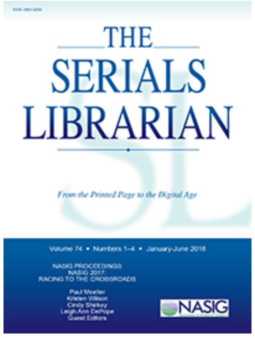
Advancing discovery science with fair data stewardship



Routledge
Taylor & Francis Group

The Serials Librarian

From the Printed Page to the Digital Age

ISSN: 0361-526X (Print) 1541-1095 (Online) Journal homepage: https://www.tandfonline.com/loi/wser20

Advancing Discovery Science with FAIR Data Stewardship: Findable, Accessible, Interoperable, Reusable

Michel Dumontier & Kathryn Wesley

To cite this article: Michel Dumontier & Kathryn Wesley (2018) Advancing Discovery Science with FAIR Data Stewardship: Findable, Accessible, Interoperable, Reusable, The Serials Librarian, 74:1-4, 39-48, DOI: 10.1080/0361526X.2018.1443651

To link to this article: https://doi.org/10.1080/0361526X.2018.1443651

THE SERIALS LIBRARIAN
2018, VOL. 74, NOS. 1–4, 39–48
https://doi.org/10.1080/0361526X.2018.1443651

Routledge
Taylor & Francis Group

VISION SESSIONS

Check for updates

Advancing Discovery Science with FAIR Data Stewardship: Findable, Accessible, Interoperable, Reusable

Michel Dumontier[a] and Kathryn Wesley[b]

[a]Presenter; [b]Recorder

ABSTRACT
This report summarizes a presentation by Dr. Michel Dumontier. It reviews innovative scientific research methods created by data science, and the need to develop infrastructure, methodologies, and user communities to advance data science. Stakeholders have proposed a set of principles to make digital resources findable, accessible, interoperable, and reusable—FAIR. FAIR principles provide guidelines, do not require specific technologies, and allow communities of stakeholders to define specific FAIR standards and develop metrics to quantify them. Libraries can be part of the new data ecosystem by providing education, data stewardship, and infrastructure.

KEYWORDS
FAIR principles; non-reproducibility; data management; data science; genuine semantic publishing; semantic web

Data science is creating innovative new scientific research methods. As libraries become more involved in data management, it will be increasingly important to be informed about and involved in developments in data infrastructure and standards. This presentation provided an overview of a number of projects, challenges facing researchers, and proposed guidelines to help address some of those problems.

Dr. Michel Dumontier, a cutting edge data scientist, began by noting that, although the audience might find it surprising, this audience was precisely the one he was interested in talking to. The reason, he explained, is "because everything is changing underneath us in terms of data and its utility, what people are doing with it. How they are advancing its state of the art in their own domains is largely now through the reuse of other people's data. I think libraries have a key role to play." He noted that he would describe some of the research his group has been doing and some initiatives they have been undertaking that will have an impact on universities and "the libraries at [their] core."

This is a new "golden age of scientific innovation," much of it driven by data science. Dumontier cited cases where companies such as Google, Facebook, and LinkedIn are using data in innovative ways, for example, "to understand social networks, to play games against experts, and to start tackling the hard problems in health—to come up with new ways of diagnosing diseases, new ways of treating diseases."

There is a problem, however, that Dumontier says "keeps me up at night." As scientists begin to perform more meta-analyses, where multiple studies are examined together in a common framework, it is being demonstrated that most experimental results cannot be reproduced. One study found that up to 64% of psychology studies surveyed had non-reproducible results. Pharmacological studies have 65–89% non-reproducible results.

There are a number of reasons for this phenomenon. First, said Dumontier, "Science is hard. One of the problems is that in science we attempt to isolate our systems or questions into interesting but perhaps complicated systems. We can't tease apart all the variables, so we are

CONTACT Michel Dumontier ✉ Institute for Data Science, Maastricht University, Maastrict, Netherlands.
Color versions of one or more of the figures in the article can be found online at www.tandfonline.com/wser.

forced to experiment with some blindsides." Another issue has to do with insufficiency of statistics. "In many cases, our data are incomplete. There might be some element of uncertainty. We don't have good representation in our samples, and some of the methods simply can't account for that," Dumontier explained. Another problem is that "biology is unruly." While there are hundreds of thousands of organisms, scientists study one at a time, so it is difficult to generalize findings across species. Dumontier noted that one of the first papers he published reported on a gene that behaved atypically, compared to how it functioned in other species.[1] Finally, medicine is complicated. Dumontier explained:

> At the heart of the work that we are doing, where we are trying to understand how we can diagnose disease, how we can treat those diseases, each and every individual is responding differently to the therapeutics that we treat them with. And so there is a genetic component. There is an environmental component. The biology of people, just one species, is already very challenging.

The problem is that "the way we have been doing science isn't good enough," giving us results we are not confident about. But data science is now creating the opportunity to do better. We can improve our confidence in experimental outcomes, Dumontier says, by looking at aggregates of reports, or by examining different lines of evidence to support a hypothesis.

An example can be found in the work of Purvesh Khatri of Stanford University, on genes involved in organ transplantation rejection. He examined genes that were up-regulated (i.e., more of their proteins are generated) and down-regulated (less of their proteins are generated) in patients for whom transplantation had been successful and those for whom it had been unsuccessful. By comparing these and examining large amounts of data available in an Open Access database on tissues from multiple organs, rather than just one, they were able to identify a pattern of specific genetic expressions that occur in transplantation rejections.[2] This work demonstrates, Dumontier said, the importance of relying on not just one experiment, but looking for global patterns across data—in this case data on not just one type of tissue, but across multiple types of tissues.

Another example from Dumontier's own work involved studying the role of genes in aging.[3] He explained that there are many hypotheses about what causes aging, and researchers use different lines of evidence to support their theories. Experimental data can support one theory, while not necessarily supporting or contradicting another theory. He looked for all data in public databases considered to be evidence supporting a theory on aging and compiled it in a data dashboard showing specific genes and features associated with lines of evidence supported for those genes. While they retrieved data supporting some lines of evidence for each gene, they also identified many gaps, where they found no results for a line of evidence with a specific gene. Dumontier explained:

> And so it turns out, that for these genes that everybody believes are strongly involved in aging, there are gaps in our knowledge. We wouldn't really know that unless we made an assessment of what is out there. … There's no problem in generating information. The problem is that we're really terrible at taking an account of what we know and what we don't know.

An additional problem identified in this study was that some genes not associated with aging shared some of the characteristics of those that are. This situation casts doubt on the association of those characteristics with aging, said Dumontier. It is an issue of confirmation bias, he explained: "You hypothesize that something is there, and then you find evidence to support it."

These cases illustrate what Dumontier calls the "grand challenge." That is, to "build intelligent systems that will help us automatically find evidence that supports or potentially disputes a scientific assertion using the totality of the information available. … How can we make use of knowledge that's been created, directly at the point that we need it to answer these kinds of questions?" Two things are required, he said. The first is data science. This includes infrastructure to manage data and create metadata for it, systems to access it, and methods to discover and retrieve it. The second need is a community to build the infrastructure and adopt it into its members' work.

One global community effort to build and adopt these methodologies and infrastructures is the Semantic Web. It is, Dumontier said, "an initiative, a set of standards to publish, to share, to query data, expert knowledge, and to interoperate between web services." The Semantic Web is built on existing web technology, which has been successful because it is a distributed, decentralized model, based on a few standards. He explained:

> We have TCP/IP [Transmission Control Protocol/Internet Protocol] to send messages between systems. We have HTTP [Hyper Text Transfer Protocol] protocol to make requests on that content. And we have HTML [Hyper Text Markup Language] standards to format the content so that we can see it. But we want to do the same for data. How do we make a web of data that we can publish to, integrate, and reuse?

Dumontier next showed an image of a "massive decentralized knowledge graph" called the Linking Open Data cloud diagram.[4] It illustrates the enormous breadth of open content available on the web from many domains, including governments, publishers, media, social networks, life sciences, and so on and the ways those data link to each other. This represents, he said, a "huge, rich network of information," including life sciences information about, for example, genes and diseases, from which new facts can be drawn.

One Semantic Web project that Dumontier has been involved with developing is Bio2RDF.[5] "The whole idea," he explained, "is to take data in its glorious heterogeneity and try to harmonize it a little bit." Using open source scripts, Bio2RDF creates linked data connecting thirty-five different biomedical databases. It provides provenance and metrics on the datasets, and make the information available through online platforms. "So we have," he said, "all the kinds of information that I talked about—information that can help us understand aging, information that can help us understand rejection in transplantation, and so on. All the stuff that biologists care about."

The "promise with linked data," Dumontier said, is that "every single thing, every single piece of data can be identified." Those identifiers use HTTP, as Uniform Resource Locators (URLs) do, and can be used to create statements about any data on the web. When placed in a browser, the identifier will often create an HTML rendering, but underneath is a machine-readable format that is standardized across the Semantic Web. Because these identifiers are interoperable, they allow data to be linked in a bidirectional way. Dumontier illustrated this point with a screenshot from Bio2RDF.

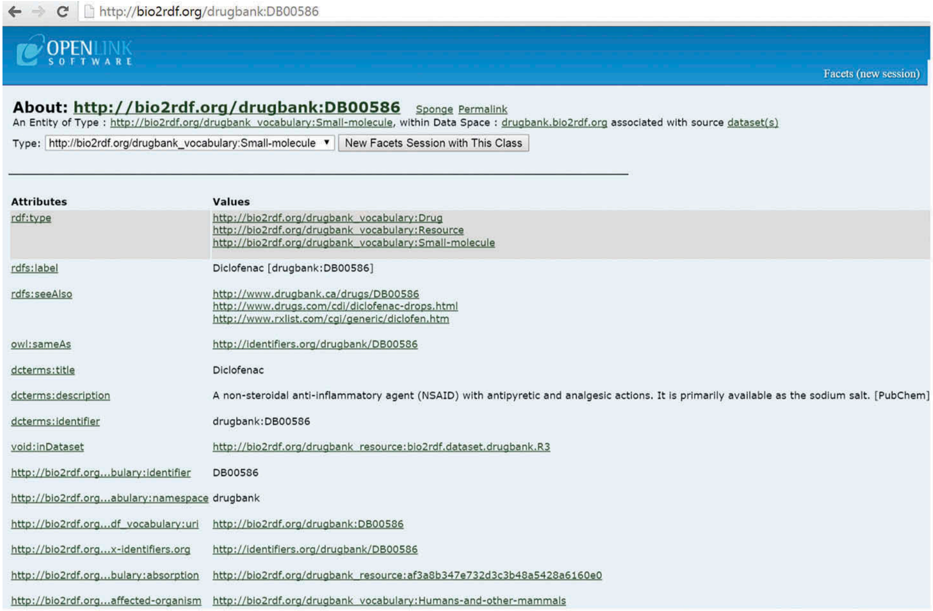
It showed an entry from a database called DrugBank for a drug called Diclofenac.[6] Some links were shown that originated in DrugBank, while others originated elsewhere and linked into the database. This bidirectionality, he said, allows navigation through the content represented in the enormous knowledge graph shown earlier.

Once data are represented in a standardized way, it becomes possible to search them for a specific purpose, Dumontier continued. For example, following the outbreak of the Ebola epidemic in West Africa in 2014, a Ph.D. student working with Dumontier built a portal that drew on information already available in Bio2RDF (Figure 1):

> [He] put together a portal which took the latest sequenced genome and connected it up through the rest of our network to see what the different functions of those genes are. What structures are linked to those genes? What small molecules potentially could be used to target this virus? As well as all the publications that are associated with it, and how they are connected up. And this is something he did in two weeks' time.[7] So, again, there is plenty of information there. You just really have to know how to pull it and build an information system of your choosing that works toward your own goals.

This idea is beginning to be noted by companies, Dumontier said. He cited the example of a Belgian company called ONTOFORCE, which has developed a semantic discovery tool with faceted search capability that searches a large and heterogeneous group of linked data sources.[8] In addition, it displays the provenance of all data retrieved, allowing for further filtering by source.

But what he is really interested in, Dumontier said, is not just information retrieval. He explained, "We want to find things that people don't know about, and so one of the things that we've been working on is how do we take this network of information and transform it into a system by which we can filter out the less informative parts of the knowledge graph and focus on the ones we're really interested in? What we're interested in is further discovery." He described a project in which he and his colleagues looked for possible new drugs to treat melanoma. They


http://bio2rdf.org/drugbank:DB00586
OPENLINK SOFTWARE
Facets (new session)
Sponge
Permalink
New Facets Session with This Class

About: http://bio2rdf.org/drugbank:DB00586

An Entity of Type : http://bio2rdf.org/drugbank_vocabulary:Small-molecule, within Data Space : drugbank.bio2rdf.org associated with source dataset(s)

Type: http://bio2rdf.org/drugbank_vocabulary:Small-molecule ▾

Attributes	Values
rdf:type	http://bio2rdf.org/drugbank_vocabulary:Drug http://bio2rdf.org/drugbank_vocabulary:Resource http://bio2rdf.org/drugbank_vocabulary:Small-molecule
rdfs:label	Diclofenac [drugbank:DB00586]
rdfs:seeAlso	http://www.drugbank.ca/drugs/DB00586 http://www.drugs.com/cdi/diclofenac-drops.html http://www.rxlist.com/cgi/generic/diclofen.htm
owl:sameAs	http://identifiers.org/drugbank/DB00586
dcterms:title	Diclofenac
dcterms:description	A non-steroidal anti-inflammatory agent (NSAID) with antipyretic and analgesic actions. It is primarily available as the sodium salt. [PubChem]
dcterms:identifier	drugbank:DB00586
void:inDataset	http://bio2rdf.org/drugbank_resource:bio2rdf.dataset.drugbank.R3
http://bio2rdf.org...bulary:identifier	DB00586
http://bio2rdf.org...abulary:namespace	drugbank
http://bio2rdf.org...df_vocabulary:uri	http://bio2rdf.org/drugbank:DB00586
http://bio2rdf.org...x-identifiers.org	http://identifiers.org/drugbank/DB00586
http://bio2rdf.org...bulary:absorption	http://bio2rdf.org/drugbank_resource:af3a8b347e732d3c3b48a5428a6160e0
http://bio2rdf.org...affected-organism	http://bio2rdf.org/drugbank_vocabulary:Humans-and-other-mammals

Figure 1. Screenshot of Bio2RDF search. Courtesy of Michel Dumontier.

started by looking at genes that have been linked to melanoma and drugs that target those genes. Then they looked at information on clinical trials testing those drugs for treatment of melanoma. What they found, Dumontier said, was that "some drugs have been approved, some of them are in various phases of clinical trials, but actually a large number of them haven't been examined for this particular indication at all. So it's not like we went out and did all these new experiments. Again, we just said, what do we know? And could any of these drugs then be new candidates for treating skin cancer?"[9]

Standardized open data are also relevant to the idea of experimental reproducibility. Dumontier described the case of a study published in 2011, which reported on a method called PREDICT, that demonstrated a high rate of accurately predicting drugs that could be repurposed.[10] He contacted the authors to ask for access to their data and tools so that his group could replicate the study. The original data were available, but the authors had used scripts from a variety of sources to work with the data. To try to reproduce the results, Dumontier's group created an open system called OpenPREDICT that will use linked data with machine learning methods.

Creating a machine learning system involves building a classifier, which is then given sets of known positive and negative examples. In the case of OpenPREDICT, the positive examples are drugs and the diseases they treat, and the negative examples are drugs and diseases they do not treat. The classifier will examine attributes of the drugs and diseases, attempting to find indications in drug/disease associations that are relevant and those that are not.

The researchers faced an immediate problem. According to Dumontier, "the set of drugs and drug indications that people use in different experimental platforms are completely different. This was extremely surprising." They used three standard data sets for drugs, diseases, and drug–disease associations and found that there was little overlap between them.[11] Only 125 drugs, thirteen diseases, and twenty-seven drug/disease associations were common across all three data sets. Dumontier concluded:

> What this tells you is that if you build a classifier with one of those data sets, you're basically just learning a part of everything that we know … you might report a very nice result in using one data set. But it's very unlikely to be predictive across any other indication or in other data sets. And that's pretty much what we found. So we

> applied machine learning methods over the different data sets, and found completely different performance. This is problematic in our field. … We're telling people that we're solving this problem, but it really just depends on which tool you use, and which data set you use. That's not really a reliable metric for advancing science. Obviously, what we need to do is create these kinds of common benchmarking data sets and for all of us to use them in our own work.

It is also possible to use data in unexpected ways to generate new insights. Dumontier related an example of another study designed to predict drug targets using data from mouse genetics.[12] He explained that drugs generally act by inhibiting a gene. In genetic studies of mice, specific genes are targeted and made nonfunctional, and the phenotypes resulting from the mutations are observed, so these two processes are similar. They reasoned that if they could match drug effects in humans to particular phenotypes in mice genetic studies, they would be able to identify the specific genes that the drugs were targeting in humans.

The researchers first had to solve the problem of reconciling the terminologies used in the different fields. Studies with mice use a vocabulary called Mammalian Phenotype Ontology, while drug studies with humans use one called Unified Medical Language System.[13] These vocabularies are similar, but not interchangeable, so an interface that could map the terms of one to the other was needed. Then, using 7,200 mice genetic characteristics, they looked for targets for almost 15,000 drugs. They were able, Dumontier said, "to recapture most of the known drug targets," using only this data. This is significant, he explained, because drug research typically looks at chemical formulas, known drug targets, and characteristics of diseases. Another important aspect of their approach is that it can help explain side effects of drugs by identifying the genes they target. He gave the example of Diclofenac, a nonsteroidal anti-inflammatory drug used to treat pain and arthritis. Their study showed that 46% of side effects, such as inflammation, gastritis, constipation, and upper gastrointestinal tract pain, could be explained by the targeting of the COX-2 gene, and 49% by the targeting of the PPARg gene. This again demonstrates, he said, the power of approaching a problem by using different kinds of data.

One further example of biomedical research using existing data took place following a rash of cases of heart damage caused by new cancer drugs. Dumontier explained that when drug companies submitted these drugs to the Food and Drug Administration (FDA) for review, they did not include any information relating to cardiotoxicity. The FDA, typically under-resourced, studied the information they were given. When incidents of cardiotoxicity began to occur, FDA officials wondered if there was existing evidence they could have found to predict it. Led by associate director for drug safety Darrell Abernethy, the FDA began to mine existing databases to find out.[14]

Dumontier noted that most of the data resources they found in this study were already included in the knowledge graph. This demonstrated a need for a system that would allow data to be gathered at the time it is needed to investigate a question (e.g., what evidence supports or disputes cardiotoxicity of a drug). He and his colleagues built a system called HyQue:[15]

> It retrieves data from a variety of different sources from our network of linked open data. It can execute different kinds of services through a Semantic Web service framework, making these ontologies to reason and to interoperate this information. And what's interesting is that you can use positive and negative finding information, and also examine the problems for this information.

When the question of cardiotoxicity of the cancer drugs was entered into HyQue, most evidence retrieved was either neutral or did not show any indications for cardiotoxicity. However, said Dumontier, one database did include data about cardiotoxicity having been recorded. But, he added, HyQue is a "fully open and transparent system. If you don't agree with what we did, you can change it. You can bring your own analytic into it, and you can try to pull out the information that is relevant for you."

All the studies described above, Dumontier said, required enormous investments of time and resources by him and his colleagues in order to work with other people's data. He continued, "Why should we have to do that? What we should expect is that people create content that is findable—that we can find it, that we can access it, that it interoperates with existing systems, and that we can readily reuse it."

These principles are reflected in the acronym FAIR—Findable, Accessible, Interoperable, and Reusable. FAIR principles have been developed and endorsed by a broad group of stakeholders—academics, researchers, scholarly publishers, funding agencies, and industry partners. FAIR applies to all digital resources and their metadata. Dumontier said, "It applies to software, to images, to data, to repositories, to web services, to scholarly publications. Whatever is digital and whatever you can put on the web is subject to this expectation."

A paper introducing the FAIR principles was published in 2016. It had fifty-four listed authors, including Dumontier, representing over forty different organizations.[16] As of May 2017, it has been cited over 100 times and garnered high altmetric rankings. In addition, the FAIR principles have been endorsed by the leaders of the G20, and a number of national and international science initiatives, including the European Open Science Cloud, Horizon 2020, the National Institutes of Health, and ELIXIR.[17] Regarding this rapid adoption, Dumontier commented, "In some sense this idea is so simple. It's not like we haven't been doing it before, but what we've got to do is raise the awareness of what we can do with data, and what we can do with these digital resources, and what has to be done in order for us to be able to rely on and use them in forward-thinking science."

FAIR includes 15 principles divided among the headings Findable, Accessible, Interoperable, and Reusable.[18] They deal with standard identifiers, metadata, how to access content, standards used, and license requirements associated with content (Table 1).

The question naturally arises of how to measure the FAIRness of data. According to Dumontier, "FAIRness reflects the extent to which a digital resource addresses the FAIR principles as per the expectations defined by a community of stakeholders." The expectations of different communities may be different. But those expectations must be made explicit, and we must be able to measure resources against those community standards.

There is movement to establish specific metrics to use in examining or quantifying the FAIRness of data. Using as an example the first FAIR principle, which says data and metadata are assigned "globally unique and persistent identifiers," Dumontier explained that digital object identifiers might be considered by some to satisfy the principle. But they may not necessarily satisfy the "persistence" attribute. He continued:

Table 1. The FAIR guiding principles.

To be findable:
F1. (meta)data are assigned a globally unique and persistent identifier
F2. Data are described with rich metadata (defined by R1 below)
F3. Metadata clearly and explicitly include the identifier of the data it describes
F4. (meta)data are registered or indexed in a searchable resource
To be accessible:
A1. (meta)data are retrievable by their identifier using a standardized communication protocol
A1.1. the protocol is open, free, and universally implementable
A1.2. the protocol allows for an authentication and authorization procedure, where necessary
A2. Metadata are accessible, even when the data are no longer available
To be Interoperable:
I1. (meta)data use a formal, accessible, shared, and broadly applicable language for knowledge representation
I2. (meta)data use vocabularies that follow FAIR principles
I3. (meta)data include qualified references to other (meta)data
To be reusable:
R1. Meta(data) are richly described with a plurality of accurate and relevant attributes
R1.1. (meta)data are released with a clear and accessible data usage license
R1.2. (meta)data are associated with detailed provenance
R1.3. (meta)data meet domain-relevant community standards

Note. Dumontier noted that the FAIR principles provide guidance on how digital content should be made available for reuse, but they do not impose any specific technology.

> In my view, persistence is represented by a commitment, a social commitment, to ensure that the resource is available in the longest term possible, and that should something catastrophic occur, you have a plan to transition that resource to another responsible owner. And so my expectation is that there is a data policy that is staying with the resource that outlines or specifies exactly how this will be done, and that all digital resources should be subject to this.

The metric for the persistence attribute, then, would be "yes" or "no" with regard to the existence of a data policy. Dumontier added that the data policy itself should be accessible and verifiable.

Metrics can be established for each of the fifteen principles and collected together into FAIRness indexes, which can be used to judge the FAIRness of any digital resource in a way that is clear and consistent. Different communities of stakeholders will be able to define their own FAIRness indexes. Dumontier showed a mockup using the EASY database from Data Archiving and Networked Services that demonstrated how FAIRness ratings might display online.[19] The mockup used a star-based rating system with a possibility of up to five stars for each of the major principles—Findable, Accessible, Interoperable, and Reusable. For example, if a resource has been assigned good metadata, including a standard identifier, it would be rated higher in the Findable attribute than a resource that has been assigned no metadata beyond a title.

All of this has broad implications for how scientific research can be done. Dumontier explained:

> This is how we've been doing science for a long time: we collect data, we do our analysis, we publish our paper. And then we don't care about the data anymore. And that has to change. The reason it has to change is we can reuse that data in a number of different contexts. We can use it to reproduce other people's work. We can use it to validate a concept. We can use it to generate new hypotheses.

Further, although data are becoming more open and accessible, scholarly publications—the articles themselves—are still inaccessible to data mining techniques because their content is not structured. Dumontier and Thomas Kuhn have been thinking about how this might change. He explained:

> What you have when you write a paper is basically a box. A box that has information, but you have to have knowledge to peer into that box, understand the content, and reuse it. In what has traditionally been called semantic publishing, you start doing some annotations, semantic annotations. What is the article about? What are the main findings? It's some structured content of what is in the paper. What we're pushing is something called genuine semantic publishing, where the content itself is already structured in a way that we can use it as a database, and it can feed right into our algorithms and serve as a knowledge graph. We can publish it, and it connects up with everything we've seen so far.[20]

These ideas are embodied in a new journal launched by Dumontier and Kuhn called *Data Science: Methods, Infrastructure, and Applications*, published by IOS Press.[21] It is Open Access, open review, and promotes the idea of genuine semantic publishing by enforcing structured data standards for all submissions. Dumontier expanded, "What that really means is that there's an emerging community of researchers that are creating new standards for publishing scientific content, for publishing those papers. We're going to try to be a platform for them. … What we ultimately want are executable studies." Citing the Jupyter Notebook, an open source project that supports "interactive data science and scientific computing across all programming languages,"[22] Dumontier noted that it provides scientists with the ability to execute studies exactly as originally conducted. That, he said, is what they want to do for papers—to have the textual content structured in a way that will allow for questioning by other researchers.

Dumontier wrapped up with a reiteration of his main points and a request to the audience. We must, he said, do something about the state of science. The solution will involve data, particularly the reuse of other people's data, discovered through new data-mining techniques. Individual studies need to be grouped together and studied in meta-analytical frameworks as soon as they are published. This will allow greater confidence about reported results. The

Semantic Web gives us a method to publish structured content that can be immediately reused, and is therefore a platform well-suited for discovery science.

Finally, Dumontier asked for the help of librarians present to promote the idea of FAIR digital resources in their academic communities. Libraries, he said, are well-situated for helping to make research data accessible for reuse. New infrastructures, tools, and approaches will be needed to accomplish this, as well as new training for "data stewards, people who understand the problem, who can guide other researchers to the solution." This will serve to build new discovery frameworks, where, ultimately, whatever content researchers create will immediately be available for reuse. Dumontier noted that he moved to Maastricht University to establish a new institute devoted to these kinds of principles. Working through multidisciplinary teams, the Institute of Data Science seeks to accelerate scientific discovery, improve health and well-being, and empower communities through data science.[23] At the heart of it, Dumontier says, is "the generation of content we can reuse," and "systematic mining using AI [artificial intelligence] technology to find those gaps, and fill those gaps."

A lengthy question and answer period followed the presentation. Audience members, clearly engaged by the topic, asked Dumontier questions both concrete and philosophical. The first question concerned transparency of data created by pharmaceutical companies. Dumontier noted that while these companies are very interested in open data, they do have be concerned with economic incentives. While their data might never be fully open, he said, they are exploring ways to collaborate. Answering a question on the role of libraries in the future of data management, Dumontier indicated he is working with a librarian at Maastricht University to develop this idea. Their goal is to create an entire infrastructure—technical, social, legal, ethical—around the researcher with libraries at the epicenter. He invited interested librarians to contact him. When asked how semantic publishing might capture and make accessible the "logical connective tissue" of journal articles, Dumontier responded with a brief overview of research being done in argumentation networks and data-driven AI. Another questioner, referring to the FAIR principle of retaining metadata even if data are no longer available, asked if data should be retained forever and whether that is even possible. Dumontier called this a good but hard question. Ultimately, there needs to be commitment by funding agencies for long-term support of data. He indicated that, while people and institutions are beginning to explore this idea, we are as yet "far away from [a] coherent framework of supporting long-term data archiving and reanalysis." Additional questions followed on the role of publishers in semantic publishing, and on garnering support outside of the research community for data standards. The final questioner asked whether FAIR principles might encourage the open sharing of clinical trial results that are not published because their findings do not support the desired results. Dumontier replied that we need access to negative results as much as to positive, that they are an intrinsic part of the discovery process. In addition, there is research value in combining these kinds of results with other data to answer new questions.

This was a very different presentation for NASIG. But Dumontier noted that we were precisely the audience he wanted to be addressing, so he was precisely the presenter we wanted to hear. The question-and-answer part of the program was more than half again as long as the presentation itself. The event generated an impressive and highly positive tweet stream. Two of the most salient tweets came from incoming NASIG President Steve Oberg:

> Kind of mind-blowing (in best sense) presentation by Dr. Michel Dumontier. Rethinking, [re-envisioning] pubs, data science, etc.

And:

> To me, FAIR principles closely align/overlap with inherent librarian values. We should be onboard.

In addition, Dumontier's openness to partnering with librarians, publishers, and commercial entities certainly resonates with NASIG tradition and values. As the organization and individual members increasingly transition from focusing on traditional models of scholarly publishing to new models, we will need to continue to connect with speakers and topics like this.

Notes

1. Michel Dumontier, Petra Höcht, Ursula Mintert, and Jan Faix, "Rac1 GTPases Control Filopodia Formation, Cell Motility, Endocytosis, Cytokinesis and Development in Dictyostelium," *Journal of Cell Science* 113, pt. 12 (June 2000): 2253–65, https://www.ncbi.nlm.nih.gov/pubmed/10825297 (accessed August 18, 2017).
2. Purvesh Khatri et al., "A Common Rejection Model (CRM) for Acute Rejection across Multiple Organs Identifies Novel Therapeutics for Organ Transplantation," *Journal of Experimental Medicine* 210, no. 11 (2013): 2205–21.doi:10.1084/jem.20122709 (accessed July 26, 2017).
3. Alison Callahan, Juan Jose Cifuentes, and Michel Dumontier, "An Evidence-Based Approach to Identify Aging-Related Genes in Caenorhabditis Elegans," *BMC Bioinformatics* 16, no. 40 (2015). doi:10.1186/s12859-015–0469-4 (accessed August 18, 2017).
4. Andrijs Abele, John P. McCrae, Paul Buitelaar, Anja Jentzsch, and Richard Cyganiak, "Linking Open Data Cloud Diagram," last updated August 22, 2017, http://lod-cloud.net/ (accessed August 24, 2017)
5. Bio2RDF, http://bio2rdf.org/ (accessed July 26, 2017).
6. DrugBank, https://www.drugbank.ca/ (accessed July 26, 2017).
7. Maulik R. Kamdar and Michel Dumontier, "An Ebola-Virus Centered Knowledge Base," *Database*, 2015, bav049 (January 2015), 10.1093/database/bav049 (accessed July 26, 2017). Bio2RDF Ebola Virus Knowledgebase, http://ebola.semanticscience.org (accessed July 26, 2017).
8. ONTOFORCE home page, http://www.ontoforce.com/ (accessed July 26, 2017).
9. James P. McClusker, Michel Dumontier, Rui Yan, Sylvia He, Jonathan S. Dordick, and Deborah L. McGuinness, "Finding Melanoma Drugs through a Probabilistic Knowledge Graph," *PeerJ Computer Science*, 2017, 3: e106, 10.7717/peerj-cs.106 (accessed July 26, 2017).
10. Assaf Gottlieb, Gideon Y. Stein, Eytan Ruppin, and Roded Sharan, "PREDICT: A Method for Inferring Novel Drug Indications with Application to Personalized Medicine," *Molecular Systems Biology*, 7, no. 1: 496 (June 7, 2011), 10.1038/msb.2011.26 (accessed July 26, 2017).
11. SIDER, http://sideeffects.embl.de/ (accessed July 26, 2017); DrugCentral, http://drugcentral.org/; and PREDICT (accessed July 26, 2017).
12. Robert Hoehndorf, Tanya Hiebert, Nigel W. Hardy, Paul N. Schofield, Georgios V. Gkoutos, and Michel Dumontier, "Mouse Model Phenotypes Provide Information about Human Drug Targets," *Bioinformatics*, 30, no. 5 (March 2014): 719–725, 10.1093/bioinformatics/btt613 (accessed July 26, 2017).
13. Mouse Genome Informatics. Mammalian Phenotype Browser, http://www.informatics.jax.org/vocab/mp_ontology (accessed July 26, 2017); National Library of Medicine. Univied Medical Language System (UMLS), https://www.nlm.nih.gov/research/umls (accessed July 16, 2017).
14. Jane P. F. Bai and Darrell R. Abernethy, "Systems Pharmacology to Predict Drug Toxicity: Integration Across Levels of Biological Organization," *Annual Review of Pharmacology and Toxicology* 53 (2013): 451–73. 10.1146/annurev-pharmtox-011112–140248 (accessed July 26, 2017).
15. Alison Callahan, Michel Dumontier, and Nigam H. Shah, "HyQue: Evaluating Hypotheses Using Semantic Web Technologies," *Journal of Biomedical Semantics*, 2, suppl. 2 (2011): 53, 10.1186/2041-1480-2-S2-S3 (accessed July 26, 2017).
16. Mark D. Wilkinson et al., "The FAIR Guiding Principles for Scientific Data Management and Stewardship," *Scientific Data* 3 (2016): 160019, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4792175/ (accessed August 24, 2017).
17. European Commission, "G20 Leaders' Communique Hangzhou Summit, 5th September 2016," http://europa.eu/rapid/press-release_STATEMENT-16-2967_en.htm (accessed July 26, 2017); European Commission, "Realising the European Open Science Cloud," 2016, https://ec.europa.eu/research/openscience/pdf/realising_the_european_open_science_cloud_2016.pdf (accessed July 26, 2017); European Commission, Directorate-General for Research & Innovation, "H2020 Programme: Guidelines on FAIR Data Management in Horizon 2020," version 3.0, July 26, 2016, http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf (accessed July 26, 2017); National Institutes of Health, Office of Strategic Coordination—The Common Fund, "Big Data to Knowledge," https://commonfund.nih.gov/bd2k (accessed July 26, 2017); ELIXIR, "FAIR," https://www.elixir-europe.org/services/interoperability/fair (accessed July 26, 2017).
18. Wilkinson et al, "The FAIR Guiding Principles for Scientific Data Management and Stewardship."
19. DANS, EASY, https://easy.dans.knaw.nl/ui/home (accessed July 26, 2017).
20. Tobias Kuhn and Michel Dumontier, "Genuine Semantic Publishing," http://www.tkuhn.org/pub/sempub/ (accessed July 26, 2017).

21. “Data Science: Methods, Infrastructure, and Applications,” https://datasciencehub.net/ (accessed July 26, 2017).
22. Jupyter, “About Project Jupyter,” https://jupyter.org/about.html
23. Maastricht University, Institute of Data Science,https://www.maastrichtuniversity.nl/research/institute-data-science (accessed July 26, 2017).

Notes on contributors

Michel Dumontier is Distinguished Professor, Institute for Data Science, Maastricht University, Maastricht, the Netherlands.

Kathryn Wesley is Continuing Resources and Government Documents Librarian, Clemson University, Clemson, South Carolina.