EUR-Lex 2012 Data Extraction using Web Services

Office des publications de l'Union européenne

DOCUMENT HISTORY

DOCUMENT HISTORY		
Version	**Release Date**	**Description**
0.01	24/01/2013	Initial draft
0.02	01/02/2013	Review
1.00	07/08/2013	Version 1.00

TABLE OF CONTENTS

LIST OF TABLES

REFERENCE AND APPLICABLE DOCUMENTS

This section contains the lists of all reference and applicable documents. When referring to any of the documents below, the bracketed reference will be used in the text, such as [R01]

REFERENCE DOCUMENTS				
Ref.	**Title**	**Reference**	**Version**	**Date**
R01	ELX-WS-Web Service User Manual-v0 01.doc	Web Service User Manual	0.01	17/04/2012
R02	CEM-EUM-End User manual-v1 02 - SL.doc	CEM-EUM	1.02	07/11/2012

Table 1: Applicable Documents

ABBREVIATIONS AND ACRONYMS

ABBREVIATIONS AND ACRONYMS	
Abbreviation	**Meaning**
WSDL	Web Services Description Language
XML	Extensible Markup Language
XSD	XML Schema
ECAS	European Citizen Action Service
OWL	Web Ontology Language
IDOL	Intelligent Data Operating Layer, http://www.autonomy.com

Table 2: Abbreviations and Acronyms

DEFINITIONS	
Term	**Meaning**
CELLAR	The CELLAR can be seen as the content repository module. This application aims to store all content and metadata needed by the Publications Office and its applications.
Ontology	A set of concepts within a domain, and the relationships among those concepts – according to a format called the Web Ontology Language (OWL).
IDOL server	The IDOL Server collects indexed data from connectors and stores it in its proprietary structure, optimized for fast processing and retrieval of data

Table 3: Definitions

1 INTRODUCTION

1.1 PURPOSE OF THE DOCUMENT

The aim of this document is to describe how to use the web service provided by the EUR-Lex 2012 system and the underlying layers, e.g. CELLAR repository. This document may be consulted by all the users interested in the usage of those web services.

1.2 SCOPE OF THE DOCUMENT

This document refers to the EUR-Lex web services functionality as defined in the Functional Specifications Document [R01] and the CELLAR WEB API services functionality as defined in [R02]

1.3 INTENDED AUDIENCE

The present document is intended to be read by the EUR-Lex 2012 registered public users who are interested in accessing and extracting data from EUR-Lex 2012 system in XML format as described in [R02]

1.4 STRUCTURE OF THE DOCUMENT

The document is organised as follows:

- **Chapter 1 - Introduction** provides an overview of the purpose of this document, and the intended audience;
- **Chapter 2 - Comparison between EUR-Lex and Cellar web API** explains how the different functionalities related to web services that are accessible from the new EUR-Lex 2012 interface and CELLAR repository.
- **Chapter 3 – Data extraction strategy** explains how to use the EUR-Lex 2012 web services together with the CELLAR WEB API services to extract the data that meets each user individual needs*;*
- **Chapter 4 – Conclusion***;*

2 COMPARISON BETWEEN EUR-LEX AND CELLAR WEB API

The new EUR-Lex 2012 provides a web service opened to external user in order to search for legal content. It is a convenient way to perform search queries on the legal content of EUR-Lex without having to use the different search forms present on the website. The search results provided are contained in a structured XML compliant with a specific schema definition.

In order to use the web service, the user must be registered in the new EUR-Lex and must subscribe to the web service.

The CELLAR web API allows performing different operations on the CELLAR. Such API encapsulates all the HTTP calls to the CELLAR and exposes convenient methods allowing the user to easily retrieve the requested content.

A quick comparison between EUR-Lex and Cellar web API is presented below.

2.1 CELLAR WEB API SERVICES

The following summarizes the specifics of the CELLAR WEB API services.

- Search service: internal, not accessible to the external users
- Metadata: access to all metadata of the branch (the display of the notice in EUR-Lex is based on the branch notice), tree or RDF notice
- Documents: the documents can be retrieved through web service
- Coverage: all contents in Cellar
- Registration mechanism: no
- Versioning: a change in the ontology is directly reflected in the web service interface.

2.2 EUR-LEX 2012 WEB SERVICES

The user has to register in order to have access to the service. An administrator will check the registration data and allow or forbid the use of the web services.

The link to the registration form is available on the web service page when the user is not yet registered to the service.

The following summarizes the specifics of the CELLAR web API services.

- Search service: based on EUR-Lex Expert queries syntax and IDOL search engine.
- Metadata: Access to all metadata of the indexation notice (the display of the list of results in EUR-Lex is based on the indexation notice)
- Documents: the documents cannot be retrieved by the WS (but can be retrieved using the URI server or Cellar WS)
- Coverage: all contents in Cellar
- Registration mechanism: yes (login/password required to access the WS)

- Versionning: a change in the ontology is not directly reflected in the WS interface.

3 DATA EXTRACTION STRATEGY

This section describes how to use the various web services available to achieve a custom-tailored data extraction that fits the particular needs of each EUR-Lex user; some examples are provided further below on how to use the web services from the different layers. The query done with the web service must respect a particular syntax. The request that have to be provided is detailed in the reference documents [R01] and [R02], mentioned in the beginning of this manual.

3.1 USING EUR-LEX 2012 WEB SERVICES ALONE

As mentioned in Chapter 2, the EUR-Lex 2012 web service provides access to the indexation notice – a sub-set of fields collected and indexed by the IDOL search engine and which are searchable. **Thus, EUR-Lex 2012 web services are recommended to perform a search on EUR-Lex content (based on Expert search queries) and retrieve list of Cellar identifiers**. Those identifiers can be used later on to extract the content in the required format(s).

3.1.1 XML INPUT

The input of the web service must respect the soap envelop structure. The soap envelop is composed of a *Header* and *Body* elements. The content of *Body* must respect an XSD defined by EUR-Lex. Moreover, the *Header* must contain information about security. The template to be used by the web service is the following:

```
<soap:Envelope xmlns:sear="http://eur-lex.europa.eu/search"
xmlns:soap="http://www.w3.org/2003/05/soap-envelope">
   <soap:Header>
      <wsse:Security soap:mustUnderstand="true" xmlns:wsse="http://docs.oasis-
open.org/wss/2004/01/oasis-200401-wss-wssecurity-secext-1.0.xsd">
         <wsse:UsernameToken wsu:Id="UsernameToken-3" xmlns:wsu="http://docs.oasis-
open.org/wss/2004/01/oasis-200401-wss-wssecurity-utility-1.0.xsd">
            <wsse:Username>${EUR-Lex username}</wsse:Username>
            <wsse:Password Type="http://docs.oasis-open.org/wss/2004/01/oasis-200401-
wss-username-token-profile-1.0#PasswordText">${WS password}</wsse:Password>
         </wsse:UsernameToken>
      </wsse:Security>
   </soap:Header>
   <soap:Body>
      <sear:searchRequest>
         <sear:expertQuery>${expert query}</sear:expertQuery>
         <sear:page>${page}</sear:page>
         <sear:pageSize>${pageSize}</sear:pageSize>
         <sear:searchLanguage>${search language}</sear:searchLanguage>
      </sear:searchRequest>
   </soap:Body>
</soap:Envelope>
```

You can find below the description of each required field:

- `wsse:Username`: Username used by the user to log in EUR-Lex. This is provided by ECAS.
- `wsse:Password`: The password received by email once the user registration to the web services has been accepted by an administrator.
- `sear:expertQuery`: The query used by the search engine to retrieve documents. The following chapter explains the syntax of the expert search.

- `searc:page:` The page of the search results, as the system uses the pagination. For instance, the user is allowed to retrieve the 10 results of the second page. It corresponds to the documents from 10 to 20.
- `sear:pageSize:` The size of the page used in the pagination.
- `sear:searchLanguage`: The search language. When using a web service client, a list of value is provided for the search language.

All these fields are required to perform a query. It is really important to respect the defined namespaces.

For further details and examples on how to invoke EUR-Lex 2012 web services please consult [R01] reference document.

3.1.2 XML RESULT

The result of the web service is a *Soap Envelope*. That envelop contains a *Body* element which contains XML elements that respect an XSD that can be retrieved from EUR-Lex.

The root element in the *Body* is the `searchResults` element. It contains:

- *numhits*: the number of results in the page;
- *totalhits*: the total number of results related to the query;
- *page*: the current page of results;
- *language*: the search language;
- *result*: the element related to a result. This element contains:
  - *reference*: the CELLAR reference;
  - *rank*: the index of the document;
  - *document_link*: links to the manifestations of the document (at word, tiff, html or pdf format), if any manifestation of the document exist;
  - *content*: the metadata of the document;

You can find below a sample of the result.

```
<S:Envelope xmlns:S="http://www.w3.org/2003/05/soap-envelope">
   <S:Body>
      <searchResults xsi:schemaLocation="http://eur-lex.europa.eu/search
http://localhost:7001/eurlex-frontoffice/eurlex-ws?xsd=3" xmlns="http://eur-
lex.europa.eu/search" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
         <numhits>10</numhits>
         <totalhits>1946</totalhits>
         <page>1</page>
         <language>en</language>
         <result>
            <reference>eng_cellar:93836665-712f-4444-a1e6-dadad5607e80_en</reference>
            <rank>1</rank>
            <content>
               <NOTICE>
                  <EXPRESSION>
                     <EXPRESSION_TITLE>
                        <VALUE>Decision on the …</VALUE>
                     </EXPRESSION_TITLE>
                     <EXPRESSION_USES_LANGUAGE>
                        <URI>
                           <IDENTIFIER>ENG</IDENTIFIER>
```

```
                </URI>
            </EXPRESSION_USES_LANGUAGE>
            …
```

The content element will contain some metadata of the document. The list of provided metadata depends on query:

- If the expert query contains a SELECT clause, the provided metadata will only be those present in the SELECT clause;
- If the expert query doesn't contain any SELECT clause, the list metadata provided will be the list defined in the default search profile of the user, in the EUR-Lex 2012 website.

For further details and examples on how to invoke EUR-Lex 2012 web services please consult [R01] reference document.

3.2 USING CELLAR WEB API SERVICES ALONE

The CELLAR WEB API allows performing different operations on the CELLAR. Such API encapsulates all the HTTP calls to the CELLAR and exposes convenience methods allowing the user to easily retrieve the requested content. They give access to:

- retrieve the tree notice of a work;
- retrieve the branch notice of a work;
- retrieve the object notice of an object (work, expression or manifestation);
- retrieve all the identifiers of a specific document (synonyms);
- retrieve the RDF/XML formatted metadata for a given resource;
- retrieve content streams of a work given a specific language and format

and how to invoke services on NAL/EUROVOC objects, namely:

- retrieve a dump;
- retrieve the supported languages;
- retrieve a concept scheme;
- retrieve the concept schemes;
- retrieve a concept;
- retrieve the concept relatives;
- retrieve the top concepts;
- retrieve the domains;

Thus, the CELLAR WEB API is very useful when one needs to access all metadata or documents displayed in a EUR-Lex notice. The EUR-Lex notice is built from the branch notice.

For further details and examples on how to invoke CELLAR WEB API services please consult [R02] reference document.

3.3 USING EUR-LEX WEB SERVICES AND CELLAR WEB API SERVICES TOGETHER

This is the recommended usage in order to get access to the complete data available in the EUR-Lex system. Data retrieval should be done in two steps:

I. Query EUR-Lex 2012 data using the EUR-Lex web services and retrieve the list of CELLAR identifiers.

II. For each CELLAR identifier invoke the CELLAR WEB API service to retrieve the required work, expression, manifestation or content stream.

3.4 DATA RETRIEVAL EXAMPLES

3.4.1 INVOKING EUR-LEX WEB SERVICES

As mentioned before, the input for the web service must respect the SOAP envelop structure. The example below returns the CELLAR_ID (internal identifier) for all the documents with CELEX numbers starting with **32013D**.

```
<soap:Envelope xmlns:soap="http://www.w3.org/2003/05/soap-envelope"
xmlns:sear="http://eur-lex.europa.eu/search">
<soap:Header>
      <wsse:Security soap:mustUnderstand="true" xmlns:wsse="http://docs.oasis-
open.org/wss/2004/01/oasis-200401-wss-wssecurity-secext-1.0.xsd"
xmlns:wsu="http://docs.oasis-open.org/wss/2004/01/oasis-200401-wss-wssecurity-utility-
1.0.xsd">
         <wsse:UsernameToken wsu:Id="UsernameToken-1">
            <wsse:Username>totomge</wsse:Username>
            <wsse:Password Type="http://docs.oasis-open.org/wss/2004/01/oasis-200401-
wss-username-token-profile-1.0#PasswordText">Qv7Z4fzkX3Y</wsse:Password>
         </wsse:UsernameToken>
      </wsse:Security>
   </soap:Header>
   <soap:Body>
      <sear:searchRequest>
         <sear:expertQuery>
                        SELECT CELLAR_ID
                        WHERE DN=32013D*
               </sear:expertQuery>
         <sear:page>1</sear:page>
         <sear:pageSize>2</sear:pageSize>
         <sear:searchLanguage>en</sear:searchLanguage>
      </sear:searchRequest>
   </soap:Body>
</soap:Envelope>
```

The response looks like shown below. The CELEX numbers for the first two (out of 17) documents are returned.

```
<S:Envelope xmlns:S="http://www.w3.org/2003/05/soap-envelope">
   <S:Body>
      <searchResults xsi:schemaLocation="http://eur-lex.europa.eu/search
http://eurlex-active-test/eurlex-frontoffice/eurlex-ws?xsd=3" xmlns="http://eur-
lex.europa.eu/search" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance">
         <numhits>2</numhits>
         <totalhits>17</totalhits>
         <page>1</page>
         <language>en</language>
         <result>
            <reference>eng_cellar:b26b0163-64ed-11e2-9693-01aa75ed71a1_en</reference>
```

```
            <rank>1</rank>
            <content>
               <NOTICE>
                  <WORK>
                     <URI>
                        <IDENTIFIER>b26b0163-64ed-11e2-9693-01aa75ed71a1</IDENTIFIER>
                     </URI>
                  </WORK>
               </NOTICE>
            </content>
         </result>
         <result>
            <reference>eng_cellar:0b692e9d-64ef-11e2-9693-01aa75ed71a1_en</reference>
            <rank>2</rank>
            <content>
               <NOTICE>
                  <WORK>
                     <URI>
                        <IDENTIFIER>0b692e9d-64ef-11e2-9693-01aa75ed71a1</IDENTIFIER>
                     </URI>
                  </WORK>
               </NOTICE>
            </content>
         </result>
      </searchResults>
   </S:Body>
</S:Envelope>
```

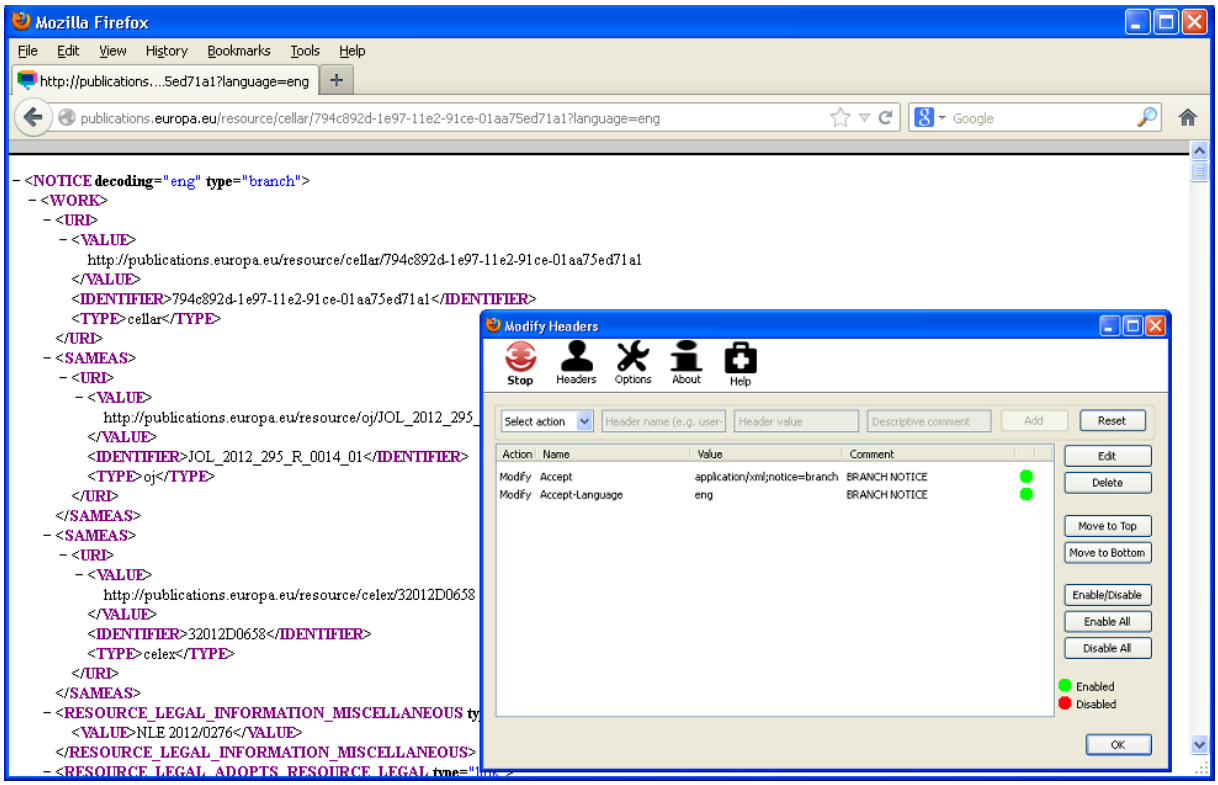
3.4.2 RETRIEVING THE XML BRANCH NOTICE THROUGH THE CELLAR API

In order to retrieve the Branch notice for each CELLAR_ID the EUR-Lex 2012 web service has returned, a GET HTTP request has to be fired to the appropriate (valid) production system and object. Below is shown an example in Firefox with the HTTP headers modified as needed. For more information please see [R02].

For instance, the URL for retrieving the branch notice for CELAR id **794c892d-1e97-11e2-91ce-01aa75ed71a1** is:

http://publications.europa.eu/resource/cellar/794c892d-1e97-11e2-91ce-01aa75ed71a1?language=eng

and the result is as shown below:



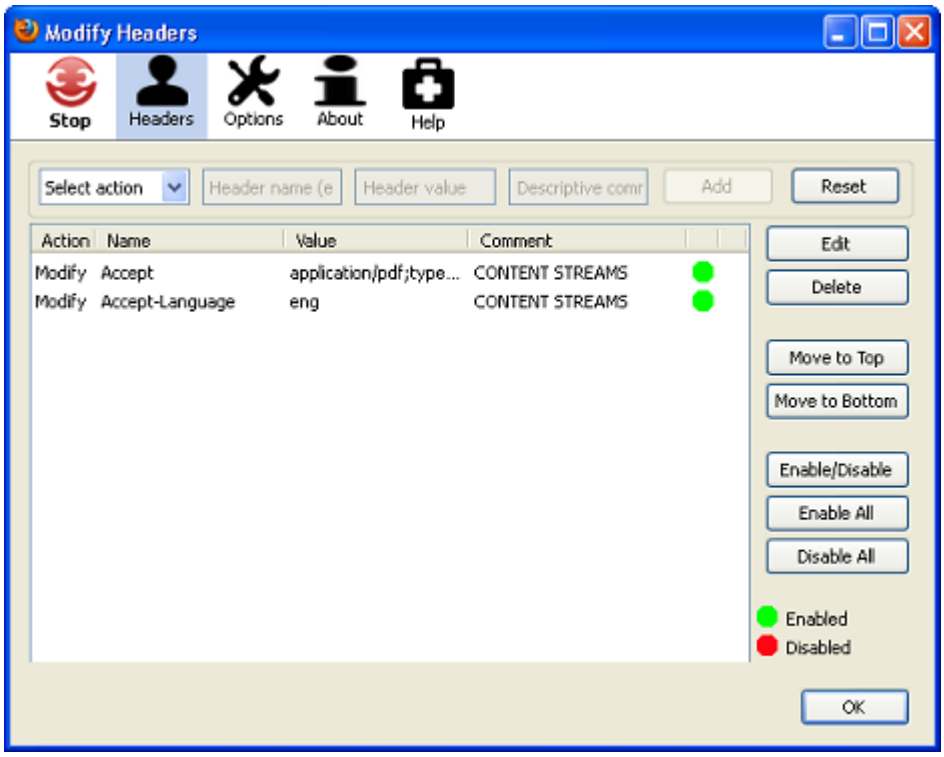
3.4.3 RETRIEVING THE CONTENT STREAM THROUGH THE CELLAR API

The content stream for each document can be obtained in a similar matter as the branch notice, via the same URL, but with the appropriate HTTP headers set.

The URL for the above document would be:

http://publications.europa.eu/resource/cellar/794c892d-1e97-11e2-91ce-01aa75ed71a1?language=eng

and the headers for English PDF version are as follow:



For more information on how to access the different types of content streams please refer to [R02]

4 CONCLUSION

The suggested approach – using the EUR-Lex web services together with the CELLAR API - combines the power and simplicity of the EUR-Lex expert search together with the better efficiency of the HTTP protocol for content download.

Even though that Web Services offer many benefits over other types of distributed computing architectures, as a general rule, they should be used with cautions, especially when downloading large amount of data. If such necessity arises, other delivery options should be sought, for instance access to pre-prepared data archives via FTP.

End of Document