Best practices for data standards

> "Data is the new oil. It's valuable, but if unrefined it cannot really be used. It has to be changed into gas, plastic, chemicals, etc to create a valuable entity that drives profitable activity; so must data be broken down, analysed for it to have value." - Clive Humby

The acquisition of data has become the gold rush of the 21st century. But as with valuable resources, those that control it tend to be more concerned with preventing access to it than sharing it. But there are some instances where sharing this resource can be more valuable than protecting it. Data has one key aspect that differentiates it from gold or oil. It can be duplicated an infinite number of times and shared. This means that if 2 people each have a piece of valuable data, and they share it, they will each have 2 pieces of data (ie. 1 + 1 = 4), generating an extra value of 2.

The two critical assumptions to mention here are that in order to encourage this sharing, (a) both parties must be interested or value the other parties' data and (b) the cost of sharing must be dramatically less than the value it generates. Unfortunately, in some cases (such as finance) where data has always been a key component of business, data is often complex, poor quality, outdated and stored in a myriad of different forms and places. In these cases, the cost of turning that data into a shareable form requires a considerable ETL process that can cost many many millions. So costly in fact, that even the 2x return in the example above is not enough. But if instead of 2 parties, you had 1000, then suddenly the return on sharing would be 1000x, which could be very worthwhile! This is where a data standard can play an immense role to generate value. Industries like transportation and healthcare have seen massive benefits from sharing data to make everyone safer and healthier.

What makes a good data standard?

When we set out to develop the FIRE data format for financial regulation, we had 3 key tenets that we deemed important to us:

1. **Community**: In our eyes, a good data standard would be freely and easily accessible to its target community. As such we wanted a data format that could be used and understood by anyone big or small working with financial data without the need for any 3rd party software licenses. Moreover, we wanted the documentation to be user-friendly, clear, concise and understood by a wide audience. For this we chose to publish our schemas on Github which hosts many community driven open source projects.
2. **Neutrality**: The format needed to capture data in a way that was not specific or tailored to a certain business model, vendor or financial institution. Hence why we could not take one bank's format and impose it on the rest of the industry (although it was offered). The one thing we found uniquely common to all firms was regulation as it aims to apply evenly across the board to all firms. Moreover, sharing of regulatory data already takes place on some level between banks and their supervisors but only high-level reports are standardised and not the underlying data.
3. **Clear Focus**: We knew the format could not be everything to everyone and hence wanted to be very clear on the objectives and primary use-cases for the format to avoid disjointed

development efforts pulling the project in different directions. Just like a startup, a new project cannot afford to not know what it is doing or it will be doomed to fail. For us, the focus was regulatory data, nothing more and nothing less. Everything in the format would require a reference to the legislation.

Best practices

Even with clear priorities, difficult decisions have to be made when developing a standard and there are best practices that can be employed in order to ensure stable development and wider adoption. In order to answer these difficult questions, we spoke with Paul Groth, Disruptive Technology Director at Elsevier Labs and co-chair of W3C Provenance Working Group, and Jeni Tennison, Deputy CEO at the Open Data Institute and previously a member of the W3C Technical Architecture Group.

Decide if you will be a standard by nature or standard by adoption

- True standards are carefully crafted by experts, such as ISO or W3C, after much deliberation and organised consultation with a panel of specialists. Other standards come into being in a more eventual manner through continuous development and gradual acceptance by market practitioners. Clarifying the development approach is an important question for potential users and the community and should be included in the project's scope. We came to the conclusion that our project is probably more suited to the latter at this stage and should be presented accordingly. As adoption increases, we can look to move to a more formal approach and share the operational overhead with other key stakeholders. That being said, it is important to not put off the formal process for too long as it can lead to a project losing focus and/or not fulfilling its objectives.

Plan on dealing with the practical realities

- Starting from a blank page, a standard has the luxury of defining an ideal solution. Practically however, the project's key principles can become compromised when faced with considerable challenges to implementation. For example, financial institutions often have gaps in their data which means a 'perfect' data schema might be impossible to implement. Dealing with issues like these requires careful assessment of the pros and cons of the compromise and must be done on a case by case basis. One of the principles of the Open Stand organisation is to "address broad market needs" and therefore it can be justified to prioritise usability over theory. Having a formal process and a board of stakeholders can really help decision making for these kinds of issues.

Think carefully about where is the best place to host your standard

- This ties in closely with accessibility and although Github is widely used, it is still a private company with its own shareholders and objectives that are not guaranteed to be in line with our projects long term needs/goals. A key risk is a change to the way they organise their domains/subdomains or reference links, which could break downstream applications that depend on those links. A more reliable solution is a dedicated project site (to also foster neutrality) or to request a fundamentally neutral organisation (like the W3C) to host it for us.

Consider how the data standard will be used

- Similar to the hosting question, it is important to understand how the schemas will be used from a practical and development standpoint. A good suggestion was to maintain the 'latest' version in the same namespace and when a new release is ready, to freeze that release with its own namespace ie. v1.0, v1.1 etc. How the standard will be used will also drive the supporting work that needs to be done such as presentations, documentation, examples and test data.

Build on what is already available

- It is almost certain that whatever you are doing, many intelligent and talented people have done related or correlated work in your field. Ignoring their results would be a huge waste of a valuable resource, particularly if you are a small team tackling a large problem. Do your research and seek out the leaders of related projects and try to engage with them and build off the work they have already done. For instance, in our case, a great deal of work has been done to develop FIX (for securities) and FPML (for derivatives) which has some interesting overlaps.

Foster growth and innovation

- The ultimate aim of a standard is to create a common, public good that can be leveraged to do even greater things. Greater things that are often not imagined until long after the standard has come into play. As such, it is important to support the community and create an open and collaborative atmosphere around the project. The standard should be interoperable and serve as a useful building block for further innovation.

Conclusions

As a result of our discussions we will seek to implement as much of the best practices we have learned. In particular, we will look to:

- Host the data schemas and related documentation on their own dedicated site, with more user-friendly links and use Github just as a development environment.
- Engage with the W3C to start a Community Group for open and organised discussion.
- Formalise a written set of guiding principles to objectively and transparently accept and reject community contributions.
- Re-factor the project to maintain a 'latest' or 'current' reference that does not change.
- Engage with SWIFT, ISDA and other organisations maintaining financial data standards.
- Think about the process, control and governance the project will require as the number of contributors, stakeholders and users grows.
