Legal Information Management, 14 (2014), pp. 244–248
© The Author(s) 2014. Published by British and Irish Association of Law Librarians doi:10.1017/S1472669614000656

Using Data to Understand How the Statute Book Works

Abstract: The statute book is a large, complex system; a vast corpus of texts dating back to the thirteenth century, now evolving at a rate of around 100,000 words a month[1]. The volume and pace of change combine with the constraints of current generation of digital tools to present a real barrier to researchers, limiting the type of research that is currently possible. The statute book is simply too big, and changes too rapidly, for any one person to easily comprehend. This situation is transformed if you view legislation as data, and then apply big data technologies and new data analysis techniques to that data. The aim of the Big Data for Law research project[2] is to do just that; applying the latest analytical techniques to legislation, making it possible to research, interrogate and understand the statute book as a whole system. An important part of the initiative is to make available the raw data for conducting this type of research, alongside new tools and methods for working with the content. In this article, John Sheridan, Head of Legislation Services at The National Archives, sets out some of the ideas that underpin the project and describes the new service that researchers can use from Spring 2015.

Keywords: legislation; statute book; big data; legal research

"I wish that the superfluous and tedious statutes were brought into one sum together, and made more plain and short." Edward VI (1537 – 1553)

The volume of legislation has always been an issue. Today no-one knows for sure how much legislation is currently in force. The Office of Parliamentary Counsel's review, "When laws become too complex"[3] estimates that there may be as many as 50 million words in the statute book, with 100,000 words added or changed every month; a rough equivalent to the complete works of Shakespeare twice a year.

Paradoxically, at the scale of the web, rather than obfuscate, size aids the discovery of new information. The quality of Google's search results is amazingly good, in large part because of how its algorithms exploit the information available. The more data a search engine algorithm has, the better its prospects of identifying a relevant resource. From a certain point of view, legislation has become just another resource on the web. As a result, search engines and web-enabled services such as legislation.gov.uk have transformed access to primary sources of law and thereby the methods used to research it. One consequence of this shift is that legislation, once the preserve of the legal professionals with access to bound volumes of statutes or to others through a good local library, is now available to almost anyone with an interest, in a couple of clicks. Thus we see that legislation is consulted and read by a much wider group of people in society than even ten years ago.

New audiences bring new needs and new expectations. The majority of the people using legislation.gov.uk are not legally trained or qualified, yet they are confronted by the volume of legislation, its piecemeal structure and frequent amendments. The Government's 'good law' initiative[4], launched at the Institute for Government in April 2013, aims to address some of these issues by making legislation more accessible and understandable for UK citizens. Yet important questions remain. If we conceive of the statute book as a large, complex and adaptive system, what do we know about it? How modular or fragmented is it? How easy to traverse? How easy are the components to modify or replace?

There has never been a more relevant time for research into the architecture and content of law, the language used in legislation and how, though interpretation by the courts, it is given effect.

BIG DATA FOR LAW

Imagine you are a historian trying to trace the use of a particular word or phrase in legislation over time, or a legal researcher exploring the effectiveness of different styles of legislative drafting. At the moment, a typical researcher can only repeatedly search legislation.gov.uk, BAILII or a

https://doi.org/10.1017/S1472669614000656 Published online by Cambridge University Press

commercial legal database. This is time consuming and very limiting in terms of what can be discovered.

Outside of a handful existing largely search-based services floated on top of legal databases, researchers lack access to large scale raw legislation data, as well as tools and the methods for undertaking research across the whole statute book. Meanwhile, the combination of low cost cloud computing, open source analytics software and new analytics methodology – the enablers of the so called 'big data revolution' – are transforming research in other fields.

There is a lot of hype around "big data". The terminology has proven very useful, creating a lot of buzz and underpinning numerous marketing campaigns by software vendors. Nevertheless, it is perfectly possible to apply the techniques often referred to under the umbrella of "big data", to legislation.

Big Data for Law, a project funded by the Arts and Humanities Research Council and The National Archives, aims to supply the raw materials for data orientated research into legislation. To do this we are delivering a new service for legal researchers that will be made available from research.legislation.gov.uk in the Spring 2015. The service will include access to bulk data to download, as well as customised tools and methods for working with legislation. Our aim is to enable a transformation in how the statute book is conceived and interrogated and researched.

In framing the project, we set three research objectives: to understand the needs and capabilities of researchers; to ensure researchers have access to as much data as possible, by designing methods and practices that will create useful new open data from closed data; and, to demonstrate what might be possible, to explore the potential for creating a 'pattern language' for legislation.

UNDERSTANDING RESEARCH USERS' NEEDS

Is the legal research community ready or equipped to make good use of big data technologies? Mindful that we are developing new capabilities largely for a non-technical (in the IT sense) legal audience, we need to understand how confident researchers are with data analysis, statistical methods and using tools to interrogate data. What research avenues could the new service would make possible? What do researchers think of our proposition or ideas for the tools that the service might include?

Working with a specialist usability consultancy, we interviewed a wide range of potential users of the new research service, both in academia as well as others such as legislation drafters, policy makers and others, such as those in think tanks who use legislation for research.

Mindful of the lack of capability to process or work with raw data, the first option we tested was providing researchers with a set of pre-packaged data analyses reported on annually, along with the methods used to produce those analyses. We described this as an 'annual census of the statute book'. The concept is to show what is possible, by pre-packaging the data analyses for researchers. We would provide detailed provenance information about the results, so that researchers could quote the analyses with confidence in their veracity.

We also tested the idea of giving researchers downloadable datasets, along with tried and tested tools and examples of data analysis methods. This could, for example, include resources for natural language processing of legislative texts that we have developed for the legislation editorial system used at The National Archives. In this option, we will not have done the research for researchers. Instead, it will enable researchers to answer their own research questions, using and adapting resources and approaches that have been proven to work. Some of these resources have taken years to develop, and this option provides a real opportunity to share the benefits of this public investment with a wider research community.

Finally, we presented the idea of giving researchers access to raw data – the entire statute book as well as other open data, created for the service, for example processing law report data for citations to legislation, or anonymised usage data for legislation.gov.uk. In this option, we will not have done the data analyses for researchers and we have no existing tools or methodologies for them to use or adapt. Providing researchers with access to raw data, in different formats, will enable them to carry out their own analyses from scratch.

A CENSUS OF THE STATUTE BOOK

Our user research revealed that the option many researchers would most like is a set of pre-packaged analyses of the data that they can easily access and use online. To meet this need the new service will provide an online census of the statute book.

When most people think of the statute book they think of words rather than numbers, yet the simple act of counting can reveal much about the law, the evolution of policy, politics, history, as well as the evolution of drafting techniques and practice. Imagine, for example, being able to count how many times a legally significant word or phrase has appeared in legislation. What might that reveal about drafting styles and trends? Imagine counting the number of internal or external cross references in legislation; how interconnected each provision is with the rest of the Act, and each Act with the rest of the statute book. What might that reveal about the complexity of legislation? Alternatively, imagine counting the number and type of amendments to understand, in aggregate, how legislation is changing over time. Are there correlations between the pieces of legislation that are subject to challenge through the courts and those with high numbers of amendments? Are there statistical indications to when a law is 'wearing out'? What might counting the number of retrospective provisions reveal? There are so

https://doi.org/10.1017/S1472669614000656 Published online by Cambridge University Press

many possibilities and yet so little about the statute book has been measured before.

With so many options, the first challenge for the project has been to decide what to count. We tested ideas for core indices with researchers during the user testing. As a result, our annual census will include indices for the number of words, the use of legally significant phrases, the frequency of amendments, the occurrence of internal and external references and the use of powers.

An important piece of feedback from the user testing was the need researchers have to know of instances, as well as counts. Users will not only be able to use these indices to discover the numbers, they will also be able to drill down into the instances. So if a word appeared 50 times in a year, for example, the researcher will be able to drill down into the data to find out where those 50 instances were; which specific pieces of legislation.

One of the challenges of measuring the statute book is knowing what to count. What are the units of measure? In the absence of fine-grained data, previous attempts to count aspects of the statute book have used course-grained units such as the number of pages of new legislation. In the Big Data for Law project we can do better, counting words, enactments and whole pieces of legislation. What you measure for which index depends, in part, on the conclusions you would like to draw from the census data. This is something the project is currently working on. Our approach is to experiment iteratively, refining our methods and approach as we learn more from researchers.

PROVIDING DATA, TOOLS AND METHODS

Despite not always having the capability, our user testing demonstrated an enormous appetite amongst researchers to access raw data for carrying out their own research, in addition to using a pre-packaged census. Some clearly had the expertise and desire to download and analyse raw data with little or no support (the growing 'law/tech' community). For this group of users we will provide legislation datasets in a wide variety of formats, including XML, HTML5 and pdf. We will also provide a number of other Linked Data sources, including a dataset of amendments to legislation, drawn from legislation.gov.uk and various Chronological Tables of Statutes. We will also be creating some entirely new datasets for people to use, such as a "clusters" dataset processed from legislation.gov.uk usage data.

Other researchers recognised the potential of the service to transform legal research, but were not at all confident that they had the technical and data analysis skills required to carry out new types of research with the data provided. For these potential users, the possibilities offered by the service are a little overwhelming.

To support these users we will need to provide a range of tools and information about methods we have used. Researchers can then use and adapt these approaches to carry out their own research. For example, we will make available some of the sophisticated tools we have developed for reconciling references to legislation, and various other capabilities for natural language processing of legislation texts.

Listening to legal researchers we have come to understand that merely providing a set of data and tools is not going to be enough to transform the possibilities for research. When thinking about how to engage and encourage researchers to use a radically new offering, the experience of the salesman pitching the 'Chop-O-Matic' came to mind. This is beautifully captured by Malcolm Gladwell in his New Yorker article of 2000[5], when he writes: "It was, after all, an innovation. It represented a different way of dicing onions and chopping liver: it required consumers to rethink the way they went about their business in the kitchen. Like most great innovations, it was disruptive. And how do you persuade people to disrupt their lives? Not merely by ingratiation or sincerity, and not by being famous or beautiful. You have to explain the invention to customers – not once or twice but three or four times, with a different twist each time. You have to show them exactly how it works and why it works, and make them follow your hands as you chop liver with it, and then tell them precisely how it fits into their routine, and, finally, sell them on the paradoxical fact that, revolutionary as the gadget is, it's not at all hard to use."

We can learn a lot from the Chop-O-Matic and the mindset of introducing revolutionary capabilities that are also very easy to use. Our aim is to introduce a service that is both innovative and disruptive to the nature and type of legal research being conducted. To gain maximum benefit from the capability we are offering will require researchers to think differently, to challenge their preconceptions about what legal research is possible and how it can be done. In a sense, our challenge, too, is to persuade legal researchers to disrupt their lives.

To achieve this, we need to explain the benefits of this new capability to researchers, to show them exactly how to use it, step by step. We need to explain clearly how these steps fit with their current research approaches and we need to make the service as easy to use as possible. We need to clearly describe the data, tools and methods we are providing, supported by plentiful examples and stories of what we have done and how the data we are making available can be used. The service will include case studies that carefully show researchers how we have taken raw data, combined and evaluated it, using tools and methods, to create the results.

CREATING NEW OPEN DATA FOR RESEARCHERS

No one information holder, or provider, has all of the data that researchers might find useful. With our partners, including the Incorporated Council of Law Reporting and Lexis Nexis, one of our research aims is to find ways of creating new open data from closed

https://doi.org/10.1017/S1472669614000656 Published online by Cambridge University Press

datasets. For example, the data we have about the users and usage of legislation.gov.uk cannot be made available as open data but we can process it to create new open datasets, identifying clusters in legislation or creating a 'recommendations' dataset (people who read Act X also looked at Act Y).

One of the new datasets we will make available through the service is N-Grams data for legislation. N-Grams data is about letters, words, or sequences of words and their frequency of occurrence. An N-Grams[6] viewer is an online tool for exploring N-Grams data. It can be used, for example, to analyse the history of language. Google introduced such a viewer back in December 2010, to analyse a very large corpus of digitised texts. The viewer makes it easy to create graphs showing the frequency of use of specific words and phrases over time, using the corpus of historical texts that were scanned and digitised as part of the Google Books project. N-Grams datasets offer far more than a search engine. Search engines simply provide a list of results, which are of limited value. With N-Grams you can find statistical outliers. Our intention is to provide a legislation-specific corpus of content that can be analysed using an N-Grams viewer tool.

This will make it very easy for researchers to look for commonly occurring words and phrases in legislation, to see how often they have been used and if this usage has changed over time. Think of phrases such as 'to or in respect of', frequently used in pensions legislation as shorthand to convey the requirement that money that should be paid directly to a person or on their behalf. Identifying when this is used, how often and what usage trends are, can provide deep insights into changing styles of legislative drafting. Or imagine plotting the use of the words 'asylum' against the use of the word 'insane' and then investigating what that reveals about changing social and linguistic norms and how these have been reflected in legislation.

N-Grams are typical of the type of new open dataset that can be computed from closed data, and can provide an invaluable and easy to use resource for researchers that provides real insight into legislative trends over time.

RESEARCHING A 'PATTERN LANGUAGE' FOR LEGISLATION

Conducting research across the statute book as a whole, using big data technologies, requires us to develop new conceptions and models for the architecture of the statute book. Big data research involves thinking differently and learning from other disciplines.

One avenue we are exploring in the Big Data for Law project is the concept of a 'pattern language' for legislation. The development and use of pattern languages has been transformative in other fields, such as architecture or software engineering over the last thirty years[7]. We think this approach has the potential to transform our understanding of the statute book too, bridging the gap between users of legislation, policy makers and drafters.

A pattern language is simply a set of design patterns. Each design pattern is framed in terms of a structured method, and generalises a particular good design practice. By naming the individual designs, the pattern language as a whole provides a common vocabulary between users and specialists. Pattern languages are most helpful when used to conceptualise and help design or manage large complex systems.

In our pattern language for legislation each design pattern consists of four elements: a name, a problem, a solution and the consequences of applying the pattern. Each design pattern is framed as a generalised solution to a commonly occurring problem. It can be applied in different ways in different circumstances – the value of the design pattern comes through the abstraction. Naming each pattern is really important, as is finding good and pithy names.

A pattern language has never been explored in the field of legislative drafting but has the potential to offer real insights into the shape of the statute book. Our pattern language for legislation will encapsulate commonly occurring legislative solutions to commonly occurring problems, such as licensing, regulation, registration or protection. We believe it has the potential to support policy makers, not trained in law, to think more clearly about what legislative solutions might be available. It may also enable us to better contextualise legislation to lay-users when presenting it online at legislation.gov.uk.

The Big Data for Law project is identifying candidate patterns for legislation in two ways: we are working with experts in law to hypothesise potential patterns that can then be tested for in the data and we will also look for patterns that arise directly from the data, for example looking for correlations in usage data.

We are currently exploring the idea of patterns in legislation with a range of individuals and organisations, ranging from commercial legal publishers to drafters and academics. To formalise the patterns rigorously we are also exploring the use of Hohfeldian jural correlatives[8]. Patterns should also help us to map the statute book and develop a deeper understanding about how the statute book is evolving. Which patterns of law making are most prevalent and how has that changed over time? We will also carry out research with the users of legislation.gov.uk to find out whether certain design patterns aid or hinder their understanding of legislation.

There is a common thread linking researchers, who we are aiding to use data to understand the statute book as a whole system, applying that data to develop methods that ameliorate complexity for lay users of services such as legislation.gov.uk, and in turn developing new strategies for those responsible for framing new legislation. Through the provision of data, tools, methods and analyses, underpinned by new conceptions of legislation, the Big Data for Law project aims to have an impact in all these ways.

https://doi.org/10.1017/S1472669614000656 Published online by Cambridge University Press

FURTHER INFORMATION

John Sheridan is the Principal Investigator for the Big Data for Law project. He also leads the team at The National Archives responsible for legislation.gov.uk and that manages the government's legislation database. The National Archives is both a non-ministerial government department and an independent research organisation. It is in this capacity that The National Archives is leading Big Data for Law project. You can find information about the project at https://www.legislation.gov.uk/projects/big-data-for-law or you can email john.sheridan@nationalarchives.gsi.gov.uk

Footnotes

[1] https://www.gov.uk/government/publications/when-laws-become-too-complex/when-laws-become-too-complex

[2] www.legislation.gov.uk/projects/big-data-for-law

[3] https://www.gov.uk/government/publications/when-laws-become-too-complex/when-laws-become-too-complex

[4] https://www.gov.uk/good-law

[5] http://gladwell.com/the-pitchman/

[6] https://books.google.com/ngrams

[7] See: A Pattern Language (Alexander, Ishikawa, & Silverstein, 1977) and Design Patterns: Elements of Reusable Object-Oriented Software (Gamma, Helm, Johnson, & Vlissides, 1994)

[8] Wesley Hohfeld, "Some Fundamental Legal Conceptions as Applied in Judicial Reasonning" (1913) 23 Yale Law Journal 16.

Biography

John Sheridan is the Head of Legislation Services at The National Archives, responsible for the team that operates legislation.gov.uk. This involves providing a range of services to government and the public, ensuring the law is accessible to all. A career civil servant, John has worked at the intersection of legislation, technology and digital service delivery for the last ten years. He is a recognised and respected leader in the field of legal information internationally. John is currently working closely with the Office of the Parliamentary Counsel helping to shape, develop and deliver the "good law" programme. He has been a regular keynote speaker about legislation, open data and linked data, at conferences and events in the UK and overseas.

https://doi.org/10.1017/S1472669614000656 Published online by Cambridge University Press