Discussion and next steps

Gary Simons,
SIL International

LSA Satellite Workshop on Sociolinguistic Archival Preparation, Portland, OR, 5 Jan 2012
Toward “engineering standards”

• The instructions to participants mentioned an ultimate goal of developing protocols and engineering standards for data sharing, but it is clear we have a long way to go

• In terms of Cieri’s model:
  – Understood > Documented > Consistent > Standard

• We are still working at the Understanding level
### Types of archivable datasets

<table>
<thead>
<tr>
<th>Type</th>
<th>Data</th>
<th>Metadata</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sociolinguistic corpus</td>
<td>Recorded speech or authored text</td>
<td>Demographic factors Situational factors</td>
</tr>
<tr>
<td>Language attitudes study</td>
<td>Measurements of attitudes</td>
<td>Demographic factors Situational factors</td>
</tr>
<tr>
<td>Language use study</td>
<td>Observations of language choice</td>
<td>Demographic factors Situational factors</td>
</tr>
</tbody>
</table>

**Recommendation**

- Start by focusing on demographic and situational factors
- Attitudes are a different kind of thing
Some demographic factors

- Gender
- Year of birth
- Age group
- Educational level
- Social class
- National origin
- Immigrant generation
- Age at arrival
- Ethnicity
- Religious affiliation
- Original religious affiliation
- Level of religious activity
- Relationship to interviewer
- First language
- Heritage language
- Language proficiency
Toward standardized factors

1. Identify all factors on which it would be possible to achieve consensus on a name and definition.

2. Develop the list of standard names with definitions (e.g. as DatCats).

3. For each factor, develop guidelines for coding the values, including specifying controlled vocabularies where possible.
Toward standardized formats

• Once standards proposals for factors are in place, the next step is to work on standardized formats to support interoperability.
• This could be the topic of another workshop in which the price of admission is a format proposal and a coded dataset that illustrates the use of that format for encoding the standardized factors.
• Participants would debate the proposals and come to consensus on best practice recommendations.