DATA CURATION INSIGHTS

JAMES CHENEY

ROYAL SOCIETY UNIVERSITY RESEARCH FELLOW

BIG
Big Data Public Private Forum

Introduction

James Cheney is a Royal Society University Research Fellow in the Laboratory for Foundations of Computer Science, University of Edinburgh, working in the areas of databases and programming languages. He has also been involved with the Digital Curation Centre and during 2008-2009 he organized a Theme Program on Principles of Provenance for the eScience Institute. James has a PhD in Computer Science from Cornell University.

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

James Cheney: I hold a Royal Society Research Fellowship, which is a multi-year research award at the University of Edinburgh at the School of Informatics. My background is in programming languages and a lot of my work over the past few years has been on the interaction between programming languages and databases. In particular, I do research related to scientific databases, data provenance, curation, annotation, data archiving and understanding the evolution of data over time. When I say "we" later in the interview I'm referring to joint work with several people, particularly Peter Buneman (Professor of Database Systems in Edinburgh).

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

James Cheney: Big Data is a popular buzzword now, especially in the database community, but the research issues it raises have always been concerns in databases. Although programming languages researchers don't seem to refer to what they're doing using the term Big Data, many people are working on topics such as cloud computing, GPU computation, database programming, or data parallelism that are relevant. The term is a little nebulous (I guess some of the later questions [in this interview] are going to address this). If I'm doing something involving databases, how do I decide if it's really about Big Data or is it only Big Data if it there is so much data that conventional database techniques don't work and you need to do something else? But certainly it seems that there is a lot of activity on something called Big Data even if nobody can quite say what it is.

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?

James Cheney: I guess I would say that the research part of my job hasn't changed much but the relative importance of emerging research topics have changed a bit over the last 12 months. Over the next 3-5 years, there is likely to be a gradual change in the curriculum for undergraduates and masters students to take more recent developments (cloud computing, new programming models, etc.) into account. This is already happening, for example others at Edinburgh teach a course on "Extreme Computing" that covers things like MapReduce, service computing and virtualization.

> "I guess I would say Big Data is not about the raw amount. Big Data is more about the rate of change; the amount vs. the resources that you need to deal with it."

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

James Cheney: I've been involved in research on data curation (what it is and how we can make it better), and a lot of our research has been motivated by looking at what scientists are currently doing in biomedical or other scientific data settings. Broadly, we've been interested in situations where people build up a scientific database through a mixture of manual and automatic processes, for example, by copying and pasting data from other databases online, or by looking at journal articles and doing manual data entry. Some of these are single-person or very small team efforts, and on the other end of the scale, there are much larger projects like EMBL/EBI [European Molecular Biology Laboratory/European Bioinformatics Institute] projects that have tens or as many as a hundred people working on building databases, and on building tools to automate the curation of these datasets. I would not claim that I know all the ways in which people approach data curation, but we've looked at a lot of different scenarios and proposed general-purpose techniques for improving curation, or at least modelling what curation is so that we can design general-purpose systems that provide better support for it.

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

James Cheney: Instead of talking about data I curate, I'm going to talk about curation projects we've interacted with, mainly in biomedical data. I've mentioned EMBL/EBI data resources. We are collaborating with EMBL and others in an EU project called DIACHRON aimed at preservation and curation for data changing over time, or *diachronic* data (as opposed to synchronic, or single timeslice data). In this type of larger project there may be 10-100 people (typically with PhDs) whose job is to curate data. This project is in early stages so we are still learning about their practices. We have also kept in touch with are two databases organized and curated by people here in Edinburgh. These are representative smaller databases. One, the Nuclear Protein Database, is largely maintained by one person and has been developed further through student projects. I guess it is around 5-10 MBs in size. But again, it was curated by one person over several

> "If the amount of money that you need to spend at data cleaning is doubling every year, even if you are only dealing with a couple of megabytes that's still a big problem."

years, so the amount of effort involved was substantial. The second, larger database is called IUPHAR-DB, the receptor database of the International Union of Pharmacology. We have worked jointly with them in several ways, including a project recently funded by NSF on data citation. IUHPAR-DB is maintained by a small team of 2-3 curators and database developers in the Queen's Medical Research Institute at the University of Edinburgh, and there are another 70-80 people worldwide serving on committees to contributing different sections of the database by sending updates to the curators.

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

James Cheney: Again, just to be clear, I'm talking about things that other people have done, and we have been interacting with them at the research level. Generally, in all of these cases, the principal goal is to build scientific resources valuable to a community. And the longer-term goal is to facilitate new kinds of research by integrating multiple sources. For example, drug discovery is an obvious motivation for the pharmacological databases, and medical or biomedical research questions for the other kinds of databases. In the medium term, the data is intended to support scientific research, and in the longer term, the hope is that this will benefit society through scientific breakthroughs, for example better understanding of how to treat rare diseases. But I think in the short term there is potential for benefit when people interested in (or suffering from) a particular medical condition can learn more about either the care they are undergoing or their treatment options, in combination with gene sequencing services. These are not really what these databases are intended for, but there is potential for people to try to use these resources for "do it yourself" personalized medicine.

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

James Cheney: So, again, though we are not the primary developers of these types of systems, I can comment on what developers seem to be doing and things that we have been picking out as potential research issues, where we see an opportunity for improvement. We are generally looking for interesting research problems that can lead to better tools for data curators, but need to make sure there is research value, which sometimes makes it more difficult to do things that will have immediate value to users. There's a chicken-and-egg problem there.

I guess there is a question on what do you mean by technologies for data curation. Do you mean technologies people are using to develop new curation tools, to host and publish the curated data, or technologies data curators are actually using on a day-to-day basis to do the curation. In the latter case, people seem to be using fairly standard tools that are not specific to data curation. Sadly, in most cases, I think, people are using things like web browser and [Microsoft] excel, word and email. I don't know if it's fair to criticize, since these are standard tools, but if you want to maintain an accurate picture of the process that is going on to build the data then you basically have to maintain it by yourself. These tools don't talk to each other at that level, and that is not going to change anytime soon.

Somebody sends the curator an update to the database, they check it and put into the databases. The database may record that you made the update, but it won't record the outer context, what changes were made between the email and the actual data that was entered which was cleaned up a little bit. We know how to support this if everyone involved would use a single, monolithic system. But nobody is really going to work that way. So how can you do this in a larger, more realistic, heterogeneous setting, with different applications communicating with each other about provenance without radically redesigning all existing applications?

In regard to which tools developers are using to build databases and websites, it's basically any popular web programming framework and any popular database system. We have seen a lot of Java or Php and MySql types of systems. Those are all stable technologies, but they are not providing direct support for the kinds of things that seem to come up over and over for these systems like history tracking or keeping track of who has done what to the data. These are all things that you can build on the top of the existing databases. But if everybody does it separately, first, you have to know what you are doing and it takes additional effort (hence, not everyone does it), and second, it is very difficult to combine this data from different sources later if everyone reinvents it slightly differently. So these are some of the issues where we think there is a benefit to research on general-purpose solutions. How should we design programming languages or database query languages with support for provenance and curation in mind, and provide efficient implementations that provide reasonable default behaviour for the provenance and archiving aspects of curation, so that people can customize a platform that does most of what is needed already instead of requiring everyone to start from ground level.

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

James Cheney: Big Data is relative, and seems to mean you have data that is larger than anything that you dealt with before, and your assumptions break down. I also see it used to describe problems that do have solutions already but the person using the term doesn't know about them. The meaning depends on the context you are in. So if you are talking about the Large Hadron Collider, astronomy data, or high-throughput sequencing, then there is quite a lot of data, and there are undoubtedly major challenges dealing with it, but any given megabyte might not be really crucial. There is an economy of scale: it may have taken a lot of effort to build the instrument but once you've built it, the data is automatically generated and you can easily get more or even repeat an experiment to recover lost data. On the other hand some data really is irreplaceable, or some operations are impossible to automate, and there is no economy of scale. For example, when you need to combine data from different sources, you generally need to find someone who knows what they are doing, to sit down and look at the schemas of the databases and design some kind of mapping and manually do some cleaning or design some scripts. That's the data integration problem, and people have been working on that for 30 years and it probably isn't going away soon. So there you are really talking about the programming effort per megabyte of high quality data is really high, so what is big about the data is the high cost to produce new information or extract value. I think that broadly Big Data seems to be about addressing challenges of scale, either in terms of how fast things are coming at you or how much it costs to get value out of what you already have.

I guess I would say Big Data is not about the raw amount. Big Data is more about the rate of change; the amount vs. the resources that you need to deal with it. If the amount of money that you need to spend on data cleaning is doubling every year, even if you are only dealing with a couple of megabytes that's still a big problem.

> “We know how to support this if everyone involved would use a single, monolithic system. But nobody is really going to work that way. So how can you do this in a larger, more realistic, heterogeneous setting, with different applications communicating with each other about provenance without radically redesigning all existing applications?”

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?

James Cheney: As my other answers have suggested, I’m really trying to call attention to what some people call the long tail of science. So you have these Big Projects that can afford to have 10 or 20 curators working on a big database. There are undoubtedly important research and technological problems there. But there are many more projects that an individual or small team work on, then suddenly their situation changes and don’t have time to work on it anymore. Or the database was supported by a three year grant and even though it’s a useful resource, at the end of the three years it is difficult to get funding for maintaining an existing resource as opposed to building a new resource. Once the database reaches a critical mass, it may be so important that NIH or MRC or other funding agencies will provide longer-term funding (though even this is rare), but there is a big gap between the initial development stage and the critical mass stage, and as a result valuable data is lost and effort wasted. I guess this comes back to the issue of designing tools that makes it easier to get something running out of the box that provides high quality curation resources as opposed to everybody having to do it themselves. Such tools should be designed with long-term sustainability in mind. This is hard.

There are lots of issues that have to do with doing something sensible with large amounts of incoming data. There are also different, important issues concerning what to do with the large number of individually curated datasets that people from various scientific disciplines are putting out there.

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

James Cheney: I would say that the things we have been working on have more to do with developing prototypes for evaluating different techniques for managing provenance or managing history of data. We can convince ourselves that they are sensible and scalable and we can write a good paper about algorithms or prototypes. That’s only the first step: there is still a lot of work needed in order to push these ideas into tools that the rest of the world is going to be willing to use, whether we define the rest of the world as bioinformatics developers or scientific database developers, or we define the rest of the world as a much larger set of people that are willing to contribute to (for example) crowdsourcing or citizen science projects. I think that one big challenge is going beyond the research prototype level, to but building better techniques for supporting data curation techniques into programming languages, or web developing frameworks, or databases, or all the other things that people actually use that don’t talk to each other. Developing these things to the point where we can test them in real situations and find out what techniques work “in the field” and pushing those into things that they are not only solved at a technical level but on the usability and architectural level. And find out what problems that we haven’t solved or even thought about yet because we haven’t been able to get the people that would use or benefit from this technology to the table. It is not that they don’t want to come to the table, it is just that everyone involved is busy and it is hard to justify slowing down and spending time on things that take you further from your main line of work. That’s equally true for people working in computer science as in other disciplines. I guess there are technical challenges and organisational challenges, resource funding and incentive challenges.

I hope that we can do better than “coping” with the problem, through some combination of efforts at the technical and research level and the organisational and usability level.

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

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