open Data on the Semantic Web for legal applications. In Finland, a series of tools and systems related to content pseudonymisation, knowledge extraction, data exploration, and data-analysis are being created in collaboration with the Ministry of Justice of Finland, the Parliament of Finland, universities, and IT companies. Finland’s main objective is to provide legal documents in machine-readable format. Linked Open Data be can used to create intelligent applications, such as the ANOPPI tool, which performs semi-automated pseudonymisation. Case law has been available online since 1997 but only for decisions made by higher and special courts. Once pseudonymisation is automated, it will be possible for Finland to publish more decisions, even those produced by district courts.

A professor at Bologna University, gave a presentation on how case-law information, once it is converted into machine-readable information, becomes usable for different purposes. For example, the speaker mentioned how this can be used to ensure open access to legal knowledge for practitioners, making justice more transparent for the citizens, reusing legal knowledge for AI and legal data analytics applications and integrating legislative sources with judgments. Robust pseudonymisation, the professor noted, is crucial to be able to comply with data protection regulation, the right to be forgotten and the special norms in each judicial domain. The speaker also stressed the importance of using the right term, "pseudonymisation", which allows for a correct normative analysis and adoption of the most suitable measures for minimising the risk of re-identification.

A professor at Vytautas Magnus University, focused on criminal decisions and the importance of developing diversified anonymisation practices. Diversification is necessary as some cases might not necessarily be considered for full anonymitisation, i.e. when there are public security issues involved.

A representative from the University of Rennes 1 (France) and University of Quebec at Montreal, shared their work on how to have both transparency and strong privacy guarantees when publishing legal documents online. Natural language documents are transformed in their semantic, statistical and structural representation, which will then be machine-readable and therefore available for searches and analytics. Great care must be given to the task of choosing which data to be considered for anonymisation, because it is now well known that the process of redaction presents issues since even after pseudonymisation elements can be easily inferred.

During the Q&A session, the experts from the academia discussed specific challenges which concern the limits of rule-based redaction when it comes to guaranteeing privacy. For this issue, a possible solution could involve discarding the one-size fits all approach in dealing with transparency and privacy, meaning that perhaps reconciliation between privacy and transparency could be achieved by distinguishing the two needs of having “precise” access or “massive” access. Opportunities could be found in the possibility of bringing together AI technology with a semantic web approach.

During the morning session, an **open discussion on national practices** on publication of judicial decisions also took place. Participants discussed among other the provision of judicial decisions as open data, approaches towards the protection of personal data, the European Case Law Identifier (ECLI) and the search engine available on the European e-Justice Portal.

To conclude the morning, **two companies** presented their solutions and products. **Consono**, a Belgium-based company, presented a solution they developed together with IBM and Oracle, first tested successfully through a pilot project with the Dutch Police and then used for the purposes of the Court of appeal of Antwerp. The company’s Dynamic Access Architecture is intended as a viable tool to tackle the main challenges courts face, such as having to handle large amounts of information, needing to recognise patterns in that information, and being able to analyse and compare data. Their anonymisation solution focuses on data on who, what, where and when.

The second presentation was delivered by **TanaCorp,** who implemented an anonymisation solution for the French Conseil d’Etat, which is currently running on Kubernetes (on Microsoft Azure Cloud & OVH). Anonymisation is provided for patronymic names, rare first names, addresses of private persons, names of legal entities, where a private person's name appears elsewhere in the document with its real name. Anonymisation is not provided for the members of the judgment panel, the clerk or secretary of the section, the lawyers, the other agents of the parties, and legal entities and their addresses.

In the afternoon, **Germany, Estonia, Romania, Latvia, Poland and Spain** presented their national solutions on anonymisation of judicial decisions.

**Germany** gave a presentation on current anonymisation practices in the country, which are indicated as an essential requirement for e-justice. Only 2% of all court decisions are published and most of them originate from last instance and high courts. There are only a few statutory publication obligations in Germany, but a general obligation to publish jurisprudence has been developed by case law and is derived from the constitution. Anonymisation focuses on the protection of natural persons and their individual rights, excluding legal persons. Both direct (e.g. names, addresses, date of birth) and indirect identifiers (e.g. profession details, academic titles, health data, descriptive information about local conditions or companies and unique features) need to be anonymised.

**Estonia** presented an ICT project on de-identification of documents through the use of AI which was used to anonymise 85K decisions. The de-identification process begun with a PDF or HTML document containing personal information and through natural language processing, and information replacement, generated a PDF or HTML document with masked personal information. The de-identification process focuses mainly on personal names and other specific identifiers.

**Romania** explained that currently anonymisation is done manually and is mainly focused on protecting personal data in civil decisions. Over the years, there have been four main attempts to improve digitalisation in the justice field. In 2005, a website was launched for all the courts, which gathers information regarding relevant jurisprudence, court meeting schedule, status of developing cases, statistical and general information about each court. Moreover, since 2010, an electronic case-tracking and management system was established. Last, a website for judges where they can access important information (EMAP) and the so-called Electronic FILE, which is a tool for parties to access documents from their judicial file, have been implemented.

**Latvia** presented a project for the improvement of investigation and litigation processes organised in collaboration with the Ministry of Justice, the prison administration, probation service, prosecutor’s office and court administration. The goal of the project was to achieve an electronic process from investigation to litigation and enforcement. For that purpose, access to electronic case materials was necessary for all parties involved, and the level of digitalisation and automation needed to be increased. Currently, there is a semi-automated solution in place when it comes to pseudonymisation. When a decision in its original form is added to the Court information system, the system then automatically imports that decision into the pseudonymisation tool, which performs text recognition and marks specific parts of the text based on templates. A user (e.g. the judge assistant, the secretary or another court employee), checks the text and replaces what needs to be changed with a pseudonym, saves the new document and publishes it on the court portal. Latvia is currently working on a new solution with machine learning technology that should provide a more accurate entity recognition, faster data processing and more customisable features.

**Poland** explained that currently judgments are imported into a system where automated anonymisation takes place. Then, a user needs to verify the automated anonymisation and is required to give the final approval for the online publication. The anonymisation process focuses on names of natural persons, identification numbers (e.g. tax identification numbers, personal identification numbers, vehicle identification numbers), and device or product names (e.g. Windows Vista operating system, HP 4100 printer, Nokia 3310 mobile phone).

**Spain** gave a presentation on their anonymisation practices for personal data in accordance with the transparency regulations of the country. The process begins with the original document of a court decision, which is received in HTML format and transformed into an XML file that carries metadata. The process of pseudonymisation involves performing several tasks, which involve structuring, field extraction, annotation, classification and finally detection and replacement of personal data.

During the **Q&A session** Romania was asked to clarify what type of data and what kind of court decisions are considered for anonymisation in the different countries. Romania said that they anonymise all decisions civil, criminal and administrative cases and that anonymization is required by law for first and last name, personal ID code, addresses, account numbers belonging to natural persons, and land registration numbers (for real estate). Romania recalled the importance of anonymising indirect identifiers as well as the protection of legal persons. Romania and Poland stress the importance of the so-called “human in the loop” who takes responsibility for validating the final documents that went through the (pseudo)anonymisation process before publication. Romania, Spain and Poland also mentioned that so far they have not had issues with particularly sensitive information that could lead, for example, to bias and discrimination. Poland mentioned that this issue will be discussed when the system will be used for prediction and analysis, but so far issues related to particularly sensitive information, bias and discrimination have not been raised.

**Day 2**

The second day of the webinar opened with a welcome statement by **Cristian Nicolau**, Head of Unit B3: e-Justice, IT and Document Management (DG JUST), who greeted all panellists and attendees, recalled the most important points made in the previous session and introduced the agenda for the morning.

The last panel saw presentations by experts from **France, Austria, Italy** and **Finland** on national solutions on anonymisation of judicial decisions.

**France** presented its current anonymisation practices and showed how its processes improved considerably following the adoption of ML anonymisation techniques. Initially, there were two levels of pseudonymisation in France, one which is done systematically for surnames and forenames, and another that is done at request of the judge in case of particular concerns for privacy or security. First, the judge of the case indicates which personal information needs to be concealed, then the decision is sent to the Supreme Courts in charge of the pseudonymisation, and lastly pseudonymisation is performed in a semi-automated manner “a posteriori” by the Court of Cassation and Conseil d’Etat. With machine learning enabled pseudonymisation, it is possible to detect a higher number of elements to be anonymised such as dates, addresses, numbers of bank accounts and license plates and so on.

**Austria** mentioned that currently, by law, only the Supreme Court is required to publish decisions. The anonymisation of court decisions is a manual task in Austria, although supported by specific word functions, but it is still a very time-consuming task. In 2018, Austria launched a project to start automating the anonymisation process and increase the number of decisions published. Austria has three core objectives: the identification of entities and metadata in court decisions, the automation of anonymisation in accordance with the respective legal requirements, and the determination of the anonymisation quality through a quality measure. To reach these goals Austria is implementing technological solutions in a combined approach.

**Italy** noted that judicial decisions contain a significant amount of personal data collected for purposes other than publication and are subject to data protection rules. Different courts have somewhat different processes. For example, in ordinary and administrative courts, anonymisation is performed ex post, while the constitutional decision are “born” with the omission of the names of the parties. The information that is considered for anonymisation ranges from personal data such as name and surname, to sensitive data such as racial or ethnic origin, political opinions, religious or philosophical beliefs, trade union membership, genetic or biometric data, heath or sexual orientation.

**Finland** said that they are now advancing legislation on opening legal documents to the public, and presented “ANOPPI”, an automatic pseudonymisation tools specifically developed for Finnish courts. The tool identifies and classifies names mentioned in documents and replaces them with pseudonyms. A user then can verify the result and make corrections if necessary.

**The open discussion** focused on a lively exchange on the complexity and significance of what seems to be a major challenge: the weighing of values at the core of the decisions, on what to prioritise when it comes to this trade-off between protecting personal data and the need to provide information to the public.

The day ended with a presentation by **Neurosoft**, the company that developed a solution for the acquisition and processing of legal information in Poland. It already had at hand useful resources, such as legal acts database, patents, company names, names of persons and geographical information. It develops technologies such as tokenization, PoS tagging and disambiguation, and effective full search, among others. Neurosoft indicated that they were able to achieve better access to case law for citizens, a more automated and consistent system than before, and an extensive experience in the challenges of anonymisation. Their future work will focus on improving the quality of anonymisation, on the creation of language models based on multilingual EU data, and the application of AI to search and reasoning tasks.

Outcomes of the event are available here: <https://ec.europa.eu/info/policies/justice-and-fundamental-rights/digitalisation-justice/conferences-and-events_en>