Semantic Web Technologies for Data Curation and Provenance

Clifford Brown,

Data Science, National Physical Laboratory, UK

Abstract

The Reproducibility issue even if not a crisis, is still a major problem in the world of science and engineering. Within metrology, making measurements at the limits that science allows for, inevitably, factors not originally considered relevant can be very relevant. Who did the measurement? How exactly did they do it? Was a mistake made? Was the equipment working correctly? All these factors can influence the outputs from a measurement process. In this work we investigate the use of Semantic Web technologies as a strategic basis on which to capture provenance meta-data and the data curation processes that will lead to a better understanding of issues affecting reproducibility.

Introduction

In the influential paper by Baker[1] presented on the subject of Reproducibility, over 1500 research scientists were asked whether they thought there was a Crisis in Reproducibility; 52% said yes absolutely, 38% agreed slightly (so that's 90% so far), 3% said No and 7% said they didn't know. So only 3% are totally confident in the data and conclusions presented in scientific journals.If that was not worrying enough, the second part of the report was probably was, more than 60% of all researchers had experienced problems in reproducing other people's results; interestingly a significant proportionadmitted having problems reproducing their OWN results.

What is causing this problem with Reproducibility?In part, failure in traceability and insufficient knowledge of uncertainty in measurement will be the main causes; but also the failure to capture and have knowledge of the full process of collecting and processing raw data will also besignificantfactors. The failure to collect sufficient data curation and provenance information means there is a lack of traceability, or audit information sufficient to reproduce the relevant science. This failure in reproducibility is a serious problem for Science and Engineering.

Like similar NMI organisations NPL generates large volumes of scientific data. These data are either directly related to the science under study (e.g. variables measured for a given scientific phenomenon) or are used to describe the experiment process (meta data).

For example, in a laboratory that studies the stress-strain relationship for a particular material, where the physical geometry, type of material and stress and strain values are direct values that are relevant to the material under study. Additionally, the temperature,

pressure and humidity of the lab environment may also be important factors that affect the data collected, and consequently the physics of stress-strain relationship for the material.

Current data curation processes at NPLlike at most scientific and engineering organisations vary considerably across the laboratory. Some laboratories such as Mass Spectrometry deploy automated data collection, analysis and storage, while others collect data in notebooks and transcribe it using various vehicles such as Excel, Matlab, LabView, Python, C# etc.

The processed data are then typically used in an output product e.g. calibration certificate, scientific paper, software, measurement device etc. Beyond this point, in general, no further use is made of the data. Two factors contribute to this: 1) the data cannot easily be discovered (no recovery), and 2) the data generation process is not documented (no reproducibility). Even if the data are found, there are major limitations to the ability to reuse this data due to lack of the data curation process. In addition to the loss of data curation information, the data provenance is also in general, lost.

Issues of Reproducibility

As the UK NMI we are champions of reproducibility, that is our fundamental reason for existing; but precisely because NPL works at the limits of measurement that science allows, we inevitably come across affects that other scientists probably would not see.So by definition of the work we do; we, like all NMIs, are also subject to reproducibility issues.

By collecting meta-data that describes, in detail, parameters for the science, we will have a better opportunity to understand why our results are as they are, understand the reproducibility issues, and potentially discover new science.

The core digital strategy behind dealing with reproducibility is what is known as the FAIR principles [2].FAIR stands for: Findable, Accessible, Interoperable andReusable. Considering these in turn.

To make our data Findable, we need Meta-Data, so for example, the name of a table and the column names are all meta-data.If you want to find a file with some data in it, you might it stored in a hierarchy of folders;each of the layers of folder-names you navigate through to get to the file are also effectively meta-data. Although this is a start, this meta-data is still insufficient; in future we need to be more descriptive about our data – and we are going to have to develop more effective meta-data to make our science knowledge more Findable. Information such as: how the data is collected, analysed, processed and stored; who did the work, when did they do it, what were the precise environmental conditions when the data was collected are typical examples of meta-data.

To make our data Accessible, in addition to being Findable, data needs to be available for use. Is it an accessible format that can immediately be used? Ideally in a machine-readable format to allow it to be downloaded and clearly understood as to what its contents mean. Data held in open standard formats such as XML or JSON provide this sort of functionality. Providing data

in proprietary formats from databases and spreadsheets will not effectively provide for Accessibility.

To make data Interoperable, in addition to being Findable and Accessible, fundamentally it needs to be understandable. In addition to being Accessible i.e. available and readable; data needs to be understood. Even in the realms of science and engineering where strict definitions of various fundamental physical quantities are maintained via a range standards mechanisms e.g. BIPM, ISO, etc. the possibility for misunderstanding, or at least misinterpreting data is still possible. What a single piece of digital data actually means even to the person who measured it is not always clear. Certainly, a short name typically associated with a column heading in a table is unlikely to be totally unambiguous in its meaning. This is where Ontologies and associated Graph database solutions have been proposed as a way forward to making data Interoperable (see Storing Scientific Knowledge and making it FAIR).

To make our data Reusable, in addition to being Findable, Accessible and Interoperable, it needs to be made available according to an acceptable licencing agreement.

Storing Scientific Knowledge and making it FAIR

How we store scientific knowledgeis crucially important to any Digital Strategy we develop that is to be FAIR compliant. Standard databases are great when you know about every piece of information you want to store, and where the definition of the data to store is not going to change. So great for information for example about contact listsor customer relation management.But, they are not good for storing new types of knowledge. Traditional database designs are based on a database schema which defines what information can be stored in the database. The schema for a particular phase of development of a product is immutable, it cannot change. It is possible for the database schema to evolve in line with new requirements for the system, but this is in general a relatively expensive operation to perform and usually requires that the application that provides access to the database to also be updated.In reality, the option to use standard database technology (SQL based) to store evolving scientific knowledge is unrealistic.

Unlike traditional database schemas that are ‘fixed’ between phases of development, linked data graphs stored as Subject-Predicate-Object triples are ‘unfixed’. This provides a storage system that can evolve dynamically and reflect changes in data structures without the need to re-define the database schema.

However, the applications and human actors that interact with the underlying graph storage system need to be intelligent enough to ‘understand’ what is meant by the data within the system – the **Semantics**. To do this requires an ‘**Ontology**’ – a dictionary of concepts for an area of knowledge and the relations between them.

In short, what does this semantic technology stack give to the metrology industry? A searchable, flexible storage environment that can grow as new knowledge is gained without the need to re-design the system from scratch every time the data structure changes.

It gives us the provision of a data curation method that is capable of capturing meta-data and data provenance information (e.g. using the WC3 Provenance Ontology [3]) in addition to the core data for a process using QUDT [2]

Conclusions

In order todeal with reproducibility issues for science and engineering knowledge, new methods for Finding, Accessing, Interacting and Reusing (FAIR principles) data need to be developed. Standard database technology will not provide the flexible storage required for ever changing science and engineering knowledge. Semantic web technologies, specifically graph storage combined with ontologies can provide this functionality.

References

1. Baker, M.; “1,500 scientists lift the lid on reproducibility”Nature 533, 452–454 (26 May 2016) doi:10.1038/533452a
2. FAIR: Wilkinson,M,D et al. “The FAIR Guiding Principles for scientific data management and stewardship” Sci.Data 3:160018 doi:10.1038 /sdata.2016.18 (2016)
3. PROV – O: The PROV Ontology http://www.w3.org/TR/prov-o/
4. QUDT: http://www.qudt.org/