Open Data on the Web: 3 Principles For Maximum Participation

A Position Paper from IBM

John M Cohn, Kelvin Lawrence, Susan Malaika, John Reinhardt, David Rook, Robert J Schloss, Biplav Srivastava, March 2013

Wikipedia describes open data as the idea that certain data should be freely available to everyone to use and republish as they wish, without restrictions from copyright, patents or other mechanisms of control. The goals of the open data movement are similar to those of other "Open" movements such as open source, open hardware, open content, and open access. [1]

IBM's experience as an open data aggregator (e.g. CityForward.org), an open data user, with doing data visualization of commercial, scientific, and government data (including spatio-temporal data), and as a middleware and solution provider to retail, travel, healthcare, financial services, education, engineering and other sectors, using data exchange and Web APIs, on the open web, on intranets, and on supply-chain-or-business-ecosystem networks, makes us hopeful that the excellent start the open data movement has achieved can grow to address the needs of those who merge open data with proprietary data in order to make decisions about how their enterprise allocates resources and what activities are performed in what ways, and of mobile web apps which embedded device/equipment/asset apps which combine open data with user or device/equipment/asset specific non-open data.

In support of the wide availability of data, and its use in decision-support applications, IBM recommends that open data, associated technologies, and conventions must be designed with 3 principles in mind:

A) *Incentives*: Open data will include a mix of free data, motivated by the desire for transparency, and data that has been made available through incentives such as government regulations. Incentives (not limited to fee-for-data) are foundational to creating the amount of open data on the web that could transform commerce, research, learning, governing and the arts.

B) *Trust and Security in Ensuring Data Quality:* Trust and Security technologies must be involved so that consuming applications know that other consuming organizations consider the provided data has quality that is "suitable for purpose".

C) *Provenance and Data Ownership*: Clear data provenance and ownership is vital to providing accountability from correct-and-authorized, correct-and-unauthorized, or incorrect open data between suppliers and consumers of the open data is clear to both parties before the data is transferred.

** Context: Experience with Open Data Aggregation, Hosting, Visualization **

Since 2010, IBM Corporate Citizenship & Corporate Affairs has sponsored City Forward, a free, web-based platform that enables city officials, academics and interested citizens world-wide to view and interact with open data while engaging in an ongoing public dialogue. The site contains open data for over 100 cities from over 50 sources around the world.

IBM worked through the experience of integrating data from many sources; we know the challenges of discovering, transforming and presenting data in a valuable way to web users, not simply to application developers – users who wish to share and contrast insights and patterns they detect in the information. IBM would like to contribute to the definition of standards that allow data consumers to more easily add value to data for the many new presentation devices and form factors we see today: phones, tablets, web sites, etc. Developers that consume open data often provide incentives by contributing increasingly limited resources to the government agencies and non-profit organizations that provide the data. Standards are necessary to provide trust in the data quality and provide the metadata to properly use and attribute open data. We have lived these issues for nearly four years and are pleased to contribute to shaping the solution for the parties involved in the open data community. IBM has also collaborated on the technology used by the web site http://DubLinked.ie .

** A: Incentives **

The benefits which people expect from Open Data only occur when there is a supply of open data, made available in ways so that its quality can be determined, no or limited additional "adapter-connector-mapper programming" is needed for a consuming application, and the data is made available in ways so that enterprises which want to use the data understand what rights they have, or need to obtain, before using it for certain purposes.

The biggest technical, social and economic advance will not come by simply adopting web-compatible technologies for open data that could have been shared, more cumbersomely, with other IT technologies. Instead, they will come because a great deal more open data will be available because incentives to provide open data are built in and those incentives are supported directly by the web technologies used by the providers, brokers, and consumers of the open data.

Some providers of data, by the nature of their mission, simply want the data they have used by others. They need no "credit", they need no "branding benefits", they need no "growth of relationships with downstream consumers". IBM urges the community to support other providers and collectors of data, who will need incentives such as "credit" "payment" "branding benefits" "relationships" "leads" as a result of providing their data to others. The tendency to "not distribute" is built into many enterprises, because there is a unknown risk to their mission when competitors, suppliers, customers, regulators are able to get their collected data and use it in unforeseen ways, often by applying sophisticated analytics which operates of "this open data" blended with unknown other data.

Data Providers must have a way (if they choose to use it) to guarantee that their (multimedia) Logo appears in front of the relevant consuming enterprise participant as a function of time, volume, and location of the consuming programmatic logic.

Data Providers must have a way (if they choose to use it) to receive identities (minimally, e-mail addresses or URIs) when their data is fetched. Current web technologies, such as Browser-held Cookies, can simplify the annoyance factor to web users of providing this identity information repeatedly.

Data Providers must have a way (if they choose to use it) to charge for the data delivered, not only as a function of the data itself, but as a function of the requester organization and the technical protocols used to transfer the data (which could vary considerably in their resources used, time to execute, etc.)

Another important incentive is relatively low cost and quick start for new providers of open data. The experience of the library and data archiving community, and of the Linked Open Data community, seems to indicate that agreement on a light weight core of metadata about the format and scope of a data source is achievable, leaving more detailed issues of metadata expression to be applied only to the kind of open data where consuming applications must have this more extensive metadata.

We can classify incentives as numerous types. Here are 3 well-known types:
1. Incentive - recognition: At this level, the producer wants simply to be recognized for the data produced and found valuable.

2. Incentive – licensing and outcomes: At this level, the data producer wants incremental share of any commercial value the consumer achieves using that open data.

3. Incentive - commercial: At this level, the data producer wants up front financial payment for usage by consumer.

We argue that incentives should be tracked as first class metadata in the open data ecosystem, because over time open data may lead to more open ways of combining the services and intelligence of numerous participants without contract-in-advance, and knowing that this is happening could be one way we document a cultural change enabled by web approaches.

** B: Trust and Security Mechanisms for Ensuring Data Quality **

While the ability to use URI's in association with various tagging, self-rating, 3rd-party-rating systems is well developed, each individual retrieval of open data could have different characteristics, which implies that the adoption of digital signature technologies in order to have clarity about the quality of the open data retrieved:

a) who collected this data (including what versions of what systems did any preprocessing of it)

b) who authorized distribution of this data

c) which version of this data was actually transferred to satisfy a particular request

d) has any assertion of "exclusive access" or "exclusive until after date-time access" be made by the data supplier

e) has any assertion of "only complete redistribution permitted" (no partial redistribution) been made

f) has any assertion been made as to the authenticity of the data and are there guarantees in place that it has not been tampered with.

While it would not be in the spirit of the Web or the Open Data movement to mandate that all data providers use trust and security technologies, we can recommend that use of these technologies always be considered, so that we do not find, 5-10 years from now, the world as full of "maldata" as it is currently full of "malware".

While regulatory protections of personally identifiable data and of data that contains owned intellectual property are emerging, we expect that in most cases, open data will normally be aggregated under an "opt-in" (explicit opt to be included) process.

** C: Provenance and Data Ownership **

The ability to use data is predicated on a clear understanding of the rights and restrictions, if any, to its use. Many enterprises collect information about people, equipment, natural systems, use of ICT systems, etc . The union of some documented "Terms and Conditions" and some "widely believed implicit contracts and purposes" is what courts use in determining if they should listen to any requests to order halts-in-data-distribution or the payment of compensating damages awards to those whose data was redistributed improperly causing some kind of harm.

In the world of plentiful information, consuming enterprises will have a choice between sources with overlapping information, and will want to choose the source which imposes the least obligations on them if in other respects the quality of the data is equivalent.

It is critical that before open data transfer time, data consumers understand the obligations with respect to later 3rd party allegations of improper distribution or use. It is important that before open data transfer time, data providers understand the obligations with respect to later 3rd party allegations of improper distribution they will need to manage because data consumers are not taking full responsibility to handle them. Realistically, we know many data providers will simply not provide their data to any consuming application which does not honor restrictions. Our goal though should be to evolve just enough technology about transferring provenance and ownership information so that the widest range of general purpose web enablers, including search engines, can list open data along with some summarized representation of restrictions on its use, perhaps starting with the work of Creative Commons on licenses [2].

** Conclusion **

Our workshop will take place in Europe, with numerous academic participants, and both communities have been in the vanguard of adopting open data practices. But IBM, as a company serving clients all over the world, in numerous sectors, and with IBM itself engaged in many kinds of services, we also want to emphasize how adopting these 3 principles for our technical approaches can address the more cautious attitude we see outside of Europe, by increasing their confidence that producing or using open data does not make them more vulnerable, but simply more effective and possibly more agile and more respected.

The hurdles to open data we have seen related to (a) lack of culture of openness even within enterprises, (b) perception of giving away control once open data is shared, especially if shared in near real-time, and (c) vulnerability and security perceptions. We hope the open data movement will both publish best practices that address the perceived security concerns heads-on, and also publish success stories aggressively to citizens, consumers, students and to small and medium organizations.

Related Materials

[1] Open Data on Wikipedia: http://en.wikipedia.org/wiki/Open_data

[2] Creative Commons: http://creativecommons.org/