Profile of Michel Dumontier,
Distinguished Professor of Data Science at
Universiteit Maastricht in the Netherlands and
32nd Annual NASIG Conference Vision Speaker

Dr. Michel Dumontier is one of three Vision Speakers at
the upcoming NASIG annual conference in June 2017.
He is the Distinguished Professor of Data Science at
Universiteit Maastricht in The Netherlands. His current
professional activities include serving on the editorial
board for Semantic Web Journal, co-chair for the World
Wide Web Consortium Semantic Web in Health Care
and Life Sciences Interest Group (W3C HCLSIG), and the
Scientific Director for Bio2RDF, an open source
biological database. You can also follow Dr. Dumontier
on Twitter as @micheldumontier.

My interview with Dr. Dumontier was completed by e-
mail on February 15, 2017.

Your background is in biochemistry and bioinformatics.
How did libraries play a role during your studies?

Libraries have long played a crucial role in my learning
and in my research. However, the nature of the
interaction has changed from when I started my
undergraduate studies to how we use the library now:
from a more physical interaction to one that is purely
virtual.

Indeed, I remember spending vast amounts of time in
the library examining different materials to glean insight
into a particular subject matter – from textbooks on
protein structure and function to peer-reviewed journal
articles detailing computational analyses of their
composition and sequence. As library collections
became searchable using the web browser, and
textbooks and journals became available in a digital
form that could be downloaded, searched, and
annotated, the need to be physically present in the
library has drastically reduced. Until fairly recently, the
role of the library in my work was largely as a
gatekeeper to the digital collections of scientific
research articles that were only realistically available
through institutional subscriptions. With the advent of
open access movement, where authors pay publishers
to make their articles available free of charge, even the gatekeeper role has been diminishing somewhat.

Nonetheless, new challenges concerning research data – description, archival, citation, and availability - are being now been taken on by the libraries. Libraries are providing key infrastructure to help researchers prepare their digital research products in a manner that meets the demands of funders, journals, and other researchers. A key part of that is by making them FAIR – findable, accessible, interoperable, and reusable. I have been fortunate to have had the opportunity to work with a diverse group of individuals, particularly from libraries, through multiple international forums to not only establish the FAIR principles, but also pursue practical implementations and their evaluation. While libraries are playing a key role in the formulation of standardized metadata, I think the future will be in the preparation and execution of resource sharing plans for digital research products.

**What attracted you to the field of data science?**

I was fortunate enough to have had a computer since I was a young child. While I used the computer mostly for playing video games, I think this helped develop my curiosity driven approach to tackling unseemingly large scientific problems. It wasn’t until graduate school did I start to program and really understand the importance of i) having access to data and ii) that it be in a form that is readily available for other uses. A large part of my work has been on how to publish biomedical knowledge such that it becomes easier to ask and answer sophisticated questions that span the molecular to the clinical. Key to this vision has been the adoption of semantic web technologies to build a large interconnected knowledge graph. More recent work has focused on using data science methods over this knowledge graph to find evidence of biological phenomena or to characterize the response of patient populations to complex therapeutic interventions.

**In your view, what are some of the possibilities of the semantic web?**

The semantic web offers the tantalizing possibility that people will be able to do more significant things with substantially less effort. At the heart of the semantic web effort is the continuous development of standards and conventions that promote the interoperability of data and services. The semantic web builds on the principles of the Web itself – a decentralized architecture that establishes the minimal behavior for maximal reuse. Just as we can use a single application to browse content the World Wide Web using HTML formatted pages and the HTTP protocol, the semantic web add new conventions to represent, link, access, query, and explore structured knowledge. The result is a massive graph that interlinks datasets across different providers and domains, removes the site-specific quirks of accessing these data, and creates new opportunities for reuse. This is a significant development because it enables users to spend less time processing data and more time focusing on the problem to be solved.

**How does data science and the Internet of Things intersect?**

Vast amounts of data are now being generated through IoT devices – the big question is how this data can be used to in the acceleration of scientific discovery, the improvement of health care and wellbeing, and in the strengthening of communities – the three pillars for the inter-faculty Institute of Data Science that we are building at Maastricht University. The massive interconnection of these devices poses substantial challenges for data scientists – primarily in how to build scalable and reliable models where data may be incomplete, inconsistent, and unlabeled. Beyond these and other technical challenges, we must also address critical social ones: consent, security and privacy. The IoT community must also grapple with issues relating to interoperability in data formats and web services – precisely the sorts of problems that the semantic web community has been dealing with over the past 20 years.
How has data mining aided your research?

We use data mining, or knowledge discovery, to uncover patterns in large and complex data. In recent work, we explored the use of graph mining methods to assess the nature and quality of linked data produced through Bio2RDF, our open source project to generate linked data for the life sciences. We are also exploring the use of data mining methods on patient electronic health record data to identify patient sub-populations that differ in their response to therapeutic interventions. We use more advanced machine learning methods to build models to predict patient response, and to understand which factors contribute the most to their response. Data mining methods provides us with the means to explore and interrogate associations of interest, but it all depends on the quality of data.

Is there an area of data science that has yet to be explored?

I stand in awe of work by Ross King and colleagues, who back in 2004, published a system called “Adam”, which they termed a robot scientist. Adam was an early form of artificial intelligence that combined logical reasoning with experimental science performed by a robot it controlled. Of late, we have heard incredible feats of complex problem solving such as answering questions on Jeopardy or winning games such as Go. These achievements have been made possible by the availability of large datasets and massive amounts of computing power that are now much more accessible than they have ever been. So, I think that the future of data science lies in a renaissance for artificial intelligence - driven by data-driven systems that ply imperfect big data and expert knowledge together using high performance statistical methods in a vastly more effective manner. This renaissance will also have major implications for human labor – we and others are exploring how people can be relieved of trivially solved problems while tackling the ones that learning methods find challenging to solve.

Do you have any additional comments?

By any measure, we live in a truly exciting time. However, the rate at which science and engineering are changing the world pose an incredible challenge in keeping up with the latest developments. A big part of what I am doing is building links across communities to ensure that we can adopt the latest technologies so as to create a virtuous circle for consumers and producers of digital resources. The more we work together, the easier it will be for all of us to share the benefits they afford.