Data Quality Initiative

At the Botanic Garden and Botanical Museum Berlin-Dahlem

David Fichtmueller

BGBM

Freie Universität Berlin

2013-10-29

Match the Country Names

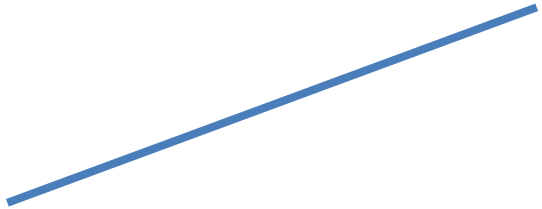
Country Name	ISO 3166-1 alpha 2 Code

Match the Country Names

Country Name	ISO 3166-1 alpha 2 Code
Италия	US
Estados Unidos	IS
Siraaliyoon	IT
アイスランド	SL

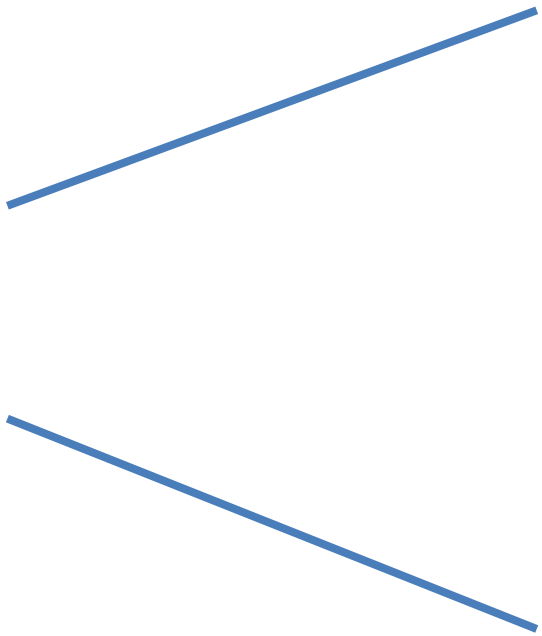
Match the Country Names

Country Name	ISO 3166-1 alpha 2 Code
Италия	US
Estados Unidos United States - Spanish	IS
Siraaliyoon	IT
アイスランド	SL

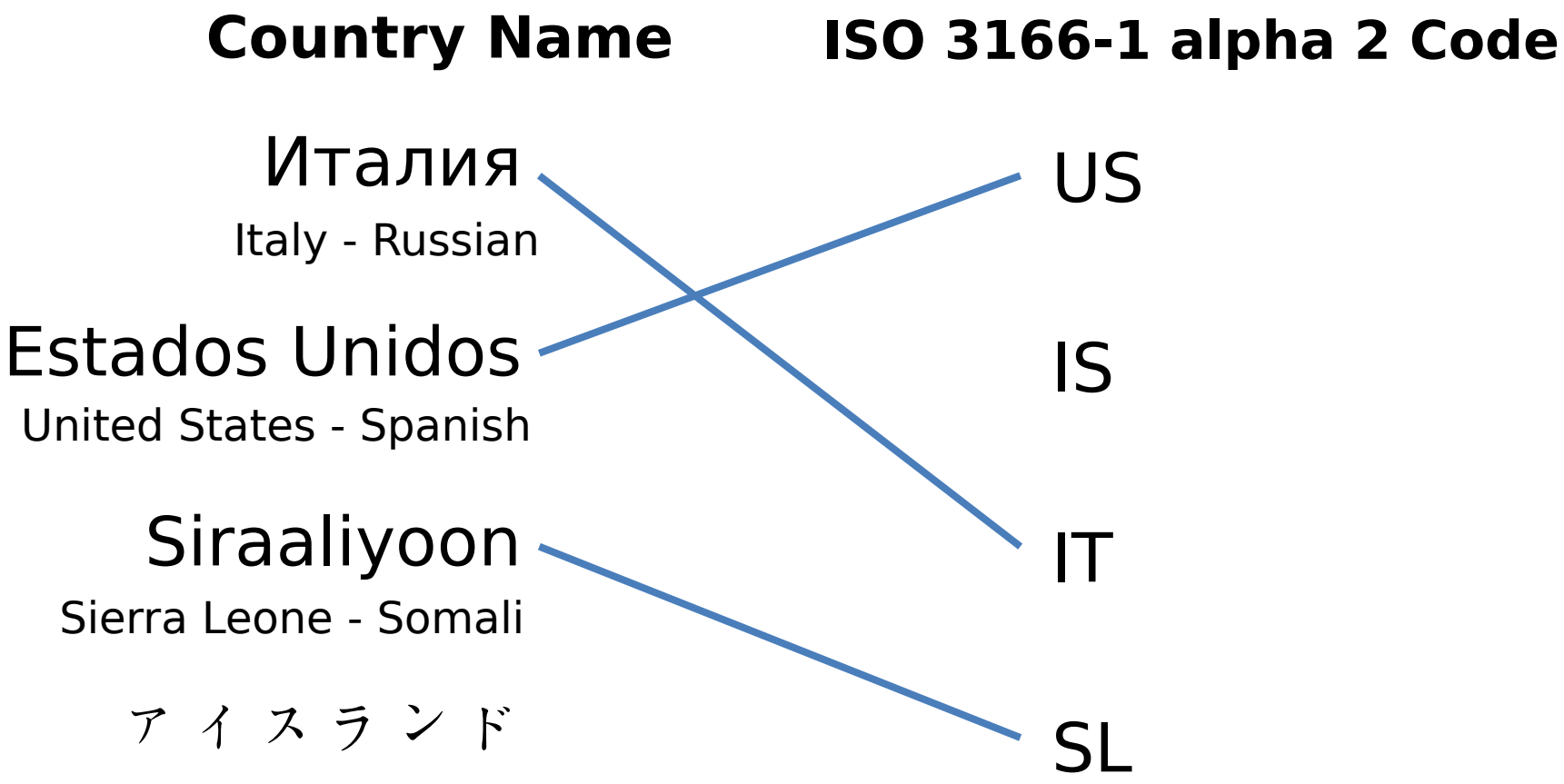


Match the Country Names

Country Name	ISO 3166-1 alpha 2 Code
Италия	US
Estados Unidos United States - Spanish	IS
Siraaliyoon Sierra Leone - Somali	IT
アイスランド	SL

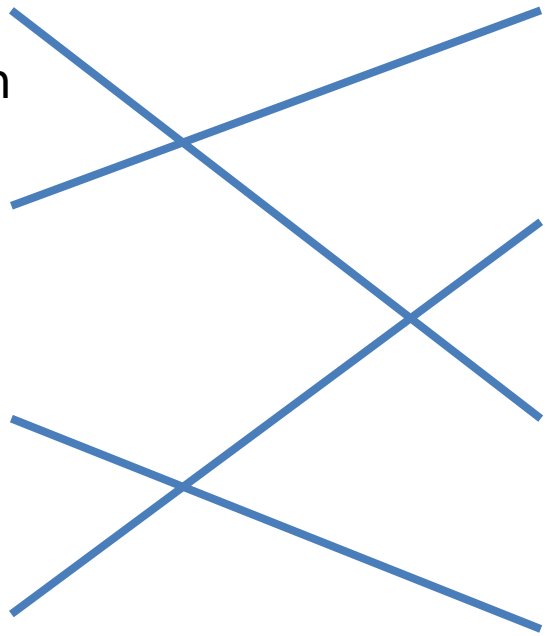


Match the Country Names



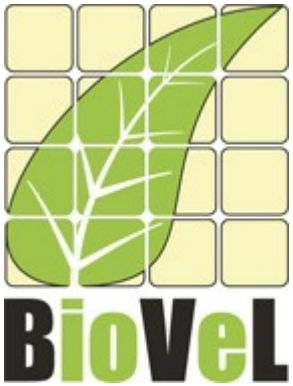
Match the Country Names

Country Name	ISO 3166-1 alpha 2 Code
Италия Italy - Russian	US
Estados Unidos United States - Spanish	IS
Siraaliyoon Sierra Leone - Somali	IT
アイスランド Iceland - Japanese	SL



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

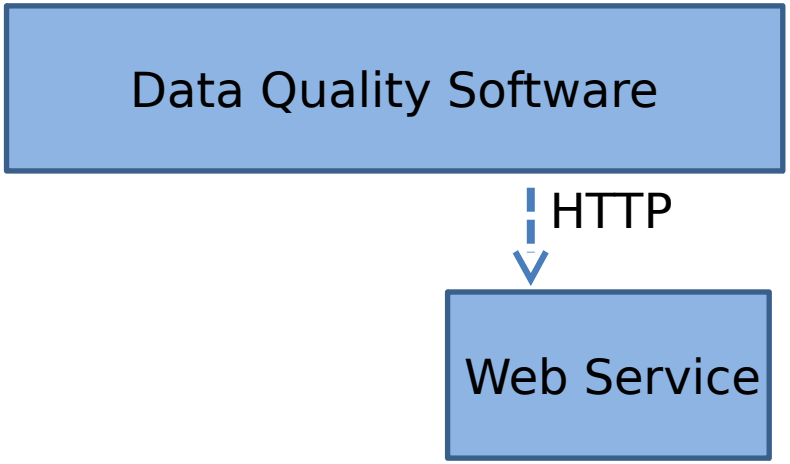
- Automated!
  - Don't bring the data to the tools, but bring the tools to the data!

How Data Quality Tools should work

How Data Quality Tools should work

The Software that accesses the research data to be checked

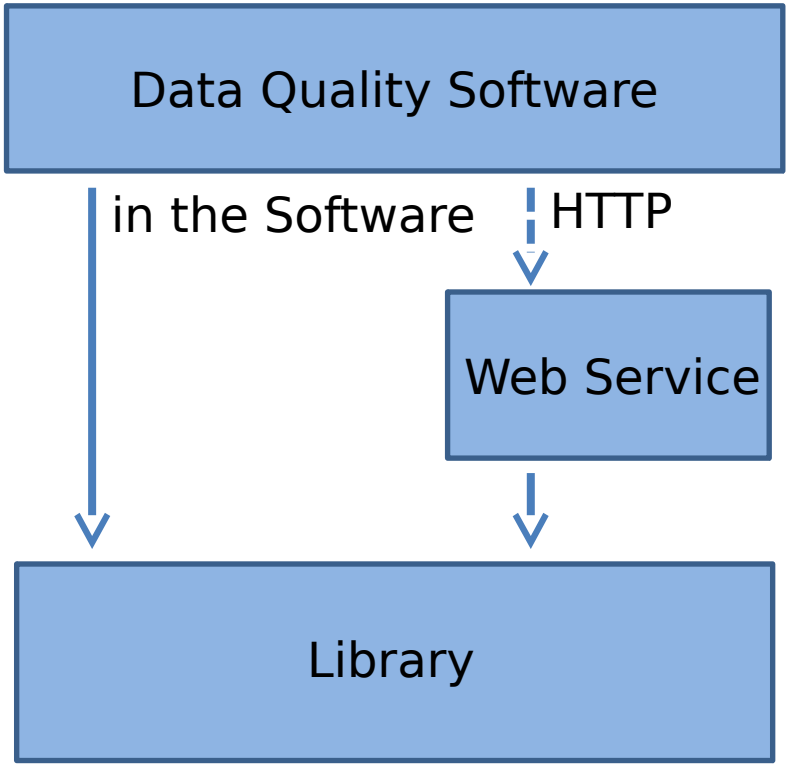
How Data Quality Tools should work



The Software that accesses the research data to be checked

Making program logic accessible via web
Example: REST-API

How Data Quality Tools should work

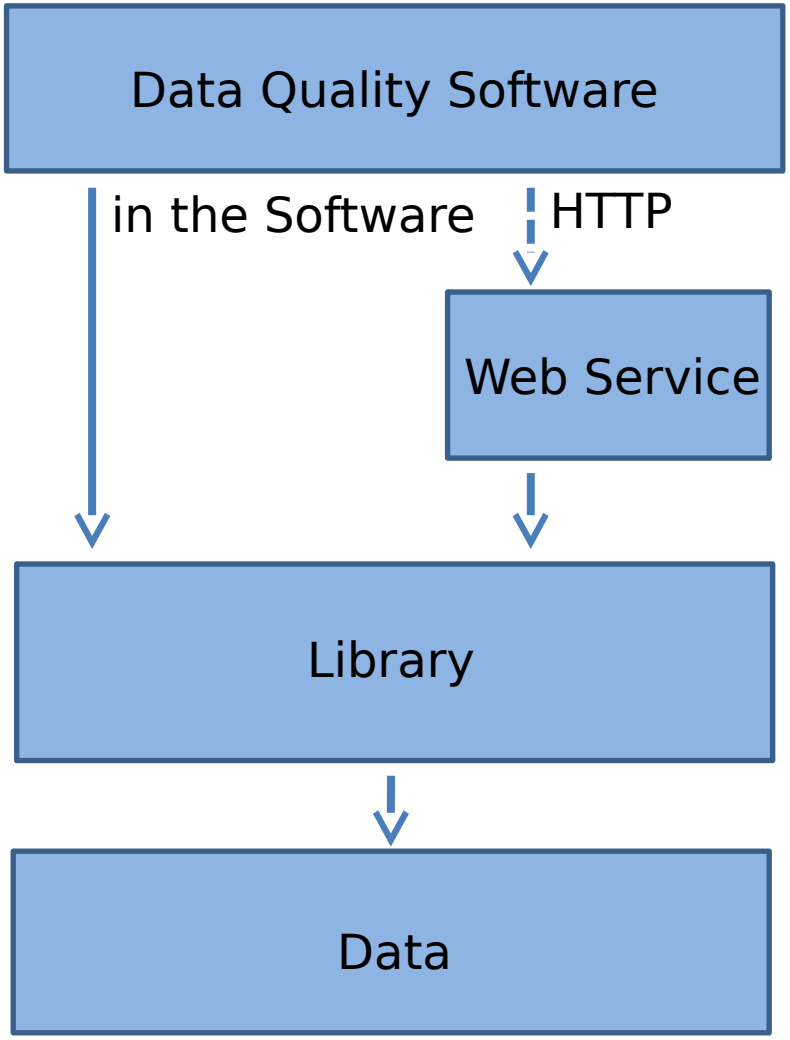


The Software that accesses the research data to be checked

Making program logic accessible via web
Example: REST-API

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

How Data Quality Tools should work



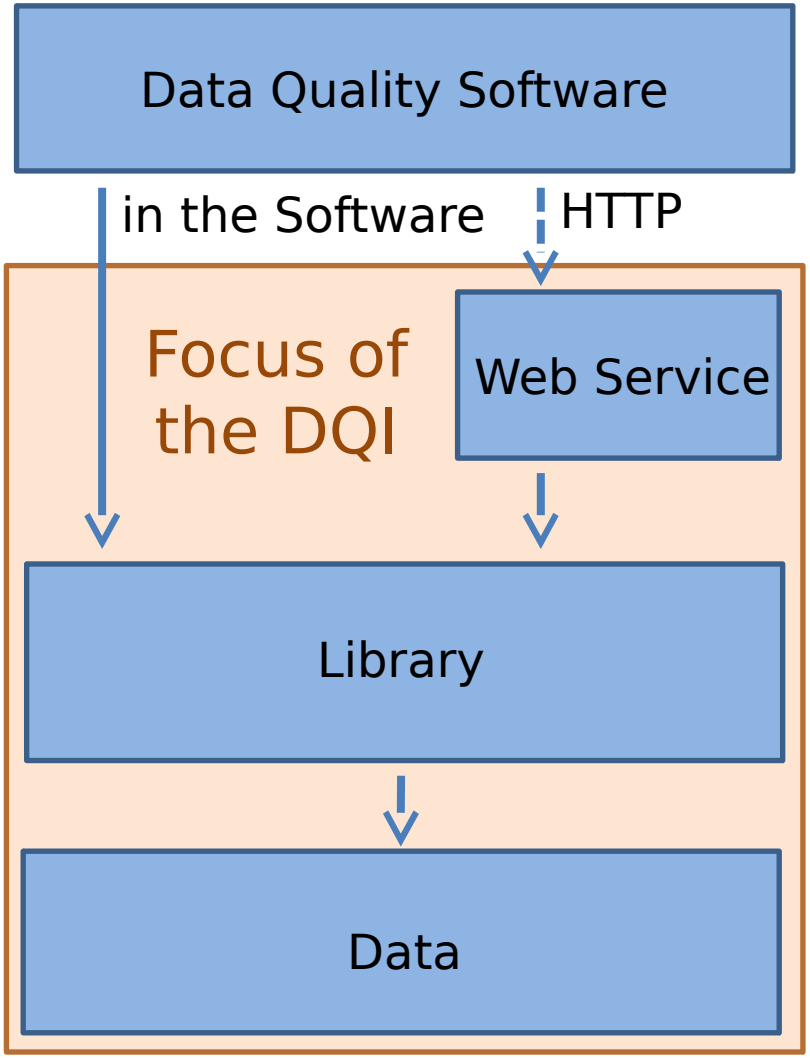
The Software that accesses the research data to be checked

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

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

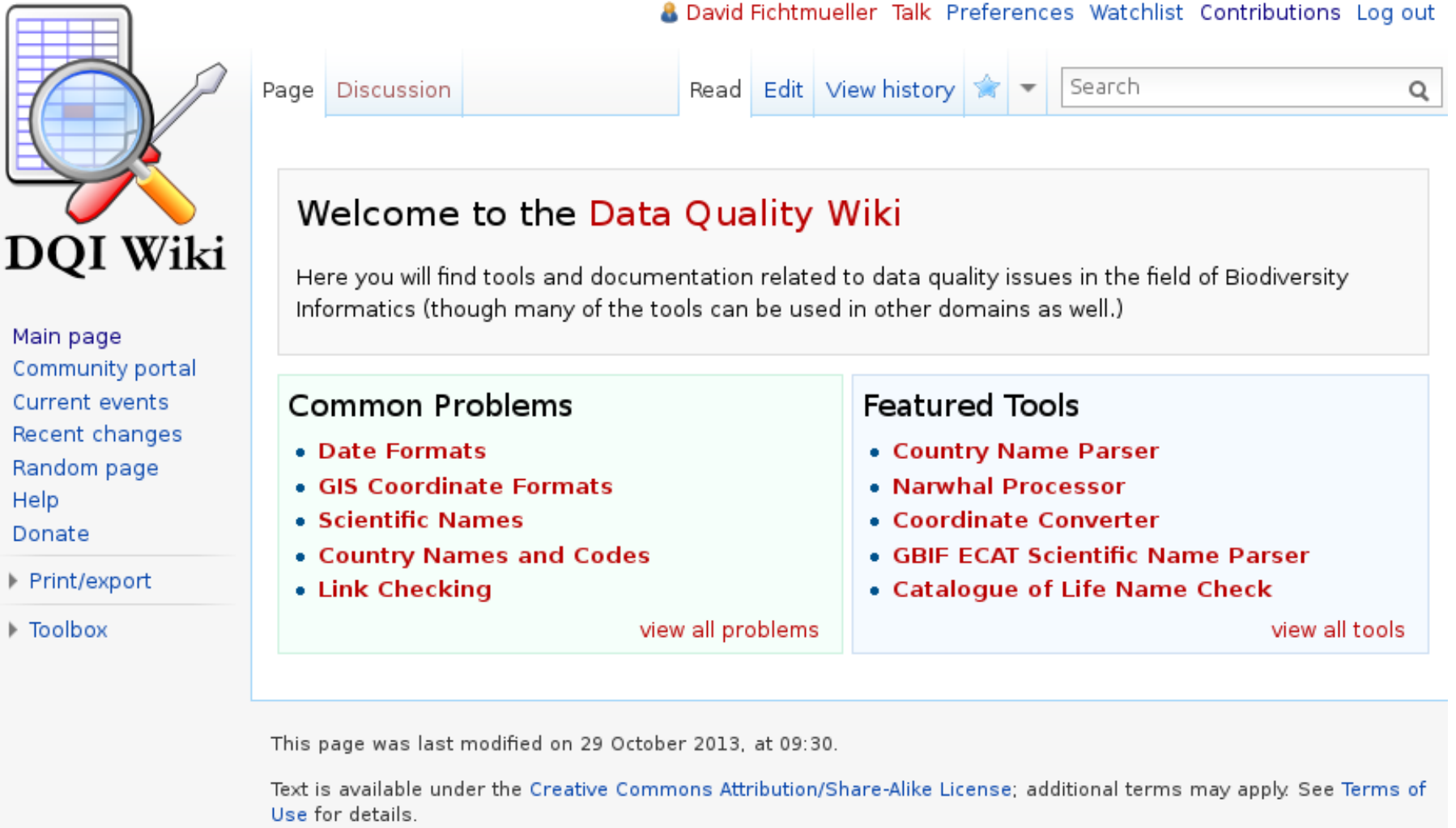
Independent of Programming Language
In a particular Format: XML, JSON, CSV, ...
Example: Dataset of Country Names

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>

Funding


OPEN
UP!


BioVeL


BiNHum
Biodiversitätsnetzwerk
des Humboldt-Rings


reBiND


DFG


DFG


ICTPSP


SEVENTH FRAMEWORK
PROGRAMME

Thank You!

Questions ?

Wiki: http://biowikifarm.net/dataquality

E-Mail: d.fichtmueller@bgbm.org