Legal data infrastructure

Open Data Institute & Thomson Reuters: discovery workshop

Background:

The Open Data Institute (ODI) is working with Thomson Reuters to explore data infrastructure in the legal sector.

In September 2016, Thomson Reuters and the ODI gathered publishers of legal data, policy makers, law firms, researchers, startups and others working in the sector for a discovery workshop. Its aims were to explore important data types that exist within the sector, and map where they sit on the data spectrum, discuss how they flow between users and explore the opportunities that taking a more open approach could bring.

Our workshop notes are open for you read and comment on. They explore current mechanisms for collecting, managing and publishing data, benefits of wider access and barriers to use. We'd like to hear your thoughts about the types of data discussed and their potential to be more openly available, or types of data we might have missed.

Workshop (2016-09-29) notes:

1. Case law data

- What are the inputs?
    - Parties, parties details, judges, dates, acts, etc
    - Are there unique identifiers and the ability to link stuff up?
    - Not enough variables recorded; are we even collecting the right data?
    - Analog data flows; only some moving to digital
- No consistent access for users in the sector
    - BAILII maintain privileged access to case law, which it then provides public access to
    - Are only select judgements sent to BAILII? Mechanism is there for data sharing, but it seems to rely on a clerk to remember to send to BAILII: can coverage be hit and miss?
    - Transcription companies produce data, which is also sent to legal publishers to add their enhancements
- Dissemination of outcomes predicated on ways of working that were born out of analog data flows
    - Law reports are crucial (but funded on print publishing models)
    - Focused on the end of the process, including transcript of what the occurrence was, details of evidence/judgement, etc
    - What are/should be the roles of court reports, judges, legal publishers, Thomson Reuters in a modern system?
- Current mechanisms for collecting, managing and publishing the data assume expertise; they should be built not to

    - Systems should anticipate that people looking for data aren't necessarily experts in the legal process
    - Has a legacy of outsourcing the technology that ultimately maintains the data resulted in convoluted/restricted flow of data in the sector?
- Who owns the copyright for data collected in court? Answer depends on who you ask…
- Opportunities for modernisation?
    - Technical skills in government; how can we ensure consideration of data infrastructure is baked into the design of our justice system (e.g. the procurement of new transcription services)?
        - Is there an opportunity for obligations to be put on external service providers from the outset?
        - Is there senior political will (e.g. based on the new MoJ Secretary of State Liz Truss' work at Defra and Sir Brian Leveson's work on modernisation)?
    - There are numerous benefits of wider access to case law data
        - monitoring of court efficiency (e.g. can it enable comparative evaluation of court performance?)
        - Support small businesses and startups to use the data to help citizens in their contact with the justice system
        - What are the business cases that will drive work on legal open data? Or should access to justice/transparency be the significant catalyst?

2. Court listings

- Magistrates' courts produce court listings every Friday (which include personal data such as date of birth, name, address, alleged charge, etc)
    - numerous PDFs generated by the MoJ's Libra computer
- de facto available to journalists (either local or those with resources to maintain presence in courts across the country) and those that can make it into the court
- Important to note that listings represent an allegation only
    - Outcome data needs to be made available similarly, otherwise does publication of court listings present an unbalanced picture?
- Thomson Reuters has its own system for RCJs, with full lifecycle of the case (including the judgement)
- How does court listing data relate to other data types? What data is needed to enables users to be able to discover/explore/search ongoing cases?
- The Criminal Procedure Rule Committee have outlined how court listings should be published, with some redacted to protect privacy and other rights
    - Data Protection Act doesn't apply to court proceedings; what about the right to be forgotten and other privacy considerations?
- Do The National Archives keep a record of the published listings?
- There is a difference between public availability and open data
    - In an analog world there was a high degree of friction related to publicly available information (i.e. PDFs put up in each court); what are the implications of open data removing this friction?

    - Data processing challenge: everything is not available in a consistent location and a consistent format; there are currently no APIs
    - BAILII is an example of friction to access; not searchable by Google
    - Should different levels of access be provided?
    - Individuals need to be involved in publication decisions, not just algorithms
    - There is a strong need to discuss more widely (e.g. with victims' rights groups and other interests)
- Challenges/barriers to open: MoJ are restricted by Libra, cannot clean list to remove data that may be in contempt of court or have other associated issues
- Publication of court listings for other courts (i.e. not just Magistrates') should also be explored
- In Scotland publication is done better: name of offender and some abstract info is publicly available
    - What can be learnt from Scotland and other nations?
    - Groups in France are currently working on how to best anonymise case law so it can be more widely used

3. Transcription data

- Audio recordings in court processed into outputs made available by BAILII and others
- Important to note that data is also generated before case comes to court (e.g. legal advice from solicitor, pre-trial docs, police details related to arrest/charge/investigation)
- For some courts, every case will contained within one audio file
- Access to transcription data depends on individual court
    - Do litigants even know it's been recorded and there's a transcript available?
- HMCTS/MoJ/probation are users of transcription data
- What is the difference/relationship between transcription data and court data? i.e. does transcription data allow for the generation of some court data variables (names, charges, etc)? Are we talking about the same thing, differing in format?
    - what are the key sources of the data that underpin the products Thomson Reuters and others offer to law firms?
- Costs can be huge (e.g. appeals court, £189 p/hr)
- Private transcription companies with outdated transcription processes
  Procedure for ensuring that nothing is disclosed in error: user must go back to the court and check nothing in transcript can't go public
- No recordings made in Magistrates' courts?
- What happens when transcripts are handed over to The National Archives? (e.g. do rights issues fall away?)
- What is going on with court transcription services in Scotland and Northern Ireland?

4. Legislation

- Is access to historic private acts of parliament provided?
- Archiving is done well, but what about access in drafting/amendment stage?
    - E.g. when a bill is in the house, could real time tabling of amendments be published (e.g. at committee stage)?
    - No way for people to understand what is going on in real time

- More demands on The National Archives by government post-brexit

5. Non-legal data

- Examples include data published by Companies House (e.g. articles of association). Land Registry (e.g. planning documents), London Gazette (e.g. insolvency records)
- Difficult to link and derive insight across datasets published by government
    - Hard to link metadata back to original documents
- Need analytics tools to properly interrogate records; can be too expensive for some users to write/access the tools to do that
- Could be better harnessed for business development and due diligence for clients
- Barriers to use
    - Quality and accuracy of data
    - Liability of data (who is responsible for its accuracy?)
    - Provenance
    - PDFs can limit utility
    - Inconsistent locations for data discovery

6. Which stakeholders are missing from today's workshop?

- HMCTS
- Transform Justice
- Victims' rights groups
- Police (and statutory users)
- Private sector clients
- EU level interests

7. Related links

- The Criminal Procedure Rule Committee publication of court lists discussion paper and recommendations
- Protocol for sharing court registers and court lists with local newspapers
- Report on the Accessibility to Judicial Decisions through Publication Standard
- Transforming our justice system: summary of reforms and consultation
- MoJ: Data principles - the right ingredients to solve the data spaghetti problem
- openlaws project deliverables on Zenodo