Deep Web vs. Dark Web

- **Dark Web**: Hidden intentionally
  - Largely to support illegal or socially unacceptable activity
  - *But legality and acceptability vary, web is trans-national and trans-cultural*
  - We won’t go here…

- **Deep Web**: Data hidden behind interfaces
  - Federated search is one answer
  - But can we crawl this data?
Does the Deep Web Matter?

- Where are the entry points?
- What is the scale?
- How “structured” is the data?
- What topics are covered?
- How well do search engines already cover this?
- What about existing specialized portals?
Size Estimate of the Deep Web

<table>
<thead>
<tr>
<th></th>
<th>Sampling Results</th>
<th>Total Estimate</th>
<th>99% Confidence Interval</th>
</tr>
</thead>
<tbody>
<tr>
<td>Deep Web sites</td>
<td>126</td>
<td>307,000</td>
<td>236,000 - 377,000</td>
</tr>
<tr>
<td>Web databases</td>
<td>190</td>
<td>450,000</td>
<td>366,000 - 535,000</td>
</tr>
<tr>
<td>—unstructured</td>
<td>43</td>
<td>102,000</td>
<td>62,000 - 142,000</td>
</tr>
<tr>
<td>—structured</td>
<td>147</td>
<td>348,000</td>
<td>275,000 - 423,000</td>
</tr>
<tr>
<td>Query interfaces</td>
<td>406</td>
<td>1,258,000</td>
<td>1,097,000 - 1,419,000</td>
</tr>
</tbody>
</table>

Search Engine Coverage

The entire deep Web

Google.com (32%)
Yahoo.com (32%)
MSN.com (11%)
All (37%)
Deep Web Components
(He, Patel, Zhang, Chang ‘07)

Challenges

• How do we know what is in a database?
  – Sample queries?
  – Search page
    • Descriptive information
    • Form fields
• How do we query it?
• How do we process results?
Can this be real?

• “General” search
  – See Google, etc.
• “Specialized” search
  – Metaquerier
  – Cazoodle