The challenges of collecting social media

Géraldine Camile – Peter Stirling
Bibliothèque nationale de France
From 2011 to today

• Boom in number of Twitter accounts collected, especially for elections

⇒ web crawls since 2011 for specific projects such as elections or Olympic games

• Today 84% of the accounts selected in 2012 are still active
Web crawl versus API crawl

• Our main goals
  – Build **representative** collections ≠ all the tweets
    • Choice of accounts and hashtags according to a collection policy
  – Render them **as they were published** on the live web
    • The links between the accounts and hashtags
    • The media (images, videos...)

• Our constraints
  – No funding for an API service subscription
  – Less development on crawl, access and preservation
The advantages

• Reproduce tweets as published on the live web

• Process uses the existing BnF workflow
  – Crawl using NetarchiveSuite and Heritrix 3
    • "websocial" harvest template: page + 1, duration 11 hours, specific filters
  – Access in the BnF web archives
The limits

• The timeline: 20 tweets per capture (40 tweets daily)

=> Representativeness?

• The depth of the capture depends on the number of accounts selected and the duration of the crawl

  • 3,690 accounts during the elections
### Two complementary approaches

<table>
<thead>
<tr>
<th>Tools for crawl, access and preservation</th>
<th>Web crawl</th>
<th>API crawl</th>
</tr>
</thead>
<tbody>
<tr>
<td>BCWeb</td>
<td><strong>BCWeb</strong></td>
<td><strong>Open source tool and libraries</strong></td>
</tr>
<tr>
<td>Heritrix 3</td>
<td><strong>Heritrix 3</strong></td>
<td><strong>Specific access and preservation development</strong></td>
</tr>
<tr>
<td>OpenWayback</td>
<td><strong>OpenWayback</strong></td>
<td></td>
</tr>
<tr>
<td>SPAR</td>
<td><strong>SPAR</strong></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Advantages</th>
<th>Render tweets as online</th>
<th>Keep the links between accounts</th>
<th>Keep the existing workflow (homogeneity and interoperability of collections, preservation)</th>
<th>Developer friendly and easy to use</th>
<th>Reliable (until now)</th>
<th>Scalability</th>
<th>Full metadata</th>
</tr>
</thead>
</table>

<table>
<thead>
<tr>
<th>Limits</th>
<th>Number of accounts (3000?)</th>
<th>Number of tweets</th>
<th>Missing videos</th>
<th>Missing short links</th>
<th>Few metadata</th>
<th>Public API limits</th>
<th>Develop and maintain specific access solution</th>
</tr>
</thead>
</table>