DATA CURATION INSIGHTS

KEVIN ASHLEY

DIRECTOR

BIG

Big Data Public Private Forum

Introduction

As Director, Kevin Ashley maps new territory as the Digital Curation Centre embarks on its third phase of evolution (2010–2013), where the accent is on enabling capacity and capability amongst the research community in matters of digital curation.

Previously, as Head of Digital Archives at the University of London Computer Centre (1997–2010), he was responsible for a multi-disciplinary group that provided services related to the preservation and reusability of digital resources on behalf of other organisations, as well as programmes of research, development and training. His group operated NDAD (the National Digital Archive of Datasets) for The National Archives of the UK for over twelve years, capturing, preserving, describing and releasing government data as part of TNA's collections. As a past or present member of numerous advisory and steering groups, including JISC's Infrastructure and Resources Committee, JIIE, the Repositories and Preservation Advisory Group, the Advisory Council for ERPANET and the Archives Hub Steering Committee, Kevin has contributed widely to the research information community.

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

Kevin Ashley: I have been in this work for three and a half years, running the Digital Curation Center. It is meant to be a national centre of expertise in the specific areas of research data management. That centre has been running around for nearly 10 years. Before I was here, I began my working career in a big medical research centre. I spent all my life, in one way or another, in service roles around research in the use of data and technology. In the medical research centre, which had about 5,000 people working there, I was involved with computing services. I initially got interested in issues around data reuse, because I was usually helping people rescuing data that had been created in a previous project and, as technologies move, or people move, people were struggling to be able to read and make use of it. I guess this all made me care about the issues on how you make sure that data is available to people in the future.

About 15 years ago, I was working in the University of London, running large scale research computing and data storage services. We set up a team to offer our skills on a commercial basis. For a number of years we had contracts with the National Archives, the British Library and others to take care of things like government databases, in some cases to make them available for reuse or in other cases protect them, as some of the data we had been dealing with were highly sensitive.

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

Kevin Ashley: The biggest changes are coming through an increasing realization that the real advances in research are made through multidisciplinary things, through people perhaps taking data that is being created in one field and reusing it in a way that the originators didn't really anticipate or expect. Because of that research, funders are increasingly placing very stringent requirements on public funded research. The material that comes out with the publications, with the data in particular, is discovered by people. That they just don't sit on their desk drawer of the person doing the research project, but that some public record of the data exists, that some of the data is available for immediate access, and that this happens in such a way where you don't need to go back to the originators to be able to interpret it. There are some fields of research that probably works that way but they are few and far between. And in general, research has often been conducted by people who protect that information because they see it as a part of their own intellectual endeavour. That is a shift in the way of thinking and it requires a shift in the way universities behave in the types of services that they need to provide to their researchers, and it is a change as well for a number of the big international data centres that work at a particular data field, that operated almost as private clubs. If you were in a particular field you had access to these resources, otherwise you are locked out, and increasingly they have to behave in a much more public way.

> "The biggest changes are coming through an increasing realization that the real advances in research are made through people taking data that is being created in one field and reusing it in a way that the originators didn't really anticipate or expect."

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?

Kevin Ashley: The job has not changed much in the past 12 months. The things that demand our attention have been defined by the shift towards openness and the shift towards bringing data together from very different fields of research. Being able to create in many cases big data out of small data. There are many different types of research where individual projects will produce perhaps a spreadsheet of data. But if you got thousands of people doing similar things around the world and you can bring all of that together, you got something very powerful, that can tell you things that none of these individual data objects can do.

People want to understand how their data can play a part on that bigger international picture and there is a lot more into coordinating the international effort to make that happen. There is a data

> “Data curation means taking care of data available for reuse in the future. It can mean adding value to it beyond some raw data collection. It may mean improving in quality in some way, maybe adding information to them so that we can easily to integrate it with other data sources. It may mean documenting it in a way that means you don’t have to be the creator of the data to make sense of it, and enabling it to survive long beyond its originators intent.”

organization, created two months ago, called international data alliance, that is backed by funders in USA, Australia and Europe to try to make global efforts in that area and that has been a big change. One important realisation is that if you get this right you can create things that are greater than the sum of their parts. You’ve got research potential in data collections that are not in one big dataset. It is one big dataset with distributed parts all over the world. For example, a dataset which maps types of fresh water fishes that could be scattered through universities all over the place. But you can pull these individual things together and you combine with something else, like soil ecology. Being able to do that easily is such a way that does not take a huge effort to integrate them. The big changes in the next three to five years are technologies to make the discovery and integration of these resources.

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

Kevin Ashley: Our works includes helping the University of Edinburgh with its own research data. We just made a proposal to commission data storage to deal with internal research data needs. Something on the range of 5 PBs. That’s by no means all the data that the university is responsible for. There is a lot more that is held on specialist infrastructures embedded on a specific department, attached to a particular lab or instrument. So that 5 PBs represents the sort of working storage for more general research uses. The dataset themselves range from a single spreadsheet, 10s KBs of data, right through to 100s of TBs.

Several thousand university researchers are generating the data we are using. The team at the centre, which is providing generic services around research data, comprises around a dozen people.

Edward Curry: What does data curation mean in your context?

Kevin Ashley: Generally, it means taking care of data available for reuse in the future, which can be done by different means. One way is to add value to it beyond some raw data collection by improving the data quality through meta information. This allows an easier integration with other data sources. Another way is to add documentation so that you don’t have to be the creator of the data to make sense of it. Both measures enable the data to survive long beyond its originators intent.

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

Kevin Ashley: One level is backing out assertions that are made by researchers. Somebody published a research paper that they claimed they discovered something: that is usually backed up data. Data gives you certain ways of challenging the assertions others make about their science, thereby limiting scientific fraud. Most cases for fraudulent research try to protect it by hiding the data.

Curated data provides a base for the researcher to build upon. Frequently you create data for one purpose but you can do other things with it. It provides value to the university as a whole by making data publicly available, by giving it some sort of permanent identifier, in the same way that publications have. It allows people to reuse the data, to cite it by itself, to give credit to someone else’s data, and that itself brings credit to the university and to the people who did the work. And to some universities this credit is directly tied to money as the amount you get depends on how your research is valued by others. Making data available also allows others to exploit it for reasons closely aligned to the original research or for reasons that can be very far removed from it.

There have been various attempts to assign quantitative values to that. There is one recent case for the British Atmospheric Data Centre that uses a methodology to assign a financial value to the reuse that happens to the data. And that shows that depending on how you assign those values, the return of investment on the cost of operating the centre is something between a factor of four to a factor of twelve. It is a strong justification for the society as a whole. There has been a bunch of studies using a similar methodology, for instance, in Australia it has been applied to archaeology data and to other areas as well.

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

Kevin Ashley: These usually varies across different domains of research, from a completely absent process, to others with highly automated workflow processes attached to a scientific instrument, that have been refined over time to ensure that you get the quality that you intend to get. Large scale collections like remote sensing and satellite data collections are becoming increasingly a problem, really driving people on big data areas of research. Sets of technologies of all sorts are becoming incredibly cheap, so that it becomes feasible to use devices which cost a few cents or a few euros each and deploy thousands of them to collect the data over a wide area, about anything from traffic movement to the earth magnetic field, seismic information, etc. But you also get other areas where data comes from individual human beings spending a lot of time looking at something and recording it. It is very difficult to characterize these things.

There is more automation in data curation and more understanding of generic techniques for doing a certain task applicable to all sorts of data curation. A certain type of data quality for instance. There has been a certain understanding on how you can create tools that can do certain checks on the data without having to be very specific to a research domain. In the commercial world this process is called data cleaning and is applied, among others, to address lists and contact lists, where you are picking out more obvious, transcription areas where people are copying things into forms or typing them incorrectly. We might have automated tools to do that but human beings are still very much involved in deciding what rules they are going to apply to those and also on interpreting what happens when your automated systems throw out anomalies, because very often we don’t have systems that we can completely rely on, and the only thing we can do is to flag out some things that a human being has to take a second look where in other cases the volume of data is such that humans can’t be involved until a much later stage. The Large Hadron Collider is a classic example, where there are two or three stages of data reduction, where data is being thrown away, in a completely automated way before human beings get involved with it. But certainly lots of the work is still people intensive and skills intensive and I think one of the concerns is that if we focus too much on technological issues we ignore the fact that without people, if we don’t train enough people with the skills to explore this technology, we’ll descent into expensive systems that are not doing us any good.

> "Another way which defines Big Data is when you do not define a specific data collection protocol in advance but when you simply use what is there which may not be what you would have collected in an ideal world, but you may be able to derive some useful knowledge from it."

Trying to embed these things into the research process, thinking from the outset when you are doing the research of imagining that your data is going to be used by someone else and therefore trying to do things correctly from the start. It requires people shift how they've done things.

If you think about the LHC for example, they deal with data at a massive scale and it is only a tiny fraction of the data that is coming from the detectors and that generates data at a such rate that it would be impossible from any technology to deal with it, so they have to embed a lot of intelligence at an early stage of the system to try to filter out what they define as interesting data from knowledge.

What is interesting is that what is noise to one person is signal to the other. There is an interesting example that I saw few years ago from a group of researchers using a weather radar where what they were interested in was measuring variables related to the size of rain drops, things related to the formation of snow particles in clouds. But it also measures a lot of stuff that it was not interesting to them and they spent a lot of time filtering out that raw data. And at the occasion of the volcano explosion in Iceland a few years back, when nobody knew a lot about what was being thrown in the atmosphere, they realised that the weather radar was probably measuring the dust coming out of the volcano, but the data was thrown away because dust doesn't look like rain drops and therefore that part of the signal was throw away. And when they look at the raw data available they realised that they had very accurate measures of where this volcanic dust was.

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

Kevin Ashley: It is built on this idea of the 3 Vs: velocity, volume and variety, that data can be big simply because there is a lot of it or that it is big simply because of the speed in which it is generated and the speed in which it needs to be processed in order to use it effectively. Also you can have big challenges not only because you have PBs of data but because data is incredibly varied and therefore consumes a lot of resources to make sense of it. Another way that defines this idea is when you do not define a specific [data collection or experimental] protocol in advance but when you simply use what is there which may not be what you would have collected in an ideal world, but you may be able to derive some useful knowledge from it.

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?

Kevin Ashley: I think it comes back again to the shift from the scenario where you carefully design the collection of data in order to answer a single question, and that in the past was what highly curated data meant to people, to an area where you think this is the data that we have and we didn't define in advance the type of questions that we are trying to answer with it. We are going to curate it in a way that makes it usable ideally for any question that somebody might try to ask it, wherever that comes from, and I think that defines a lot what characterizes the Big Data world today, where you take what's there and you see what you can do with it. And I think we have got a lot of more people with those types of skills, the data science type of people, who can think and work in that way.

Lots of the interesting ways to make use data in the research context is by bringing in together lots and lots of little bits of data that collectively make something big. And because they were not produced by the same processes, and you don't have the time and effort to go back and force them all to be entirely the same as each other, you've got different technological challenges of dealing with data which are from variable quality, which might have different types of observations and have being produced in a different context. For example in the biomedical and life sciences, where even the names for a chemical element in the brain might diverge: the way a cell biologist talk about these things is different from the way of a pharmacologist, but for a particular piece of research you may need to integrate data from both fields and from something else as well. We need to be able to do that sort of thing in a way that is machine supported, such that we can get a rigorous idea if they are all measurements of the same process and therefore you can use them together.

Provenance is one area, also crossing terminologies or descriptions from different domains. Also issues related to intellectual property and attribution. It is one thing to reuse a dataset which was collected and another thing to get credit. When you are putting together data from thousands of different sources there are many people that question, whether or not it will make sense to give attributions or credit to all of those sources. Also from the licensing perspective, if this is the thing I'm going to do, what are the data sources that I can use?

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

Kevin Ashley: Technologies related to ontologies and in particular being able to make machine actionable assertions out of things which are actually expressed in natural language is something that I cross again and again, often because we are dealing with things at the moment where lots of descriptions of information around data was designed to humans to read. If you are integrating thousands of sources you don't have time to read thousands of descriptions, so you need the machine to read the description for you and to tell you something about what is there.

There are areas in which crowdsourcing have been incredibly successful and where we have seen a task where you bring originally raw data collections that at the moment aren't amenable to machine analysis because our techniques aren't advanced enough. The intriguing thing about crowdsourcing is that there is a lot we need to learn there about what makes people work and contribute. And there are lots of examples of projects that failed to generate interest and people didn't buy it. That power of crowdsourcing, getting the intellectual power of lots of human beings together and in the end that you can be pretty confident about the quality that you, where you smoothed all the problems, the variation in quality, the fact that some people may deliberately try to sabotage it just because they think that's funny … There are technologies there, but I think there is something to learn about the [social] dimensions of the problem.

Also technologies related to mobile technologies to generate and collect data at scale that previously it wasn't feasible.

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

SEVENTH FRAMEWORK PROGRAMME

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