Empowering Information Systems and Fostering Metadata Driven Data Management

Pascal Heus, December 2020
pascal.heus@gmail.com | https://www.linkedin.com/in/pascal/

Introduction

For over 20 years, I've had the privilege to collaborate with national statistical agencies, data archives, international organizations, research centers, users, and other groups around the globe. My initial background and passion is in information technology, but I became over time an expert in data management, in particular data production, publication, sharing, quality, privacy, and of more importantly metadata. Most of all, I came to understand the importance of and need for data for the greater good of our planet, societies, and individuals, as a fundamental instrument to drive research and innovation, support evidence based decision making, assess impact of policies and actions on the ground, and measure the health of our nations.

Below brief thoughts and recommendations on three intertwined topics that I feel are fundamental to the modernization and future of data research infrastructure and practices.

The incredibly fast pace at which technology has evolved in the past 30 years has had a tremendous impact on the data world. Many organizations and statistical systems have struggled to adjust and keep pace, particularly in the public sector which by nature is slow to adjust to change. This will likely become easier In the coming decade, as we pass the management baton to the next generation of data scientists and information technologists, born with a natural affinity with our new environment, and less constrained by the fear of the unknown. Our current role and responsibility are to support and facilitate this transition.

Empowering Information Systems

A crucial aspect of modernization of DRI is to improve access to data for computer applications. While one might think this is already the case, it is far from the truth. Data is traditionally stored in either proprietary formats (e.g. databases, SAS/Stata/SPSS, Excel, and the likes) or in ASCII text files. The firsts require specialized and often paid licensed software to access, while the latter is openly readable, but carries very limited information about the data. Both lock away web developers whose environment and technology are not well equipped to read either format. This situation results in significant data / code wrangling and ineffectiveness.

Software is no longer just a handy tool for producers, researchers and users. The emergence of advanced algorithms, machine learning, and artificial intelligence has raised the role of computer applications from utility to intelligent agents and companions, that can take over several tasks, and commonly outperform their human counterparts. By transferring the burden of repetitive tasks, discovery, and analysis to information systems, we can foster automation, greatly reduce data and code wrangling, improve overall data quality (particularly accessibility and timeliness), and overall strengthen production or research capacity and effectiveness.

Addressing this issue is relatively simple: deliver data as a service over industry standard based application program interfaces (APIs). Service oriented architectures (SOA) have been proven effective in many other sectors, and instrumental to the success of the Internet and social networks. Hand-in-hand with this is the use of metadata and the adoption of standards, which is further discussed below. Note that this does not take away any of the traditional delivery mechanisms, and on the contrary facilitates it, as a web service can always provide data for retrieval in many formats, for interactive or offline reuse.

Unifying data and metadata

Working with data requires comprehensive knowledge about its meaning, structure, characteristics, provenance, and many other aspects. Unfortunately, such information is rarely available alongside the data, if at all. Further, when present, it is commonly in human friendly formats, such as unstructured documents (if not just in people's heads), putting it outside the reach of information systems.

One reason behind this situation is that data management tools are commonly metadata poor and do not provide functionalities to store such information. Other barriers include a lack of good documentation practices or simply limited budget allocated to such tasks.

This is where metadata comes into play. In this context, I refer to metadata as human knowledge and unstructured documentation about the data converted to machine friendly formats (e.g. JSON, XML, RDF), so it can be understood and processed by information systems.

Metadata enables software and tools to act on the data, and is essential to empower information systems. As emphasized above, there is no loss in doing this, the documentation can be redelivered to users in their favorite formats, so..

While this is a widely known and talked about issue, progress has been slow to materialize. Increased efforts are necessary, with a focus at the data collection and production levels. Too often this task is pushed downstream to data archives or even researchers. Early capture ensures comprehensive documentation and benefits producers as much as end-users.

In the short to medium term, we need to:

- Invest in metadata capture and production. This is often left out of the data production project's plan or budget.
- Support the implementation of metadata management tools. While standards and best practices have emerged, the lack of tools is a significant barrier to adoption.

- Foster a (meta)data culture (further discussed below)

In the longer term, we should support R&D towards the creation of new (meta)data collection, production, publication, and analysis tools and platforms capable of storing and managing both concurrently. This will likely require new database technology and applications, with the latter potentially being integrated into next-generation platforms such as Python and R. API-driven services likewise will play a major role in the area.

Fostering a (meta)data management culture

Data and metadata should not be perceived as two different things. One should not live without the others, be captured at different points in time, or stored and published apart. Unfortunately this is not the way things have been done in the past, and many agencies and individuals seem to be strongly entrenched in traditional practices.

While there is a broad recognition that metadata is needed, an adoption struggle seems widely present. Leveraging metadata requires not only the right tools, but a change of mindset. The resistance generally seems to be coupled with both a lack of familiarity and understanding of metadata-driven data management and a natural resistance to change.

Several actions can be taken to alleviate these issues:

- Emphasizing the benefits of metadata, preferably by providing practical examples, engaging in pilot projects, and sharing success stories.
- Establishing non-intrusive integration strategies. It is important to acknowledge that day to day operations need to continue, and therefore changes must be introduced over time, not overnight. The fear factor is greatly reduced once individuals understand the new way to operate will not impair their ability to do their job, but rather make it easier.
- Documenting cost-benefit analysis. This can be done from a return on investment perspective, both at the organizational and global levels. A small investment in metadata upstream can save hours of work for hundreds of users and researchers downstream. It is also important to look at the current situation and assess the costs of not leveraging metadata.
- Taking an incremental and approach. Do not try to do it all at once. Differentiate between future and past data production. One can start right away with new data, and fill historical metadata over time.
- Training individuals and organizations on metadata-driven data management.

Conclusions

The above are just three of the many challenges surrounding DRI, such as linked data, social data, cloud technology, safe open data, privacy and disclosure control, machine learning, sensor data, geospatial data. Modernising and strengthening infrastructure by empowering information systems is however fundamental and supports progress across the board. This requires software to have access to the same knowledge as their human counterparts through service oriented architectures and comprehensive metadata. Improving the situation upstream has high impact downstream by reducing burden and wrangling for many users.

Technology, standards, tools, best practices, and experts are available today to make this happen. Managing change is typically the biggest challenge and barrier. Approaching this in an incremental, AGILE, non-intrusive manner is desired as day to day operations need to continue.

There are naturally several ways NDRIO can help, particularly in regards to fostering awareness, educating, and supporting pilot and R&D projects. Close collaboration and dialog with other national initiatives, such as Portage, technology savvy and metadata aware communities, such as the Research Data Alliance, CODATA, national entities like Statistics Canada, CIHI, and academic partners is essential to success.

The above have been central to my work and research, reflected through the projects I have been involved in, or our recently released Rich Data Services platform (https://www.richdataservices.com). I hope this helps and inspires others to engage in the modernization of DRI, and welcome comments, feedback, suggestions, or opportunities to collaborate..

About the author

Pascal is an IT specialist and data scientist with over 30 years of experience. An expert in socio-economic, health, education, official statistics, and research data management, he leads major research and development projects, and collaborates with agencies around the globe on the implementation of innovative data management solutions surrounding statistical and scientific data production, archiving, dissemination, and analysis.

Pascal is a member of the Research Data Canada (RDC) Standards and Interoperability National Committee (SINC), and recently joined the CODATA Canada National Committee. He has in the past been extensively involved as a technical expert in the design and adoption of the Data Documentation Initiative (DDI) metadata specification and related standards.

Based in Calgary, AB, Pascal operates through Metadata Technology North America (USA), Integrated Data Management Services (Canada), or as an individual consultant. His early data career was with the World Bank, where he fostered statistical capacity, particularly in the Africa region, and supported the launch of the International Household Survey Network.

Aside from his passion for technology and data, he has strong personal interests in gamification, quantum information science, quantum physics, astrophysics, and music.