Untangling the data debate: definitions and implications

Data is a hot topic right now: from big data, to open data and linked data, entrepreneurs and policy makers are making big claims about 'data revolutions'. But, not all 'data' are the same, and good decision making about data involves knowing the differences.

	Definitions	**Potential implications**
Big data	Data that requires 'massive' computing power to process (Crawford & Boyd, 2011). Massive computing power, originally only available on supercomputers, is increasingly available on desktop computers or via low cost cloud computing.	Companies and researchers can 'data mine' vast data resources, to identify trends and patterns. Big data is often generated by combining different datasets. *Digital traces from individuals and companies are increasingly captured and stored for their potential value as 'big data'.*
Raw data	Primary data, as collected or measured direct from the source. *Or* Data in a form that allows it to be easily manipulated, sorted, filtered and remixed.	Access to raw data can allows journalists, researchers and citizens to 'fact check' official analysis. Programmers are interested in building innovative services with raw data.
Real-time data	Data measured and made accessible with minimal delay. Often accessed over the web as a stream of data through APIs (Application Programming Interfaces).	Real-time data supports rapid identifications trends. Data can support the development of 'early warning systems' (e.g. Google Flu Trends; Ushahidi). 'Smart systems' and 'smart cities' can be configured to respond to real-time data and adapt to changing circumstances.
Open data	Datasets that are made accessible in non-proprietary formats under licenses that permit unrestricted re-use (OKF - Open Knowledge Foundation, 2006). Open government data involves governments providing many of their datasets online in this way.	Third-parties can innovate with open data, generating social and economic benefits. Citizens and advocacy groups can use open government data to hold state institutions to account. Data can be shared between institutions with less friction.
Personal/ private data	Data about an individual that they have a right to control access to. Such data might be gathered by companies, governments or other third-parties in order to provide a service to someone, or as part of regulatory and law-enforcement activities.	Many big and raw datasets are based on aggregating personal data, and combining them with other data. Effective anonymisation of personal data is difficult particularly when open data provides the pieces for 'jigsaw identification' of private facts about people (Ohm, 2009).
Linked data	Datasets are published in the RDF format using URIs (web addresses) to identify the elements they contain, with links made between datasets (Berners-Lee, 2006; Shadbolt, Hall, & Berners-Lee, 2006).	A 'web of linked data' emerges, supporting 'smart applications' (Allemang & Hendler, 2008) that can follow the links between datasets. This provides the foundations for the Semantic Web.

Tim Davies: Open Data Impacts research note – http://www.opendataimpacts.net
Version 0.1 – March 2012 – tim@practicalparticipation.co.uk

More dimensions of data:
These are just a few different types of data commonly discussed in policy debates. There are many other data-distinctions we could also draw. For example: we can look at whether data was crowd-sourced, statistically sampled, or collected through a census. The content of a dataset also has important influence on the implications that working with that data will have: an operational dataset of performance statistics is very different from a geographical dataset describing the road network for example.

Crossovers and conflicts:
Almost all of the above types of data can be found in combination: you can have big linked raw data; real-time open data; raw personal data; and so-on.

There are some combinations that must be addressed with care. For example, 'open data' and 'personal data' are two categories that are generally kept apart for good reason: open data involves giving up control over access to a dataset, whilst personal data is the data an individual has the right to control access over.

These can be found in combination on platforms like Twitter, when individuals choose to give wider access to personal information by sharing it in a public space, but this is different from the controller of a dataset of personal data making that whole dataset openly available.

A nuanced debate:
It's not uncommon to see claims and anecdotes about the impacts of 'big data' use in companies like Amazon, Google or Twitter being used to justify publishing 'open' and 'raw data' from governments, drawing on aggregating 'personal data'. This sort of treatment glosses over the difference between types of data, the contents of the datasets, and the contexts they are used in. Looking to the potential of data use from different contexts, and looking to transfer learning between sectors can support economic and social innovation, but it also needs critical questions to be asked, such as:

- **What kind of data is this case describing?**
- **Does the data I'm dealing with have similar properties?**
- **Can the impacts of this data apply to the data I'm dealing with?**
- **What other considerations apply to the data I'm dealing with?**

Bibliography/further reading:
See http://www.opendataimpacts.net for ongoing work.

Allemang, D., & Hendler, J. A. (2008). *Semantic web for the working ontologist: modeling in RDF, RDFS and OWL*. Morgan Kaufmann. Retrieved from

Berners-Lee, T. (2006, July). Linked Data - Design Issues. Retrieved from http://www.w3.org/DesignIssues/LinkedData.html

Crawford, K., & Boyd, D. (2011). Six Provocations for Big Data.

Davies, T. (2010). *Open data, democracy and public sector reform: A look at open government data use from data. gov. uk*. Practical Participation. Retrieved from http://www.practicalparticipation.co.uk/odi/report

OKF - Open Knowledge Foundation. (2006). Open Knowledge Definition. Retrieved March 4, 2010, from http://www.opendefinition.org/

Ohm, P. (2009). Broken promises of privacy: Responding to the surprising failure of anonymization. *Imagine*. Retrieved from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=1450006

Shadbolt, N., Hall, W., & Berners-Lee, T. (2006). The Semantic Web Revisited. *IEEE intelligent systems*, *21*(3), 96–101.