Opening Up the Natural History Heritage for Europeana

COMPONENT REPORT

Project Acronym: **OpenUp!**

Grant Agreement No: **270890**

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

C7.2.2 Analysis of documentation, gaps and plan for needed additional documentation

Revision: 5a (Final)

Authors:

Janno Jõgeva	UT-NHM
Boris Jacob	MRAC
Patricia Mergen	MRAC

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	2012-03-08	J. Jõgeva; B. Jacob	UT-NHM; MRAC	Initial ideas and testing of automated analysis procedures.
2	2012-04-20	J. Jõgeva; B. Jacob	UT-NHM; MRAC	Update with findings of C7.2.1 Identification of existing documentation.
3	2012-05-31	J. Jõgeva; B. Jacob	UT-NHM; MRAC	Incorporation of new documentation, fixing of broken hyperlinks.
4	2012-08-29	J. Jõgeva; B. Jacob	UT-NHM; MRAC	Final Draft, incorporation of findings of 1st Intermediate Review Report.
5	2012-08-29	P. Mergen	MRAC	Revision
5a	2012-09-20	Coordination Team (A. Michel/ P. Böttinger)	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This deliverable contains original unpublished work except where clearly indicated otherwise.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
Work Package Leader WP7 (Patricia Mergen, MRAC)	2012-08-29	5	Yes
Project Coordinator (W. Berendsohn, BGBM)	2012-09-27	5a	Yes

Table of Contents

1 INTRODUCTION

This component gives an update on the work done for *C7.2.1 Identification of existing documentation* in the chapters *methodology* and *documentation sources published.* Based on this work an analysis of the identified documentation has been done, gaps have been identified and a plan for additionally needed documentation is presented.

2 UPDATE ON C7.2.1 – IDENTIFICATION OF EXISTING DOCUMENTATION

2.1 Methodology

In Component C7.2.1 a set of required metadata was established[1], which has been enhanced for this component to allow a better checking and gap analysis of the documentation. The four additional fields are:

1. **No.:** Unique number for every record.
2. **Audience:**[2] Users to whom this document is aimed to: Scientists; Collections Manager / Technicians; PR Department / Communication / Education; Legal Department / Legal Advisor; Direction / Policy Maker.
3. **Relevance:** Indicates the relevance for the OpenUp! project participants, indicated by the values "Direct" and "Indirect". The decision is based on an analysis of the content of the documents, taken into account the experience of the Helpdesk team in answering questions within the OpenUp! project and their broader experience from the work in the Biodiversity Informatics domain.
4. **URL updated:** Indicates if the URL has changed since it was first registered. Possible values "yes" and "n/a".

2.2 Documentation Sources

In Component C7.2.1 we identified 14 different sources of documentation, the following three are an update to this list.

2.2.1 BDTracker

The Biodiversity Service & Application Tracker is a database of software, tools and resources for taxonomists. The site was set up during the EDIT project and is now maintained in the Drupal infrastructure at the Royal Museum for Central Africa at http://bdtracker.cybertaxonomy.africamuseum.be/

[1] See C7.2.1 Identification of existing documentation, p. 3f.

[2] The categorisation of users is based on Annex I: Description of Work, Part B, p.29: European Natural history collections, incl. botanic gardens; Scientists (specifically in taxonomy, ecology, biodiversity conservation, and sustainable land management) and its further discussion in C7.1.2 Network Helpdesks Strategy and Coordination, p.5f.

2.2.2 ICT Policy Support Programme

The ICT PSP is part of the Competitiveness and Innovation framework Programme (CIP) at the European Commission. Documentation can be found in the "library" of the homepage at http://ec.europa.eu/information_society/activities/ict_psp/library/index_en.htm.

2.2.3 PESI

Public project deliverables from the Pan-European Species directories Infrastructure can be accessed at http://www.eu-nomen.eu/pesi/remository?func=select&id=75.

3 ANALYSIS OF DOCUMENTATION

The table below shows the analysis of the documentation between the publication of C7.2.1 in April 2012 and the end of June 2012. The documents are registered in a database on the OpenUp! Helpdesk at http://openup.helpdesk.africamuseum.be/documentation, and as spreadsheet on Google Drive[3].

Table 1: Analysis of documentation (as of June 2012)

Theme / Aspect	**Amount April 2012**	**Amount June 2012**
Number of resources	273	316
Number of pages (Where applicable)	5782	6844
Relevance direct/indirect	n/a	78/238
Published in (Language)*		
English	n/a	289
Spanish	n/a	11
French	n/a	11
German	n/a	3
Italian	n/a	2
Romanian	n/a	2
Polish	n/a	2

[3] https://docs.google.com/spreadsheet/ccc?key=0Ak5RuKB4k4wjdDY3MlRXdkpUSE8yM3Y0eXdFNUowcmc&pli=1#gid=0 the older version from C7.2.1 (April 2012) is accessible at https://docs.google.com/spreadsheet/ccc?key=0AoIrOUcsTAbidGVma0tTLWZBVEZqZWRIUU1NTDll%20TlE#gid=0

Theme / Aspect	**Amount April 2012**	**Amount June 2012**
Bulgarian	n/a	1
Chinese	n/a	1
Greek	n/a	1
Hungarian	n/a	1
Lithuanian	n/a	1
Portuguese	n/a	1
Target user group (Audience)*		
Collections Manager / Technicians	n/a	157
Direction / Policy Makers	n/a	88
Legal Department / Legal Advisor	n/a	80
Scientists	n/a	80
PR Department / Communication / Education	n/a	80
Published as (Type)*		
PDF	202	243
Website	24	26
Video	13	13
Slide presentation	10	10
Wiki	6	6
Resource Centre	2	2
Code Repository	2	2
Blog	1	1
Themes*		
Projects and Networks*	206	275
EDIT	48	48
Europeana	38	50

Theme / Aspect	Amount April 2012	Amount June 2012
BHL-Europe	30	51
GBIF	25	56
STERNA	24	36
OpenUp!	22	40
European IPR-Helpdesk	11	11
PESI	n/a	9
BioCASe	6	4
TDWG	4	4
CETAF	3	3
FP7	3	4
Natural Europe	2	2
ICOM	2	2
Species 2000	1	1
USEandDIFFUSE	1	1
ENBI	1	1
ICT Policy Support Programme	n/a	2
Technical questions*	77	135
Database systems	33	43
General	14	17
Data exchange schemas and protocols	14	14
BioCASe Provider Software	13	13
Multimedia content	3	3
Copyright and IPR questions	65	70

3.1 De-duplication of records in the list of documents

The de-duplication algorithm is based on URL comparison. And one result of this check was, that the overall final number of documents as of April 2012 has – with 273 instead of 300+ documents – been lower than originally stated. However, after adding 43 more documents to the list of documentations, the final number of documents is 316.

3.2 Checking and correction of hyperlinks

From the 273 documents mentioned in C7.2.1 an amount of 44 (~16%) were no longer available under the URL first registered. This is a fairly high amount and would be bad news for the sustainable access of documentation. But it seems to be more a singular event than a systematic problem of access. All broken or misleading hyperlinks were from Europeana documentation, which by the end of last year introduced http://pro.europeana.eu/ as a new service has been set up where users can find documentations from Europeana and other projects related to Europeana. A lot of documents which have previously been accessible *via* the Europeana v.1 website are now accessible *via* europeana.pro. In general we can state that over a period of two months the majority of the hyperlinks and providing documentation sources stayed stable.

3.3 Extraction of PDF metadata

The majority of the identified documentation is available in the Portable Document Format (PDF). For the automated extraction of the metadata out of PDF documents, scripts were implemented. The only meaningful data that could be extracted on a large scale was the number of pages. Different metadata, like title, version, institution, and author(s), has also been extracted automatically from the title page of the documents, in case of standardised layout. This data had to be checked manually afterwards. The work done showed that there is high potential of automated extraction and analysis of data from those nine documentation sources if the internal PDF metadata fields like title, author, subject, keywords, etc. are used, and if more meaningful information is presented on the title page in a human readable and structured way, so that it can easily be interpreted automatically as well.

4 GAPS ANALYSIS AND PLAN FOR NEEDED ADDITIONAL DOCUMENTATION

Documentation, both existing and newly produced documentation, needs to be freely, openly, and easily accessible for everyone online. It has to be stable and accessible in the long term, so that the source can be referenced and the document consulted continuously. The documentation itself has to describe the subject sufficiently and give up-to-date information about it. It should be written in a language that the target audience understands.

4.1 Accessibility, stability

Open, free and easy online access to the documentation is a major requirement. Furthermore documentation needs to be stable and accessible on the long term. All documentation that has been identified in C7.2.1 has been accessible online openly and for free, there were no barriers in form of a login required.

There has been an issue with the stability of hyperlinks to documentation from Europeana as described in chapter 3.2 “Checking and correction of hyperlinks”. The hyperlinks were updated, but the general problem of long term availability to documentation stays. Most of the documentation from the ENBI project, for example, were no longer accessible. For the production of new documentation in the OpenUp! project that means that relevant external information needs to be incorporated into the OpenUp! documentation, rather than just linked to, in case the original source is no longer available.

Another aspect of accessibility to documentation relates to the file format. The vast majority of the documentation is available in PDF format (see table 1), which is a de facto open standard. All documentation from the OpenUp! project is also provided in PDF format. There is also information on websites on the Helpdesk, but this will be updated in the PDF document of the OpenUp! Guidelines. One of the most important sources of information in the OpenUp! project is the BioCASe documentation on The PyWrapper Wiki[4]. As it is a website it should for sustainability reasons be mirrored or made available as a PDF version as well.

To ensure long term availability the documentation created in OpenUp! needs to be mirrored and stored on several servers outside of the OpenUp! Website and the OpenUp! Helpdesk. This also supports the performance indicator No. 6 “Mirror the aggregation site as well as the documentation to at least three sites across Europe to ensure performance and availability”. There are several possibilities for that like the documentation system on Europeana Pro[5], the GBIF Resource Center[6], or the free reference manager and social network Mendeley[7].

4.2 Topic

In regard to the OpenUp! project topic gaps can happen if tools, workflows or agreements made in the course of the project or by third parties are not documented in time. At this point in the project, the BPS and mapping of the data sources is functional for zoological and botanical data sources[8], we have not identified any gap in this regard. However, the technical development on the OpenUp! architecture and related technologies and schemata is not yet finished, that means that some documentations are in the process to be changed in the future. A complete documentation of OpenUp! services in different languages will be available by the end of the project[9].

[4] See http://wiki.bgbm.org/bps/index.php/Main_Page

[5] See http://pro.europeana.eu/

[6] See http://www.gbif.org/orc/

[7] See http://www.mendeley.com/

[8] See “D12 Local zoological providers software and metadata mapping functional for all content data sources”, and “D13 Local botanical providers software and metadata mapping functional for all content data sources

[9] C7.2.5 Availability of complete documentation of OpenUp! services in diff. languages on the helpdesk website (M36)

The BioCASe documentation on the PyWrapper wiki is up to date with the last BioCASe Provider Software version 3.2. The ABCD schema is well documented and the documentation on the EFG extension is published as an OpenUp! component. The gap in documentation on the Europeana Data Exchange Agreement (DEA) has been closed end of last year with the publication of the Europeana Licensing Framework. And the upcoming change in the metadata model by Europeana from Europeana Semantic Elements (ESE) to the Europeana Data Modell (EDM) is documented in several technical and overview documents by Europeana, which will be adapted to OpenUp! specific needs in the future.

4.3 Up-to-datedness

The OpenUp! documentation and documentation on related tools, workflows and/or agreements need to reflect the state of the art. With the existing documentation no gap has been identified. As mentioned in chapter 4.2 "*Topic*", however, tools are still in development and need to be documented in the process of the project. The OpenUp! Helpdesk is working closely with the partners within the OpenUp! project to ensure that the documentation meets the requirements set up in this document. There is also close collaboration with Europeana, and the helpdesks from BioCASe and GBIF, as well as with TDWG to identify possible up-to-datedness gaps and address them.

4.4 Language

A vast majority of the documentation identified is available in the English language. For the course of the project itself and with the original project partners this is not a gap. In regard of the sustainability and re-use of the project results by future partners or technology users, this could be a gap, which will be addressed by the publication by the end of the OpenUp! project (C7.2.5). The plan is to have documentation with key information for new partners available in different languages and refer then to the original documentation in English.

5 OUTLOOK

For month 32 (October 2013), newly produced documentation will be made available on the helpdesk website (D21/C7.2.4), and at the final stage of the project in month 36 (February 2014), the complete documentation of OpenUp! services will be available on the website in different languages (C7.2.5). In these documents we will also address the suggestions made during the first OpenUp! review meeting, that

> "The project outcomes on the metadata, tools for quality assurance and enrichment and ingest to Europeana as well as costs analysis of the whole process could be helpful for institutions which would like to apply the tools and services in the future."

In this task the Helpdesk will continue the close collaboration with the partners within the OpenUp! project, as well as with Europeana, and the helpdesks from BioCASe and GBIF, the GBIF nodes, and TDWG. Especially with Europeana and GBIF discussions will take place to make available the OpenUp! documentation on their servers, to meet the performance indicator No. 6, to mirror the documentation to at least three sites across Europe, to ensure long time accessibility also after the end of the OpenUp! project.

6 LIST OF REFERENCES

- C7.1.2 - Network Helpdesks Strategy and Coordination. Pere Roca Ristol, Boris Jacob, Anne-Sophie Archambeau, Franck Theeten, James Davy, Patricia Mergen: http://open-up.eu/sites/open-up.eu/files/u2/C7%201%202_Network%20Helpdesks%20Strategy%20and%20Coordination.pdf
- C7.2.1 - Identification of existing documentation. Janno Jõgeva, Boris Jacob, Anne-Sophie Archambeau, Pere Roca Ristol, Franck Theeten, Hanna Koivula, Patricia Mergen: http://open-up.eu/sites/open-up.eu/files/u2/C721_Identification_of_existing_documentation_incl_annex_UT-NHM_v4a.pdf
- D11 - D7.2.3 Enriched + checked existing documentation in orig. language (website): http://open-up.eu/sites/open-up.eu/files/u2/D11-D723_Enriched_%2B_checked_existing_documentation_in_orig_language_%28website%29_MRAC_v3ac.pdf
  & Annex: http://open-up.eu/sites/open-up.eu/files/u2/D11-D723_Annex_Enriched_%2B_checked_existing_documentation_in_orig_language_%28website%29_MRAC_v3a.pdf