PAUL GROTH

ASSOCIATE PROFESSOR

VU VRIJE UNIVERSITEIT AMSTERDAM

BIG
Big Data Public Private Forum

Introduction

Paul Groth is an Assistant Professor (i.e. Universiteitsdocent) at the VU University Amsterdam in the Web & Media Group and also a member of the Network Institute. He researches approaches for dealing with large amounts of diverse contextualized knowledge with a particular focus on the web and e-Science applications. Specifically, his research covers data provenance, Web Science, knowledge integration and knowledge sharing.

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

Paul Groth: Currently I'm an assistant professor at VU Amsterdam at the Web and Media Group. I did my PhD at the University of Southampton in distributed systems with the focus on data provenance (i.e. where the data come from, how is it processed, etc.) in large scale grids. Then, I moved to do a postdoc at the Information Sciences Institute at the University of Southern California. There we were looking at workflow systems, again for large scale data processing. We were trying to use intelligence to help people design these workflows using background knowledge. Afterwards, I moved to Amsterdam and I moved towards large-scale data management with a focus on provenance but for doing Web Science type analysis (analysis of data on the Web). I'm currently working on two major projects, one is called Semantically Mapping Science and this is where we are looking for data sources on the Web, things like blog posts, social networks, paper networks, aggregating all that data together so we can understand the processes on how scientific innovation happens. And another project we have is called OpenPhacts, and here again we are looking at how do we take all this data up there, which is public pharmaceutical data, integrate it and provide a coherent integrated dataset to people, but still knowing its provenance. This is really the theme around large-scale data. When you integrate all this data you get something you can work with, but then you ask: where does it come from? This work on provenance led to my co-chairing of a standardization committee at W3C for interchanging provenance data.

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

Paul Groth: It is interesting to answer that question because I can try to answer it from my computer science researcher hat and then I can also answer from my project with the domain scientists that I work with. I will start with the domain experts. I work with organisational scientists who are trying to study the scientific system, and data is really transforming the way they do their science. Before we had this theory on how for example scientists become independent from their PhD supervisors. To do that, you usually look at small scale databases or you call somebody and do an interview. But now we can look at social networks, we can look on Twitter, we can analyse all the papers at once, we can analyse all call for papers at once. This ability to observe what people are doing is really changing the way they do science, it is interesting to see them trying to grapple with the technology. That's why I'm involved with them, because the technology is still not there, and also to understand how does that change methodologically what they do. So they are going from a theory-driven kind of science to a more observational science, now that they have the data. This is from the social sciences domain experts.

> "people are going to start to see that we need to do data curation, we need to adopt systematic methodologies"

I'm also working with biomedical people, in particular in the domain of pharmacology and this is a project called OpenPhacts. What we see there is that they don't know what to do with all the data they have. They have so much data, and the quality of the data has varying degrees. They are having trouble doing discovery, and they can't manage the data that they have, so that they need to use easy-to-use kind of workflows to do that kind of analysis. There what you see is that they think they have this kind of potential for all the data that they collected, but getting insight out of the data is still really difficult.

As a researcher I think Big Data is making us to think a little bit across our domains. Instead of being a database person or an AI person, in the computer science area, we are starting to see people saying: "to deal with this data I need to start thinking about everything from how fast my database is working, to can I distribute this?, can I use the tools out there?, can I develop algorithms that take advantage of large-scale data?". It has been a challenge doing that, but it has being exciting for Computer Science. For example, I do not consider myself a database person, I consider myself more of a data integration person, AI or Knowledge Representation person, but lately we have been looking at databases around NoSQL, just because we need to deal with a lot of data to do what we want to do on the data integration side.

> "…we are looking at how do you deal with data from multiple messy sources and make it available in a cohesive fashion while preserving the information about where it came from."

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?

Paul Groth: My central job hasn't changed in the past 12 months: it is doing research around this area. What I have noticed is that there is a lot of more excitement and there are lots of people now interested in Data Science. Here in Amsterdam we are doing a Data Science research center. I've been to many events where people are talking about Big Data, Data Science. The conversation is exploding around this area. But in respect to my research, there were not big changes since I was already more or less in the space.

In respect to the next five years what I'm hoping is that the systems perspective keeps coming, so we will be able to talk more about the systems perspective and start publishing a lot more on it instead of have to chop up everything, in terms of publishing papers for each individual thing. I also think that for computer science research at least the industrial world is really pushing us to be even bolder that we otherwise might have done in the past. Because they move so fast and they do not need to explain what they do. As researchers we often need to explain why you do a system like this, here are the algorithms, why is this important. While in industry they just iterate. So in five years out as a research community, we need to figure out ways we can keep ahead of the industry game so that we can even more rapidly iterate than they can, while still trying to be principled in our approach.

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

Paul Groth: Talking in the context of the OpenPhacts project. For us data curation really means data provenance. Data curation is the manipulation and integration in a cohesive fashion. That is really what we are trying to do. When you take data from a single or from multiple sources and make it available in a coherent fashion to somebody else. In our work in OpenPhacts, for example, we are looking at how do you deal with data from multiple messy sources and make it available in a cohesive fashion while preserving the information about where it came from. I think that is a central challenge to the aspect of merging all this data together.

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

Paul Groth: We integrate Chemistry databases, for example the ChemBL, Chemspider. We also integrate protein databases (for example Uniprot), drug databases (for example DrugBank), pathways databases (in particular WikiPathways), and the interesting thing is that each of these datasets are in fact already curated datasets. Uniprot for example is a curated dataset for protein and gene information and they actually employ 50 fulltime curators that are reading the literature and putting that information into a database. ChemBL is a chemistry database that is built from the literature. So these are massive databases built from people reading from the literature. This is really interesting. So we get these databases that are built from the literature and then they become massive. For example Uniprot is something like 6 billion triples. Our databases for example, after we do some selections, are around about 2 billion triples. We are already scoped to go to roughly 10-30 billion triples for the OpenPhacts project. We have these curated databases that we are integrating. We have a couple of users who actually help us to decide how we can actually create links among these datasets. We have 3-4 users helping us to figure out how we can connect the Uniprot database with the ChemBL database and what we've realized is that, even integrating these databases, the linksets between those databases have to be domain specific. Different users have different perspectives on what a quality link is, so we have to expose on how we are doing that linkage across these datasets, which are already curated from the biomedical literature.

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

Paul Groth: A central issue within OpenPhacts and essentially within Drug discovery is how we navigate between potential targets within humans and potential compounds that can do something to those targets and how we can subdivide that space. These two, compounds and targets, live in different worlds: one is essentially Biology and the other is essentially Chemistry, so by integrating the data you are essentially able to bridge between these two worlds and what we have done is to provide an API which exposes that data, so that you can build tools around that to allow users to navigate that space in different ways, depending on their use cases. We have a bunch of different tools from chemistry browsers, to workflow systems, to user display tools that let you interrogate this common space and let you bridge between those things.

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

Paul Groth: We make sure to use standards for describing our datasets. We use the vocabulary of interlinked datasets (*VoID*). We use the *Prov* standard to describe how those datasets were created. Then when we integrate those datasets syntactically using RDF. Then we use the VoID linkset specification to tell us how are we linking across these datasets. And those linksets also have descriptions of the provenance on how they were created. Everywhere in the pipeline, we are able to track where all of our datasets and linksets come from, and essentially the provenance of everything. In Big Data, in particular, we forget that data curation isn't just users or individuals deciding on things. We often do integration with our tools and this is also a curation in itself: we are making decisions about how to integrate data. That's a crucial insight. Even technologists who are not aware that they are doing data curation, because they are not domain specialists, are in fact doing some curation decisions. Also, in terms of specific technologies, we use Wikis and we use other community curation aspects, where people can go in and, if they see particular things, they can comment on that or change it, and then we track the revision history as well. We do that in particular for synonyms.

The data integration pipeline is pretty much of a mess. There is not a lot of tooling available and not a lot of mechanisms focused on ensuring that when you take data from multiple places, that is all integrated, that you can track back to the original data. I think that a lot of Big Data is a lot of small data put together. What happens is: we push all of the small data into one thing and then we forget where it comes from. And there is no infrastructure out there that is really good in doing that. I have been impressed with the Wiki-style curation, what that supports. That is a very good tooling for data curation and you are seeing things like that, like Wikidata.

> “We often do integration with our tools and this is also curation in itself: we are making decisions about how to integrate data.”

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

Paul Groth: Big Data is about velocity, volume and variety. If you have a lot of data coming in really fast, that is velocity. If you have lots of data, then that’s volume. And I would define lots of data as forcing you outside what you can essentially store in RAM. If you have to go outside of RAM or you have to go to another box - you are probably doing Big Data.

Variety is a really overlooked aspect of Big Data related to data curation. Most of Big Data is not a uniform big block. Specially, when you are looking at the Web or at these biomedical databases that I was talking before. This stuff is messy, there is lot of it, and there are lots of small pieces of mess that you take and shove together. Another example outside the biomedical domain is, if you look something like common crawl, like a crawl of the Web, the data there is necessarily Big Data, but each data there is very small and very messy, and a lot of what we are doing there is dealing with that variety.

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?

Paul Groth: Big Data is going to do data curation even more necessary. We had this idea that we can have all of this data: let’s get this twitter dataset, let’s download that tweetstream. What we are going to see is that to get really good insights out of this massive data, we need to spend a lot of time preparing the data. And that is what, in most of the time, data scientists do. They spend horrendous amount of time in data preparation. And data preparation is data curation because you are making choices. Data curators always think about the choices that they are making. But people dealing with Big Data specially in the Data Sciences sense, they ignore these choices. They make these choices so that they can get on with their analysis. And I think what is going to happen is that people are going to start to see that we need to do data curation, we need to adopt systematic methodologies, we need to know what we are doing. Essentially we are going to move towards where data curation is going to be thought as really necessary. So the impact of Big Data on data curation is essentially driving up the demand for data curation.

This idea that we can pay people for doing our data curation is essentially not going to work at the scale that we have. We can’t curate the Web. Yahoo didn’t work, right? But there is the need to automate some of the procedures for data curation and the need to understand what that automation is doing. How can we deal with these data curation steps without having the human on the loop but still being aware of what that curation is doing. Data preparation is this kind of art now in Big Data and that is where we need some help. Automation of data curation pipelines is central.

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

Paul Groth: Algorithms for understanding data shapes. So I’ve got this graph data and I need it to become a table and vice versa, how do they fit together. Essentially coping with that heterogeneity. I am seeing a lot of people dealing with streams. I was talking to people from Astronomy with masses of streams. There are lots of people focused on how to do good algorithms to do data analysis, but that’s all assuming that you’ve got the data in the first place. I think we are running out of the easy data. What happens when you have in a Pharma company 500.000 spreadsheets? You can probably get insights if you put all these spreadsheets into a single place and they are all cohesive, but how do you do that? Automating, getting all data and integrating that data.

Research on data integration is one that it is missing, specially, domain-specific data integration. A lot of what we have done in classical data integration ignored domain knowledge. Many times, when we do data integration, we are making domain decisions. Management environments for data, how do we make pipelines to do analysis from the data preparation stage all the end to the data analysis and being able to repeat those things. I think pipelines and reproducibility are aspects that we need to focus on Big Data.

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

SEVENTH FRAMEWORK PROGRAMME

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

http://big-project.eu

TECHNICAL WORKING GROUPS AND INDUSTRY