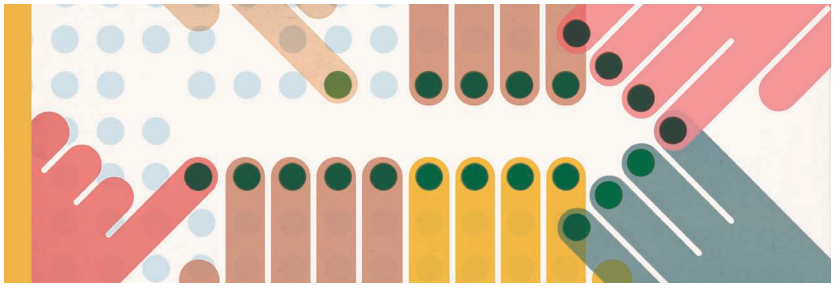
# Solving an Open Legal Data Puzzle With an Interdisciplinary Team

**Lőrinc Thurnay, Bernhard Riedl, Anna-Sophie Novak, Verena Schmid, and Thomas J. Lampoltshammer**, Danube University Krems

*// To get on a mutual ground as a team of computer scientists and legal experts, mapping open legal data, we had to shift our perspectives, dive into foreign concepts, and collaborate closely. //*



**UNDERSTANDING A COMPLEX** architecture not only poses a challenge but can also be a rewarding experience when properly untangled. In this article, we share our experiences of how we, a team of computer scientists and legal experts, worked together on a research project to map open legal data from several sources to one standard schema. It was clear from the beginning that we needed a multidisciplinary team[1] consisting of legal and tech experts following an agile software development process.[2] Still, we had not realized just how complex and ambiguous legal databases are for software engineers to navigate.

*Digital Object Identifier 10.1109/MS.2021.3117728*
*Date of current version: 23 December 2021*

To get on the same page, we had to look at our domains through our colleagues' eyes, formulate our questions and answers in a common language, and, to some extent, learn our way around the other, foreign domain. We finally succeeded at our task and also seized emerging opportunities for data enrichment. To explain why this metadata mapping exercise required such close collaboration with legal experts, we would first like to provide some context about the data we worked with.

## Context

The data sources to be mapped were Austrian open legal data, published with custom metadata schemata, under the *CC BY 3.0 AT* license[3] by the Austrian Legal Information System[4] (henceforth RIS[5]) and the Austrian Parliament,[6–9] cataloged by data.gv.at, the open government data portal of Austria.[10] The target schema consists of well-defined subschemata (DCAT,[11] ELI,[12], and Akoma-Ntoso[13]), all widely used standards in their domains.

### Austrian Open Legal Data

Legal documents connect with each other in a range of implicit and explicit ways, such as by amendments, references, or topics. Managing related documents in large data sets is typically a task of database engineers; however, the structures and linkages of legal documents are designed by legal documentarists and thus are not always a clean fit in database management systems. While arguably the two domains have been on convergent trajectories for a few decades, we measure the history of legal documents in centuries. When managing old but still legally relevant documents, database engineers must bow before legal tradition.

With texts referencing back as far as the year 1772, the standards of documents in our primary source RIS have been changing as different forms of governments, political systems, international alliances, and occupations came and went. The roots of RIS go back to 1983, it has been publicly available since 1997, and it is the only authentic source of federal law gazettes published since 2004. It manages legal documents from all 10 jurisdictions in our project's scope: the Federal State of Austria and its nine federal provinces, each sourced from different, independent authorities. An ever-growing government information system with such historical, organizational, and contentual breadth is a massive ship to navigate.[14] As a result, the metadata of its JSON/REST application programming interface (API) show some age and inconsistency. Unfortunately, API documentation is limited to the usage of endpoints and does not cover the structure and meaning of its contents.

Our secondary data source, the Austrian Parliament's API, provides data from the last hundred years. Currently, they are a collection of RSS feeds extended with some custom fields. They are consistent in structure but provide limited valuable data.

In the following, we introduce examples of typical challenges that working with open legal data may pose. We take these examples to describe our respective experiences and takeaways regarding interdisciplinary collaboration on open data.

## Need for Tighter Collaboration

To establish a reliable data mapping strategy, one needs to have a solid understanding of the source data structure. In several cases, RIS metadata are seemingly incorrect or contradictory.

For example, we worked on establishing a chain of different versions of texts through time. We did not have explicit references in metadata between these versions, so we relied on date fields: if one document entered into force the day after the other went out of force, we established a link. We assumed that documents would always come into force before going out of force but found thousands of examples doing the opposite. We also found different versions of consolidated sections (documents summarizing legal texts for a particular point of time) that were in force simultaneously.

The legal experts concluded that both seemingly paradoxical cases were legit (the former are cases of retroactive legislation, the latter are outliers having to do with the distinction between formal and material derogation). RIS does not provide detailed enough metadata to capture all nuances of these relations, and the simplification introduced these apparent anomalies.

### Experiences: Tech Experts

Working with RIS data sets was somewhat cumbersome. Not having specifications of the metadata you are interfacing with is never ideal, but you can usually make reasonable assumptions based on labels, value formats, and emerging patterns. Pair programming helped us build trust in each other.[15] Trust was important, as we often had to admit to ourselves that we kept hitting roadblocks. More and more, we had to ask questions of our legal colleagues.

### Experiences: Legal Experts

We work with Austrian law and RIS every day, so we tried to support the technical development as much as possible. Initially, we did not know what our tech colleagues expected of us, so we focused on their specific questions and answered them to the best of our legal knowledge. Our colleagues were inclusive, and we all tried to be honest about frustrating uncertainties, which helped create a trusting atmosphere.

## Shifting Perspectives

We needed to interface with three APIs to gather the metadata required for the target metadata schema. Legally binding texts (law gazettes) and documents that consolidate these texts for better readability (consolidated sections) are provided by RIS, while the parliament provides metadata about parliamentary procedures that precede law gazettes. To map these sources into one target schema, we had to join three databases. However, the APIs provide no foreign keys necessary for database joins, forcing the data into silos. We had not anticipated this obstacle because the same linkages did exist as hyperlinks on the RIS user interface, and we had assumed the APIs would follow the same design.

### Experiences: Tech Experts

Finding out that these databases could not be joined was a blow. We set out to explore the possibility of establishing linkages based on other metadata values, but most metadata in RIS are intended to be human-readable. We had to learn more about the metadatas' syntax and semantics, so we turned to our legal colleagues again.

For such a collaborative exploratory task, we needed an efficient process. First, we identified fields that looked promising to become the basis of joining databases. Then, we had to illustrate our technical problem to our legal colleagues in an accessible way. Our colleagues then analyzed the nature of the correlations between the

fields we had identified. The structure of foreign keys must be deterministic, so we needed our colleagues to anticipate any variations or exceptions in these fields. At the same time, we performed sanity checks with the same intent. When a promising approach emerged, we implemented it, then verified the results with the links on the RIS website.

Discussing issues across domains is a tricky thing. You try to anticipate your colleague's point of view without understanding it yourself. We knew each other well, so we communicated smoothly, but when we got a response, sometimes we felt that it did not exactly answer the question we had asked. The most challenging aspect of this exercise was understanding if we had asked an incorrect question, if our point had been lost in translation, or if the answer was correct and we just did not understand it.

### Experiences: Legal Experts

At times working together was quite frustrating because we kept underestimating how much effort seemingly simple exchanges would take. Interdisciplinary cooperation, especially in the beginning, needed time to get going.

Many of the questions our tech colleagues asked us caught us off guard. These were simple questions about the law but coming from a perspective we had never taken before. We often found ourselves improvising, asking follow-up questions, and having long discussions and debates. It was difficult to accept when we did not have answers. Even though our tech colleagues reached out to us for help, we finally found ourselves most productive by assuming the role of a student—without inhibitions to ask any questions to understand what their concerns were really about.

## Concepts, Perspectives, Language

Law gazettes' metadata make references to documents of parliamentary procedures that led up to the law gazette's publishing. These parliamentary documents may be different types of government bills, followed by committee reports and records of plenary meetings of the Austrian National Council and Federal Council. They offer valuable context to how a law came about, and thus, we included them in our target metadata schema. This required joining the databases of law gazettes and the parliament.

Figure 1 illustrates the process that let us join these databases without

RIS law gazette metadata — BGBl. III Nr. 93/2018

```
"NationalCouncil":
  "NR: GP XXV RV 1670 AB 1732 S. 190.",
"FederalCouncil":
  "BR: AB 9848 S. 870."
```

Parliamentary step metadata — State Treaty 1670 d. B. (XXV)

```
<item>
  <date>07.06.2017</date>
  <link>https://www.parlament.gv.at/↩
          PAKT/VHG/XXV/I/I_01670/index.shtml</link>
  <title>Mehrseitiges Übereinkommen zur[...]</title>
  <type>State Treaty</type>
  <id>XXV/I/I_01670</id>
</item>
```

Decision table

| Doc. types in law gazette metadata | Doc. types in parliamentary metadata | Index | Document ID schema |
|---|---|---|---|
| Government Bill | | RV | {period}/I/I_{number} |
| | Art 15a Constitutional Agreement | | |
| | EU Decision or Initiative | | |
| | Federal Council Motion | | |
| | Popular Initiative | | |
| | State Treaty | | |
| Private Members' Bill | | IA | {period}/A/A_{number} |
| National Council Committee Report | | AB | {period}/I/I_{number} |
| | Committee Motion | | |
| National Council Plenary Meeting | | S. | {period}/NRSITZ/NRSITZ_{number} |
| Federal Council Committee Report | | AB | BR/I-BR/I-BR_{number} |
| Federal Council Plenary Meeting | | S. | BR/BRSITZ/BRSITZ_{number} |

**FIGURE 1.** The process of joining Austrian law gazette metadata with parliamentary metadata.

foreign keys. In law gazette metadata, we see semistructured strings of abbreviations and numbers, which we learned were law gazette publications' traditional reference style to parliamentary documents. In parliamentary documents' metadata we found and verified a <link> field containing the document's URL with a unique ID embedded in it. As seen in this example, a Roman numeral XXV (legislative period, GP) and an Arabic number 1670 are present in both metadata. These common values served as the basis for joining the databases and, as such their presence was crucial to our approach.

As the next step, we established that each abbreviation in the law gazette's metadata refers to a specific parliamentary document type (e.g., RV to government bill) followed by the number of that document. Regarding parliamentary documents, we found that each type has a different document ID schema. With this, we created a decision table, allowing us to pick the suitable ID schema for each part of a law gazette reference. Treating the abbreviations of a law gazette's references as indices, we could extract the relevant numeric values and recreate the referenced parliamentary document's ID. Embedded in the parliamentary document's <link> field, we extracted the document's ID using regex and persisted it as <id>. We could now join law gazettes with all of the parliamentary steps they referenced.

Figure 1 illustrates one of the benefits of this database join: under <type>, parliamentary documents' metadata provides a more detailed typology of **document types** than the references in law gazettes' metadata. We only see a reference to a government bill from the law gazette, but by linking with the Parliament's API, we can assert that the document in question is indeed a state treaty.

### Experiences: Legal Experts

The whole process of interacting with our tech colleagues felt similar to speaking to foreigners learning our native tongue. Surely, they had the flip side of the same experience! We put a lot of work into explaining legal vocabulary and principles, but given how much back-and-forth this complex task needed, it was necessary to establish a common ground.

Teaching your own language requires you to revisit the basics and regain awareness of elementary concepts that you use instinctively. Our colleagues asked us fundamental questions, and our goal was to answer with all of the necessary details but as understandable and uncomplicated as possible. In general, our collaboration was a valuable exercise in teaching. An interesting result of our collaboration is that our tech colleagues are now bona fide specialists in a narrow field of the law.

### Experiences: Tech Experts

Collaboration meant not only learning each others' jargon but also familiarizing ourselves with the other domain's unique syntax and media. We had to get comfortable with legal documents' notation, their hierarchical segmentation levels, and many referencing styles. On the other hand, to avoid preparing spreadsheets every time we needed our colleagues to verify a change we implemented, we sent them Python source code to look at. The fact that they did not even flinch we took as a sign of a real cohesive interdisciplinary team.

## Reflections and Benefits

It is clear that throughout the project, trust within the team was a crucial factor in our eventual success, facilitated by pair programming, open and inclusive feedback cycles, and prior acquaintance.[2] One can see the typical development stages of a small team[16] play out in our collaboration. We went through *forming*: we established roles and goals and started a superficial collaboration. We experienced *storming*: the complexity of the project became evident, and we faced the challenges of interdisciplinary communication. In the *norming* phase, we gained an understanding of the other discipline's concerns and embraced a shared language. Finally, we were *performing*.

In many cases, what seemed to be data errors from a technical perspective proved to be errors from the legal perspective, too. When we found suspicious data, we would perform quality triage together to decide: is it an error, a curious legal phenomenon, or a result of database design constraints? If an error, is it possible to fix it, work around it, or provide a fallback value? In a legitimate case, can we map it to the target schema? Do we have resources to deal with it, or can we justify a system boundary for it?

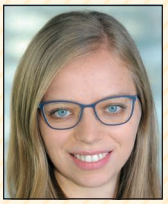
From sanity checks and unit tests that we created during data mapping, we ended up implementing a standardized data quality reporting tool, covering a few dozen types of errors. We had already been in touch with the data publisher, RIS, so as another (originally unintended) collaboration aspect of the project, we established a process to provide them with data quality reports periodically. In a few months, RIS already published some of the fixes to these errors. In effect, the meticulous process of making sense of undocumented open data led to the emerging opportunity to create

## ABOUT THE AUTHORS

**LŐRINC THURNAY** is a research fellow and software engineer at the Center for E-Governance at Danube University Krems, Krems an der Donau, Austria. His research focuses on open data applications. Thurnay received his master's degree in e-government technologies and services from Tallinn University of Technology, Estonia. He can be contacted at https://lorinc.thurnay.hu or loerinc.thurnay@donau-uni.ac.at.

**BERNHARD RIEDL** is a senior scientist at Danube University Krems, Krems an der Donau, Austria, and an agile evangelist responsible for the transformation of several organizations to modern forms of collaboration. His research focuses on techniques and processes in agile software development as well as on security and data protection in information systems. Riedl received his Ph.D. in technical sciences from Vienna University of Technology, Austria. He is a Senior Member of IEEE and a member of Scrum Alliance. He can be contacted at https://www.bernhard-riedl.com/ or bernhard.riedl@ieee.org

**ANNA-SOPHIE NOVAK** is a predoctoral research associate at the Center for E-Governance at Danube University Krems, Krems an der Donau, Austria. Her research focuses on the legal implications of emerging technologies and of digitizing and simplifying administrative processes. She can be contacted at https://linkedin.com/in/anna-sophie-novak-1a7a2082/ or anna-sophie.novak@donau-uni.ac.at.



**VERENA SCHMID** is a research fellow and legal scholar at the Center for E-Governance at Danube University Krems, Krems an der Donau, Austria. Her research focuses on the intersection between law and technology in public administration as well as in legislation processes. She can be contacted at https://linkedin.com/in/verenaschmid/ or verena.schmid@donau-uni.ac.at.

**THOMAS J. LAMPOLTSHAMMER** is an assistant professor and deputy head of the Center for E-Governance at Danube University Krems, Krems an der Donau, Austria. His current research focuses on the domain of data governance, the effects of ICT application in a connected society, and the effects on a data-driven society. Lampoltshammer received his doctoral degree in applied geoinformatics from the University of Salzburg, Austria. He can be contacted at https://linkedin.com/in/tj-lampolt.shammer/ or thomas.lampoltshammer@donau-uni.ac.at.

a feedback loop with the publisher and thus increase the quality of open data itself, achieving the open data life cycle.[17]

### Experiences: Legal Experts

This project was a chance for us to gain practical insight into another domain. Certainly not how to code, but we learned a lot about the tech way of thinking and the nature of the problems our colleagues face in their everyday work. The more we worked

together, the better we became at gauging tech questions and the sort of answers our colleagues needed from us.

This was a valuable learning experience, as with the emergence of synergic fields, like smart contracts, even closer collaboration between fields will be needed. We are certainly happy to have dipped our toes into the tech world, and we look forward to collaborating more with legal-curious tech experts.

### Experiences: Tech Experts

Working across domains has been a rewarding experience. When you reach that place of clarity with four people finally in agreement—talking about the same thing, meaning the same thing—it is an exhausting but gratifying moment and is very much worth the effort.

We collaborated so closely because of data silos as well as the inconsistencies and lack of documentation of our data sources. However, these issues are not specific to open legal data; they occure frequently in any kind of open data.[18] Therefore, based on our experiences, we conclude that a similar agile interdisciplinary collaboration strategy is generally useful in complex open data projects.

### Acknowledgments

We thank Günther Schefbeck of the Austrian Parliament who was a legal expert in the project, continuously providing us with key legal advice regarding our work described in this article, whereas the experiences described in this article are those of the authors. For brevity, we simplified some technical and legal details that are not relevant to the focus of this article. The work described here was part of the ManyLaws project (INEA/CEF/ICT/A2017/1567047).

### References

1. H. Takeuchi and I. Nonaka, "The new new product development game," *J. Product Innovation Manage.*, vol. 3, no. 3, pp. 205–206, 1986. doi: 10.1016/0737-6782(86)90053-6.
2. M. Cohn, *Succeeding With Agile: Software Development Using Scrum*. London: Pearson Education, 2010.
3. "Attribution 3.0 Austria (CC BY 3.0 AT)," Creative Commons. https://creativecommons.org/licenses/by/3.0/at/deed.en (accessed May 2021).
4. Ministry for Digital and Economic Affairs. data.gv.at https://www.data.gv.at/katalog/dataset/ris2_5 (accessed May 2021).
5. "RIS–Information about the Legal Information System," Ministry for Digital and Economic Affairs. https://www.ris.bka.gv.at/UI/Erv/Info.aspx (accessed May 2021).
6. "Katalog–Regierungsvorlagen und Gesetzesinitiativen," data.gv.at, Feb 23, 2017. https://www.data.gv.at/katalog/dataset/0d207682-c031-4cd0-af4d-9d40d229b57e (accessed May 2021).
7. "Katalog–Ausschussberichte," data.gv.at, Feb 23, 2017. https://www.data.gv.at/katalog/dataset/53dd46f6-23d6-4e17-a911-d9dadde3cfb2 (accessed May 2021).
8. "Katalog–Plenarsitzungen Nationalrat," data.gv.at, Feb 23, 2017. https://www.data.gv.at/katalog/dataset/1d5a7724-33ae-461b-a2c8-dd43ce24619c (accessed May 2021).
9. "Katalog–Plenarsitzungen Bundesrat," data.gv.at, Feb 23, 2017. https://www.data.gv.at/katalog/dataset/71434e72-1a72-4c07-a5d3-672ffefa5895 (accessed May 2021).
10. "Open Government Data (OGD)," Federal Ministry Republic of Austria – Digital and Economic Affairs. https://www.bmdw.gv.at/en/Topics/Digitalisation/In-administration/Open-Government-Data.html (accessed May 2021).
11. "Data catalog vocabulary (DCAT) version 2," World Wide Web Consortium. https://www.w3.org/TR/vocab-dcat-2/ (accessed May 2021).
12. "European Legislation Identifier (ELI)," Publications Office of the European Union. https://op.europa.eu/en/web/eu-vocabularies/eli (accessed May 2021).
13. M. Palmirani, R. Sperberg, G. Vergottini, A. Ntoso, and F. Vitali, "Akoma Ntoso Version 1.0, Part 1: XML Vocabulary OASIS Standard," OASIS, 2018. http://docs.oasis-open.org/legaldocml/akn-core/v1.0/os/part1-vocabulary/akn-core-v1.0-os-part1-vocabulary.html
14. A. Alexandrova, L. Rapanotti, and I. Horrocks, "The legacy problem in government agencies: An exploratory study," in *Proc. 16th Annu. Int. Conf. Digit. Government Res.*, May 2015, pp. 150–159.
15. L. Williams and R. Kessler, "All I really need to know about pair programming I learned in kindergarten," *Commun. ACM*, vol. 43, no. 5, pp. 108–114, 2000. doi: 10.1145/332833.332848.
16. B. W. Tuckman, "Developmental sequence in small groups," *Psychol. Bull.*, vol. 63, no. 6, pp. 384–399, 1965. doi: 10.1037/h0022100.
17. Y. Charalabidis, A. Zuiderwijk, C. Alexopoulos, M. Janssen, T. Lampoltshammer, and E. Ferro, *The World of Open Data* (Public Administration and Information Technology Series). Cham: Springer International Publishing, 2018.
18. M. Beno, K. Figl, J. Umbrich, and A. Polleres, "Perception of key barriers in using and publishing open data," *JeDEM-eJ. eDemocracy Open Government*, vol. 9, no. 2, pp. 134–165, 2017. doi: 10.29379/jedem.v9i2.465.