A realistic look at open data

Sharon Dawes
www.ctg.albany.edu

Using Open Data Workshop
Brussels, June 19-20, 2012
“I think you should be more explicit here in step two.”

Sidney Harris, 2012
Sources of data problems

Conventional wisdom
Provenance
Practices

Consequences of the problems
Underuse
Misuse
Non-use

Shifting costs and responsibilities
Where do open data come from?

- Administrative systems
- Embedded in program operations
- Governed by specific policies and laws
- Gathered in particular contexts for certain internal purposes
- By people with different kinds and levels of knowledge and expertise
Case 1: Give me shelter
Case 2: Cadastral records
Case 3: Where does the money go?
data
technology

data
"I'm right there in the room, and no one even acknowledges me."

The New Yorker, 9/18/06
Data quality = fitness for use

• Matters most from the user’s point of view
• Depends on the user’s purpose
• Four types of quality:
  – Intrinsic
  – Contextual
  – Representational
  – Access-related
• Usually involves trade-offs
  – e.g., timeliness vs. completeness

<table>
<thead>
<tr>
<th>Dimension</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Accessibility</td>
<td>Extent to which data is available, or easily and quickly retrievable</td>
</tr>
<tr>
<td>Appropriate Amount of Data</td>
<td>Extent to which the volume of data is appropriate for the task at hand</td>
</tr>
<tr>
<td>Believability</td>
<td>Extent to which data is regarded as true and credible</td>
</tr>
<tr>
<td>Completeness</td>
<td>Extent to which data is not missing and is of sufficient breadth and depth for the task at hand</td>
</tr>
<tr>
<td>Concise Representation</td>
<td>Extent to which data is compactly represented</td>
</tr>
<tr>
<td>Consistent Representation</td>
<td>Extent to which data is presented in the same format</td>
</tr>
<tr>
<td>Ease of Manipulation</td>
<td>Extent to which data is easy to manipulate and apply to different tasks</td>
</tr>
<tr>
<td>Free-of-Error</td>
<td>Extent to which data is correct and reliable</td>
</tr>
<tr>
<td>Interpretability</td>
<td>Extent to which data is in appropriate languages, symbols, and units, and the definitions are clear</td>
</tr>
<tr>
<td>Objectivity</td>
<td>Extent to which data is unbiased, unprejudiced, and impartial</td>
</tr>
<tr>
<td>Relevancy</td>
<td>Extent to which data is applicable and helpful for the task at hand</td>
</tr>
<tr>
<td>Reputation</td>
<td>Extent to which data is highly regarded in terms of its source or content</td>
</tr>
<tr>
<td>Security</td>
<td>Extent to which access to data is restricted appropriately to maintain its security</td>
</tr>
<tr>
<td>Timeliness</td>
<td>Extent to which the data is sufficiently up-to-date for the task at hand</td>
</tr>
<tr>
<td>Understandability</td>
<td>Extent to which data is easily comprehended</td>
</tr>
<tr>
<td>Value-Added</td>
<td>Extent to which data is beneficial and provides advantages from its use</td>
</tr>
</tbody>
</table>

Pipino, et al., 2002
Data quality “tools”

• For providers
  – Appreciate data as an asset, a source of value
  – Adopt information policies to preserve and enhance value
  – Create and maintain metadata to support unknown users
  – Adopt stewardship practices

• For users
  – Be skeptical, ask questions
  – Understand the nature and context of the data
  – Use data sets with caution
  – Combine data sets with great caution
© 1996 Ted Goff