DATA CURATION INSIGHTS

ANTONY WILLIAMS

ROYAL SOCIETY OF CHEMISTRY, CHEMSPIDER

RSC Advancing the Chemical Sciences

BIG

Big Data Public Private Forum

Introduction

Antony John Williams, is a British chemist and expert in the fields of both nuclear magnetic resonance (NMR) spectroscopy and cheminformatics at the Royal Society of Chemistry. He is the founder of the ChemSpider website that was purchased by the Royal Society of Chemistry in May 2009. He is a science blogger, one of the hosts of the SciMobileApps wiki,a community-based wiki for Scientific Mobile Apps, and is a book author.

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

Antony Williams: My role at Royal Society of Chemistry is to run the e-Science team and the projects that we are involved with. My history is, as a chemist I came up through academia, then through government labs, then through Fortune 500 labs, working at Eastern Kodak company where I got into informatics side of managing of data, then moved into a startup company where we were building software tools for processing and managing data and then I started a project called ChemSpider which some would say is dealing with Big Data of Chemistry. ChemSpider is a website for chemists which has millions of chemical structures and associated information in databases but also links out on the Web, we are literally going to make the Web searchable by chemistry. My role at the RSC is the head of team that builds that part of the project. Not only do we manage large database of chemistry, we are also involved in delivering on a set of grants that we have won. The Big Data one is the OpenPhacts project which is a Semantic Web project integrating chemistry and biological data across multiple public datasets that exist in the domain. Another one is the chemical database service for the UK, so it is a national service that will be around for 5 years where we have to deliver access for academics to chemistry related data from commercial systems as well as some commercial tools. The biggest part of the project is for us to build a data repository to host the various forms of chemistry data that exists. So we have to figure out how to deliver that. The project started in January, we are only in April, it would start out in June in terms of building the repository aspect.

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

Antony Williams: There are actually two answers to this question, depending on which hat you want me to wear. So, I work for a publisher (RSC is a publisher). We publish tens of journals for the scientific domain of chemistry. So there are multiple changes going on in the publishing industry. But my role is to build the chem-informatics platform that supports chemistry data for the community. The biggest change in the publishing industry to push for open access. The age of open science is upon us, open access

> "In my understanding, Big Data can be understood as an explosion in data to the point where it becomes really hard to manage by humans alone and requires machines to process and digest the information."

open data, open source, open standards, there are increasing demands from funding agencies to make publications open access, there is an increasing number of chemists who expect data to be fully available to them to use and repurpose as they see fit. That's quite challenging right now in terms of funding open access, as the funding agencies may be asking for it but, are they providing enough funds to port publishing into open access format? Also there are so many conversations going on in the domain, from the scientist side and from the publisher side, whether the open access model will work. Is it sustainable on long term? Who will pay for it? Is it really a benefit? Mostly because of a lot of the open access publishers that are coming around are setting themselves as very small businesses and quality of what they do is questionable. There are some excellent open access publishers, but there is a whole of them coming online who blip up for few months and then disappear. So there are many questions around the future of open access.

The open data side is very interesting because of the data that is generated on the public funds is increasingly becoming available through government. Over here in the States where I live, the government is insisting that a lot of data should become public. Chemistry is little different than Biology and Physics. It is generally true that you can monetize chemistry in a very different way than you can monetize Physics and Biology. For example, I would say that the Protein Data Bank is all public and the GenBank was also made public easily. In the world of chemistry, if you discover a single chemical entity, you can monetize it for billions of dollars through the Pharmaceutical industry. So chemistry is very patentable, that is not to say that physics and biology aren't. It is just that more and more of their research becomes available online very publicly and very openly. So we are up against challenge where some chemists want to see it go open and many chemists want to see it stay closed behind doors until it is published. So there is a tension in the community around what should be open and what shouldn't.

So, now there are many open standards that will support the data exchange. There are lots of open source efforts that support what is going on in the domain of chemistry. The tool sets are mostly ready. The data standards are mostly ready, to choose when people are ready to free up the data.

In terms of data, there are some really good standards that can be used. None of them are perfect but all of them can be tweaked. Things like ontologies, they are definitely usable and they can be extended as necessary. We have got a really good head start on things, I would say.

If you look online there are so many ways to access data. Primarily people use the term 'google' as a verb rather than site today. There is so much data in the chemistry domain to find easily. We are coming to be amenable to contributing data to that. So as publishers if we don't allow search engines to index us, we are going to be less discoverable. We are definitely contributing data but we are also consuming from it. What we do at the e-Science side of building up platforms, we pull data out of those search engines very effectively, by using very specific searches. There is an enormous data cloud out there that we can tap into at this point. Big Data is playing a role in our ability to deliver.

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?

Antony Williams: To do more with what is becoming available. The tool sets that are available to tap into the online data continue to change and continue to expand in capability. We are running to stay caught up. I would say that we are leading many parts of the Chemistry community. But there is so much going on the area that we are also always catching up.

Certainly more chemistry data and more expectations in terms of what could be done with it. Certainly we are going to continue our move into dealing with Semantic Web technologies. Contributing open web data and being amenable to sharing what we have been up to in order to stay relevant.

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

Antony Williams: So I am probably one of those people who actually would get shot in front of some audiences because certain audiences think of data curation as a long term storage and availability, which we also consider an aspect of that, but when we think of data curation we also use the term primarily around validation of data, and identifying what is more valid information than noise. In the world of online chemistry for sure, you have no problem finding answers, what is more likely to be right than wrong, what's junk, what's disposable, what's valuable. So we developed the process where we provided tools to the community to contribute to data curation, in this case meaning data validation, so we hold a lot of information resources from over 400 data sources and the community contributions. But for us data curation means validation, checking, assuring that is high quality, it also comes with the longevity of information, so that we were holding it in a manner that will be accessible in the long term.

In terms of data standards it means that the data that are generated from certain instruments needs to be converted from its binary form into something that is represented in a data standard that is extensible, otherwise the data generated three years ago, if the binary format changes, there is no way to open it without access to the original process and platform, so we have to think about that ahead of time.

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

Antony Williams: So in term of chemistry and the chemistry focus the data sets we are hosting right now is 28 millions chemicals with a lot of associated information: chemical names, analytical data, patents, publications. In terms of physical size, a couple of Terabytes at most. It can be surprising for people how much the 300.000 publications we have in our entire RSC publishing archive back to 1841 could be carried around on a single hard drive. So that said, people would not think about that really as Big Data, but in our domain it is very significant.

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

Antony Williams: In our domain curated data takes us down the path of being able to serve data of value to the chemistry community. For example, if somebody gave you a boiling point for a chemical and it happened to be wrong, a chemist that did a distillation with the wrong information, then you got an explosion on your hands. It can be life saving or life threatening depending on the way you look at it. It also assist in development of models so with good data you can develop very good predictive models and therefore you could cut costs, because you could model the majority of property you might had to measure. Some properties might cost ten bucks a measure and others cost hundreds. The pharmaceutical industries did that over the years, the predictive technologies are very good in some areas. So good quality data means good models. Means faster science, better science. So it can definitely improve science.

The curation of certain aspects of our datasets allows us to do a very good job of linking up the Web. So the validation of chemicals against their various chemical names allows us to link into Big Data clouds such as the ones served by Google. A drug might be called taxol. If I do a search on google for taxol it would find everything on taxol, using taxol as the search string. But taxol happens to be called many other things. When we aggregate all the different names, bring them all together, and use the different names to search simultaneously across different datasets we bring the users closer to complete dataset. Again it leads you in the path to better science.

Edward Curry: I just want jump up back to question five for a second. I just want to clarify the fact, you have 28 millions chemical entries, and you are basically managing information for that 28 million entities across every type of form that you find relevant to these data (e.g. patents, publications). Do you have other data sets there as well?

Antony Williams: Yes, there are other datasets as well, data from the Pharmaceutical industry, assay data, biological data that we can link to. If you look at the other side of what other data do we curate we have very large publishing archives, chemistry articles that were published, but large means large enough to fill a hard drive. In our domain this is a very significant size. There is a significant amount of content in there.

Going back into what the future will look like in a couple of years. The publishing industry will likely to be pushed, dragged into a publishing revolution called nanopublications and this is the smallest unit of knowledge that can be extracted. So, a single mass

"…with good data you can develop very good predictive models and therefore you could cut costs, because you could model the majority of properties you might had to measure. Some properties might cost ten bucks a measure and others cost hundreds.

publication a 10-15 page PDF file, can contain thousands of nanopublications and a nanopublication is a very small unit of knowledge, for example chemical A has melting point X. This drug interacts with this protein and binds to this level. Those are distinct units of information that when scientists read papers they are after those specific pieces. They don't really care about the entire paper. They read the paper to pull out these valuable nuggets of information. We are definitely going to be into many millions of nanopublications. They will be represented in XML or RDF which can create some very large datasets. Considering the Openphacts project that we have been working on, by the time we map together the biology and chemical entities on the public domain we are expecting something around 30 billion triples.

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

Antony Williams: We use a mixture of computational approaches some of which are house built. For example, we have built our own tool called CVSP (chemical validation and standardization platform) which can be used to check chemicals and tell quickly whether or not they are validly represented or if there is any data quality issues. We use standard technologies for hosting data, for example, SQL server as a backend and standard web platform ASP. NET. We are now moving more towards Semantic Web type platforms.

From the performance perspective: because of the nature of the searches that we have to deliver to the community, we are more optimized around standard querying rather than Semantic Web querying. We are also using a wiki-like approach for people to interact with the data, so that they can annotate it, validate it, curate it, delete it, flag it for deletion, all of this can be done via wiki. We are just about to roll out in the next couple of weeks a rewards and recognition system that will recognize the contributions of the scientists to the data validation and engagement process. Their contributions will be measured and made publicly viewable and accessible. This is aligned with initiatives such as altmetrics which aims at providing alternative metrics for scientific contributions, such as contributions to datasets, code, etc.

Edward Curry: Specifically on the size of the community that you are engaging with. How many scientists are actively involved with?

Antony Williams: In terms of the usage 13-14.000 unique users a day and it is growing probably 30-40% a year. In terms of annotators, on a daily basis we are talking about a couple of dozens. If we take another platform as an example, such as Chemical Wikipedia there are around 10.000 articles and around 6 people doing the majority of the work. Lots of users but very few givers.

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

Antony Williams In my understanding, Big Data can be understood as an explosion in data to the point where it becomes really hard to manage by humans alone and requires machines to process and digest the information. For example, let's take the example of Google, which have devised high performance algorithms and approaches to pull information from enormous data which is out there on the Web. All this information cannot be handled by humans alone and so we require machines which do these and other computational tasks for us.

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?

Antony Williams: Algorithms are becoming more and more intelligent and I believe we will get to the point where algorithms will contribute to the "likesonomy" we are in. Data can be validated, approved and annotated in fractions of a second by an algorithm which might take hours for a human to do the same task. In other words, algorithms will be able to decide the quality of data using data validation against reference standards. In fact some of the search engines are actually doing some sort of validation by checking data across the Web.

From the technological perspectives, making things more mobile and interactive is incredibly important. Let say, interfaces and ease of actions is highly desired. We are probably going to provide accessibility in a better way so that people will be using mobile platforms more than they are using today, hence making it a lot easier to engage with the data.

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

Antony Williams We should come up with ways to encourage participation. Because many of the people that I talk to have the skills, some do have the time to contribute and participate but I commonly get the question "What is in it for me?". The rewards and recognitions approach and the associated technologies, delivering possible ways to contribute, having measures of their contributions coming out of the system, that are traceable back to them, I think will be a big one.

With the appropriate APIs mashing together these systems (crowdsourcing platforms, scientific contributions metrics) and with big players behind them want to use them, I think this thing will move forward. If you take the example of Google scholar you can think of scientists not being measured by how many publications or how many times they are cited. If Google Scholar also start taking into account contributions such as the number of Wikipedia Articles, datasets, scientific blogs you would see bigger contributions on this space. A fundamental step for this is the adoption of a scientific researcher ID (such as ORCID). ORCID is like a SSN for individual researcher. Publishers are going to start to use ORCID on their publications. Scientists are going to use their ORCID number to link up to their datasets. This will allow a much more fine-grained understanding of the scientific contributions.The funding agencies and the university bodies are working together on the UK (I'm not sure when this is going to be approved), to give an ORCID number for every researcher and every student.

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)