Component Report

Project Acronym: OpenUp!

Grant Agreement No: 270890

Project Title: Opening up the Natural History Heritage for Europeana

---

C2.1.1 – Collections Data Quality toolkit prototype

Revision: Final

Authors:

Anton Güntsch (BGBM)
Andreas Kohlbecker (BGBM)
Felix Hilgerdenaar (BGBM)
With contributions from the OpenUp! Technology Management Group

Project co-funded by the European Commission within the ICT Policy Support Programme		
Dissemination Level		
P	Public	X
C	Confidential, only for members of the consortium and the Commission Services	

REVISION HISTORY AND STATEMENT OF ORIGINALITY

Revision History

Revision	Date	Author	Organisation	Description
1	2011-06-01	Anton Güntsch & TMG	BGBM	1st design documentation for the Data Quality Toolkit on the TMG Scratchpad Site including a page for the compilation of integrity rules
2	2011-08-15	Anton Güntsch, Andreas Kohlbecker, Felix Hilgerdenaar & TMG	BGBM	Full specification of the prototype including the User Interface design.
2a	2011-08-26	Coordination Team	BGBM	Minor editing.
2b	2011-08-28	Coordinator	BGBM	Minor editing.

Statement of Originality:

This deliverable contains original unpublished work except where clearly indicated otherwise. Acknowledgement of previously published material and of the work of others has been made through appropriate citation, quotation or both.

Distribution

Recipient	Date	Version	Accepted YES/NO
TMG	2011-08-18	2	YES
Work Package Leader	2011-08-25	2	YES
Project Coordinator	2011-08-28	2a-2b	YES

C2.1.1 – Collection Data Quality Toolkit prototype

Please note: The following description has been copied (with minor changes) from the OpenUp! Scratchpad site of the Technology Management Group (TMG). The deliverable itself is the prototype, which is publicly available at http://services.bgbm.org/DataQualityToolkit.

Overview

The Data Quality Toolkit is an open web-based application for OpenUp! data providers and BioCASe (Biological Collection Access Service) providers in general performing data quality checks on their data. The system integrates a set of distributed quality services into a single and consistent user interface, thereby hiding the complexity of the individual services. In particular, the Quality Toolkit will integrate the Zoological and Botanical Quality services developed by WP 4 and WP 5 as well as the Data Integrity Services developed within WP 2.

The Data Quality Toolkit operates directly on a given BioCASe provider service installation. It pages through a subset of records (collection units) specified by the given user query (e.g. for a specific Genus) and applies a set of user-selectable quality testing procedures.

The result comes as an annotated ABCD-document containing all unit-records with one or more quality problems. (ABCD is a community XML standard for Access to Biological Collection Data.) Annotations explaining the problems are directly placed in the form of structured comments next to the elements they refer to. Using ABCD as a reporting format has two advantages over a proprietary format: 1) the connection between data and their annotations is directly visible and does not have to be "explained" using a different structure and 2) using ABCD opens opportunities for future developments of software components which automatically re-insert annotated data into provider databases.

Prototype & UI

The current Data Quality Toolkit prototype is available at http://services.bgbm.org/DataQualityToolkit.

It has a simple HTML User Interface offering fields for i) specifying the BioCASe Provider Installation to be analyzed, ii) selecting a set of data quality rules to be applied, and iii) filtering the subset of unit-records to be analyzed:

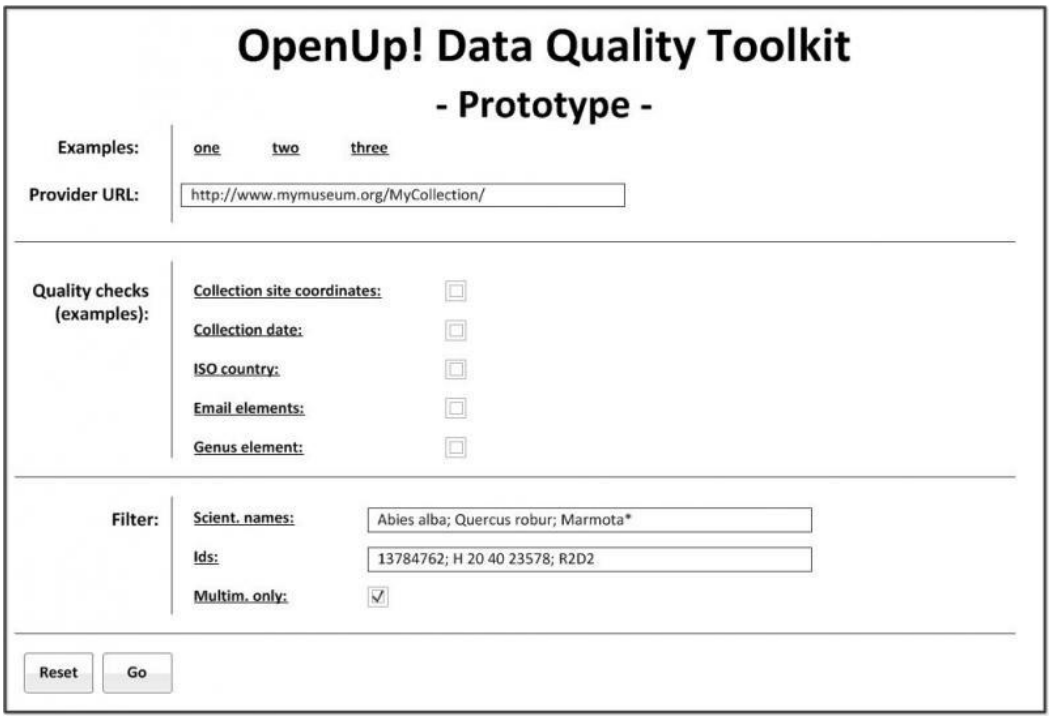

Fig.: Data Quality Toolkit User Interface (prototype)

The present state of the implementation is of limited usefulness because zoological and botanical quality services are not yet available and could not be integrated at this stage. Also, the set of rules implemented into the data integrity service is still very limited. However, the basic functionality of the system (construction of queries, paging through ABCD records, applying quality rules, compilation of the response document) is fully functional and demonstrated with the prototype.

Implementation

The Data Quality Toolkit implementation is based on Node.js, running on top of the Google V8 JavaScript Engine. This has some interesting implications as JavaScript is the only programming language needed for both the client and the server. Programming is done 'asynchronously' using 'callbacks' and non-blocking IO. This leads to a highly effective programming of concurrent processes. The individual software modules are:

- **HTTP server** provides basic functions of an HTTP server and controls the workflow.
- **HTTP client** communicates with the BioCASe providers and other servers.
- **XML parser** builds data structures from XML data.
- **Validator** uses these data structures to apply the data quality rules.
- **Rules module** contains the definitions of the individual rules in the form of JSON objects.
- **Config module** is used for rules mapping, paging, and annotations parameters.

Individual ABCD XML elements can receive multiple annotations from the application of several rules. The paging process for BioCASe provider installations is configurable with regard to the page size (the number of unit records per page).

Rule set development

The set of rules implemented in the Toolkit is still very limited. The rules will be continuously developed during the project. The latest state of the set will always be documented on the TMG website at http://open-up.eu/content/data-quality-toolkit-integrity-rules.

TO-DOs

- Thorough system testing.
- Add asynchronous access to the annotated result-documents (see ODIS specification at http://open-up.eu/content/openup-data-integrity-service-odis).
- Include botanical and zoological quality services.
- Add more rules for integrity tests (use ABCD-ESE mapping as one source of inspiration: http://open-up.eu/content/wp3-abcd-ese-mapping-workshop-june-21-22-bgbm).
- Adding a capabilities request to ODIS and generate HTML form for the client dynamically
- Doll up the design of the Data Quality Toolkit user interface.