Interoperability in Legal Data Spaces

Christian Dirschl
Chief Content Architect
Wolters Kluwer Deutschland GmbH

Background

The European data strategy is about the creation of a single data market that ensures Europe's global competitiveness and data sovereignty. Common European data spaces will allow that more data becomes available for broad use, while keeping the companies and individuals who generate the data in control at the same time. They bring together on the one hand relevant data infrastructures and on the other hand governance frameworks in order to facilitate data pooling and sharing. These common European data spaces are currently foreseen in nine different areas, but more is needed. Although not in the current list, a data space for legal data is highly important, since legal aspects are needed in any industry data application, most obvious of course in the Public Administration common European data space. One core aspect of a prosper and successful data space is an environment that is interoperable, both from a data as well as from a tool point of view.

Interoperability is an important cornerstone of a product or system to work properly with other products or systems. Semantic interoperability means to automatically interpret the information exchanged in a meaningful and accurate fashion in order to produce useful results. To achieve semantic interoperability, both sides must refer to a common information exchange reference model. This means that the content of the information exchange requests are unambiguously defined: what is sent is the same as what is understood.

Legal data is very domain specific. It has its own language style and even distinct vocabularies. Moreover, even words that are used in day to day language have a very specific and often different meaning in a legal context (e.g. difference between rent and loan).
In addition, the legal industry has very strict rules and a restrictive culture when it comes to data, digitalization, privacy, security, liability, and collaboration in general. It e.g. goes way beyond the typical GDPR discussion and can e.g. also be about a very strict confidentiality.
Legal business and data is also very local. So each country and even region has its own legal tradition and data. Therefore, topics like multilingualism or cross border business pose a specific challenge to this industry.
All these aspects mentioned above have an impact on what interoperability in Legal Data Spaces really means.

Interoperability in Legal Data Spaces

Therefore, on a practical level it comes down to the following main characteristics that need to be addressed: seamless workflow and transparency.
The main challenge here is that both characteristics must go along with each other. Processing steps can only be executed when e.g. consent for doing so is given. In the past, the focus was on technical aspects of workflow processing and governance and transparency was just an overarching phenomenon that was dealt with upfront or at the end. The upcoming data economy needs more granularity and this is where built-in legal considerations come into play.

Of course, consent can still be clustered for several process steps, yet the application of consent is on step level. But what is a relevant step in this regard? Mainly, a step is executed when either data is manipulated in a way that could have governance issues (e.g. several datasets from different sources are processed or one dataset is processed with a language model) or that data is transferred from one application to another. So not all workflow steps are necessarily relevant steps from a legal perspective.

Addressing seamless workflow

Apart from the technical requirements to have interoperable tools in place that communicate e.g. via APIs, we need in addition a common understanding on the process itself. Since we are talking about data-driven AI applications, the semantics of the data needs to be available in a machine-readable form. This also addresses some other requirements like the multilingualism issue, the local flavor issue and is also the basis for the second aspect, which is transparency – because only when things are known they can be made visible and be explained.

This semantic model therefore is the core challenge of the interoperability task. It covers many different aspects and needs to be split up in several modules, each represented in a standardized format like SKOS. On the one hand we need a model of the legal domain, so that the tools processing legal data have a basic knowledge about what they are doing. This covers things like legal entities, legal document types, legal processes, etc. On the other hand, we need to have processual legal knowledge and facts for enabling users to express consent on workflow steps that need to be taken. This covers e.g. information about copyright, licenses, privacy, security, and liability.

All this semantic information needs to be deeply embedded in the technology architecture, because finally it triggers and drives the applications themselves.

A specific challenge lies in the execution of consent statements by the users. We foresee a wide variability of user types and in order to express consent they need to understand what they express consent for. As a starting point we see two major user groups (application users with sufficient basic legal knowledge like employees at public administrations and application users with advanced to expert legal knowledge like lawyers, judges and legal counsels). These two user groups do not only differ in their capability of understanding legal matters, they also differ in their expectations of what information they should be presented with. The basic professional wants mainly an exception handling. As long as everything is green, he will express his consent. Only if contradictions appear that require a decision he basically wants to know what kind of problem there is and decides on that basis (or forwards it to a legal expert for further analysis). The legal expert however always wants to have full control over the process and wants to dig into all details when he thinks it is necessary, so even matters that are classified as unproblematic by the governance layer of the application.

This leads to the following draft architecture/process:
- The user defines a workflow
- The machine builds the workflow and highlights the consent decision steps
- A dashboard gives an overview on these consent decision steps and their status (red, yellow, green) with respect to the data processed in that workflow
- The user can click on each consent decision step and first gets an information on the status and what (in the case of red and yellow) problem category is requiring the decision
- If the user needs more detailed information then he can click on an additional explanation button (key here is that this explanation is for clarification also a workflow overview like which datasets are currently processed, but the main purpose of this explanation is the legal

impact that the consent decision has (e.g. a violation of a license or the loss of copyright with mixed datasets)

The interesting question is of course what happens when consent is not [given](). This could lead to an alternative workflow, the termination of the workflow or a continuation of the workflow on a "test" basis in order to determine if the result is so beneficial that measures to avoid the "non consent impact" are worth it.

Addressing transparency

The example of user consent above has already introduced some aspects of transparency: the notion of user group specific explainability, the power of visualization and the necessity to give access to transparency on a very granular level of the process.

The main challenge on transparency is that all information to make an informed decision by a specific user needs to be available in an easily consumable way. On the one hand we need an overarching knowledge model covering the required information and on the other hand we need to introduce mechanisms that make sure that this information is available, so either given by the data/tool providers or by deriving it from general assumptions.

The incompleteness of metadata is of course not a new phenomenon and there are solutions in place how to deal with it. In this specific case however, the impact of noise or errors or incompleteness is very severe since the whole application flow is depending on it. It is like when you have a geo application and data on all public transport data is missing.

This leads to the following flow:

- Collect all information from data and tool providers with respect to legal considerations based on the proposed semantic model
- Validation of input and generation of quality report
- Introduce generic business rules and best practices to augment the legal landscape
- Execute the defined workflows with the legal impact described in the previous flow description
- Enable legal information update mechanisms

Summary

We have shown what an Interoperability for a Legal Data Space means in concrete terms. The main message is in order to be able to run actually any, but for sure a Legal Data Space, legal requirements and rules need to be included in the architectural design of the platform as a core component. An overarching semantic model needs to be defined and adhered to. Only then a data economy based on data spaces can emerge and receive the trust from all stakeholders that it requires.