DATA CURATION INSIGHTS

HELEN LIPPELL
METADATA MANAGER

PRESS ASSOCIATION

BIG

Big Data Public Private Forum

Introduction

Helen Lippell is an experienced information professional with extensive experience of managing data and metadata for organisations across the media, financial, government and entertainment sectors. Her passion is building knowledge organisation systems such as ontologies to support a wide range of business applications such as search, content publishing and business intelligence.

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

Helen Lippell: I am working at the Press Association (PA) which is a partner in BIG. It is a very well established news agency for the UK and Ireland. My job title is metadata manager, which means introducing metadata decriptors into our core content. It is a fairly new role, but it basically means applying more semantic information to content in order to package it more flexibly. The traditional product for the PA is the Newswire. There is a need to use that data more to describe it better to curate it better.

I have been working about 15 years in this industry. I started working in media and financial publishing organizations including The Financial Times, BBC. I would describe myself as an information professional. I had different j ob titles but it always boils down to make sense of data and content, structuring things describing things.

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

Helen Lippell: I think it is related to the sheer increase in volumes of data that can be captured. Also it is related to the fact that now everyone expect to watch and consumer whatever they want, when they want on whatever device they want, al of which can generate enormous amounts of data. I guess the big media organisations and PA is one of them, is working out on how they fir into that landscape, because the traditional business model might not always be effective.

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?

Helen Lippell: I just moved to a new role that is still shaping. I would say more thinking about the future of the information professional role. I come out from a traditional librarianship world which means describing and structuring things for tousands of years. But, now we've got online data. I think understanding how these roles are changing, in partuclar for data curators. One change is the need for analytical skills: trying to make sense of massive information, trying to extract value from it. Extracting meaning from large volumes of data.

> "We understand that it is not about a single tool but a whole suite of tools is needed to make sure that the data curation suits into the whole lifecycle effectively."

How can this be applied to business need (how they can make money) or how governments can govern effectively. So bridging that gap between technological and statistical. The role is now becoming more technical.

These roles have been in the industry for a while but sometimes they get obscure job titles fro some reason.

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

Helen Lippell: Having some processes in place to actually make sense of this enormous fire-hoses of words that media companies produce. And being able to respond to customers' requirements. If someone brings up a bunch of content, well Google search cannot deal with them due to noise. So data curation means in this context tagging the content appropriately, building semantic graphs to find what is related to what and then applying those concepts to the content, in order to satisfy those customer needs.

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

Helen Lippell: I am curating data that is newsworthy like people, organizations, events, places, and using a mixture of public domain information like Freebase, with BI kind of knowledge and insights, and exporting to an RDF triple store to be used by journalists when creating content. The data we are producing are stories, images, multimedia and the identification of key concepts within that.

This is one part of the story. PA is an extremely diverse organization. For example we have a subsidiary that does weather forecasting: they could be dealing with enormous volumes of data as well.

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

Helen Lippell: Broadly speaking PA provides media and content to large a range of customers. Newspapers, BBC, sport sites, niche websites, local newspapers, etc.

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

Helen Lippell: PA uses a metadata management tool which has in-built functionality and APIs which crawls data from the public domain and injects into our system. Later, unique identifiers are assigned to the public content and data curators check and validate the data before pushing it into the production system.

Journalists will progressively getting involved into this process.

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

Helen Lippell: In my opinion, Big Data is an exponential increase in data which unprecedent in the human history. It is not just about the exponential growth in data but also about the technology that goes along with that. It is related to the technologies that can cope with the exponential growth of data. It is closely related to cloud computing which offers the flexibility, scalability and cost advantages that make large-scale data processing viable.

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?

Helen Lippell: The amount of digital information and the volume of data we have today, it requires data curation to be looked at very different ways. We have got millions of information resources and extracting the value from all of these requires a different kind of information literacy. I think the whole Big Data movement will provide some sort of applications or algorithms which data curators will be able to use in order to find the real value hidden deep inside millions of information resources.

From my perspective things like text analytics are core technologies. Text analytics have been around for a while, but only now it has gone into mainstream, enabling data curators to work with massive corpora, to find what they need.

.

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

Helen Lippell: We understand that it is not about a single tool but a whole suite of tools is needed to make sure that the data curation suits into the whole lifecycle effectively. All the stuff around the community, review cycles and how to integrate it with the rest of the architecture is central. Tools need to be flexible to export the data in whatever format is needed.

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.

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

http://big-project.eu

TECHNICAL WORKING GROUPS AND INDUSTRY