DATA CURATION INSIGHTS

JENI TENNISON

TECHNICAL DIRECTOR

Open Data Institute

BIG

Big Data Public Private Forum

Introduction

Jeni Tennison is the Technical Director of the Open Data Institute. She originally trained as a psychologist and knowledge engineer, gaining a PhD in collaborative ontology development from the University of Nottingham. She went on to work as an independent consultant and practitioner, specialising in open data publishing and consumption, including XML, JSON and linked data APIs. Before joining the ODI, Jeni was the technical architect and lead developer for legislation.gov.uk, which pioneered the use of open data APIs within the public sector, set a new standard in the publication of legislation on the web, and formed the basis of The National Archives' strategy for bringing the UK's legislation up to date as open, public data.

"The challenge of open data is how to best get that data published in a way that other people can get insights from it. Privacy of data is also a key issue."

Edward Curry: Please tell us a little about yourself and your history in the industry, role at the organisation, past experience?

Jeni Tennison: I am technical director at Open Data Institute with responsibility for all the technical side of Research and Development. I am interested in making sure we have right the technology that helps people to produce open data, from a single figure such as a VAT rate to fast moving large data sets. I am also interested in technology that helps in discovering data, in linking data together data. From a Data Curation angle my interest is in how we help people to collaborate over datasets that are open. Before I joined ODI I worked for legislation.gov.uk as a technical architect where we had many challenges around curating large databases.

Edward Curry: What is the biggest change going on in your industry at this time? Is Big Data playing a role in the change?

Jeni Tennison: I am in the open data industry and the biggest change is simply adopting, and understanding the role and value of open data. Big data plays a role in that a lot of companies and public organization recognise their data has value. The challenge of open data is how to best get that data published in a way that other people can get insights from it. Privacy of data is also a key issue.

Edward Curry: How has your job changed in the past 12 months? How do you expect it to change in the next 3-5 years?

Jeni Tennison: It is odd question for me because I just started working in my current job last October. My job didn't exist 12 months ago. Currently, the way open data is tackled focuses on individual dataset and the conversation tends to focus on the legal questions surrounding if the dataset can made open . What I think we will see happening in next 3-5 years is we get to the stage where there is sufficient amounts of open data available so that conversation will be more about how we can link data together to get much better insight. I think we will be getting into more technical questions about how do we best create linkages between dataset published by different organizations and which technologies we can use to link data together.

Edward Curry: What does data curation mean to you in your organisation?

Jeni Tennison: legislation.gov.uk presents a really interesting challenge around the curation of data. Partly because its unstructured data which is legal text rather than figures and numbers which is what we think of as Big Data. Within legislation.gov.uk we have all laws that were enforced in 1991. Those date back to around 1287 for the first laws. Curating that data is a matter of managing two sets of changes that happen. One set of changes is the new legislation that gets brought into the dataset and the other set of changes is as new legislation comes into force. The new legislation effectively overrides, or amends, the old legislation. Curation of the older legislation is a matter of identifying where those changes happen and writing that into the content of the legislation itself.

Edward Curry: What data do you curate? What is the size of dataset? How many users are involved in curating the data?

Jeni Tennison: Legislative texts which are semi-structured documents with parts and paragraphs that are numbered. The size of the dataset is around 70,000 items of legislation that vary in size from 2 pages to 500 pages approximately. There is a huge range of different sizes of legislation. The thing that makes it most interesting is that we try to capture each single point of time in those legislations. So every time a legislation changes occurs, we take a new snapshot of that item of legislation. While some items don't change that much, others can change hundreds of times over the course of their lives. We are taking in tens of gigabytes of documents. It is published as the raw XML data, also as HTML pages that people can navigate around and to create PDFs that people can print out.

Edward Curry: What are the uses of the curated data? What value does it give?

Jeni Tennison: People look at legislation in order to tell what the law is currently, so that they know what they need to do. They often also

"Open data helps people feel that they have ownership over the data that is being produced and brings up all of these challenges on how we manage data curation in that kind of environment."

want to know what the law used to look like, in particular if you imagine a court case taking place several years after the initial offense, knowing what the law stood at that time is important.

As well as the legislation itself, there is a whole lot of data about the changes of legislation that have happened over time. The data gets updated constantly day by day as new legislation comes in, that also has value in that if you are an organization that provides legal advice then you need to know when those changes happen. It also gives value to the UK as a whole for anybody who needs to access legal information including lawyers, campaigners, ordinary people, and whole range to people.

The biggest challenge we have had with the legislation is simply keeping it up to date, when there are multiple changes coming in over time. So we are running about 6 years behind in terms of the legislation that is on the site and how it was about 6 years ago rather than how it is now. This situation that has been the state of affairs for years and years. It is never actually been fully up to date. A large part of my efforts and the effort of the team at National Archives has been around redoing the processes around managing and curating that data, so that the legislation could be brought up to date. Because it is a huge manual task, just in terms of understanding what effects mean in terms of change in the text in legislation and then actually making those changes all currently involve human intervention. A lot of the changes that we are doing was to make them in such a way that they would either not involve human intervention or to lessen the requirement of human editing. But that work has been more or less done. There are lots of editing work that we still need to do but from the technical level then managing the changes and helping people to edit them has been more or less done.

When I look at the future of legislation and what I think the next challenges are from a data perspective, it is about trying to extract the extra information that is held within legislation that could really form a bedrock of understanding of the way that the UK thinks about itself and the way it works. For example, legislation has loads of definitions of terms like company, academy, or motorway, and loads of definitions of terms that are important from a legal perspective and important from the perspective that how we in the UK understand the ways UK operates. There are also lots of lists of things, for example one particular piece of legislation that is produced every year has a list of all of the government owned departments, agencies that should be listed as part of government accounts. So that list of government organizations is an important source of information about how government thinks about itself. I think is 3-5 years time we will be looking at how to extract that extra value out of legislation itself, how to analyze the natural language that is used there in order to produce references for definitions of terms and references for lists of things that would useful in lots of other context as well.

Edward Curry: What processes and technologies do you use for data curation? Any comments on the performance or design of the technologies?

Jeni Tennison: The old process was kept completely in house and involved people basically mapping out the changes by hand on hard copies of legislation. This was then translated into excel spreadsheets which were merged into bigger excel spreadsheets of changes and then went into a set of transformations to get into the website. The new process that we are aiming at involves a lot more automation. Automatic markup of the legislation, automatically persisting those changes into a database of changes, which then is published into the website. Helping people by providing much more structure to their work, providing online editing. But perhaps the most fundamental change in how the data was managed was instead of it being managed by a small group of people within the national archives, the curation of the data was opened up for other people to contribute to its management. This includes private sector organizations, includes other government departments, academics, all of them being motivated by having the laws up-to-date.

Providing the technologies for enabling people outside the organization to be able to make the same kind of changes from people inside the organization was for me a fundamental shift in how data gets managed and curated. However this opening brings together many challenges. The review process is an important part of the workflow, where changes are reviewed and validated by a well-defined process, being able to determine and evolve a trust relation with individuals outside the organization.

Open data helps people feel that they have ownership over the data that is being produced and brings up all of these challenges on how we manage data curation in that kind of environment.

Edward Curry: What is your understanding of the term "Big Data"?

Jeni Tennison: Large datasets and rapidly changing datasets. However, the term seems to have been overused. I have seen people using "Big Data" for things that they can cover in a spreadsheet.

Edward Curry: What influences do you think "Big Data" will have on future of data curation? What are the technological demands of curation in the "Big Data" context?

Jeni Tennison: One of the things I'm exploring here is this idea of collaboration over datasets which is an important aspect of how we should be managing data in the future. From a Big Data perspective the challenges are around finding the slices, views or ways into the dataset that enables you to find the bits that need to be edited, changed. From the Big Data perspective we need an approach to manage targeted small changes on large datasets.

Edward Curry: What data curation technologies will cope with "Big Data"?

Jeni Tennison: From my point of view, if I look at the requirements for data curation technologies then I will be targeting the ones that enable collaborative approaches to curating data. An approach that enables multiple users to change the same dataset in a simple manner with some level of protection so that people can review the changes as they come in. These will be the types of technologies that will be most interesting in order to manage the collaborative management of huge data sets.

My experience with legislation was that this had to be done in a very specialized manner around the particular domain. For example, if we look at open street map that needs to be specialized around geographic data. I don't know whether the solutions and the technologies are going to be those kind of specialized technologies or whether there will be more generic solutions.

About the BIG Project

The BIG project aims to create a collaborative platform to address the challenges and discuss the opportunities offered by technologies that provide treatment of large volumes of data (Big Data) and its impact in the new economy. BIG's contributions will be crucial for both the industry and the scientific community, policy makers and the general public, since the management of large amounts of data play an increasingly important role in society and in the current economy

TECHNICAL WORKING GROUPS AND INDUSTRY

http://big-project.eu

CONTACT

Collaborative Project
in Information and Communication Technologies
2012 — 2014
General contact
info@big-project.eu

Data Curation Working Group
Dr. Edward Curry
Digital Enterprise Research Institute
National University of Ireland, Galway
Email: ed.curry@deri.org

Project Coordinator
Jose Maria Cavanillas de San Segundo
Research & Innovation Director
Atos Spain Corporation

Albarracín 25
28037 Madrid, Spain

Phone: +34 912148609
Fax: +34 917543252
Email: jose-maria.cavanillas@atosresearch.eu

Strategic Director

Prof. Wolfgang Wahlster
CEO and Scientific Director
Deutsches Forschungszentrum für Künstliche Intelligenz

66123 Saarbrücken, Germany

Phone: +49 681 85775 5252 or 5251
Fax: +49 681 85775 5383
Email: wahlster@dfki.de

Project co-funded by the European Commission within the 7th Framework Programme (Grant Agreement No. 318062)