Legal Information Management, 24 (2024), pp. 267–271

doi:10.1017/S1472669624000598

Access All Areas: Breaking Down the Barriers to Legal Information

Abstract: The decisions of courts and tribunals, and the statutes that societies live under, are the building blocks to the rule of law. Access to this information for legal professionals can be difficult while access for the general public is often impossible. This article by **Gavin Sheridan**, the co-founder and CEO at Vizlegal, considers a first principles approach to the barriers to access, how the legal industry has adapted, how technology companies can address and improve it and how the future of legal information involves open access, open standards and innovation built on top of legal data.
Keywords: legal information; rule of law; legal information providers

INTRODUCTION

Legal information is difficult. It is also often something of a blind spot for both legal practitioners and academics. This is something this author has had direct experience of, having built a legal technology company from scratch over the past several years. This process revealed challenges, but also opportunities that exist throughout the world for improving this situation.

It is difficult because legal information is hard to get, difficult to organise, cumbersome to connect and tricky to update on an ongoing basis. Add to those problems the worlds of soft copy *vs* hard copy, digital formats – be they open or proprietary – general searchability and quality, and you end up with a mess in most jurisdictions in the world.

To say seeking to solve these problems is a challenge would be a severe understatement. Since, not only are you seeking to solve problems for the industry you aim to serve, but you are also trying to solve the problems of the source of much of that information – public bodies, including the courts.

Legal practitioners and academics often have a blind spot in this area for one main reason – they never directly experience it. Often, they study and then have professional careers where access to legal information via paid vendors is a foregone conclusion. Confronting the issue of open access to legal data is something that rarely, if ever, arises – despite this being the way most people in a society access legal information.

Coming from a background of both a social media and an open source intelligence startup,[1] while also having experienced the courts as a lay litigant and journalist involved in access to information issues, this author felt motivated to tackle it.

During an Access to Information on the Environment[2] (AIE) case over the course of five years, that ultimately led to an Irish Supreme Court judgment[3] it was certainly interesting to see how courts functioned, how things like court lists operated, and how practitioners appeared to have a supernatural ability to track it all manually – all while being able to navigate information on specific public websites with apparent ease.

DEVELOPING A PRODUCT

When starting to address a problem in any industry one usually will sit down with future potential customers and ask them a series of questions about the types of work and tasks they perform on a routine basis. A curious interviewer will probe deeply with the so-called 'Five Whys', or a repeated asking of the question 'why' to each response received. This tunes the interviewer to get to the nub of potential problems, as well as attempting to get to the core of how or why certain key tasks are carried out. Starting this process entirely from scratch is always fascinating.

Having spent many hours interviewing practitioners in this way it became clear that opportunities existed that had remained largely unaddressed by existing technologies – sometimes for reasons of arbitrary distinctions. For example, practitioners would classify the tasks that sit within the 'legal research' category in one way, while talking about 'case management' in another. Meanwhile 'knowledge management' as a concept lived within several buckets, while 'practice management' or 'matter management' seemed to exist within a different silo.

Ultimately, though, to an outside observer many of these distinctions appear artificial. They often merely described jobs (or tasks), or in a software development framework, something akin to 'Jobs to be Done' (JTBD),[4] rather than entire 'vertical' stacks of software. It appears

these silos were developed by accident, or perhaps via inertia over the past several decades.

Legal research in and of itself is a task that sits within an array of wider tasks that need to be performed at a law firm. As does 'current awareness' – which is slightly different, but also very similar as it applies to the broader concept of ongoing 'knowledge'. All these tasks lead to outputs that ultimately serve the client, or ultimately the progress or growth of the law firm itself. Research at law firms produces work products that serve the firm in the short or longer term, while current awareness fuels the firm's ability to stay on top of how the law is constantly changing and evolving, and in turn how to best advise its clients.

Ultimately though, software that seeks to address these problems in isolation can find themselves being perceived as limited 'lookalike' products, too like some that already exist, or as the industry says, they "suffer from a lack of differentiation". In the early days this would usually come in the form of "but how is this different from Westlaw / Lexis?" *et al*.

Starting a legal technology company from scratch is therefore a liberating experience. You are not tied in to pre-existing notions of industry verticals, but rather you can explore what it is that lawyers do, and then how you might be able to help improve their lives with new software. This first-principles approach was the one taken at Vizlegal.[5]

TACKLING THE INDUSTRY ONE LAWYER AT A TIME

For those who might not be familiar with the customer development and product development processes, they essentially boil down to relatively simple concepts. First, interview potential users / customers about the problems they face daily, how they currently might solve them, and drill into if a solution might be something they are willing to pay for. Second, once the product is in their hands, continue to ask those questions, both related to the existing solution they are using but also to other adjacent problems they may be having in their daily work.

If you can get both of these techniques right, you might end up with a product people are willing to pay for because it solves real problems they face on a daily basis.

There are some caveats, however. Users lie to you – or rather sometimes they tell you what you want to hear. This is sometimes out of a wish to be nice, so would not pass what is sometimes called "The Mom Test".[6] It's important to not necessarily listen to what the user is telling you, but sometimes what they're *not* telling you.

Once you've started putting software into the hands of your new users, you can begin to tweak and improve it, a process in software development called 'iterative' development. This process is essentially never ending. If your company survives long enough you will continue this process over and over again, most likely initially 'vertically' or within the same problem-space, and later, horizontally, identifying other problems to solve that are adjacent to ones you're already on the way to solving.

SOURCING DATA, AND ISSUES ARISING

Of course, this is all okay in principle, but to build a product you need a developer or development team, and in this space you will also need data if your chosen problem is one that bedevils many practitioners and companies – access to legal information. And this is where you may run into problems.

The problem that infects the entire legal information world is paper. Not just the physical form, but the paper transformed into its digital equivalent – the PDF (Portable Document Format). Since legal information first started being put online in the 1990s this has invariably been the format used to publish information such as statutes and judgments. The reason for this is that the PDF, being an un-editable format and most closely approximating the A4 page, is deemed to be a suitable facsimile of the physical paper that the world is used to. There are good reasons why this is the chosen format, but many of those reasons have fallen away over the past couple of decades.

When canvassing the world for the approaches taken by various courts and public bodies, the answer was almost always the same: information was made available in PDF, and almost exclusively so. There are difficulties with this when data structure is one of your key needs. Unless the data is structured, you will invariably have to spend resources making PDFs more searchable, more structured (for example, separate data fields for citations, parties, legal representatives and judges), to help your users to be better able to find and track what they are seeking.

In terms of best practice, though, it was the European Union institutions that had – from both a case law and legislative position – created one of the early standards in how to publish legal information in structured, open, accessible formats. The EU publications office in Brussels developed EurLex,[7] a repository of EU law, and CELEX,[8] a code system for the different sectors of EU law. But they did not just copy the text of the judgments being delivered by the Court of Justice, they structured them in Extensible Markup Language (XML) (see *Fig 1*.).

This meant that all EU case law could be copied, structured and queried and, importantly, PDFs of the decisions could be generated rather than served (or both). The XML of CJEU (Court of Justice of the European Union) judgments will not just include each language of the Member States, but will include – encoded – the statutes being interpreted by that judgment, and the other judgments it cites, down to the paragraph number in the cited decision.

This detail enriches the judgment itself, with metadata that is important to understand that judgment at a point in time. And not just that; all changes across the corpus

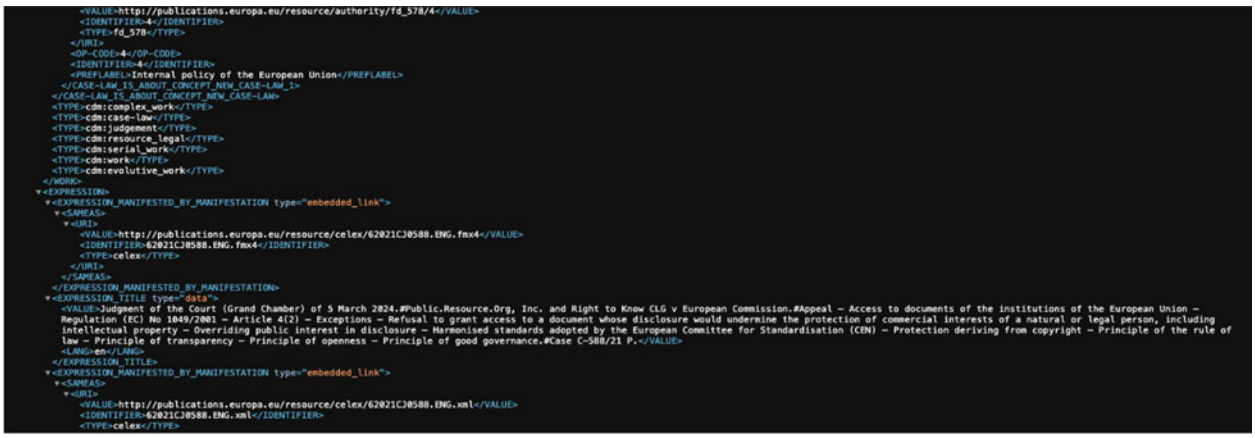

Figure 1: An example of a CJEU judgment encoded in XML

are updated as each case and law appears, and EurLex has a web service, or API (application programming interface), that informs re-users of the data of changes across their own corpus.

In a serious way, this is judgments as *data*, not simply as text.

In recent years the Case Law project[9] of the UK National Archives has taken a broadly similar approach to structuring court judgments and tribunal decisions, and earlier the consolidated data of primary legislation[10] – and this is to be welcomed. But across the world, this remains a rare way to receive law, which is unfortunate both for the public and for companies working in the area.

There are opportunities for public bodies and courts to engage in the move for access to justice, to make law available to everyone, for free, in open, accessible, structured formats, rather than proprietary and closed formats like .docx and PDF that we are all familiar with. This will lead to both commercial and open source projects that will benefit all users, not just legal professionals.

MAKING IMPROVEMENTS FOR USERS

But commercial providers cannot sit around and wait for this change to happen, so many will do much of this work themselves. This involves large ETL[11] (extract, transform, and load) systems for the automated pulling of data from public sources. Once the data is structured – often a painstaking process – it gives startup commercial providers the ability to give new capabilities to their customers, based on customer interviews about potential needs. It also allows for the widening of the scope of data, since there are so many niche sources of data that might be ignored by the larger incumbent commercial providers.

When speaking to customers it was a common refrain that many sources of legal information that are free and publicly available are a) difficult or impossible to search b) contained filters that were either incomplete, filters that probably should exist as an option but did not, or did not function as expected or c) responsiveness to searching and loading pages was slow or sometimes websites might crash, and d) when results did load, they sometimes were incomplete, or did not reach the expectations in terms of the numbers of results the user expected.

Solving these problems for users of a commercial platform led to genuine user engagement and the much sought after 'user happiness'. Some concrete examples are outlined below.

1. In one instance an important public resource for searching legal documents was removed, ostensibly for GDPR (General Data Protection Regulation) reasons. The PDFs themselves could be found and loaded, but the ability to search for keywords within them disappeared. This was fixed on the commercial side by ingesting all documents and redacting pages that could potentially cause a GDPR issue.
2. One public website with thousands of important quasi-judicial decisions covering some 16 years disappeared – for reasons to do with the creation of a new public body website. The website had had some ways to filter those decisions, but did not have some other obvious filtering options. In this case all decisions were available via the author's commercial platform, and additional filters were made available to make searching easier.
3. An ongoing problem with many court and public body websites is the publishing of scanned documents, meaning the text within the documents themselves is not searchable for keywords. An obvious solution for a commercial provider is to subject these documents to Optical Character Recognition (OCR), thus making them useable again.
4. A courts website publishes legal diaries in long HTML (Hyper Text Markup Language) format in all capital letters, and legal firms need to check if their individual cases are listed, sometimes multiple times a day or week – meaning valuable human hours were

being spent on manual effort to check these. A solution was to ingest all legal diaries and detect the names of all firms listed on every diary, and in turn insert those detected dates into an automated calendar for the firm.

5. A quasi-judicial body publishes its decisions in PDFs but the names of the official making the decision are embedded as an image with a signature onto the PDF. In order to allow users to filter by names of officials it was necessary to OCR thousands of images within the PDFs, and manually fix spelling errors. This enables users to search within the decisions of given officials.
6. Because decisions of lower tribunals can be appealed to higher tribunals, and in turn to the courts, users would have to do multiple searches on different websites to find each decision in the sequence of appeals, to provide greater context. No single system was developed to unify these decisions using a common identifier. A solution was to map the relationship between thousands of decisions and their underlying dockets, and then to run algorithms on lower tribunal websites to match their titles backwards, thus allowing the creation of case 'Timelines' across multiple appeals, websites and databases.
7. Similarly for references to the CJEU in Luxembourg, no system has yet been designed to standardise how to connect respective references to the court to their underlying cases in Member States. So a solution for Ireland was to semi-automatically discover connections between Irish cases and their references, and in turn add them to Timelines (see Fig 2). The user goal here is to *reduce* the need to search.
8. Court Rules and Practice Directions can in some jurisdictions both be unconsolidated, or consolidated, but with a time lag. This can make the lives of users more difficult, as the current version of a Court Rule is required to perform key functions. A solution is to use Git[12] technology to first consolidate all Rules, and then update as amendments appears, while also alerting users to those changes.

WHERE NEXT?

Of course, besides the issue of standards, formats, common IDs, searchability, APIs and everything else, is one other key obstacle: missing data. In many jurisdictions thousands of decisions are simply missing, having never made the jump from hard copy to digital. The only way to get them is from a book, or from an archive. Layered on top of this is the additional issue of copyright on 'reported' judgments – where the reported version is the only digital version available.

These can create real issues for the public and are perhaps something of a headache for many jurisdictions. There generally are few efforts to address this issue often, it seems, because it straddles several domains including data ownership and the involvement of different public bodies, such as national archives, the judiciary, the courts and government ministries.

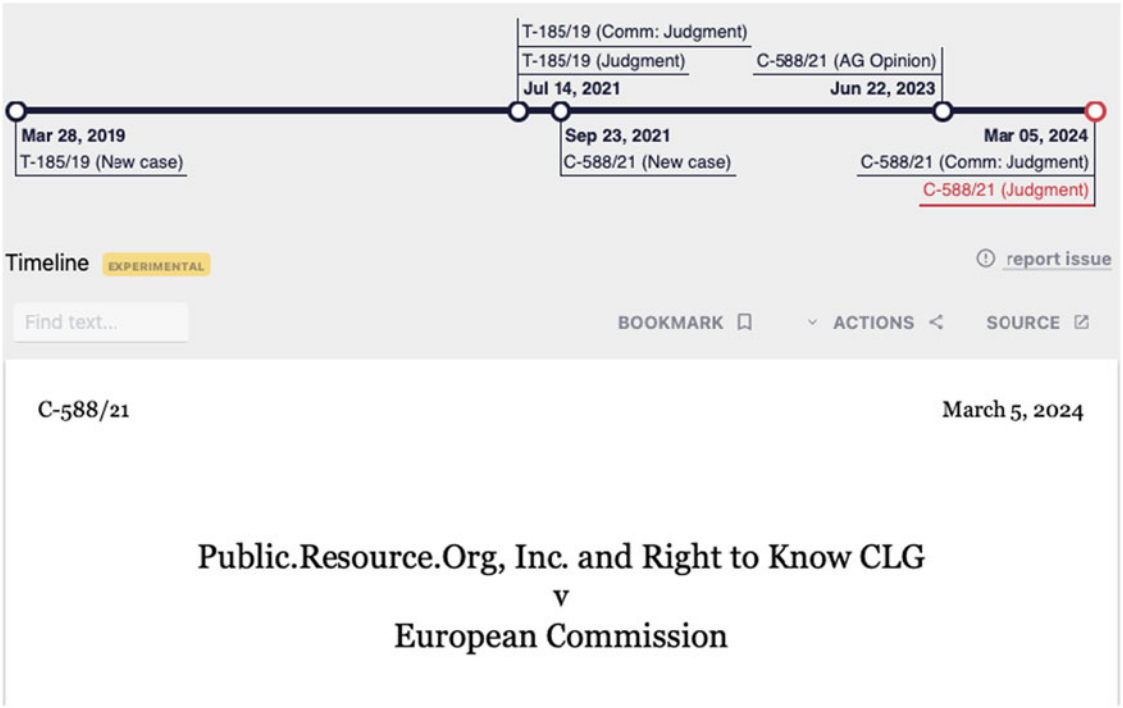

Figure 2: A timeline of events in the CJEU case C-588/21

In an ideal world all case law would be available to the public, for free, ideally as structured data and failing that in open or proprietary digital formats – including all historical decisions. These decisions would also be 'official' in the sense that they are the official copies made available by the jurisdiction and can therefore be relied upon by legal professionals and the public alike. The concept of quasi-official, but private, copyrighted 'Reported' judgments is understandable in the context of the time in which they were created, but it also poses difficulties for the public where those private repositories are the only digital copies available – and are behind a paywall.

Which leads to the inevitable Artificial Intelligence (AI) questions. If some data is copyrighted, and other data is unavailable since it was never digitised in the first place, can any reasonable AI system or model give adequate answers in a world where only partial data was used to train the model? Given that there is now ongoing litigation in multiple jurisdictions about the use of copyrighted material to train large language models, to what degree have copyrighted versions of case law – such as Reported judgments – been used to train them? And how do we know of the gaps that may exist in any of the current models – or will we ever know – simply because models have never been exposed to them because they were never digitised (or does it even matter for older documents)?

And to think about the potential future, in order for any legal specific language models to be created for, say, specific jurisdictions or geographies, it would seem to this author to be self-evident that for the best outcome, some steps should be carried out first, including:

- The collecting and digitisation of all case law and the further structuring of that data – with the work carried out by or on behalf of public bodies that are custodians of that data
- The licensing of data as open, so that it can be used for research and development in relevant AI fields, with the simultaneous publication at the earliest opportunity of those archives for public benefit, both historically and on an ongoing basis
- That any models trained using that data would have to declare the sourcing in a transparent manner

In so doing, the public benefits from the widest possible access online to judicial and quasi-judicial decisions, while public bodies and companies benefit from the research and development that can be carried out on that data, without fear of breaching privacy principles or copyright.

Endnotes

[1] *Storyful.com* <https://storyful.com/>
[2] *Access to Information on the Environment* <www.gov.ie/en/organisation-information/1e52cb-access-to-information-on-the-environment-aie/>
[3] *National Asset Management Agency -v- Commissioner for Environmental Information* [2015] IESC 51
[4] *Know Your Customers' "Jobs to Be Done"* <https://hbr.org/2016/09/know-your-customers-jobs-to-be-done>
[5] *Vizlegal* <www.vizlegal.com/>
[6] *The Mom Test* <www.momtestbook.com/>
[7] *Eur-Lex* <https://eur-lex.europa.eu/content/tools/TableOfSectors/types_of_documents_in_eurlex.html>
[8] *CELEX Numbers* <https://eur-lex.europa.eu/content/help/eurlex-content/celex-number.html>
[9] *Find Case Law* <https://caselaw.nationalarchives.gov.uk/>
[10] *Legislation.gov.uk* <www.legislation.gov.uk/>
[11] *ETL* <https://en.wikipedia.org/wiki/Extract,_transform,_load>
[12] *Git* <https://en.wikipedia.org/wiki/Git>

Biography

Gavin Sheridan is the co-founder and CEO at Vizlegal – a widely used legal intelligence platform in Ireland – and co-founder at Right To Know – an NGO engaged in access to information litigation. He was previously Director of Innovation at Open-Source Intelligence company Storyful. Right To Know, along with Public.Resource.Org, were successful in a recent significant CJEU case on the right of the public to freely access the Harmonised Technical Standards of the European Union (Case C-588/21).