Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

Semantics & Knowledge Learning for Chemical Design
01/05/2019
AI^3 Science Discovery Network+
Solent Conference Centre

Dr Samantha Kanza
University of Southampton

09/07/2019

AI3SD-Event-Series:Report-10

Semantics & Knowledge Learning for Chemical Design
AI3SD-Event-Series:Report-10
09/07/2019
DOI: 10.5258/SOTON/P0010
Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery
This Network+ is EPSRC Funded under Grant No: EP/S000356/1
Principal Investigator: *Professor Jeremy Frey*
Co-Investigator: *Professor Mahesan Niranjan*
Network+ Coordinator: *Dr Samantha Kanza*

Contents

1 Event Details

Title	Semantics & Knowledge Learning for Chemical Design
Organisers	AI^3 Science Discovery Network+
Dates	01/05/2019
Programme	Programme
No. Participants	17
Location	Solent Conference Centre

2 Event Summary and Format

This event was run by AI^{3}SD and was designed to explore the different aspects of using semantic web technologies in the chemical space, and to promote discussions about why these technologies are so important and the different ways in which they can be used. The event was a full day, hosted at the Solent Conference Centre in Southampton. The programme was made up of a number of presentations, starting with the importance of provenance of data and a history of the semantic web over the last 20 years with some details of where we are now. The event then progressed to some more specific talks about using semantic web technologies in the chemical domain, to make predictions and to model different aspects of chemistry in ontologies. These presentations were all run consecutively so it was possible to attend each talk, and the final order of the day was an expert panel made up of the speakers. There was plenty of time for networking, as there was both a lunch and drinks session included as part of the day.

Figure 1: Solent Conference Centre

3 Event Background

Designing chemicals, discovering new drugs, discovering materials and indeed all aspects of scientific discovery are all tasks that are highly data driven, and Semantic Web technologies are key to enabling researchers to deal with high levels of data in a useful and meaningful way. Semantic technologies facilitate representing data in a formal, structured, and interoperable way, and enable data to be reasoned over to infer potential relationships. AI^{3}SD organised this workshop to explore the ways in which Semantic Web technologies can be used to drive predictions in chemical design, including using Machine Learning and other AI techniques to exploit semantic links in knowledge graphs and linked datasets. This event forms part of the AI^{3}SD Event Series, which aims to bring people together around important areas of using Artificial Intelligence for Scientific Discovery.

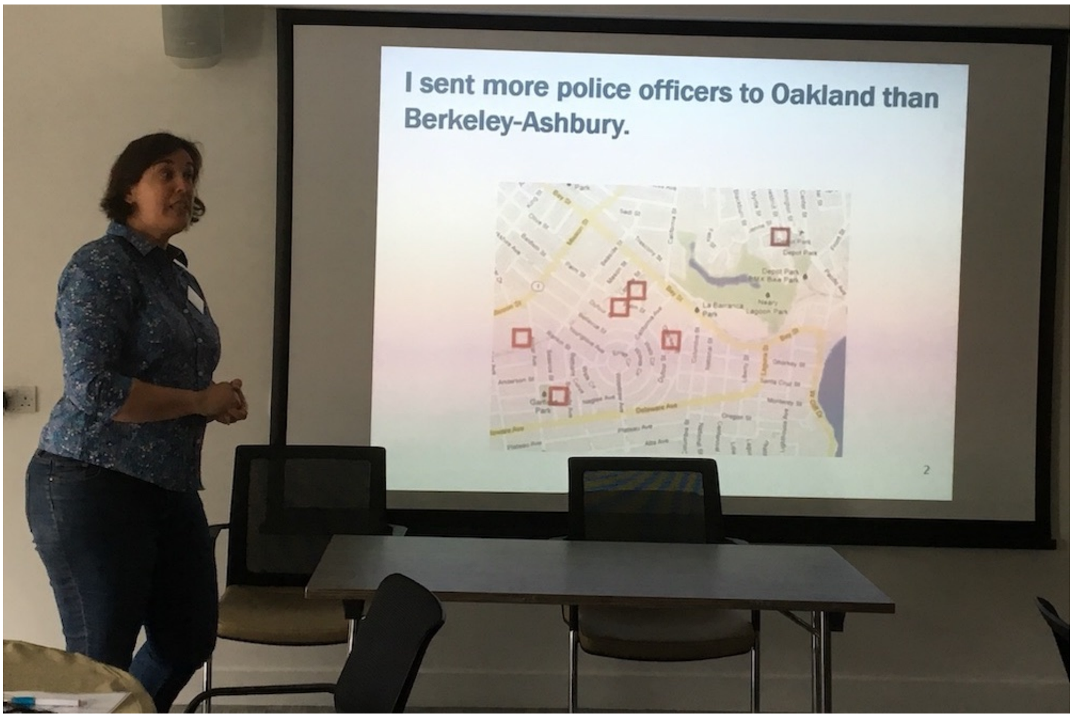
4 Talks

Figure 2: Professor Jeremy Frey

The session of talks was opened by Professor Jeremy Frey giving a short introduction to the Network+ and the workshop format. In his talk Jeremy stated that the Semantic Web is the missing link between data and AI. This workshop was organized to explore the different ways in which Semantic Web technologies can be used for chemical discovery, and one of the intended outputs of this workshop is to create an article or white paper about this event. Jeremy cast our minds back to the AI3SD Launch meeting held at the SCI back in December 2018, noting that whilst we had lots of fantastic discussions and presentations about Machine Learning, AI and scientific discovery, there was a warning note sounded by our advisory board member Tony Hey. The Government is backing AI very heavily, and have made a big bet on it with high expectations of its prospects, If we fail to deliver more than just an incremental change, then this could be more than the third disaster for AI and Science, and acceptability of science. We need to deliver interesting and important things, and thus far people have had many misconceptions about the Semantic Web and we need to put those right.

5 You did WHAT? - Dr Age Chapman

Our first talk was by Dr Age Chapman, an Associate Professor at the University of Southampton, and it centred around the importance of provenance. Age begins by saying, there has been



I sent more police officers to Oakland than Berkeley-Ashbury.

Figure 3: Dr Age Chapman

a lot of work done to explain what is going on in Machine Learning, but in reality, this is a great black box and there aren't really explanations about what is going on behind the scenes. Age makes these points with some real -world examples of headlines of stories, that mask a lot of information behind their titles.

Example 1: "I sent more police officers to Oakland than Berkeley-Ashby' - A recommender system suggested that more officers should be sent to Oakland than Berkeley-Asbury, but why? Underneath this system is a machine-learning algorithm that determines where to send police forces. Using a non-racist machine-learning algorithm, this system will consider information about the prevalence of recent crime among certain postcodes, and make recommendations about which postcodes should have more police officers sent to them. This is an explanation of sorts, but in no means tells the whole story. It's also arguably not really a machine-learning algorithm. It's a mathematical equation for predicting where events will happen based on previous events. If a series of events get to a certain level then it is activated and triggers a response, in this case to send police officers to a specific place. But is this actually an accurate representation of crime? Does it mean that these areas will have more crime? What data is actually being put into this algorithm? Seemingly data from the police forces is being used, which means it will have a systemic bias. We already send more police to certain areas, which raises the potential for uncovering crime -based activity in these areas, which in turn means that they will hit the activation level sooner. The transparency required here isn't just about understanding the algorithm, it's about understanding where the data for the algorithm came from, and what was done with it.

Example 2: "I decided to investigate the effect of human/mouse MAPI on AB-mediated inhibition of LTP" – This is a loaded title where a lot of decisions have been made to ascertain what study to conduct. The backend searches of the right area to be investigated here relied on the knowledge representation. But who did this? And how or why should we trust it?

Example 3: "I issued a recall on burgers in school meals" – Why did you do this? How did this happen?

These three examples all bring up the question of why? How and why did we get to these answers? This leads us to provenance. This is a record of the history of creation and modification of data. It is modelled as a series of entities or artefacts. We want to know who is using our data, and if there is a problem with the data. Provenance is very important to the Semantic Web. Take the third example of the burger recall from above. What is the chain of process? How is it done now? Does this process occur from when somebody gets sick? Do we trace back to records from the last inspection to see when this might have occurred? If we know that something goes back to a certain place – what is the risk of ending up with an issue? And what is the origin? Wouldn't it be nice to know where to put the sensors, rather than waiting for someone to get sick? This would be a better way to track and trace food and assess risk up and down the food supply. But we need to capture what information is needed at each place. By law this should be the way things are done, however realistically even in the best-case scenario this information can just about be pulled into a digital record. If we can trace the food going around the schools, then we can assess the risks and the certainty going forward. By using risk models we know what the potential issues to look out for are, e.g. pathogens in the food such as E.coli. However, in order to be able to do this we need to be able to go up and down the food chain.

Going forward we need to be asking how and why did I get to these particular results? If we are using AI we need to use the answers to understand the uncertainty and whether the information should be trusted or not. We need provenance. There is a formal standard for provenance (W3C). There are different ways of representing provenance information depending on what you are looking to do. But, how do you capture the provenance and keep the chain together? There are challenges when it comes to provenance such as identity issues. For example, if you do a calculation and end with some data which is put on a shared drive. If you then make a copy or email it to somebody or re-name it because you are working on the next version, then how can people know if they are working on the same thing? You can use URIs to specify different versions so you can make sure you work off the same document. This would require a handshake so you can establish provenance.

To decide how to use provenance you need to think about your use cases, and what do you care about using it for? For the food chain example, you need to care about where to put the sensors and where contamination enters the food chain. It is important to have information about the food, and storage temperatures etc. Whereas if we were looking to care about health of sensors, we would want to consider the battery life on the sensors and capture information at different points. Essentially, to successfully implement provenance you need to understand the flow of information and how many people it goes through. This will allow you to determine the capture points. You also need to decide what instruments you are using for this information capture and how that information is going to be stored. Provenance can be used as a new data stream, the metadata can be captured and record where a source or process is tainted. We need to consider where the data is actually coming from and who has touched it along the way.

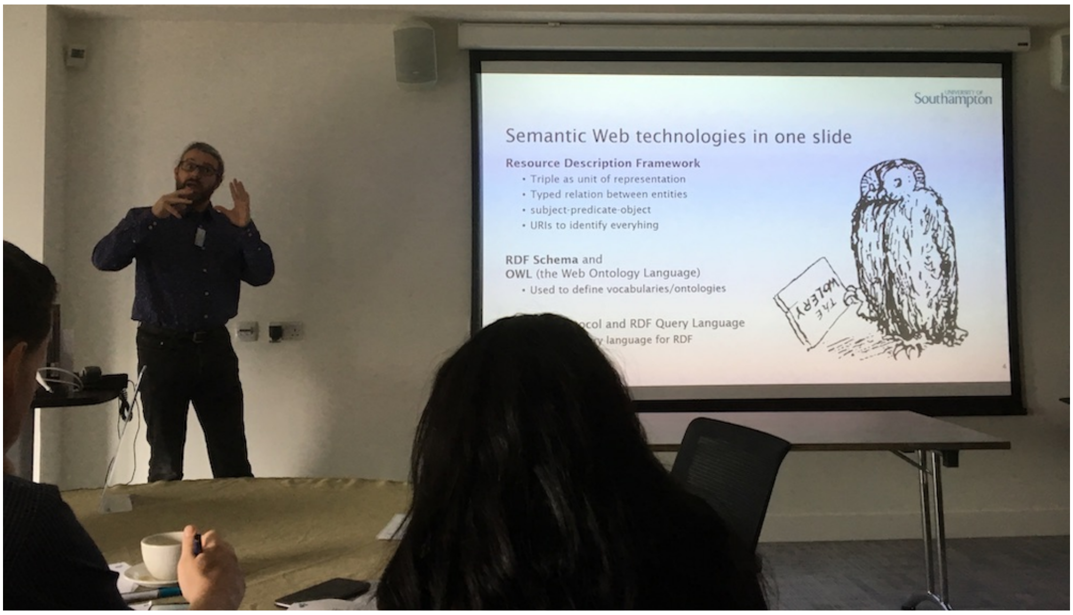
Age's gave us two final take homes from this presentation:

1. Unfortunately provenance suffers from us not eating our own dog food, as does the Semantic Web. There are many advocates of this technology but unfortunately it is frequently not used in practice. Equally there are some systems that lend themselves really well to provenance, and others do not, which means that they frequently get disregarded

with respect to recording provenance information. We need to be using the technologies that the community produce. Furthermore, when publishing standards it shouldn't be a case of publish and forget, we need to active and critical consumers of the things we produce.

2. Secondly, it is exciting to do AI for science, and it is important to get better techniques and know what is going on in the black box with the AI but we need to know what is being fed into the AI, what is the AI used for and what did the data look like and what happened to the data? It's complicated enough to follow a process that exists, let alone one you are unable to envisage. In scientific discovery we do things for one reason and create data and we don't know what it is going to be used for, and we do our best to describe it. We need a tool to teach best practices and enable us to do this better.

6 The Semantic Web at 20: Lessons from two decades of developing Linked Data Applications: Dr Nicholas Gibbins



UNIVERSITY OF Southampton

Semantic Web technologies in one slide

Resource Description Framework

- Triple as unit of representation
- Typed relation between entities
- subject-predicate-object
- URIs to identify everyhing

RDF Schema and
OWL (the Web Ontology Language)

- Used to define vocabularies/ontologies

ocol and RDF Query Language
y language for RDF

Figure 4: Dr Nicholas Gibbins

Our second talk of the day was presented by Dr Nicholas Gibbins, an Associate Professor at the University of Southampton who has a long history of working with Semantic Web Technologies. He was actually on the team itself that produced OWL and part of the team who won the first Semantic Web Challenge.

Nick began his presentation by reminding us that it has been twenty years since the first RDF standard was published. In 1999 Tim Berners-Lee outlined his vision for the web which links to the Semantic Web. We weren't there in 1999 and realistically we still aren't quite there now, and it isn't 100% clear how we are going to get there. Realistically the Semantic Web is symbolic AI (as opposed to non-symbolic AI such as Machine Learning) and AI has been through several blunders over the years. The 1993 report killed off all AI funding except for projects that didn't specifically explain that they were AI such as expert systems and neural networks. There have been fashions with AI and Symbolic AI came into fashion in the 70's.

So knowing that the Semantic Web is Symbolic AI, what is the Semantic Web? The Semantic Web brings context and meaning to data, and has three main strands of technologies.

- RDF [1] (Resource Description Framework) which represents the data in triples of the form (subject → predicate → object)
- RDF Schema [2] (RDFS) and OWL [3] (Web Ontology Language) which are languages used to create ontologies
- SPARQL [4] which is the query language of the Semantic Web, which uses an SQL esque syntax

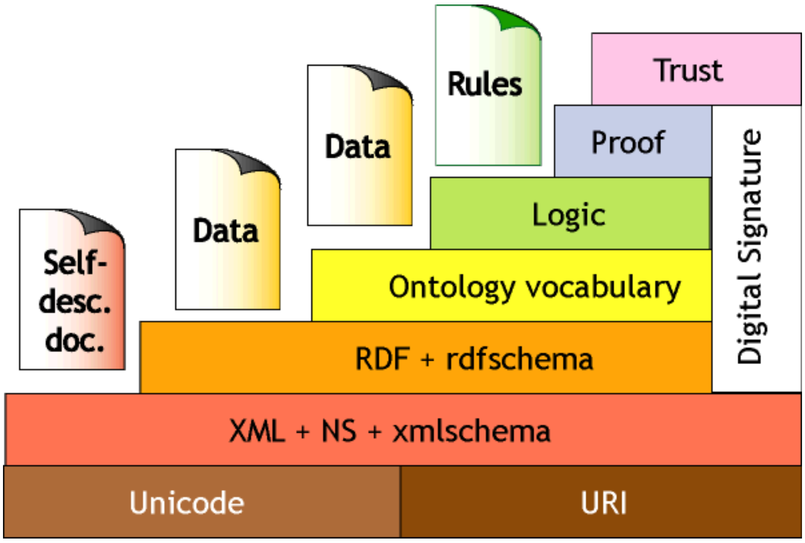


Figure 5: Semantic Web Layer Cake [5]

It has been 20 years since RDF was first published, but has it grown since then? Tim Berners-Lee said in his vision that there should be a lot of data on the web. There is a project in Germany that tries to track data on the web published using Semantic Web technologies. They have shown that the Web of Data was quite small in 2007, but has grown significantly in the last twelve years. However, there are definitely some domains that are more prevalent than others in this Web of Data. Figure 9 demonstrates that lots of data is being published, particularly by domains such as the life sciences but is it actually being used? And if so why not?

Nick has put together nine lessons for the Semantic Web, focusing on some of the missteps that have been taken so far, and providing recommendations for enhancing the usage of Semantic Web technologies going forward.

Nine lessons for the Semantic Web

1. **Beware of the Hype** – Early publications raised an unrealistic hype, such as Tim Berners-Lee and Jim Hendler's famous article in Scientific American in 2001 [7]. This is extremely unrealistic. The Semantic web isn't a silver bullet and it isn't good for everything, for example, it isn't particularly good for quantitative data, it can be represented but it's not natural. The Semantic Web is not good for reasoning with uncertainty. However, it is good for representing qualitative data, combining heterogeneous data and facilitating interoperability. *Lesson: We need to be aware of what the Semantic Web can do.*

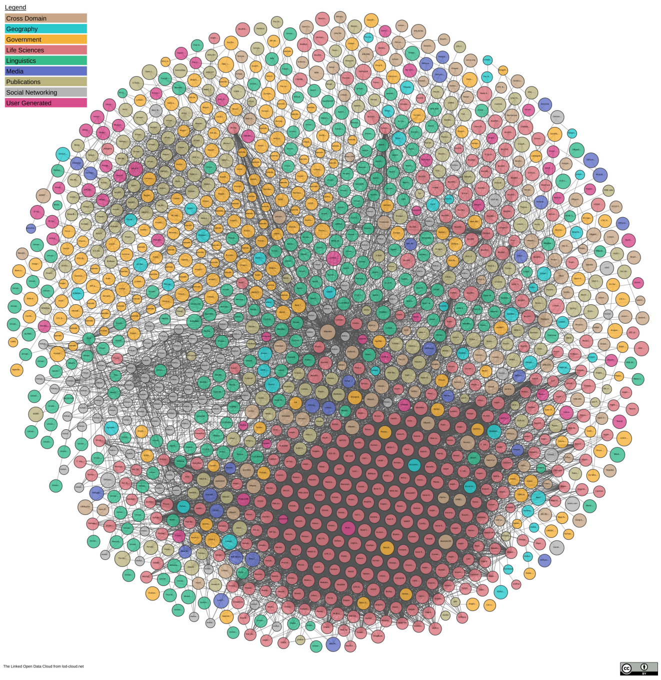


Figure 6: Web of Data [6]

2. **A Little Semantics Goes a Long Way** – The semantics in the Semantic Web should be used with a lowercase s. It is important to remember that a small amount of semantics goes a long way. We mostly only need a light touch of reasoning, and reasoning at scale with expressive languages is expensive. The initial Semantic Web standards were expressive ontology languages where OWL and RDFS complicated each other with their presence and these technologies weren't user centred; it was a solution driven project which made things too complicated. Nobody asked the users what they wanted, and it is imperative to include the people. *Lesson: Going forward we need to play to the strengths of the technologies when creating ontologies. For a lot of things you may only need RDFS reasoning, and you won't always need OWL reasoning. Furthermore, if you have a very large dataset then you aren't going to want to perform OWL reasoning over the entire dataset.*

3. **It's the triples, stupid** – Early Semantic Web development included an XML push, and realistically RDF-XML wasn't a great idea and it wasn't particularly well behaved or usable. It was arguably the worst of both worlds. It was a verbose format that didn't play well with any XML tools or formats that other people were using. Because RDF-XML wasn't overly useable, other members of the community kept coming up with new formats such as N3, Ntriples, Turtle, RDFa, JSON-LD. This resulted in wasting time on arguments about the right format to use and promote, when in reality it is irrelevant. What really matters is that RDF has the underlying simple triple model. We are representing the triples and these formats are just about how we write this stuff down, and there are multiple libraries for conversion if necessary. *Lesson: If you are going to use SW you need to be format agnostic.*

4. **The Myth of "build it and they will come"** – Just because you put your linked data online doesn't mean anyone will use it. There was an early enthusiasm for Semantic Web with Government Open Data, but people didn't use it. Complex technologies typically have high barriers to adoption and usage. Realistically it takes significant effort to publish semantic open data and yet people don't' seem to be seeing the benefits. So then is it

worth it? *Lesson: To see a return on this you need to understand how and why users will use your data. Be supportive of your users, make your data accessible and your processes sustainable.*

5. **The Semantic Web is part of the Web** – Tim Berners-Lee's vision was about links between data, but seemingly people forgot about this for the first 10 years. The early standards focused on representation and query languages like OWL and SPARQL. We use URIs to identify SW resources. But what do they mean and what can you do with them? What happens if you put a URI that identifies a person into a web browser? These people won't appear so what should? This opens up philosophical discussions about what should happen here, and these questions weren't thought about at the beginning. *Lesson: Need to follow the hard won modern best practices, which is still a difficult thing to do!.*

6. **It matters HOW you publish!** – It's not enough just to put your data online, but how do you do it? Does it go all in one big file? Should you break it down? That would be easier for consumers who only want to access specific sections of the data, but it is more effort for the publisher, and additionally more effort for consumers if they do wish to retrieve all of the data. You could also offer query interfaces so you can get what you want, but SPARQL queries are expensive and also require an understanding of SPARQL. *Lesson: When you publish consider how your users will want to interact with the data, and model how it is going to be used.*

7. **Context Matters!** – In the early days of the Semantic Web, the Semantic Web Layer Cake was created (see Figure 5), and trust was at the top of these layers. What do we mean by trust now? Who said something (digital signatures) and evidence that they were telling the truth? Tim Berners-Lee was talking about mathematical proofs but now we should be talking about provenance. If you publish something online, why should users use your data or trust it? *Lesson: You must publish data with metadata and give them enough information to use it! Who published it? When was it published? Where did it come from? Use the Provenance ontologies (Prov-o).*

8. **'Standard' is not 'Mature'** – There are countless standards in W3C, and not all of them are mature or independent; therefore just because something exists as a standard, it doesn't mean that it will always be worth using. W3C demand that you publish standards with examples of implementations. Which is good, but even if they've come through this process they may not scale or have been tested widely, and quite a lot aren't used. This is the chicken and egg issue, which perpetuates in Semantic Web technologies. There is no point in adopting technology without an implementation, but there is equally no point in implementing technology without users. *Lesson: Make pragmatic decisions when choosing technologies. Do you have access to robust mature implementations? Do your users?.*

9. **Eat your own Dog Food** – Similarly to Age's presentation the final point here is to practice what you preach! We produce and promote technologies that we don't actually use and we are good at producing and promoting data that we don't use. Don't publish and forget! You can't expect your users to put up with what you won't put up with yourself. *Lesson: We need to make sure that we are active and critical consumers of what we make.*

So what are our future directions? The Semantic Web is twenty years old and has pretty fixed foundations with respect to it's technologies. RDF, OWL and SPAQRL haven't been updated briefly over the years but a realistically pretty fixed and are now at a quite robust stage. Some of the technologies seem to be at a dead end such as Semantic Web Services, and Rule Interchange Format. There is currently no industrial use or academic research around these technologies.

Some of the more recent standards work at W3C has been to create a linked data platform to make the Semantic Web part of the Web finally. There has also been work done towards representing the context of published data using prov-o and derived ontologies. Finally they have been supporting the use of Semantic Web technologies in specific domains, with domains specific ontologies and community ontology endeavours such as schema.org, which is used to semantically represent websites so they can be better searched. As shown by the Linked Data Cloud in Figure 9, the life sciences community is very active in this area and the life sciences and chemistry communities have been conducting a lot of this work themselves, whilst W3C tries to work out the bridges between these communities and identify where more things could happen. To conclude, we have the technologies; we now need the human endeavour to push them forward.

7 Challenging Chemistry: Solving Molecular Problems - Professor Jonathan Goodman

Figure 7: Professor Jonathan Goodman

The third talk of the day was given by Professor Jonathan Goodman, who is a Professor of Chemistry and Director of Studies of Chemistry at the University of Cambridge where he also serves as Academic Dean. The focus of Jonathan's talk was around solving molecular problems. He begins by pointing out that AI has already begun to change the way that chemistry is done, and there have been promises that AI will "revolutionise" chemistry, but there are still many things left to solve. So what is holding us back?

Chemistry is quantitative, uncertain and complicated, which isn't ideal for semantic analysis. There is a plethora of chemical data and a lot of it isn't well understood. Provenance is very important in this area as when data comes out of a machine, scientists want to know where it came from and what assumptions went into creating it, and the theories that underpinned those assumptions. We need the tools to achieve this but the currently available tools aren't suitable, and there aren't many people working on the types of tools that are needed.

Taking the periodic table as an example, when it was first conceptualised it looked somewhat ragged, with the impression that something was missing and that it was clearly a work in progress. Realistically today it is still a work in progress but additional work has been done to make it look tidy, which could be perceived to be a solution but in reality, this does not change the fact that there is still a lot of unknown information surrounding it.

Work has been done to try and use Machine Learning on the periodic table to discover which elements are connected to which other elements using chemistry papers. This approach used the Bag-of-Words Model [8]. This Machine Learning algorithm identified that Palladium was very similar to rhodium, but was substantially less similar to silver.

This type of unsupervised learning can elicit information that makes sense, and even though there are likely to be errors in singular papers, it could be hoped that these would be averaged out using a larger set of papers. If surprises occur then these should be questioned and investigated. Realistically we should be able to figure out what is going on, even with uncertainty. For example if you have a star diagram, and you identify a new star, you can be fairly confident that it will fit into that diagram, and you can be equally confident about where there are unlikely to be any stars.

However, as discussed above, chemistry is a really difficult and complex domain. Creating new molecules or transformations is hard. We know about 10^8 molecules, and about 10^{200} small molecules that are possible. But does every molecule we know about help us to explain the 10^{192} molecules? This is a very complex problem, and arguably an even more complex problem is chemical reactions, we know about 10^7 reactions, but how many more are there?
We have a lot of data about the molecules that we do know about, and potentially have computer science tools that will help us analyse something. We can try and use successes and surprises in predicting how molecules should behave to try and discover where our knowledge is limited and uncover new ways to solve molecular problems.

Jonathan then considers some areas in which Machine Learning could be used to further molecular research. For example, take maitotoxin, the largest single molecule that has been attempted to be synthesized. It has a ‘repeating’ pattern expect for a single ring which is arguably ‘odd’. How can we know this is the right structure? This is the sort of molecule you could look at through one shot learning than through traditional Machine Learning ways, to understand this you need to generate all the molecules like it and put into a neural network or Machine Learning algorithm. Or you could just look at this molecule and consider the interesting points and find exceptions. Machine learning is good at looking for patterns in the data and finding exceptions so this could be a helpful tool in this process.

A second example was to consider a smaller molecule: C168H338. This is an interesting but complicated molecule. Even though it looks simpler than maitotoxin, there are more possible isomers than particles in the universe, and even though many of these are impossible to make, there are still countless different potentials. This raises the question of what is the smallest molecule that cannot be made? This motif could be checked for in the possible isomers, to identify if it is disallowed, but the ratio of allowed to disallowed drops to almost zero when looking at all the possible isomers. The chemical space is extraordinarily complicated, as it you make a tiny change in a molecule it can vastly change the properties of the molecule. And in addition to these complexities, it is still very difficult to get hold of the necessary chemical data to perform this research. We need to identify how computers can help in these instances. Chemistry knows and understands a lot of these points, but how can this knowledge be represented in an ontology? We need data to be accessible in a way that it can be made use of by computer science tools.

Accessibility is a key issue here, to both data and research. We need access to data to be able to discover the peculiar links that will further research. Things are changing with open access publishing in ways that we don't understand. Universities are struggling to afford the expensive journals, and are starting to move more towards open access publishing, although that is still a costly endeavour. However, hopefully that should on average make more chemical information available that we currently have. Additionally, there are some chemical data providers such as the International Union of Pure and Applied Chemistry (IUPAC) [9], The American Chemistry Society (ACS) [10] and The Royal Society of Chemistry (RSC) [11]. However, we cannot hope to make any real steps forward without computational help, and to do this we need the data!

This data also needs to be in a format that can be consumed by the Semantic Web and AI technologies. The patterns that chemists are looking for aren't actually that complicated but the data that holds this information needs to be able to be formatted in a way that makes it computationally accessible, and currently chemical data is currently not in a format that could be easily handled using these technologies. For example, if we are creating something dangerous such as hydrogen cyanide, how can we get this across in a research paper? Highlight it? Could we write a program to put this into RDF and automatically make these connections? Realistically we should be able to answer this now. There is probably a lot of data that could make products safer but we need to work together with computer scientists to put the data into the correct semantic formats to actually do this, as the chemical data isn't there yet.

This could be addressed by putting molecules into a usable format such as InChI Keys [12] which are a text encoding of a molecular structure, but this could cause clashes as there are less InChI keys than there are molecules. Further, InChI keys don't work for everything, they don't work well for organometallics or reactions yet either. We should be making things FAIR [13] and if this is achieved we should be able to handle our molecular data better. All molecular data comes from a complicated process. Do we have the right structure? We often do, but can't always be certain that what we think we have is what we actually have. Certainty is less 100% but tools are being developed to increase this. If we do this well we can make more of the data we already have. We should be able to get a computer to pull out patterns in a large group of papers. However, again a significant barrier to this is getting the data available in a convenient form and being legally allowed to re-use it.

A key conclusion here that reverberates around every Machine Learning discussion, is that Machine Learning needs data! Tools are being developed to try and ensure that data doesn't leave the laboratory without being appropriately captured and now we need to work together to get chemical data into an accessible reusable interoperable form so the powers of Semantic and AI technologies can truly be realised to make significant progress in the areas of molecular discovery.

8 Semantics vs. Statistics in Chemistry – Dr Colin Bachelor

The penultimate talk of the day was presented by Colin Bachelor, a theoretical chemist by training who works at the Royal Society of Chemistry. Colin took us through semantics versus statistics in the world of chemistry. This talk comprised five main sections, beginning by discussing the different approaches of Semantic Web technologies and statistics, highlighting some important questions about these areas in 2011, discussing the notion of whether you can tell if something works or not, followed by detailing the work that has been done at the Royal Society of Chemistry in this area, and finishing with some tentative conclusions.

Colin begins by introducing the notion of semantics vs statistics, and noting that this talk could also have been entitled "rationalism vs empiricism in chemical data". So what are these two

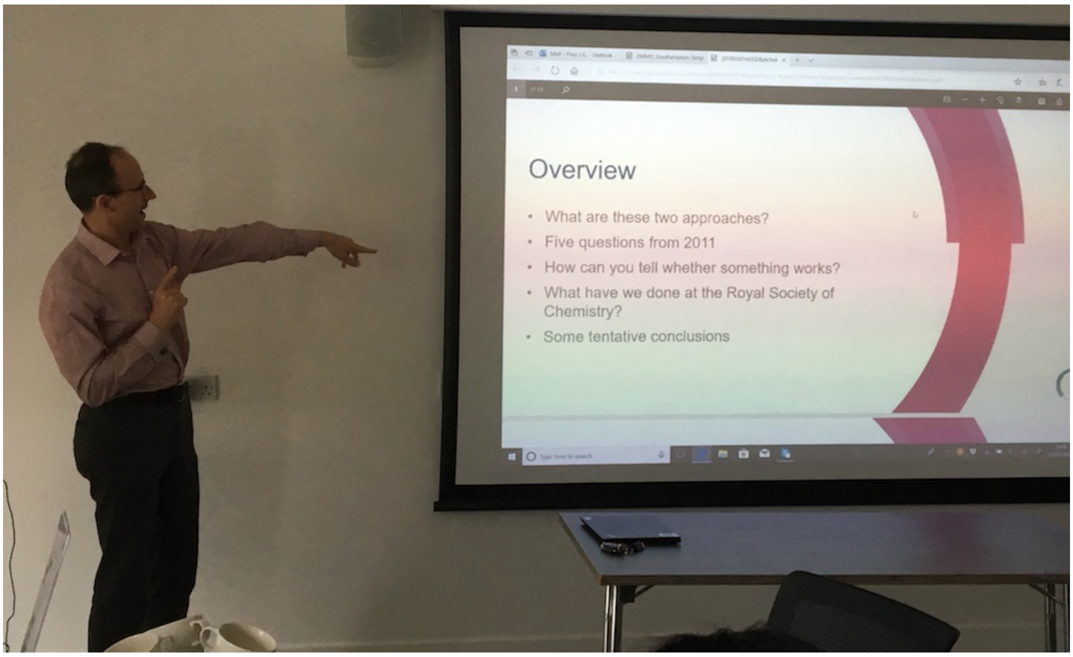

Overview
What are these two approaches?
Five questions from 2011
How can you tell whether something works?
What have we done at the Royal Society of Chemistry?
Some tentative conclusions

Figure 8: Dr Colin Bachelor

approaches? We have language as a formal, logical system; or language as a use. We have gotten quite a long way with the first one.

Colin introduces us to two important quotes about the notion of language. Firstly, one by Karen Spärck Jones, who created the search engine and coined the quote "words stand only for themselves". Secondly J.R. Firth, a linguist who stated "You shall know a word by the company it keeps". This notes that we need to know about the words around a words, e.g. cat can mean different things depending on what words it is next to. Realistically, this is a statement about context, and the Semantic Web puts things into context and gives them meaning. Whereas statistics can be used for deep learning.

These two approaches use different types of classifications. The Semantic Web approach is Aristolian, using hierarchies and ontologies; whereas for statistics the approaches are unsupervised such as clustering, but they identify a set of features. These can be combined by using formal concept analysis and essentially creating tick boxes for all the different features and creating a diamond lattice of your ontology.

Despite how long the Semantic Web, and indeed AI technologies have been around, there are still many questions surrounding how to use these technologies and around what their true capabilities are. The questions that were posed around these areas in 2011 were:

1. What can you achieve through letting words or images stand for themselves? - What are you actually trying to do? Is it solution led and you have the capabilities of OWL2 etc and you look for a problem to apply to it.
2. What do you hope to achieve through semantic markup?
3. What kinds of non-textual internal data do you have access to?
4. What can you do with your non-textual data?
5. What external data do you have access to?

We have suffered some AI Winters, for example in the 1970's and 80's, but we are now living in an AI Summer, and these questions are still relevant, particularly questions 3 to 5. Which leads us to the big question of: How can you tell if something works or is valid? There are typically two ways of approaching this, intrinsic criteria where a system is evaluated with respect to its overall objective and whether that objective has been met, or by extrinsic evaluation whereby a system is evaluated based on its functions in a specific context, such as usage or usability.

Colin then took us through three projects that have been worked on by the Royal Society of Chemistry in the area of semantic technologies.

Project Prospect: This project involved semantic annotation of chemistry articles using manual annotation and text mining to produce enhanced HTML, RSS feeds and open source ontologies. It is already possible to extract chemical structures using molecule names as these are a structured language in themselves, and equally code can be written to extract chemical structures from binary files such as ChemDraw files. Therefore, the rationale behind this project was to understand what hand created ontologies could add to chemistry articles beyond the tried and tested methods of indexing and searching by chemical structures. As part of this project, the Named Reactions Ontology (RXNO) was created [14]. This was done by getting two experts with synthetic organic chemistry PhDs to annotate documents individually to classify batches of reactions, and then see how their classifications compared to each other. The differences were discussed and the guidelines were put into a flowchart. Unfortunately, this does have limitations as much as RXNO was classified by hand and as such automatic reasoners cannot copy very well with it. This project was evaluated partially by usage and partly by impact. In this instance there wasn't entirely solid evaluation criteria as there was uncertainty about what to look for. More information on this project can be found in [15].

Subject Categories for RSC Advances: The idea behind this project was to divide up megajournal (RSC Advances) [16] into browsable chunks using bag-of-words and bag-of-cited-journals [8] and output a contents page. If it is citing new combinations of journals then that is worth noting. More categories mean that it takes longer to calculate topics, e.g. if you have 100 categories then 80% of these will potentially look interesting and the other 20% won't be worth using. These categories will be created by taking suggestions from editors and selecting a seed article and iterating until you are satisfied with the generated list. This project was evaluated partly on usage and partly via governance. Realistically uptake has not been significant. This could be because of the design of the website or because it isn't something people actually want to use. A potential lesson here is that whilst a research board of a project might be in favour of the outputs, doesn't mean the users necessarily will.

OpenPhacts: This third project was to use shared identifiers to describe disparate data of interest in drug discovery, and output large RDF files for input into industrial systems. Two dozen properties were calculated for $> 10^6$ molecules using the CHEMINF ontology [17]for cheminformatics, the QUDT Ontology [18] for units and measures, and ChemSpider [19] for the molecule IDs, This project was evaluated based on whether it worked and then its take-up by industry, which was a difficult and anecdotal way to measure success.

Colin then took us through some of the aspects of the core technologies that make up the Semantic Web.

So what is RDF? Everything is binary and there are no variables, things are divided into predicates, individuals and classes. This has similarities to First-Order Logic, although this only has predicates and no classes. Although OWL can use logic of this ilk for reasoning and inferencing. OWL however provides sophisticated class handling and facilities handling classes and predicates together. This suggests that RDF subjects and objects (in the subject → predicate

→ object formation of a triple) should be the ones to do the heavy lifting, not the predicate. Hence in OWL ontologies a lot of the logic is attributed to the classes rather than the predicates (which are referred to as object properties and data properties in an ontology).

When you are creating RDF you can choose what your triples look like as there are a number of representation options:

- **Dublin core** – everything lives in the predicate, and this has a large amount of predicates. OWL is good at handling classes but less good at handling predicates.
- **Ad hoc Semantics** – write down sentences and make links based on English. Here there is no clear division of labour between predicates/subjects/objects.
- **Event Semantics** – this reduces the work done by predicates.

In conclusion. Semantics can mean all kinds of things. It can have explicit meanings or abstract meanings, and there are a wide variety of types of semantics such as event semantics or ad hoc semantics as detailed above. There is no one set definition, in fact the more meanings of the word semantics you come across, the more lenses you have to look at a given problem. Furthermore, data for semantic representation can come in all different forms such as tables, pictures, protocols etc, and you need lots of information to be able to make the correct decision about which semantics to use, and generally this will depend on the data available and the overall aims of what you are trying to achieve with the data as to which method or methods is the most suitable.

9 EMMO (European Materials & Modelling Ontology): semantic knowledge organisation for applied sciences - Dr Alexandra Simperler

Our final speaker of the day was Dr Alexandra Simperler from Goldbeck Consulting. Alexandra works on the EMMO Project, which stands for European Materials & Modelling Ontology [20]. This project is between Goldbeck Consulting, Access, SINTEF, The Fraunhofer Institute IWM and the University of Bologna.

Alexandra begins by giving us an introduction into ontologies. Ontologies are for if you need to connect data, or discover new things, or discover new materials. Thus far the Far East is dominating with ontologies to generate materials ontologies. However, most of these are unfortunately short lived. Typically someone will decide that they need an ontology for a specific purpose, create it, and then it will disappear over time.

There is a drive for interoperability, people want to work together and unify and standardise. Examples of this can be found everywhere, such as the Allanthrope foundation. Typically when you buy different equipment for analysis, these pieces of equipment would all use different file formats, which didn't interface with other pieces of equipment. A set of companies got together to address this issue and created the Allanthrope foundation to facilitate unification. Another example is the ISO Standards which are used across industry. These standards require at the very least, taxonomies of terms.

Ontologies are also created and used in a bid to facilitate interoperability. Alexandra talks through the Semantic of Knowledge Organisation Systems (as shown in Figure **??**. Starting from the bottom you have a list, and need to consider what should go into the list and how it can be handled. Next comes the informal hierarchy such as a table of contents or xml, which then go into a thesaurus to check for things that have the same spelling but different meanings.

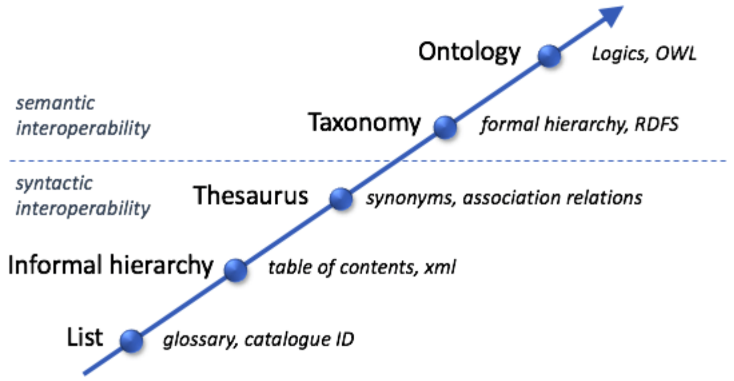


Figure 9: Web of Data taken from: [20]

This is followed by syntax interoperability. To reach semantic interoperability we have to start with a taxonomy (formal hierarchy). The final piece in the puzzle is OWL, which Alexandra compared to a spiders web as it can weave many complex relationships between your taxonomies. These can be made as complicated or as simple as is necessary.

There is a big issue across ontologies with respect to speaking the same language. There is a language issue as people speak about the same things in different ways and there are different terms for the same things, and the same terms for different things across different disciplines. With respect to reviewing materials modelling, we want models, where the models are optimistic models with a chemistry equation (e.g. Schrodinger equation), and then the materials equation, which together form the optimistic model. When people talk about multiscale modelling they dispute over how they do this. Looking at the granularity the model goes for could do this, or potentially one could take a picture of material and run a quantum chemistry equation on it.

In April 2018 the CEN (European Committee for Standardization) Workshop Agreement CWA 17284 was made with respect to "Materials Modelling – terminology, classification and metadata". The Modelling Data (MODA) is formed from the following components:

1. Use Case
2. Start translating this into the model – different questions about this model
3. Solver
4. Raw Data → need to do some post processing to get further

This requires formalisation and reasoning, and poses the question what should the EMMO (European Materials & Modelling Ontology) be able to do? Should it be able to take a use case from the real world, create an ontology to provide philosophy around everything that plays a role in this use case? This would involve involves software, model, and measurement techniques etc.

With respect to the EMMO there are only four primitives: the taxonomy (classification), mereotopology (parthood and slicing), semiotic (philosophy), and the set theory (membership). In the ontology world, physical entities are used to describe other physical entities, there is a user world and a physical world. One can begin with an abstract concept of terms and progress to the geometric/topological level and then to the physical level. EMMO gives users a way to communicate their interpretations but is not an automatic connection between the real and the ontological world.

EMMO material entities are defined by a hierarchy of parthood relations, and can maintain its granularity. Axioms are important! For example, birds can fly, but penguins are birds and they can't fly. We need axioms to stop us coming to the wrong conclusions.
The core of EMMO can be broken down into three layers:

- **Abstract Conceptual Layer** – clear separation between the set (set theory) and item (mereotopology)
- **Geometrical/Topological layer** – this module works with space and time, and can be time sliced in time, space or time and space
- **Physical Level** – real world entities

EMMO also contains semiotics, which are needed to determine the correct context, as translating between different languages and terms is more than just going word for word. As discussed above different disciplines use different terms to describe the same thing, and the same words to describe different things and therefore context is imperative.

The University of Bologna (a partner in this project) has a strong philosophy department and the need for a philosophical input into the ontology was cited, although equally it was noted that neither a physicist not a philosopher by themselves would be best placed to create the ontology, and therefore they decided to work together.

EMMO is continuing to be developed and there are plans for its usage over the next few years in materials modelling and ontologies. It can be used in many different fields including science, modelling, AI and Big Data Analysis, Characterisation and Industry.

So where do we go from here? The Materials Modelling Council funding comes to an end in august. This project has laid the foundations and they have been looking for things to make this council go further, and it's important to note that whilst the project has the word European in it, it is for everyone to get involved with.

10 Discussion

The final discussions at this meeting were towards forming an action plan. Professor Frey started off our meeting by discussing the eScience projects and how at that instance the core pieces were there but the required tools were not. At that point it was possible to sell the idea of what could be done with linking data, but the tools and the triple stores did not work appropriately at that time. The first version of a semantic ELN at the University of Southampton was on a tablet for portability in the lab purposes [hughes2004semantic], but the memory and network speed meant that the RDF graph of an experiment was not held on the tablet as that wasn't feasible. There were no specific boundaries to running software on the tablet so when it realised that there were connections missing it added them, but then when it resynced with the triple store after the experiment had been conducted it raised duplication issues. Realistically now this would be less of an issue, but at the time this was a considerable problem. We need good practice for creating these pieces of software.

Professor Frey queried the workshop group to discuss the pain points and where issues arise in the current world of using Semantic Web technologies. A major pain point that is brought up at every meeting we run in one way or another is regarding data. In this instance, the pain point is that we spend a lot of time reconfiguring data from different places. Could we do something here? Can the Semantic Web help? We consume data and produce it, and we know about the issues regarding bringing data in but we often magnify the issues. What would we have liked someone else to have done to make the input data better?

Suggestions included:

- We could make use of the Software Sustainability Institute (SSI) tool [21]. This tool will let you know if your software is sustainable and provide tailored advice on whether it is worth preserving and how to improve it.
- We need software data tools that can answer questions such as: Is my data shareable? Is it worth putting onto a database or an international database?
- Making sure that any data that is published comes with some sort of analysis tool so people can make better use of it.
- Making ontologies available: Lots of academic papers detail the creation of ontologies but don't give any tangible link to where others can access this ontology.

This raised the question of what is reasonable to ask people to do with respect of their data? Asking people to add a lot of metadata can prove lengthy and take a lot of time. We need to identify the minimum subset of data requirements to make our data more useful. For example, there are issues with producing the right metadata for models. Less information is needed to describe the model than the terabytes of simulations, but it requires a lot of work to figure out what is important, and some of the issues may only surface years later. It is also very difficult to represent uncertainty in data; heterogeneous data can be represented but this doesn't allow for calculating the uncertainty. Potentially this needs to be calculated elsewhere and inserted in.

Another pain point that was referenced was how can one compose a patchwork quilt of different Semantic Web approaches and toolkit. There is currently no unified representation. Additionally, there is uncertainty about how to remove information from the Semantic Web. For example, if information about a molecule is made publicly available, and that information is then found to be wrong, how does that information get removed? A further pain point is also the nature of OWL property assertions, as OWL2 can have negative property assertions and this causes inconsistencies, this is a known issue and potentially this is up to provenance to sort out.

Following on from the issues related to removing incorrect information from the public domain, there was a pain point issue raised in relation to journal papers, publications and reviews. Lots of bad papers are currently slipping through the journal submission and review process and this needs to be addressed. It is often not possible to check the data and code that is submitted alongside these papers (assuming of course that it is submitted in the first place).

11 Action Plan & Conclusions

Overall, we need to set some examples of good practice. Semantics need to be appreciated for that what they actually are, and not underappreciated because they have been given unrealistic promises to live up to. We need to find ways to minimise the effort and demonstrate the real benefit that can come from using semantic technologies. Additionally, systems that use Semantic Web technologies don't always need to be really complicated, and we don't always need to use every high-level complicated feature available. Technologies should potentially only be used in the correct context for their use and we need to decide on a minimum level for this, and also ensure that the remits of these technologies (like the Semantic Web) are made clear so there are no misinterpretations, and ensure that they aren't made to sound unnecessarily overcomplicated in presentations such that others perceive them to be too difficult to use.

The idea of plug and play with ontologies is a lie, we need to push for good upper level ontologies that can be expanded out into domain specific ontologies. We need to provide users with an easier way in, and whilst we can use features during creation of these tools we cannot rely on our users to make use of them, and we should be making our tools clear and simple to use.

We need to establish best practices. From the perspective of data, we need to decide on a consistent standard and minimum set of requirements for datasets. Ideally there would be some uses cases for where data can be tagged and or represented semantically. As we have so much data and it is constantly changing, capturing this for reproducibility is hard, and this approach is the only way that will make it possible. We need to set examples of good practice and in the nature of eating our own dog food, provide provenance for each piece of data (or perhaps create a tool that can do this automatically).

Similarly, for software we need to tune how we publish this to ensure that it is done in the most useful way. It isn't enough just to make software open source or publicly available, you need to publish explanations and guidelines alongside the code. For journal papers we should be making archives available of well curated data as demonstrators, to demonstrate the gold standard that the scientific community needs. There could also potentially be a new role in this process for data editors or reviewers, who review journal papers purely from a data perspective.

Finally, we need networks! There are people working with AI in many different areas, but we need to expand the Semantic Web community! AI needs Semantics and the AI community should be taking up these technologies both to enhance their own research and work, and to make improvements in the usage and accessibility of semantic systems.

12 Participants

This was a small workshop attended by 17 individuals from industry and academia and a range of disciplines. The majority of the participants were from academia, but there was still an industry representation, and the participants expertise ranged across different aspects of both Chemistry and Computer Science, including those with expert knowledge on Semantic Web Technologies.

13 Related Events

For those who are interested in getting involved with the Semantic Web community and attending related events there are some additional events that cover similar areas of interest.

- Semantics 2019 in Karlsruhe, Germany (9th-12th September 2019)
- ISWC – International Semantic Web Conference in Auckland, New Zealand (26th-30th October 2019)
- ESWC – Extended Semantic Web Conference in Heaklion, Greece (31stMay-4th June 2020)

Each of these conferences are for presenting Semantic Web based research, innovative technology and applications.

Upcoming events of interest can be found on the AI3SD website events page.
http://www.ai3sd.org/events/ai3sd-events
http://www.ai3sd.org/events/events-of-interest

References

[1] W3C. W3C RDF. World Wide Web Consortium; 2014. [Online: Accessed 09-Jul-2019]. Available from: http://www.w3.org/RDF/.

[2] W3C. W3C RDF Schema. World Wide Web Consortium; 2014. [Online: Accessed 09-Jul-2019]. Available from: http://www.w3.org/TR/rdf-schema/.

[3] W3C. Ontologies. World Wide Web Consortium; 2015. [Online: Accessed 09-Jul-2019]. Available from: https://www.w3.org/standards/semanticweb/ontology.

[4] W3C. W3C SPARQL. World Wide Web Consortium; 2013. [Online: Accessed 31-July-2014]. Available from: http://www.w3.org/TR/sparql11-overview/.

[5] E Miller. Layer Cake;. [Online: Accessed 09-Jul-2019]. Available from: https://www.w3.org/2001/09/06-ecdl/slide17-0.html.

[6] lod-cloud net. The Linked Open Data Cloud; 2019. [Online: Accessed 09-Jul-2019]. Available from: https://lod-cloud.net/.

[7] Berners-Lee T, Hendler J, Lassila O, et al. The Semantic Web. Scientific American. 2001;284(5):28–37.

[8] Zhang Y, Jin R, Zhou ZH. Understanding bag-of-words model: a statistical framework. International Journal of Machine Learning and Cybernetics. 2010;1(1-4):43–52.

[9] International Union of Pure and Applied Chemistry. IUPAC; 2019. [Online: Accessed 09-Jul-2019]. Available from: https://iupac.org/.

[10] American Chemical Society. ACS; 2019. [Online: Accessed 09-Jul-2019]. Available from: https://www.acs.org/content/acs/en.html.

[11] Royal Society of Chemistry. RSC; 2019. [Online: Accessed 09-Jul-2019]. Available from: https://www.rsc.org/.

[12] Heller SR, McNaught A, Pletnev I, Stein S, Tchekhovskoi D. InChI, the IUPAC international chemical identifier. Journal of cheminformatics. 2015;7(1):23.

[13] Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data. 2016;3.

[14] Royal Society of Chemistry. RXNO: reaction ontologies; 2019. [Online: Accessed 09-Jul-2019]. Available from: https://github.com/rsc-ontologies/rxno.

[15] Royal Society of Chemistry. Chemical Named Entity Recognition and Literature Mark-up; 2008. [Online: Accessed 09-Jul-2019]. Available from: https://www.w3.org/wiki/images/6/68/HCLSIG$$F2F$$2008-10_F2F$1_Prospect-20081021-HCLSIGkeynote.ppt.

[16] Royal Society of Chemistry. RSC Advances; 2019. [Online: Accessed 09-Jul-2019].

[17] Hastings J, Chepelev L, Willighagen E, Adams N, Steinbeck C, Dumontier M. The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web. PloS one. 2011;6(10):e25513.

[18] QUDT org. QUDT; 2017. [Online: Accessed 09-Jul-2019]. Available from: http://www.qudt.org/.

[19] Royal Society of Chemistry. ChemSpider; 2019. [Online: Accessed 09-Jul-2019]. Available from: http://www.chemspider.com/.

[20] The European Materials Modelling Council. EMMO: an Ontology for Applied Sciences; 2019. [Online: Accessed 09-Jul-2019]. Available from: https://emmc.info/emmo-info/.

[21] Software Sustainability Institute. Online Sustainability Evaluation; 2019. [Online: Accessed 09-Jul-2019]. Available from: https://www.software.ac.uk/.