Data Quality Initiative
At the Botanic Garden and Botanical Museum Berlin-Dahlem

David Fichtmueller

2013-10-29
Match the Country Names

| Country Name | ISO 3166-1 alpha 2 Code |
### Match the Country Names

<table>
<thead>
<tr>
<th>Country Name</th>
<th>ISO 3166-1 alpha 2 Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Италия</td>
<td>US</td>
</tr>
<tr>
<td>Estados Unidos</td>
<td>IS</td>
</tr>
<tr>
<td>Siraaliyoon</td>
<td>IT</td>
</tr>
<tr>
<td>アイスランド</td>
<td>SL</td>
</tr>
<tr>
<td>Country Name</td>
<td>ISO 3166-1 alpha 2 Code</td>
</tr>
<tr>
<td>----------------------------</td>
<td>-------------------------</td>
</tr>
<tr>
<td>Италия</td>
<td>US</td>
</tr>
<tr>
<td>Estados Unidos</td>
<td>IS</td>
</tr>
<tr>
<td>アイスランド</td>
<td>SL</td>
</tr>
<tr>
<td>Siraaliyyoon</td>
<td>IT</td>
</tr>
</tbody>
</table>

United States - Spanish
Match the Country Names

<table>
<thead>
<tr>
<th>Country Name</th>
<th>ISO 3166-1 alpha 2 Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Италия</td>
<td>US</td>
</tr>
<tr>
<td>Estados Unidos</td>
<td>IS</td>
</tr>
<tr>
<td>Siraaliyoon</td>
<td>IT</td>
</tr>
<tr>
<td>アイスランド</td>
<td>SL</td>
</tr>
</tbody>
</table>

Италия
Estados Unidos - Spanish
Siraaliyoon - Somali
アイスランド
Match the Country Names

<table>
<thead>
<tr>
<th>Country Name</th>
<th>ISO 3166-1 alpha 2 Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Италия</td>
<td>US</td>
</tr>
<tr>
<td>Estados Unidos</td>
<td>IS</td>
</tr>
<tr>
<td>Siraaliyoon</td>
<td>IT</td>
</tr>
<tr>
<td>アイスランド</td>
<td>SL</td>
</tr>
</tbody>
</table>

Италия
Italy - Russian

Estados Unidos
United States - Spanish

Siraaliyoon
Sierra Leone - Somali

アイスランド
Match the Country Names

<table>
<thead>
<tr>
<th>Country Name</th>
<th>ISO 3166-1 alpha 2 Code</th>
</tr>
</thead>
<tbody>
<tr>
<td>Италия</td>
<td>US</td>
</tr>
<tr>
<td>Italy - Russian</td>
<td></td>
</tr>
<tr>
<td>Estados Unidos</td>
<td>IS</td>
</tr>
<tr>
<td>United States - Spanish</td>
<td></td>
</tr>
<tr>
<td>Siraaliyoon</td>
<td>IT</td>
</tr>
<tr>
<td>Sierra Leone - Somali</td>
<td></td>
</tr>
<tr>
<td>アイスランド</td>
<td>SL</td>
</tr>
<tr>
<td>Iceland - Japanese</td>
<td></td>
</tr>
</tbody>
</table>
Data Quality Initiative (DQI)

4 Projects at the Botanic Garden and Botanical Museum Berlin-Dahlem (BGBM) about DQ
Goal

• Avoid Duplicate Work
• Create Better Tools
• Share Knowledge
• Make Tools/Knowledge public
  – Open Source Software License
What are Data Quality Tools?

• Any Software that helps improve Data Quality
  – Detect Errors
  and/or
  – Correct Errors

• Automated!
  – Don't bring the data to the tools, but bring the tools to the data!
How Data Quality Tools should work
How Data Quality Tools should work

Data Quality Software

The Software that accesses the research data to be checked
How Data Quality Tools should work

Data Quality Software

The Software that accesses the research data to be checked

HTTP

Web Service

Making program logic accessible via web
*Example: REST-API*
How Data Quality Tools should work

- **Data Quality Software**: The Software that accesses the research data to be checked.

- **Web Service**: Making program logic accessible via web. *Example*: REST-API.

- **Library**: Contains program logic, API. Depending on Programming Language. *Example*: **Jar-File for Java-Library**.
How Data Quality Tools should work

Data Quality Software

in the Software

HTTP

Web Service

Library

Data

The Software that accesses the research data to be checked

Making program logic accessible via web
Example: REST-API

Contains program logic, API
Depending on Programming Language
Example: Jar-File for Java-Library

Independent of Programming Language
In a particular Format: XML, JSON, CSV, ...
Example: Dataset of Country Names
How Data Quality Tools should work

The Software that accesses the research data to be checked

Making program logic accessible via web
*Example: REST-API*

Contains program logic, API
Depending on Programming Language
*Example: Jar-File for Java-Library*

Independent of Programming Language
In a particular Format: XML, JSON, CSV, ...
*Example: Dataset of Country Names*
Current Focus

• Occurrence and Collection Data
• Correction on individual values or combination of values of one individual
• No group validation
  – Outliner Detection
  – Duplicate Detection
• Programming Languages: Java and JavaScript
What can the DQI do for you?

Public Wiki:  http://biowikifarm.net/dataquality
What can you do for the DQI?

• Let us know about good data sets / libraries / web services
• Spread the word, join the discussion
• Bundle your tools in a library
• Improve existing tools
• Turn a library into a web service
• Suggest new tools
• Port a library to a different language
Future of the Data Quality Initiative

- More and better tools
- Fill the Wiki
- Code Hosting and Bug Tracking
- One DQ-Library to rule them all
- Hosting for Web Services?
- <Insert your idea here>
Thank You!

Questions?

Wiki:  http://biowikifarm.net/dataquality

E-Mail:  d.fichttmueller@bgbm.org