Opening Up the Natural History Heritage for Europeana

D07 - DELIVERABLE 2.1.2

Project Acronym:	**OpenUp!**
Grant Agreement No:	**270890**
Project Title:	**Opening up the Natural History Heritage for Europeana**

Collections Data Quality toolkit production version

D07 – Deliverable 2.1.2

Revision: Final

Authors:

Anton Güntsch **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	

0 REVISION AND DISTRIBUTION 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, Felix Hilgerdenaar & TMG	BGBM	Full specification of the prototype including the User Interface design.
3	2012-02-17	Anton Güntsch, Felix Hilgerdenaar	BGBM	Collections Data Quality toolkit service up and running. Service URL distributed to TMG.
3a	2012-02-23	Coordination Team	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This deliverable contains original unpublished work except where clearly indicated otherwise.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 (G. Malarky, NHM)	2011-08-25	2	YES
TMG	2012-02-17	3	YES
Work Package Leader (G. Malarky, NHM)	2012-02-20	3	YES
Project Coordinator	2012-02-23	3a	YES

TABLE OF CONTENTS

0 REVISION AND DISTRIBUTION HISTORY AND STATEMENT OF ORIGINALITY 2
1 OVERVIEW 4
2 USER INTERFACE 4
3 OUTLOOK 5

Please note: The component itself is the Data Quality Toolkit User interface and service, which is publicly available at http://services.bgbm.org/DataQualityToolkit/.

1 OVERVIEW

The OpenUp! Data Quality Toolkit is a web based interface to a growing set of data quality services helping data providers to identify potential syntactical and semantic problems in their Metadata. By providing the tool as a web application on a central server hosted by the BGBM, we are minimizing problems linked to distribution and versioning of software.

Presently, the system includes quality checks for date elements, collection site coordinates, syntax of email and ISO country elements, zoological and botanical names, as well as multimedia object elements. Additional rules are continuously added on a website maintained by the OpenUp! Technology Management Group (TMG) and successively implemented as services and integrated into the User interface.

Results of quality checks are returned as annotated ABCD-documents with the annotations directly following the ABCD-elements they refer to. Using ABCD as the response format will allow us to generate different response formats using style sheet transformations. Being machine readable, it also opens the perspective to be able to feedback annotations automatically back into collection databases in the future.

The system has presently no User restrictions and is open to BioCASE providers who are part of the OpenUp! consortium as well as external providers. The URL of the User interface will be published on the OpenUp! helpdesk site by end of February 2012.

2 USER INTERFACE

The Data Quality Toolkit User interface is simple html-form (see fig. 1) with the following three sections for User entries and settings:

The **Provider settings** section has a field for entering the URL of a BioCASE provider software installation to be quality-checked. In addition, Users can select synchronous or asynchronous access to response documents using the “Sync” option. When using synchronous access, the quality report will be directly returned to the User once all quality checks have been performed. For large record-sets it is better to select asynchronous access which returns a URL to the response document before all quality checks have been performed. Users can then retrieve result documents later without having to wait for their display in the browser window used for sending the query.

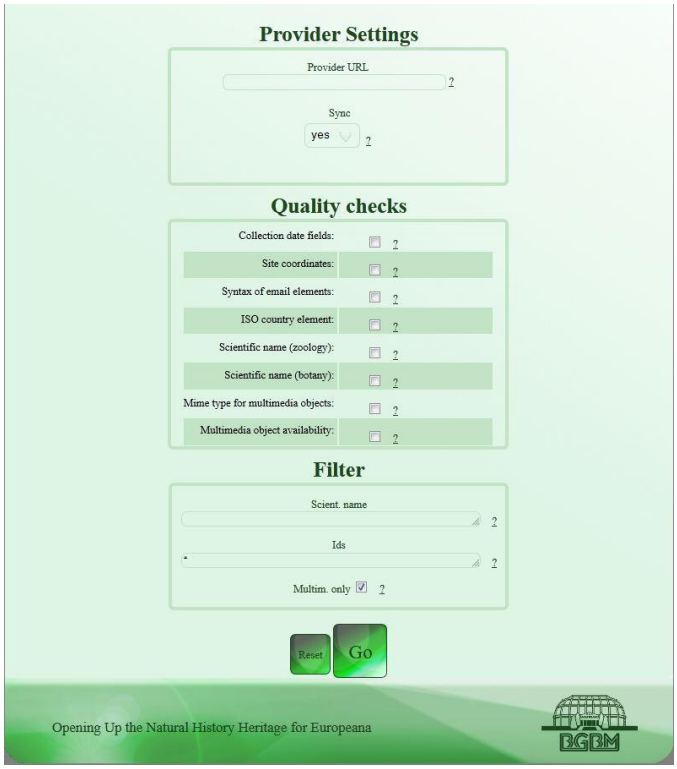

Fig. 1: Screenshot of the main form of the Data Quality Toolkit.

The **Quality checks** section allows Users to select specific quality checks they would like to perform on their collection Metadata. Presently, eight different quality services can be selected ranging from simple syntax checks for email elements to complex semantic checks for scientific names.

Finally, the **Filter** section is used to restrict quality checks to a subset of records available through a given BioCASE provider software installation. The “Scientific name” field can be used to restrict the checks to a particular species or genus and the “IDs” field restricts the quality checks to individual unit records.

Help texts have been implemented directly into the User interface. Each User field has an associated help text explaining its concept and giving examples where appropriate. In addition, response documents contain a header indicating whether the quality checks have been performed successfully and how many collection units have been analysed.

3 OUTLOOK

Both OpenUp! data quality services and toolkit will be continuously extended with additional rules over the course of the project based on new requirements from the OpenUp! provider community as well as the

availability of quality services which can be integrated into the infrastructure. Collaborations with similar endeavours will be of great help. The 7th framework project BioVeL (http://www.cs.cf.ac.uk/biovel/) dealing with workflows for Biodiversity Sciences has the development of quality workflows for primary biodiversity data as one of its crosscutting activities. OpenUp! and BioVeL will work together on a common data quality service architecture helping both projects to benefit from each other's developments. The reBiND project (http://rebind.bgbm.org) funded by the German research Foundation has a component for the development of mechanisms for correcting invalid ABCD-documents, which will partly rely on automatic detection of integrity problems in collection Metadata. OpenUp! and reBiND will develop integrity services in a way that services can be used mutually. We are highly optimistic that these synergies will allow us to sustain the development and maintenance of quality services that are of great benefit to the OpenUp! and the BioCASE provider community in general.