Component Report

Project Acronym: **OpenUp!**

Grant Agreement No: **270890**

Project Title: **Opening up the Natural History Heritage for Europeana**

C3.3.1 – Data Integrity Service Up and Running

Revision: Final

Author:

Anton Güntsch (BGBM) and the BGBM Biodiversity Informatics Team

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	2011-03-28	Anton Güntsch & TMG	BGBM	1st compilation of integrity service requirements and a diagram showing the basic information flows on the TMG scratchpad-site.
2	2011-06-20	Anton Güntsch & BDI Team	BGBM	Full specification of the data integrity service mock-up
3	2011-11-21	Anton Güntsch & BDI Team	BGBM	Revised interface specification available on TMG scratchpad-site. Data integrity service up and running.
3a	2011-11-22	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This deliverable contains original unpublished work except where clearly indicated otherwise.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	2011-11-21	3	YES
Work Package Leader		3	YES
Project Coordinator	2011-11-22	3a	YES

C3.3.1 – Data Integrity Service Up and Running

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 functional service, which is accessible at http://open-up.eu/content/openup-data-integrity-service-odis*

Overview

The OpenUp! Data Integrity Service (ODIS) is a rule-based system for checking the syntactical correctness of ABCD data accessible *via* a given BioCASE provider service installation. It can be used as a stand-alone service and will also be part of the Collections Data Quality Toolkit together with services for botanical and zoological names. The set of rules will be continuously developed at http://open-up.eu/content/data-quality-toolkit-integrity-rules.

The access point to the ODIS service is http://services.bgbm.org/odis/.

A user friendly query form is available at http://services.bgbm.org/DataQualityToolkit

ODIS requests

ODIS requests can be sent by using HTTP Post with the following f key/value pairs:

Key	Value	Description
ProviderURL	URL	The access point of the BioCASe provider installation to be analysed
Sync	true/false	Decides whether the client expects a synchronous response for the given request. If Sync=true is submitted, the service will return a full ABCD document containing ABCD unit-records containing annotations encoded as XML-comments. If Sync=false, the service will return a URL from which an annotated ABCD-document can be downloaded.
Rules	number as hex string	Selects the rules used for validating. If bit n of the number equals 1, rule n will be selected. Otherwise rule n will not be used. The available rules are listed on http://open-up.eu/content/data-quality-toolkit-integrity-rules. Notice that rule 0 is not defined. Example: rules 1 and 2 are selected → bin: 110; hex: 6 → the string "6" will be sent to the service.
BioCASeRequest	BioCASe-Protocol 1.3 search request	The BioCASeRequest value is used to restrict quality control to a reasonable number of records, for example by considering a specific genus only. As the BioCASe protocol request format has already the necessary mechanisms to filter records in a SQL-like XML-encoded syntax, we are using this very protocol syntax here as well. For a detailed description of the BioCASe protocol see http://www.biocase.org/products/protocols/.

ODIS responses

The ODIS response is an annotated ABCD-document, which can be delivered directly or as a downloadable file, depending on the value of the Sync-argument in the ODIS request. If sync=false has been requested, the client receives the URI of the ABCD document to be downloaded. In this case, trying to fetch the result-document before it is completely available for download will produce an HTTP server error (temporarily not available).

For storing annotations generated by the integrity checker service we use XML comments (<!-- comment -->) with the assumption that an annotation is always directly following the opening tag of the element it belongs to. The following example shows an annotation of a scientific name belonging to a determination in an ABCD 2.06 document:

```
<?xml version='1.0' encoding='UTF-8'?>
<DataSets xmlns='http://www.tdwg.org/schemas/abcd/2.06'>
<DataSet>
  <TechnicalContacts></TechnicalContacts>
  <ContentContacts>
    <ContentContact>
      <Name>John Smith</Name>
      <Email>j.smith@NaturalHistoryCollection.org</Email>
    </ContentContact>
  </ContentContacts>
  <Metadata>
    <Description>
      <Representation language='en'>
        <Title>herbarium collection</Title>
      </Representation>
    </Description>
    <RevisionData>
      <DateModified>2001-03-01T00:00:00</DateModified>
    </RevisionData>
  </Metadata>
  <Units>
    <Unit>
      <SourceInstitutionID>BEBOP</SourceInstitutionID>
      <SourceID>HerbCol</SourceID>
      <UnitID>1136</UnitID>
      <Identifications>
        <Identification>
          <Result>
            <TaxonIdentified>
              <ScientificName>
                <FullScientificNameString>Calendula arvensis (Vaill.) L.</FullScientificNameString>
              </ScientificName>
            </TaxonIdentified>
```

```
      </Result>
    </Identification>
    <Identification>
      <Result>
        <TaxonIdentified>
          <ScientificName>
            <FullScientificNameString>
              <!--
                    <Annotations>
                      <Annotation>
                        <Context>OpenUp</Context>
                        <ISODateTime>2011-03-29T12:24.321Z</ISODateTime>
                        <MethodOrAgent>ODIS V0.7</MethodOrAgent>
                        <Type>Warning</Type>
                        <Message>Scientific name string seems to be malformed</Message>
                        <Suggest>Calendula incana Willd.</Suggest>
                      </Annotation>
                    </Annotations>
                    -->
                Calendula Incana Willd.</FullScientificNameString>
            </ScientificName>
          </TaxonIdentified>
        </Result>
      </Identification>
    </Identifications>
  </Unit>
 </Units>
</DataSet>
</DataSets>
```

The precise syntax of an annotation is defined with the schema annotation.xsd. Each annotation consists of the following 6 elements:

Element name	Description
Context	A short but meaningful string indicating the context of an annotation. For annotations in the context of OpenUp! this will simply be "OpenUp". The context can be used for example for searching for annotations that are relevant for a specific purpose.
ISODateTime	ISO8601 encoded date and time of generation of this annotation
MethodOrAgent	An indication of who or what created an annotation. This might be a software application, a service, or a person. In the context of OpenUp, this will be something like "ODIS V1.7".
Type	Type of annotation: "comment" or "warning".
Message	A free-text representation of the annotation (for human consumption).
Suggest	In some cases the service might be able to suggest a correction of the annotated element.

If the client chooses asynchronous access (sync=false), the server/service will return the HTTP status code 303 ("see other") and the HTTP header "Location" with the new URI for download. The body will also contain this URI as an html-encoded link. If a client tries to access the response file before it has been put up for download, the server will return the http status code 503 ("service unavailable") and the http header will contain the field "retry-after" set to the number of seconds to wait before the response document can be expected (currently ODIS returns a constant value). For a detailed list of HTTP status codes used by OpenUp! validation services please refer to A common API for name data quality services.

Implementation

To facilitate event-based programming, ODIS is completely implemented in Javascript using node.js (http://www.nodejs.org/).

Documents may contain several DataSets which in turn contain multiple Units. Based on a given request, the service fetches the data from a particular BioCASe provider installation and propagates response pages directly to the Validator-Component of the data quality service. The service then creates rule instances for each ABCD unit. Rules will then be processed asynchronously. Once all rules have been processed, the response ABCD document will be assembled and either directly returned or put up for asynchronous download (depending on the sync parameter passed to the service).

Next steps

The Data Integrity Service is stable and already used by the prototype data quality toolkit. Further developments of the service will mainly focus on the incorporation of additional quality rules as well as performance and improvement of the user-friendliness of the data quality response documents.