## 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.

<table>
<thead>
<tr>
<th>Definitions</th>
<th>Potential implications</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Big data</strong></td>
<td>Data that requires ‘massive’ computing power to process (Crawford &amp; Boyd, 2011). Massive computing power, originally only available on supercomputers, is increasingly available on desktop computers or via low cost cloud computing.</td>
</tr>
<tr>
<td><strong>Raw data</strong></td>
<td>Primary data, as collected or measured direct from the source. <em>Or</em> Data in a form that allows it to be easily manipulated, sorted, filtered and remixed.</td>
</tr>
<tr>
<td><strong>Real-time data</strong></td>
<td>Data measured and made accessible with minimal delay. Often accessed over the web as a stream of data through APIs (Application Programming Interfaces).</td>
</tr>
<tr>
<td><strong>Open data</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Personal/private data</strong></td>
<td>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.</td>
</tr>
<tr>
<td><strong>Linked data</strong></td>
<td>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, &amp; Berners-Lee, 2006).</td>
</tr>
</tbody>
</table>
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.