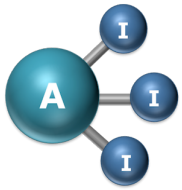

Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

The Physical Sciences Data-Science Service (PSDS)

Patterns

Failed it to Nailed it: Data Standards and Guidelines
05/11/2020
Online Event

Dr Samantha Kanza & Dr Nicola Knight
University of Southampton

10/12/2020

AI3SD-Event-Series:Report-20

Failed it to Nailed it: Data Standards and Guidelines
AI3SD-Event-Series:Report-20
10/12/2020
DOI: 10.5258/SOTON/P0033
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*

An EPSRC National Research Facility to facilitate Data Science in the Physical Sciences: The Physical Sciences Data science Service (PSDS)
This Facility is EPSRC Funded under Grant No: EP/S020357/1

Principal Investigator: *Professor Simon Coles*
Co-Investigators: *Dr Brian Matthews, Dr Juan Bicarregui & Professor Jeremy Frey*

Contents

1 Event Details

Title	Failed it to Nailed it: Data Standards and Guidelines
Organisers	AI^3 Science Discovery Network+, Patterns Journal & Physical Sciences Data-Science Service
Dates	05/11/2020
Programme	AI3SD Event Programme
No. Participants	47
Location	Online Event
Organisation / Local Chairs	Dr Samantha Kanza & Dr Nicola Knight

2 Event Summary and Format

This event was the second of the 'Failed it to Nailed it' online data seminar series. The event was hosted online through a zoom conference. The event ran for approximately three hours in an afternoon session.

There were three talks given on the topics of metadata, open data and linked data covered both from a domain agnostic point of view and with a specific focus on the bioscience and humanities domains. These talks were followed by an interactive panel session with experts from each of these topics talking about their experiences with data standards and guidelines.

3 Event Background

This event is part of the 'Failed it to Nailed it' data seminar series. This event series, currently comprised of four online events is a collaboration between AI^3 Science Discovery Network+, Patterns & the Physical Sciences Data-Science Service (PSDS). This event series follows on from a data sharing survey that was undertaken earlier in 2020. Each event in the series handles a different aspect of dealing with data aiming to educate and inform researchers about how to work well with their data, as well as encouraging discussion along the way. Following on from these events the organisers hope to be able to organise more face-to-face events in 2021 which will expand this event series.

Understanding the different data standards that are available is absolutely essential to any data-driven research, irrespective of the type of data. In these events we want to encourage researchers to consider this as a fundamental aspect of managing, presenting and organising their data rather than an afterthought, or something they wish they did but never got round to. This event aimed to provide an introduction to the different types of data standards, alongside examples of their usage and suggestions for best practices.

4 Talks

4.1 Metadata

This section will provide a general introduction to Metadata and FAIR, followed by a detailed summary of the talk relating to Metadata.

4.1.1 Introduction to Metadata

The basic definition of metadata is quite simply data that describes other data. Metadata can provide context around a piece of data, such as: the machine used to generate it, or the temperature at which a measurement was taken. The variety of possible metadata is extremely wide and it can be used in a number of ways, from describing authorship of a document through to describing the size of an image. In essence though, it helps data to be better searched, organized and understood [1, 2].

4.1.2 Introduction to FAIR

The FAIR principles, published by Wilkinson et al. in 2016 [3] are a set of guidelines designed to improve the Findability, Accessibility, Interoperability and Reusability of data and other digital research objects. In particular these principles increase both the human and machine readability of data. This should not be confused with the similarly named fair data certification[1] that applies to companies' handling of customer data. Regarding FAIR data as defined by Wilkinson et al., for each of the F,A,I and R strands there are several principles that should be implemented for data to be FAIR. The living document for the FAIR guiding principles can be found at GO-FAIR [4] and their website is also a useful information resource. Many organisations are working on FAIR and the implementation of FAIR within research disciplines, some other resources for FAIR include: FAIRsFAIR[2], CODATA[3], RDA[4], EOSC[5] & FORCE11[6].

4.1.3 Data generation, data standards, and metadata capture in drug discovery - Dr Martin-Immanuel Bittner (Arctoris)

https://orcid.org/0000-0001-8279-6900

Figure 1: Martin-Immanuel Bittner

[1] https://www.fairdata.org.uk/principles/
[2] https://www.fairsfair.eu/
[3] https://codata.org/
[4] https://www.rd-alliance.org/
[5] https://www.eoscsecretariat.eu/working-groups/fair-working-group
[6] https://www.force11.org/fairprinciples

The full video of Martin's talk can be viewed here: https://youtu.be/t3R6BV_8XGI [5]

Dr Martin-Immanuel Bittner is the Chief Executive Officer of Arctoris, the world's first fully automated drug discovery platform that he co-founded in 2016. He holds both a medical degree and a DPhil in Oncology. Martin has extensive research experience covering both clinical trials and preclinical drug discovery and is an active member of several leading cancer research organisations, including EACR, AACR, and ESTRO. In recognition of his research achievements, he was elected a member of the Young Academy of the German National Academy of Sciences in 2018.

Martin's talk focused on how data is currently being generated and captured in the drug discovery space and ways in which this process can be improved to increase the reproducibility of research. As scientists and researchers, our aims are to make research more robust, reliable and arrive at meaningful insights that can benefit society. Being able to rely on high quality data is what makes it possible for us to progress in biomedical research.

The talk began by looking at the current research mode and how it compares against other industries. Martin highlighted that pharma R&D productivity is declining [6], measured as drugs produced per research dollar invested in research. The term Eroom's Law has been coined to describe this inverse relationship to information technology's Moore's Law. A key issue in the decline in research productivity is the quality of the data that is produced. It has been shown that only 10-20% of peer reviewed findings are actually reproducible [7].

Another issue surrounding data generation is the method by which this data is collected. Much of the approach to data generation is still fundamentally the same as 50 years ago. Much of researchers' time is taken up with manual lab work rather than discussion, reading and hypothesising. Although we may have new pieces of kit and better microscopes, the methodology has not kept pace with technological innovation that we can see in many other areas. Martin commented that other industries have adopted technologies such as robotics, advanced electronics and cloud analytics and by doing so completely transformed the ways in which they operate.

Our current capabilities in data capture are just the tip of the iceberg compared to the possibilities that already exist. Currently, datasets are frequently isolated and unlinked and are stored in non-standardised formats which contain ambiguous descriptions of methodologies. An example given by Martin shows how a scientific protocol asking a researcher to 'mix' a sample could mean stirring, shaking, vortexing etc. and does not contain sufficient detail to actually enable another researcher to reproduce the exact experiment. This results in varying experimental outcomes and affects how reliable the data is in the end. But looking ahead, Martin talked about how we are already able to capture significantly larger quantities of data across the experimental lifecycle. This data is built on standardised protocols, linked metadata, verified reagent provenance and audit capabilities to provide a much richer, more robust data source. Martin discussed how these steps lead to improvements in the biomedical research landscape.

In the life sciences, one area of particular importance is reagent provenance. We want to make sure we only use reagents that we can be sure of what they are and where they came from. Martin highlighted cell line identification as an example where it is currently estimated that 50% of cell lines in research are either contaminated or misidentified. This casts doubt on years if not decades of research that was carried out with these cell lines.

Data generation has to be thought about in a completely different way to improve research

outputs. A key part of this is the associated metadata that is critical to make data accessible, both for humans and machines. Martin showed an example of different types of note taking[7] using lab books, spreadsheets and then fully linked and machine-readable data. We should be moving towards representing data in an interconnected, fully machine readable form, such as RDF, where we can capture and harness the power of rich metadata.

FAIR Data

Within the life sciences, and other disciplines, the concept of FAIR has garnered a lot of attention over the last few years. This is the concept of making data **F**indable, **A**ccessible, **I**nteroperable and **R**eusable. Martin explained the concepts of FAIR through use of a layered diagram[8]; at the core is the 'Digital Object', with other layers signifying 'Identifiers', 'Standards & Code' and 'Metadata'. Martin explained each of these areas a little more in detail.

Digital Object: This contains the actual data or code within it. To enable FAIR it must be given in a standard format rather than an obscure or custom file format. The other layers around the digital object enhance the data to allow more reuse.

Identifiers: Each digital object should be assigned a unique and persistent identifier to allow it to be unambiguously identified and tracked. There are several different persistent identifier (PIDs) schemes[9] that exist for digital objects, examples include: DOIs[10] (digital object identifiers) and URNs[11] (Uniform Resource Name). This should also extend to using PIDs in other areas, Martin highlights several examples; to identify authors (ORCIDs[12]), projects (e.g. RAiDs[13]) and funders / associated research resources (e.g. RRIDs[14]). These can help immensely in many applications to ensure that it is clear exactly who or what is being referred to.

Standards & Code: Implementing standards within your digital objects aims to make them as widely accessible as possible and allow their re-use a long time into the future, when other file formats may no longer be supported.

Metadata: This provides the context for the data or digital object. While basic metadata may help in finding the relevant data, richer metadata about the process through which the data is obtained is very helpful to understand the full context and allow more meaningful analysis. Examples given by Martin include: temperature, humidity and gas concentration at point of data collection.

Martin outlined the principles of FAIR [3] which are presented in Table 1 and commented that FAIR data need to be considered within every single research project.

- **Findability:** In addition to points already captured above for findability, Martin highlighted the need for metadata to be searchable and queryable so the records can actually be found.
- **Accessibility:** Looking at the accessibility guidelines the need for standardised protocols for metadata is clearly outlined, along with the need for these to be open and continually available.

[7] This can be found in Martin's talk at 08:00 - https://youtu.be/t3R6BV_8XGI
[8] This can be found in Martin's talk starting at 09:28 - https://youtu.be/t3R6BV_8XGI
[9] See https://www.dpconline.org/handbook/technical-solutions-and-tools/persistent-identifiers
[10] The International DOI foundation https://www.doi.org/
[11] See https://tools.ietf.org/html/rfc3986 although the URN term is largely deprecated except in narrow terms.
[12] Researcher Identifier https://orcid.org/
[13] Research Activity Identifier https://www.raid.org.au/
[14] Research Resource Identifiers https://scicrunch.org/resources

- **Interoperability:** The interoperability points relate to data being able to be integrated with other systems and data. Important considerations in this area are the use of vocabularies, terminologies and ontologies when describing the data.
- **Reusability:** To allow reuse the data should include accurate and rich metadata along with provenance records so researchers know where the data originated from and allows verification of original data and tracking of attributes such as authorship. It is also important to ensure that any data meets the relevant community standards. In some communities these data standards may already be well developed, but in other communities this standardisation is still a work in progress.

The data should be **F**indable	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 they describe F4. (Meta)data are registered or indexed in a searchable resource
The data should be **A**ccessible	A1. (Meta)data are retrievable by their identifier using a standardised communications protocol A1.1 The protocol is open, free, and universally implementable A1.2 The protocol allows for an authentication and authorisation procedure, where necessary A2. Metadata are accessible, even when the data are no longer available
The data should be **I**nteroperable	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
The data should be **R**eusable	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

Table 1: Listing of the FAIR principles - adapted from the FAIR guiding principles [3, 4]

Martin then presented a case study[15] from Arctoris within the context of a biomedical experiment. The goal was to create a single process that follows an experiment from idea to completion, allowing full metadata capture across the whole lifecycle and gathering information in a FAIR-compliant way throughout. The FAIRification process (making existing data FAIR) has gained a lot of attention in recent times, but it is not straightforward. It is better to ensure data is already generated in a FAIR-compliant way.

One stage of the process that Martin gave examples of was the AEFF (Arctoris Experiment File Format) and how it can be used to improve experiment reproducibility and reusability. The AEFF contains unambiguous experiment descriptions in a FAIR manner, in particular with detailed specifications and validators to ensure interoperability across languages while ensuring strong performance when utilised. Alongside the description of data sources, each experiment file contains a variety of metadata, such as structural metadata, acquisition metadata and statistical metadata. The AEFF is also used in conjunction with widely recognised ontologies.

The impact on two areas is worth highlighting in particular; this includes aspects relating to reproducibility and reuse.

Improving Reproducibility:

- Better defined experimental protocols ensure experiments are run in the same way each time.
- Automated systems can benefit from versioned optimisations to the protocol before an experimental run.
- Regular 'control experiments' ascertain the platform state and background variations in measurements outside of the data collection experiments.

Improving Reuse:

- Experimental protocols can be reused both within Arctoris and by outside parties, ensured via the interoperability measures included with the AEFF.
- Experimental data can be reused for data and metadata mining.
- Intermediate results can be processed and reprocessed in an experimental pipeline.

An example given by Martin contrasted a biochemical experiment conducted in a traditional experimental approach with the fully automated approach taken by Arctoris[16]. In the automated approach the amount of metadata captured includes information about the equipment, experimental conditions, the users among many other things. In addition to richer metadata, the system can produce a much more comprehensive dataset with a significantly larger number of data points for analysis. Martin highlighted that the capture of metadata has huge implications for the reuse of research results and thus their value for the whole research ecosystem.

There are a number of initiatives that recognise the need for work to be undertaken in data standards. In particular, Martin mentioned the Pistoia Alliance[17] which is a global not-for-profit organisation working within the life science and healthcare areas to lower the barriers to innovation. Members of the Pistoia Alliance include large pharmaceutical corporations, start-up companies and academic researchers. Together they are united by the recognised need to introduce data standards in biomedical research, as the current methods of generating data are

[15]The case study begins at 16:05 in Martin's video - https://youtu.be/t3R6BV_8XGI

[16]An example of the data captured in the experimental file is given in Martin's talk at 22:40 - https://youtu.be/t3R6BV_8XGI

[17]Pistoia Alliance https://www.pistoiaalliance.org/

not sustainable.

In conclusion, Martin noted that the topics he covered are just scratching the surface of the areas of data, data standards and data stewardship. It is an area that is evolving rapidly. There is work being undertaken in many different areas, including FAIR data implementation, AI in drug discovery, ontologies, lab of the future and unified data models to name just a few areas. We need to move from just looking at isolated datasets to looking at the whole data landscape with a focus on high quality, accessible, shareable data to allow us to make new insights.

Questions following the presentation:

Q: There is a difference between catalogue data and record data in terms of metadata, is there work to separate out different layers of metadata e.g. bibliographic and catalogue data separate from domain metadata?
A: *In the past people have just thought data is data, but it is clear now that there are many different layers to this data and metadata. Martin suggests good use of PIDs may help in this area to make sure you can actually find the object once you know the identifier.*
Follow up comment: In the UK we have the equipment data database[18]. This provides a database of research equipment, and can provide shared IDs for equipment and details about it. Now we can start tagging experiments with the identifiers for the equipment. In the future hopefully we can start to utilise this data in much better ways.

Q: What about dublin core[19] as a method of describing datasets? (Directed to Chris & Martin)
A: ***(Chris)*** *Dublin core is much better than nothing at all, but it really needs to have other systems and standards used in conjunction with it.*

Q: How did you come up with your data capture model for the experimental lifecycle, i.e. from specific experience or from studying existing research on experimental/research lifecycles across life science companies?
A: *When we started out, looking at automation and towards providing a platform for end-to-end experiment automation, we were among the first working on this. There are a handful of other companies (e.g. Synthace, Transcriptic) that had published information on description and capture of automated experiment processes. This information was examined in conjunction with other best practice information that could be found and we used our knowledge and experience to build on top of that. End-to-end automation is still a young field in the life sciences so there are still many areas that need to be expanded upon.*

Q: Would a different life science company be collecting the same metadata, or would they need to capture different metadata?
A: *There are some key metadata that can be captured across all processes that significantly aid reproducibility (e.g. temperature, timings). If these are captured within a laboratory then you are already making large steps towards experimental reproducibility. As of now, only a very small percentage of experiments worldwide are being conducted on automated systems, and the remainder are still largely done manually. Most (semi-)automated laboratories will largely be trying to collect the same types of metadata but the depth to which the metadata is collected will vary tremendously between companies.*
Over the next few years hopefully we will see improvements in the amount and depth of metadata being captured, as what we currently have is not enough.

[18] https://www.jisc.ac.uk/rd/projects/equipment-sharing-made-easy
[19] https://dublincore.org/

4.2 Open Data

This section will provide a general introduction to open data, followed by a detailed summary of the talk relating to open data.

4.2.1 Introduction to Open Data

Open data is data that is freely available to others to access, use, and share as they wish, for any purpose, without any legal restrictions [8]. There are a lot of different terms out there relating to data, big data, shared data, linked data etc. However, whilst the term open data can be applied alongside any of those terms, in itself the "open data" bracket purely refers to the notion of data that anyone is free to use, share and access, whereby the only potential requirement alongside that is to attribute the original source of the data, or to sharealike (copies or adaptations should be shared under the same license as the original). The Open Knowledge Foundation have produced a full open definition on what it means for data to be open, which can be viewed on their website[20].

The most important aspects of open data are:

- **Availability and Access:** Data should be available in their entirety, in a convenient and modifiable form, and should at most hold a reasonable cost required for reproduction.
- **Re-use and Redistribution:** Data must be allowed to be re-used and re-distributed, including merging with other datasets.
- **Universal Participation:** There should be no restriction on who can use, re-use or re-distribute. For example, there shouldn't be a restriction that denies commercial usage.

To find out more about open data we recommend visiting the Open Data Institute (ODI) Website[21] and reading the Open Data Handbook[22] produced by the Open Knowledge Foundation.

4.2.2 Giving your open data the best chance to realise its potential - Mr Chris Gutteridge (University of Southampton)

https://orcid.org/0000-0001-9201-5987

Figure 2: Chris Gutteridge

[20] https://opendefinition.org/od/2.1/en/
[21] https://theodi.org/
[22] https://opendatahandbook.org/guide/en/what-is-open-data/

The full video of Chris's talk can be viewed here: https://youtu.be/3SvkOjEOCgc [9]

Chris Gutteridge is a Systems, Information and Web programmer, part of the IT Innovation team in the School of Electronics and Computer Science at the University of Southampton. He is an advocate for open data, linked data and the open web and was the lead developer of EPrints. He has won several awards including The Times Higher Award (for the Southampton open data service) and the Jason Farradane Award "in recognition of outstanding contribution to the information profession".

Chris was involved in the web and open data from the early days; which meant that he saw problems occur before others were aware of them. Some of his earlier projects in this space involved setting up the University of Southampton open data service[23] and Equipment.data[24], a portal for sharing, harvesting and aggregating data about institutional facilities and equipment across the UK.

Through his extensive work in this area, he was able to gain a deep understanding of the problems of early adopters for open data, and learn from mistakes that had been made in previous projects. A key reoccurring issue of trying to persuade people to make their data open was the reluctance to do so, and the multitude of reasons that were given for that reluctance. This lead to the creation of "Open Data Bingo".

Terrorists will use it	We'll get spam	It's too big	It's not very interesting
Thieves will use it	I don't mind, but someone else might	We will get too many enquiries	Lawyers want a custom License
There's no API	Poor Quality	There's already a project to...	We might want to use it in a paper
It's too complicated	Data Protection	People may misinterpret the data	What if we want to sell it later

Figure 3: Open Data Excuses Bingo: Taken from Chris Gutteridge's Slides, based on the Concerns about opening up data Google Doc

This bingo sheet is a compilation of all of the reasons given by people not to share their data. However, despite all of these reasons, some people are still making their data available, so it logically follows that there are some steps that can be taken to overcome these concerns. Chris worked with Alex Dutton, a programmer from Oxford University, to create a Google Document: Concerns about opening up data[25] that contains a great deal of advice on these issues based on their combined experience. This project led to the creation of a second document: Getting more value from open data publishing[26] about how to get more value out of the data you have already created and shared. The main body of this talk is based on this document.

[23] https://data.southampton.ac.uk/
[24] https://www.jisc.ac.uk/rd/projects/equipment-sharing-made-easy
[25] https://docs.google.com/document/d/1nDtHpnIDTY_G32EMJniXaOGBufjHCCk4VC9WGOf7jK4/edit
[26] https://docs.google.com/document/d/1vd8yOagTPPDcZsHpNMh-bIiKL3tjPDb4xrbXbP1hrdM/edit

Chris presented the 'dataset reuse hygiene factors', which is a list of reasons why people might not want to use your data, and how to potentially mitigate these. A key point about these hygiene factors is that it matters significantly more how bad they are, as opposed to how good they are. Each factor that scores poorly incrementally decreases the likelihood of re-use; however, there is a trade off of effort per quality factor vs the number of users who won't use your data due to this factor, and at some point that will reach an equilibrium. Therefore it is significantly more productive to double your effort in the area you are performing worst at, rather than increasing your effort in an area that you are already performing well in. Chris presented a summary of the main hygiene factors, noting the simplest and cheapest methods of addressing these.

Value of dataset to audience: This really matters if you are considering this prior to data generation; however, once you have already collected the data this is out of your control. If you are being funded to create useful data, then obviously it is imperative that you consider how to create useful data.

Potential audience size: Your data can have a very varying audience size depending on the relevance and topic. For example, data about the opening times of coffee shops in a specific city, would potentially only have the audience of that city who were interested in coffee shops. However, you can use standards which mean that your data could be used with other datasets, thus increasing your audience size that way.

Ease of discovery: You need to make sure people can find your data, as it doesn't matter how good it is, if it can't be found then it won't be used. You can increase your ease of discovery both by getting it listed in catalogues that have good descriptive metadata such that it can be found, but also by word of mouth. It is worth tweeting about it and ensuring that people know it exists and where it is. You can test how findable your data is by searching for it yourself using keywords you think others might use, and seeing if you can actually find it.

Ease of grasping the value of the dataset: You need to describe your data clearly, even if the data itself is well produced and it is easily findable, a poor description may mean that someone finding it cannot see its value. Make sure you have a human readable title and description that details what the data does and why it is valuable. Another way of improving this factor is to produce a blog post about the data, and then cross link the blog post to the metadata and vice versa. This will provide more entry points to the data and provide further clarity.

Ease of exploiting dataset: Your data needs to be easy to use! Even if it does hold value and can be located, if it is difficult to use you might lose your potential audience at this stage. There are three key aspects to this:

- **Publishing:** When you publish your data you should ensure that your data is clean, clear and adheres to standards where possible. Ideally it should be provided in multiple formats and use IDs that link to well understood data catalogues.
- **Documenting:** Show people how to use your data, provide examples for them to work through and if your data depends on specific software or libraries then include references to these to demonstrate how to properly use it and get the most out of it.
- **Communicating:** Enable people to communicate with you about your data! If you put contact details (e.g. your twitter handle or email address) into the documentation then people can get in touch with you if there are issues. It is also useful to have the dataset on GitHub and allow people to raise issues against it. This way it can be iteratively improved if necessary.

Perceived quality and reliability (trust and provenance): Perception is everything, people MUST be able to trust your data. If something goes wrong and stops working, e.g. there are errors in the data, or it becomes unavailable for a period of time, people will stop trusting it. You need to make it clear that it is trustworthy, and that any errors will be updated and fixed in a timely manner. Further, the provenance (origin) of the data should be clear.

Perceived neutrality (AKA "not invented here" syndrome): There can be social factors that leave researchers reluctant to use data from "rival" organisations, or even just organisations that aren't their own. Neutrality needs to be achieved. Firstly, there's no reason you can't share data from other institutions if it is useful! Secondly, you can improve neutrality towards your data if you publish it in a way that isn't overtly linked to your institution.

Conclusions: This isn't an exhaustive list, there are always going to be other factors that you haven't considered that prevent people from using your data. Something really important to consider therefore is, looking at your data and if people aren't using it, have a think about what aspects you aren't touching on. If you aren't addressing one of these factors, work on addressing them, rather than expending additional effort on factors you have already addressed. Further, if you feel you have addressed all of these, then ask others to look at your data as it will always be easier for an external pair of eyes to identify a new hygiene factor that you hadn't considered. The documents referenced in this talk are living documents, and Chris is happy to add to them so please get in touch with him if you wish to discuss this.

Questions following the presentation:

Q: If you were looking for open data sources, are there any repositories that you would recommend for people to use?
A: *Usually Google, or ask peer groups. Asking people in extended communities is usually the best way to find repositories that are relevant to what you are doing. You need to be able to trust the datasets; however, it is important to know who is paying for it, updating it etc. If you are pulling something together for a hobby project it is less of a concern but if you are making a long standing project and you are using an API, it is well worth checking these things to ensure that it will still be working in years to come.*

Q: I think the issue of building up trust is very important, you addressed some of the issues but would you like to say anything more on how to build up trust as a source of data?
A: *A lot of it is about reputation with your organisation, but ultimately people tend to trust things that are funded. Having an end of life strategy for your data also helps to build trust. Typically projects that finish result in their webpages and potentially subsequent datasets slowly degrading over time and eventually disappearing. A strategy we have found is to archive them and make them the responsibility of the library. You should be planning an end of life strategy for your dataset as soon as it is born. It is also advisable to provide a full data dump so that even if APIs or applications that lie on top of that data may not always be able to continue, the data will remain. To make people trust them, show people what your long term commitment is to the data even once your funding finishes, e.g. regular backups and archives that can be publicly accessed.*

Q: One of your open data bingo cards says "its too big" and I looked at the accompanying document and it says "it's not necessarily as big as you think it is" but we had some comments in the previous data event in this series (Data Tips & Tricks) about the size of some of the scientific data that some of our scientific researchers are dealing with, particularly in crystallography, we have a lot of computational chemists in the Network and they produce

large quantities of data and work with big data, so would that still come under "it's not as big as you think it is"?
A: *No, that is definitely as big as you think it is. Sometimes you can't make the data "open open" due to the sheer size. There are other options, such as sending it via a hard drive, or making it available in chunks.*

Q: Over the years I have found neutrality very important in terms of getting broader engagement. I completely agree that toning down branding helps here - this is probably a generational issue as you suggest, but perhaps also a national culture (e.g. we get lots of overseas use of institutional repository content)?
A: *I think a key aspect here is the use of standards. If your data uses standards then there will be less of a concern about using it, as long as the people using it won't get into trouble for using it. As long as your data is from a respectable institution, it is more likely to be used if you are using multiple datasets from different institutions, but if you are the only one from an external institution and it is a lot of effort to use your data e.g. the standards (or lack thereof) in your data aren't compatible with the rest of the data they want to use, then they are unlikely to go to the effort of using it.*

Q: How do institutions balance not fanfaring vs needing to show outcomes/impact (e.g. for REF)?
A: *I think don't fanfare it to the people you want to reuse it. DO fanfare it in the REF and to funders!*

4.3 Linked Data

This section will provide a general introduction to Semantic Web Technologies, followed by a detailed summary of the talk relating to Linked Data.

4.3.1 Introduction to Semantic Web Technologies

The Semantic Web was conceptualized by Sir Tim Berners-Lee, the inventor of the world wide web [10]. The main goal of the Semantic Web was to create machine readable data such that machines could better interpret information [11]. The Semantic Web provides a set of core standards to facilitate the creation of rich machine readable datasets with embedded content and meaning. The semantic toolkit has a range of tools and technologies available, but this introduction will concentrate on the three core technologies at the heart of the Semantic Web: RDF, Ontologies, and SPARQL.

The core data format of the Semantic Web is called Resource Description Framework (RDF) [12]. This uses a graph model to store the data in triples, subject → predicate → object, whereby the predicate defines the relationship between the subject and the object. Triples can be linked together by then making the object of one triple the subject of another, which enables us to break down datasets into this linked data format. The subject, predicate and object are often formed of persistent Uniform Resource Identifiers (URIs), meaning that once a term has been defined with a URI, this URI can be referred to again if you wish to refer to the same specific instance.

RDF enables data to be modelled and stored in this linked graph format, but it alone is not sufficient enough to represent the required domain knowledge that provides the context and meaning behind the data. In order to achieve this we need ontologies. Ontologies are essentially taxonomies / controlled vocabularies that allow us to create formal definitions of the common terms used within a specific domain. They enable us to describe the hierarchy of classes used to define the relationships and restrictions of different concepts within the ontology. By doing this, reusable terms can be built up for use within other systems that employ the same terminology. Ontologies are typically written in the Web Ontology Language (OWL) [13] as it has substantially more expressive capabilities, but simple ontologies can also be written in RDF Schema (RDFS) [14]. By combining linked data and ontologies it is possible to create knowledge graphs, which are essentially graph network structures to describe real world entities and their relationships. These interconnected knowledge graphs of linked entities are powerful structures and enable software agents known as reasoners to navigate graphs and infer implicit info from explicitly defined facts.

Simple Protocol and RDF Query Language (SPARQL) [15] is the Semantic Web Query Language. It is based on SQL and has a similar structure of selecting / inserting / deleting, except in a graph form rather than a tabular database form. SPARQL enables you to search on concepts, so you can add context to your search, and can be used to make complex queries over combined datasets to pull out new, previously unexplored, connections.

4.3.2 Linked Data - Examples and Heuristics - Dr Terhi Nurmikko-Fuller (Australian National University)

https://orcid.org/0000-0002-0688-3006

Figure 4: Terhi Nurmikko-Fuller

The full video of Terhi's talk can be viewed here: https://youtu.be/Q8l8Y-44fqo [16]

Due to the time zone differences Terhi's talk was pre-recorded and played to the audience during the event.

Dr Terhi Nurmikko-Fuller is a Senior lecturer in Digital Humanities at the Australian National University. She holds degrees in Ancient History, Web Science, Cuneiform and Near Eastern Studies and Museum Studies, in addition to a PhD on examining the potential of linked data as the method for publishing Assyriological data. Terhi's research involves interdisciplinary experimentation into the ways digital technologies can be used in the Humanities, Arts, and Social Sciences. She is a member of several Australian Government groups including the Linked Data Working Group, and is on the Steering Committee of Linked Pasts, an international consortium for Linked Data in the Humanities.

Terhi's talk began by laying out the challenges of converting messy, incomplete and highly heterogeneous data into machine parsable data, and why she believes that linked data technologies are highly suited to this task.

Data can be stored in a multitude of different types, particularly if you are working with historical data. It could exist in RDF, or at least in a structured format such as a relational database. It might be unstructured but still electronic such as tabular/CSV data, or it might even be handwritten notes, annotations of a book, or analogue data. This in itself is a huge challenge with respect to converting data into a machine readable format. Additionally, datasets are rarely simple or consistent, there can be subtle nuances and challenging ambiguities in the data that need to be well understood before a conversion can occur; further, datasets are often incomplete and it is not feasible just to "fill in" missing data. These nuances and ambiguities need to be retained within the data irrespective of how it is processed, and this is where linked data process can be extremely useful by capturing the tacit knowledge from the minds of domain experts.

Terhi's talk ran through three case studies, each describing a different linked data project she has worked on, explaining the different approaches used and noting the different practical considerations that she has learned.

Case Study 1: ElEPHãT: Early English Print in the HathiTrust a Linked Semantic Worksets Prototype [17–20]

Description: This project looked to combine information from a set of independent sources into a single richer dataset that could be analysed and accessed from a single point of entry. These sources came from Phase 1 of the Early English Books Online Text Creations Partnership (EBBO-TCP)[27] and a subset of the Hathi Digital Library[28]; both datasets contained publications from the late 1400's to 1700's. The key objective of this project was to generate RDF to represent the metadata extracted from EBBO-TCP dataset and combine it with RDF produced by the Hathi trust to enrich the metadata layer. This was achieved by using appropriate ontologies to markup the data, with a prototype user interface that enabled scholars to investigate the linked dataset.

Workflow: This project consisted of several separate steps.

- Consultation workshops to identify the research questions that needed to be solved.
- Technical analysis of metadata requirements, and ontology selection. This was a combination of re-use of existing ontologies where appropriate: (Bibliographic Ontologies: MODS/MADS RDF[29] and BIBFrame[30], the Provenance Ontology PROV-O[31] the Research Object Ontology RO[32]) and some custom classes and properties that were generated as part of the project.
- Selection of tools to generate RDF: Predominantly used a tool called Web Karma[33] which required a great deal of initial work but produces very high quality RDF.
- Storing RDF: Used the Virtuosa triple store, and populated it with the RDF produced by EBBO-TCP. To give an idea of scale, this dataset comprised 1137502 triples.
- Implementation of the system architecture and prototyping the user interface.

Lessons Learned:

- **Data can be staggeringly diverse:** The data was based on original historical documents, including texts, volumes and pamphlets written between the late 1400's-1700's. These documents were produced at a time where metadata standards had not been standardised and as such there was a staggering diversity of expression. In just under 25,000 records, there were 22 distinct types of information in the headers alone, and 1510 distinct expressions of publication date that were parsed into 54 different types.
- **Institutional policies can impact Linked Data projects:** Institutional policies dictate that only EBBO data is available to the public, which means that even RDF produced from Hathi trust metadata cannot be made public, it can feature in the analysis and will affect the search results that are displayed in the ElEPHãT project user interface, but the specific underlying material is not available, demonstrating the affect that institutional policies can have on these types of project.

[27] https://quod.lib.umich.edu/e/eebogroup/
[28] https://www.hathitrust.org/
[29] http://www.loc.gov/standards/mods/modsrdf-primer.html
[30] https://www.loc.gov/bibframe/
[31] https://www.w3.org/TR/prov-o/
[32] https://wf4ever.github.io/ro/2016-01-28/
[33] https://usc-isi-i2.github.io/karma/

Case Study 2: In Concert: Towards a Collaborative Digital Archive of Musical Ephemera: [21, 22]

Description: This project involved bringing together diverse datasets in the area of digital musicology: The datasets involved were: i) Calendar of London Concerts 1750-1800; ii) 19th Century London Concert Life (1815-1895) and iii) OCR data of the British Musical Bibliography. This project demonstrated first hand the challenges that can be presented by working with legacy datasets that don't necessarily facilitate linking and referencing to from external sources.

Workflow: There were several important design decisions made as part of this project which influenced the workflow. These were made to reflect the requirements of the project and due to timing/scoping decisions:

- **Iterative design of the data:** This project used an iterative design process whereby the datasets and ontologies were being continuously updated. This meant that as further opportunities for semantic enrichment were identified throughout this process they could be incorporated.
- **Bottom up approach:** The new RDF data created was extensively based on the existing data structures.
- **Specify rather than transform:** This means that the data would be kept as close as possible to its original form, with additional specifications added only to capture and reveal semantics. In the case of calendar of london concerts, the classes of the underlying ontological structure were directly aligned with information types of the original data. Classes of the ontology contain, as their instance data, the content of a particular column from the original table. This means the workflows from the original processed data should be reproducible.
- **Selection of tools to generate RDF:** Similarly to the previous project, Web Karma was used for transforming the tabular data and producing the ontological mappings. For the relational database data, the ontological mapping was performed using a tool called D2R.

Lessons Learned:

- **Initial time investment = more high quality triples:** Workflows that require a greater investment of user time initially produce a higher quality set of triples. Even if you need to use an automated system there will be a requirement for some user input at some point in the transformation process.
- **Find the tools that suit you:** Try and find tools that support the prior knowledge of the developer/user/RDF producer and that rely on existing systems and known existing levels of competence.

Case Study 3: Jazz Cats (Jazz Collection of Aggregated Triples) [22–25]

Description: This project involved bringing together three separate datasets related to jazz music, one in a tabular/CSV form, one in a relational database, and one dataset that was already in RDF. This project was worked on after the previous two projects, and as such used the lessons learned and tools identified from those.

Workflow: There were several important design decisions made as part of this project which influenced the workflow. These were made to reflect the requirements of the project and due to timing/scoping decisions:

- **Ontological Mapping:** Identified re-use of classes and properties from existing ontologies, and enriched them with bespoke project specific classes and properties (which were later published as part of the Jazz Cats Project Ontology on GitHub).
- **Project specific URIs:** Rather than simply utilise URIs from external authority files such as VIAF or wikimedia, we chose to implement our own project specific URIs and align those with the equivalence from these aforementioned external authority files.
- **Selection of tools to generate RDF:** Similarly to the previous project, Web Karma was used for transforming the tabular data and producing the ontological mappings. For the relational database data, the ontological mapping was performed using a tool called D2R.
- **RDF Storage:** The pre-produced RDF was ingested into the triple store. This was to ensure that if the RDF dataset was moved, or was periodically unavailable, then the project could still continue to function.

Lessons Learned:

- **Embrace your ambiguities:** Don't try and remove ambiguities or uncertainty, choose to capture it. Hence, implementing project specific URIs to capture this aspect of the data.

Conclusions:

- **It's not just about the technology:** There are political, institutional, legal and socio-cultural considerations to every linked data project. It is of fundamental importance that these are taken into consideration from the onset of the project and not simply added as a box ticking exercise at the end of a completed workflow.
- **One tool doesn't fit all:** The best tools and systems to recommend are not simple and universal, choose those that most clearly and effectively align with the existing skills, functional requirements and necessities of a particular project, researcher or workflow. The tool that is best for you might not be the best tool for someone else, it really depends on your skillsets and specific aims. There is also real value in testing out different options and evaluating choices to settle on a tool that works the best for you.
- **Sometimes practicality needs take priority over philosophical ideology:** In the Jazz Cats project, ingesting the triples from the existing RDF dataset into the triple store was implemented as a solution not because it was necessarily the most theoretically or ideologically suitable solution, but because it provided a degree of safety to the project.

5 Panel Session

Following on from the talks a panel was convened, the panel members comprising of the two speakers that were present at the event and Dr Samantha Kanza stepping in for Dr Terhi Nurmikko-Fuller due to time zone differences.

The panel members were:

- Dr Martin-Immanuel Bittner - Arctoris
- Mr Chris Gutteridge - University of Southampton
- Dr Samantha Kanza - University of Southampton

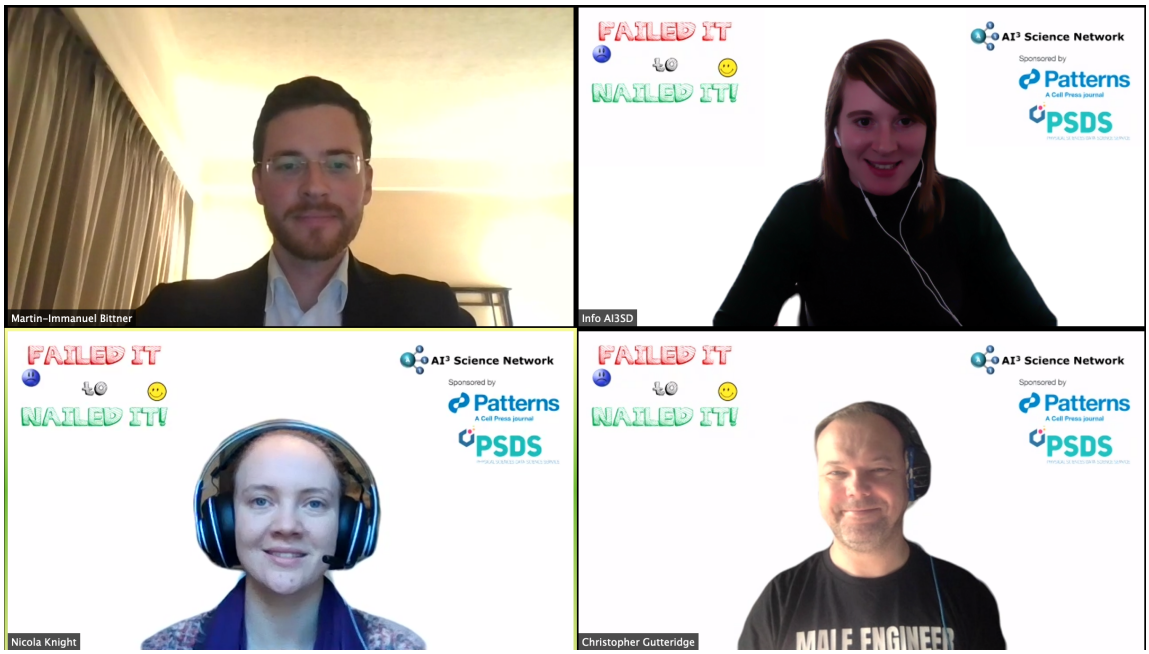

Figure 5: The Panel members

The panel was chaired by Dr Nicola Knight. The questions asked to the panel members were a mixture of pre-prepared questions and questions asked by members of the audience. The report content below outlines the questions asked and summarises the discussion and responses that followed these questions.

Q1 - If you could make sure everyone knew a key thing about meta/linked/open data - what would it be?

- Make sure that people know that metadata exists! Just having your data alone is not enough.
- You need to think things like metadata through before you start your work, trying to add them as an afterthought is very difficult.
- Unless you really know the data you are working with, it is tricky to prove a negative. You only know what the system believes is true.
- Assumptions become magnified when you join datasets. When using other people's data their assumptions may be very different to yours.
- When creating an ontology, you need to have a really clear idea on the use case for the ontology. You need to work backward from your end goal to create it.
- You need to think about your data conversion (e.g. into linked data) as well before you create your ontology.

- Datasets with missing data can be very difficult to deal with! Especially when converting into linked data. You should consider a strategy for how you will handle missing data.

Q2 - What excites you the most about the future of meta/linked/open data?

- The application of AI to semantically enriched data. When we have well-organised, machine readable, semantically rich data we will be able to do so much more with machine learning than we are currently capable of.
- AI will only work well if you feed it with robust, structured and valid data. Application of the FAIR data principles can enable us to have data of high enough quality to make large advances with machine learning technologies.
- The current systems aren't really anywhere near as good as they could be. If we can capture data at source it would be significantly better and produce much more valuable research, so there is significant room here for known improvement.
- Semantic annotation formats (including JSON-LD[34]), there is a lot of power in annotating data semantically without having to convert your entire dataset into linked data.
- Integrating data with published papers so that it is easily accessible and able to view in different contexts. This would enable easier initial examination of data when thoughts are sparked by the curiosity of researchers.

Q3 - Do you have any advice for people who are complete beginners with data standards?

- Don't be afraid to jump in and play around with datasets and coding. Get dummy datasets[35] and try out things with them. You can always try creating ontologies for things you are familiar with outside of your work.
- Places like DBpedia[36] are useful to explore and get your head around datasets and queries. Learn to read before you can write!
- Make sure you have an understanding of the key concepts in the Wilkinson Paper [3] and then you can branch out from there. But make sure you are always willing to try things out.
- You won't learn by just reading about things, you actually need to try things out and look at examples. Good places to learn: Pizza Ontology Tutorial[37], data.southampton[38]
- You will benefit from querying already built ontologies and datasets when you're learning. When you are starting from scratch it may be your query or your ontology that is wrong and you won't know where the error lies.

Q4 - What is a frequent mistake that you see when trying to implement these data standards?

- Implementing a system because it is the current 'big thing' e.g. doing RDF, but only implementing it at the end of a project and not really understanding what it was for. This results in datasets that people haven't tested and will likely never be used. Have an idea of how your data would be used and aim to have at least one person outside of your team who will use your data! (if you intend for it to be reused)
- Simply not starting to use the standards at all. Yes things may go wrong and need to be fixed but you need to start somewhere. You shouldn't just pass it down the line, it's too important. Ask for help if you need it from other people who have done it before so you can implement it properly.

[34] https://json-ld.org/
[35] http://data.totl.net/
[36] https://wiki.dbpedia.org/
[37] http://mowl-power.cs.man.ac.uk/protegeowltutorial/resources/ProtegeOWLTutorialP4_v1_3.pdf
[38] https://data.southampton.ac.uk/

- You can really overcomplicate your ontologies by putting in unnecessary functionality. Think clearly about what the use cases are.
- It is good to look at existing ontologies and use them where it is possible, but if you are looking to create a new ontology really think carefully about how it will be used and by who. Don't just make an ontology for the sake of it. But equally don't try to crowbar your data into something which isn't suitable.
- Think about what you need your data to look like at the end before you design the system. Otherwise you will likely need to rework a lot of things.

Q5 - Are there any other resources that you would recommend to help people learn more?

- Prefix.cc[39] - a namespace lookup service for RDF where you can search commonly used abbreviations and popularly used namespaces. You should probably familiarise yourself with the top items on this list to see what is out there.
- FAIR guiding principles in the Wilkinson paper [3]
- Pistoia Alliance documentation - particularly within the biomedical space.
- Property matrix plugin in Protege[40] (ontology IDE) - allows you to open up and view your relationships and you can drag and drop to create links.
- R2RML library[41] - R2RML [26] is a mapping language, you can link files like CSVs to linked data. This library allows use of javascript functions to improve your mapping.

Q6 - How can linked data be applied to scientific and more numerical data?
There are many applications within the scientific domain and within scientific research.

- Semantic annotation and markup of electronic lab notebooks (ELNs) and experimental records to provide much richer metadata around the research. E.g. linking to pieces of equipment, chemicals, who wrote it, related projects etc.
- SEED Project[42] - Semantic Enrichment of ELN data
- There are many other areas using Semantic Web within scientific research. In particular this paper about Semantic Web tools [27] shows examples of Semantic Web uses in drug discovery. It is beginning to be used within machine learning projects and research to explore undiscovered links.

Q7 - How do we try to get more people on board with implementing standards in their data?

- We need to further demonstrate the benefits! It can be a lot of work to implement, so people need to be able to weigh up the time cost/benefits.
- Tools and technologies to make it easier for people who don't have a strong background in areas like computer science or coding.
- Get buy in from large bodies, such as funding bodies. Both from the education point of view, but also potentially part of a funding requirement in the future.
- Grass roots initiatives like the networks, engaging with these can really help to get people on board and pass on messages to future researchers.
- Standardisation within a community domain, e.g. subject specific metadata.

[39] http://prefix.cc/
[40] https://protegewiki.stanford.edu/wiki/Matrix
[41] https://github.com/chrdebru/r2rml
[42] https://www.pistoiaalliance.org/projects/current-projects/seed-project/

6 Participants

Participants attended from a wide variety of backgrounds due to the online nature of the event. While the majority of attendees were from the UK, there were a number of registrations from other countries.

7 Conclusions

This event captured a wealth of information about different types of data standards to produce a set of lessons learned and best practice recommendations based on significant experience. A common recommendation from experts in all three standards is that implementing these data standards is something that should be thought through before commencing a project. It is obvious that the amount of data that is currently being generated and will be generated in the future is growing immeasurably, and we need to think carefully about the data we generate before we generate it.

Significant time is wasted in cleaning, adapting and converting data to adhere to standards, and to create useful metadata for it; a situation that can often occur because people are unwilling to open up their data if it is not in a state to share, but equally didn't create it as such in the first place. This is not something that should be fixed as an afterthought, we need to consider what data standards we want our data to adhere to **before** we start generating it. However, this does not mean that if this hasn't been done that a project is unsalvageable, there is always room for improvement and we strongly encourage all researchers to revisit their data and see if there are improvements that can be made to how they are conforming to data standards.

Further, there will always be concerns, or blockers, or reasons to not standardise or open up your data, but we should be focussing on mitigating these concerns and figuring out how to work around them, rather than using them as excuses not to standardise or open up our data. Finally, this is not just a technical or data-based endeavour, it is also a human one. All three talks demonstrated the socio-cultural, institutional and logistical issues that can be involved with implementing data standards. These need to be taken into account just as much as the technological and data aspects. Creating standardised open linked datasets with good metadata can only be achieved through co-operation and collaboration on a mass scale.

8 Related Events

Details of the other events in the Failed it to Nailed it data seminar series can be found here:
https://www.ai3sd.org/ai3sd-online-seminar-series/data-seminar-series-2020/
Each of these events will have video recordings and a report associated with it.

Details of other AI3SD events and events of interest can be found on the AI3SD website events page:
https://www.ai3sd.org/ai3sd-events/
https://www.ai3sd.org/events/events-of-interest/

References

[1] Digital Curation Centre: What are Metadata Standards [Internet]; 2007. [cited 2020 Nov 19]. Available from: https://www.dcc.ac.uk/guidance/briefing-papers/standards-watch-papers/what-are-metadata-standards.

[2] Day M. Metadata [Internet]. Digital Curation Centre; 2005. [cited 2020 Nov 19]. Available from: https://www.dcc.ac.uk/resources/curation-reference-manual/completed-chapters/metadata.

[3] 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(1):160018. Available from: https://doi.org/10.1038/sdata.2016.18.

[4] GO FAIR: FAIR Principles [Internet]; 2020. [cited 2020 Nov 19]. Available from: https://www.go-fair.org/fair-principles/.

[5] Bittner MI. AI3SD Video: Data Generation, Data Standards and Metadata Capture in Drug Discovery;. AI3SD, PSDS & Patterns Failed it to Nailed it: Getting Data Sharing Right Seminar Series 2020. 2020. Available from: http://dx.doi.org/10.5258/SOTON/P0064.

[6] Scannell JW, Blanckley A, Boldon H, Warrington B. Diagnosing the decline in pharmaceutical R&D efficiency. Nature reviews Drug discovery. 2012;11(3):191. Available from: https://doi.org/10.1038/nrd3681.

[7] Begley CG, Ellis LM. Raise standards for preclinical cancer research. Nature. 2012 Mar;483(7391):531–533. Available from: https://doi.org/10.1038/483531a.

[8] Open Data Institute (ODI). What is 'open data' and why should we care? [Internet]; 2020. [cited 2020 Nov 19]. Available from: https://theodi.org/article/what-is-open-data-and-why-should-we-care/.

[9] Gutteridge C. AI3SD Video: Giving your Open Data the best chance to realise its potential;. AI3SD, PSDS & Patterns Failed it to Nailed it: Getting Data Sharing Right Seminar Series 2020. 2020. Available from: http://dx.doi.org/10.5258/SOTON/P0065.

[10] Berners-Lee T, Hendler J, Lassila O. The Semantic Web. Scientific American. 2001;284(5):34–43. Available from: https://www.jstor.org/stable/26059207.

[11] Shadbolt N, Berners-Lee T, Hall W. The Semantic Web Revisited. IEEE intelligent systems. 2006;21(3):96–101. Available from: https://doi.org/10.1109/MIS.2006.62.

[12] W3C. Resource Description Framework (RDF) [Internet]. World Wide Web Consortium; 2014. [cited 2020 Nov 19]. Available from: https://www.w3.org/TR/rdf11-concepts/.

[13] W3C. OWL 2 Web Ontology Language Document Overview (Second Edition) [Internet]. World Wide Web Consortium; 2012. [cited 2020 Nov 19]. Available from: https://www.w3.org/TR/owl2-overview/.

[14] W3C. RDF Schema 1.1 [Internet]. World Wide Web Consortium; 2014. [cited 2020 Nov 19]. Available from: https://www.w3.org/TR/rdf-schema/.

[15] W3C. SPARQL 1.1 Query Language [Internet]. World Wide Web Consortium; 2013. [cited 2020 Nov 19]. Available from: https://www.w3.org/TR/sparql11-query/.

[16] Nurmikko-Fuller T. AI3SD Video: Linked Data – Examples and Heuristics;. AI3SD, PSDS & Patterns Failed it to Nailed it: Getting Data Sharing Right Seminar Series 2020. 2020. Available from: http://dx.doi.org/10.5258/SOTON/P0066.

[17] Jett J, Nurmikko-Fuller T, Cole TW, Page KR, Downie JS. Enhancing Scholarly Use of Digital Libraries: A Comparative Survey and Review of Bibliographic Metadata Ontologies. In: Proceedings of the 16th ACM/IEEE-CS on Joint Conference on Digital Libraries. JCDL '16. New York, NY, USA: Association for Computing Machinery; 2016. p. 35–44. Available from: https://doi.org/10.1145/2910896.2910903.

[18] Nurmikko-Fuller T, Page KR, Willcox P, Jett J, Maden C, Cole T, et al. Building Complex Research Collections in Digital Libraries: A Survey of Ontology Implications. In: Proceedings of the 15th ACM/IEEE-CS Joint Conference on Digital Libraries. JCDL '15. New York, NY, USA: Association for Computing Machinery; 2015. p. 169–172. Available from: https://doi.org/10.1145/2756406.2756944.

[19] Khan NJ, Nurmikko-Fuller T, Page KR. BABY ElEPHãT: building an analytical bibliography for a prosopography in early English imprint data; 2016. Available from: https://doi.org/10.9776/16588.

[20] Page K, Willcox P. ElEPHãT: Early English Print in the HathiTrust, a Linked Semantic Worksets Prototype; 2015. Available from: http://hdl.handle.net/2142/79017.

[21] Nurmikko-Fuller T, Dix A, Weigl DM, Page KR. In Collaboration with In Concert: Reflecting a Digital Library as Linked Data for Performance Ephemera. In: Proceedings of the 3rd International workshop on Digital Libraries for Musicology. DLfM 2016. New York, NY, USA: Association for Computing Machinery; 2016. p. 17–24. Available from: https://doi.org/10.1145/2970044.2970049.

[22] Nurmikko-Fuller T, Bangert D, Dix A, Weigl D, Page K. Building Prototypes Aggregating Musicological Datasets on the Semantic Web. Bibliothek Forschung und Praxis. 2018 Jun;42(2):206–221. Publisher: De Gruyter Section: Bibliothek Forschung und Praxis. Available from: https://doi.org/10.1515/bfp-2018-0025.

[23] Nurmikko-Fuller T, Bangert D, Abdul-Rahman A, et al. All the Things You Are: Accessing An Enriched Musicological Prosopography Through JazzCats; 2017. Available from: https://openresearch-repository.anu.edu.au/bitstream/1885/204569/2/01_Nurmikko-Fuller_All_the_Things_You_Are%253A_2017.pdf.

[24] Nurmikko-Fuller T, Bangert D, Hao Y, Downie JS. Swinging Triples: Bridging Jazz Performance Datasets using Linked Data. In: Proceedings of the 1st International Workshop on Semantic Applications for Audio and Music. SAAM '18. New York, NY, USA: Association for Computing Machinery; 2018. p. 42–45. Available from: https://doi.org/10.1145/3243907.3243914.

[25] Bangert D, Nurmikko-Fuller T, Downie JS, Hao Y. Jazzcats: navigating an RDF triplestore of integrated performance metadata. In: Proceedings of the 5th International Conference on Digital Libraries for Musicology. DLfM '18. New York, NY, USA: Association for Computing Machinery; 2018. p. 74–77. Available from: https://doi.org/10.1145/3273024.3273031.

[26] W3C. R2RML: RDB to RDF Mapping Language [Internet]. World Wide Web Consortium; 2012. [cited 2020 Nov 19]. Available from: https://www.w3.org/TR/r2rml/.

[27] Kanza S, Frey JG. A new wave of innovation in Semantic web tools for drug discovery. Expert Opinion on Drug Discovery. 2019;14(5):433–444. Available from: https://doi.org/10.1080/17460441.2019.1586880.