History

<table>
<thead>
<tr>
<th>Version</th>
<th>Date</th>
<th>Reason of change</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>2014-06-10</td>
<td>Document created (e.g. structure proposed, initial input...)</td>
</tr>
<tr>
<td>2</td>
<td>2014-07-17</td>
<td>Input ZBW included</td>
</tr>
<tr>
<td>3</td>
<td>2014-07-22</td>
<td>Version for internal review</td>
</tr>
<tr>
<td>4</td>
<td>2014-07-28</td>
<td>Final version</td>
</tr>
</tbody>
</table>

Impressum

Full project title: Enhancing Europe’s eXchange in Cultural Educational and Scientific reSources
Grant Agreement No: 600601
Workpackage Leader: Thomas Orgel, Jr-DIG
Project Co-ordinator: Silvia Russegger, Jr-DIG
Scientific Project Leader: Michael Granitzer, Uni-Passau

Acknowledgement: The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement no 600601.

Disclaimer: This document does not represent the opinion of the European Community, and the European Community is not responsible for any use that might be made of its content. This document contains material, which is the copyright of certain EEXCESS consortium parties, and may not be reproduced or copied without permission. All EEXCESS consortium parties have agreed to full publication of this document. The commercial use of any information contained in this document may require a license from the proprietor of that information. Neither the EEXCESS consortium as a whole, nor a certain party of the EEXCESS consortium warrant that the information contained in this document is capable of use, nor that use of the information is free from risk, and does not accept any liability for loss or damage suffered by any person using this information.
Table of Contents

1 Executive Summary .............................................................................................................. 1

2 Introduction ......................................................................................................................... 2
  2.1 Purpose of this Document .............................................................................................. 2
  2.2 Scope of this Document .................................................................................................. 2
  2.3 Status of this Document ................................................................................................... 2
  2.4 Related Documents ......................................................................................................... 2

3 Integration of a New Data Provider ...................................................................................... 3
  3.1 EEXCESS Architecture ................................................................................................. 3
  3.2 Define a Mapping ............................................................................................................. 4
  3.3 Create a new PartnerRecommender .................................................................................. 6
  3.4 Configure the PartnerRecommender – ConfigFile ............................................................ 6
  3.5 Extend the FederatedRecommender .................................................................................. 7
  3.6 Creation and Deployment of the war-File ......................................................................... 7

4 Metadata Extraction and Enrichment .................................................................................. 8
  4.1 Approach ........................................................................................................................ 8
  4.2 Proof of Concept with imdas pro ..................................................................................... 8
    4.2.1 Geonames Web Service .............................................................................................. 8
  4.3 Implementation in the EEXCESS PartnerRecommender ................................................. 10
    4.3.1 Implementation Europeano .......................................................................................... 10
    4.3.2 Implementation KIM.collect ....................................................................................... 11
    4.3.3 Implementation Mendeley ........................................................................................... 11
    4.3.4 Implementation wisenmedia ....................................................................................... 11
    4.3.5 Implementation ZBW ................................................................................................ 11
  4.4 EEXCESS Data Enrichment ........................................................................................... 11
    4.4.1 Examples of Enriched Data ....................................................................................... 11

5 EEXCESS Social Data Enrichment ...................................................................................... 22
  5.1 Current status .................................................................................................................. 22
    5.1.1 EconCrawler Documentation ................................................................................... 22
  5.2 Outlook ............................................................................................................................ 23

6 Data quality assurance ...................................................................................................... 25
  6.1 State of the Art ............................................................................................................... 25
  6.2 Approach and Planned Work ......................................................................................... 28
    6.2.1 Assessing Input Data Quality ...................................................................................... 28
    6.2.2 Assessment of Enrichment Results .......................................................................... 28
    6.2.3 Assessment of Mapping Results ................................................................................. 28
    6.2.4 Next Steps ............................................................................................................... 29

7 Conclusions ....................................................................................................................... 30

8 References .......................................................................................................................... 31
9    Glossary .......................................................................................................................... 32
1 Executive Summary

EEXCESS aims to bring cultural, scientific and educational content to the user. This means that the user gets additional information within the environment the user is using and working with. EEXCESS aims to unfold the treasure of cultural, scientific and educational content to improve the interconnectedness between these different domains.

One important achievement of EEXCESS is to provide not only software components but working prototypes. One prototype is e.g. a Chrome browser plugin that enables to list additional cultural and scientific information while looking at and editing a Wikipedia article.

To get the results of different institutions and therefore different information systems listed the following tasks have to be fulfilled by content provider:

1) Create a mapping of local data to the EEXCESS data format
2) Include the data in the EEXCESS Framework by providing a PartnerRecommender on the local system that is integrated in the Federated Recommender of EEXCESS

The mapping creation process is necessary to match data in a way that similar information is presented in a similar way. This is not only necessary for federated data presentation but almost more important for the distributed search in different systems within the PartnerRecommender. The user sitting in a library for instance doesn’t know – will even not be interested in – how, where and in which data fields the information is explicitly stored. The user is simply interested in getting additional useful information otherwise the user will not use EEXCESS features anymore.

Within EEXCESS this mapping of the partner systems is already done. But as a second achievement of EEXCESS is to increase the numbers of content provider joining the EEXCESS network it is necessary to provide the mapping structure and a software component to create this mapping without having implantation skills. This so called ConfigTool that was already used to create the mapping of the actual content provider in EEXCESS is made public in conjunction with D4.2.

The above listed second task needs to be done to get the data of local system explicitly listed within the EEXCESS features – e.g. the Chrome browser plugin. In order to enlarge the performance it is foreseen to have a PartnerRecommender on the local systems that performs the explicit queries coming from the FederatedRecommender. The FederatedRecommender merges the results coming from the local system depending on user profile information. This includes for instance ranking of the results depending on the interests of the user. The FederatedRecommender is not part of this deliverables as it is covered in work package 3.

The following description is a first “handbook” to get the ConfigTool and the PartnerRecommender running. It answers questions like “How can the mapping of my local data to EEXCESS data format be done?” or “What has to be done to get my results listed and ranked as well?”.

Aim of EEXCESS is not only to list results from local systems – although the listing of federated and ranked results is a first important step – but to enrich this information depending on the content itself. So if a user e.g. is looking for information regarding the “Vorfriede von Leoben”, a peace treaty between Austria and France in the 18th century he might also be interested in similar events in Leoben, Styria at a similar time. EEXCESS will enrich the results depending on keywords, structured vocabulary – as for instance location terms in the GeoNames thesaurus- or blog entries. First results of this enrichment are also described within the present deliverable.
2 Introduction

2.1 Purpose of this Document

This deliverable provides the first prototype for the integration of cultural, scientific and educational data sources as well as the enrichment of those data sources from social media channels.

2.2 Scope of this Document

Deliverable 4.2 will give an overview on the different components of the EEXCESS Framework for data provision. It will also list and describe the necessary tasks that have to be fulfilled by different content provider to get data included within the results EEXCESS provides.

In the domain “Metadata Mapping, Extraction and Enrichment” the document describes the development and implementation based on the requirements on D1.1 and the analysis described in D4.1. As Task 4.4 regarding the data quality assurance started later the present deliverable gives an overview and outlook on what is actually done and planned for the next prototype.

2.3 Status of this Document

This is the final version of D4.2. The project internal review was done by KNOW.

2.4 Related Documents

Before reading this document it is recommended to be familiar with the following documents:

- D1.1 First Conceptual Architecture and Requirements Definition
- D4.1 Integration and Enrichment Specifications and Analysis
3 Integration of a New Data Provider

This section provides a guide how to integrate a new data provider into the EEXCESS Framework. Explaining the architecture of the system components makes it easier to explain and understand the tasks that have to be fulfilled to get data integrated in EEXCESS.

The description helps to get local data listed in the result list of EEXCESS as well.

In the following sections we provide a guide how a new data provider can be integrated into the EEXCESS Framework. To act as a data provider for EEXCESS it’s necessary to have an API that provides a function to search in the data and retrieves the results of this search.

3.1 EEXCESS Architecture

To understand how we can integrate the new data provider into the EEXCESS Framework we give a brief overview over the EEXCESS architecture. To explain how the EEXCESS Framework works, we start at the user and his client. For our first use case in EEXCESS the “client” is a Chrome-browser, but this special client can of course be replaced by any other type of client (e.g. learn management system). The following list gives an overview on the different tasks tight from the start of a user driven search process to the final result presentation.

- The client detects an information need and creates a query.
- The client sends this query to the Privacy Proxy. The Privacy Proxy invokes the Federated Recommender. This component has the information which data provider are actually available.
- The Federated Recommender calls all the PartnerRecommender
- PartnerRecommender sends results back.
- The results are integrated in one combined view and are sent back to the client via the Privacy Proxy.

A more detailed description of the components and their interaction is provided in D1.1.

To add a new data provider we need to build a new PartnerRecommender which queries the partner data store and returns the results in the EEXCESS data format to the Federated Recommender.

The PartnerRecommender needs to run on a Tomcat or another Java Servlet Container.

The following figure gives an overview on the EEXCESS component as specified and described in detail in D1.1 First Conceptual Architecture and Requirements Definition.
3.2 Define a Mapping

As a first step we create a mapping from the service response to the EEXCESS format. Therefore it’s helpful to have an example response of the data providers service. D4.1 provides a detailed description of the so called ConfigTool and how a mapping could be created with the ConfigTool. The ConfigTool is available at:
The ConfigTool is a tool that allows the creation of a mapping from a local data structure to the EEXCESS data format without implementation skills. It offers user interfaces for all steps of mapping creation right from opening a mapping project to the creation of the XSL transformation file.

Here we give a brief introduction to the ConfigTool:

1) Create a new project with the ConfigTool, while entering a name for the project and select a Source Schema and a Target Schema.

A “project” here is a very special mapping from one local data format (Source Schema) to an output data format (Target Schema, here: EEXCESS data format).

2) Create a concept in the Source Schema for every field which needs to be mapped.
3) Define mappings from the new Concepts in the Source Schema to the MEON-Concepts. The MEON-Concepts already contain the definition of the EEXCESS data format in which the local data should be mapped.

4) When all mappings are defined Transformation-Files can be created directly with the ConfigTool (Menu Transformation -> Create XSL). This Transformation-File is used in the PartnerRecommender later on.

3.3 Create a new PartnerRecommender

The PartnerRecommender and the FederatedRecommender are written in JAVA. The source code is hosted by the Know-Center and is accessible at the KnowCenter-SVN (https://knowminer.at/svn/eexcess/) which is actually not public yet.

This is the folder-structure:

- Code
  - Partners
    - EUROPEANA
    - KIM.collect
    - Mendeley
    - wissenmedia
    - ZBW
  - Recommender
  - Reference
    - Partner-data-layer
    - PartnerRecommender
    - Partner-web-service

These structure and folders are used to build PartnerRecommender. In the folder “partners” implementations and configurations for the existing data providers are all already available. In the folder “reference” the API’s from WP3 und WP4 and der REST-Service of the PartnerRecommender can be found.

3.4 Configure the PartnerRecommender – ConfigFile

A configuration file exists in the PartnerRecommender.

```
{
  "partnerConnectorClass": "eu.eexcess.zbw.recommender.PartnerConnector",
  "queryGeneratorClass": "eu.eexcess.zbw.recommender.ZBWQueryGenerator",
  "searchEndpoint": "https://api.econbiz.de/v1/search?q=${query}&xml=true",
  "transformerClass": "eu.eexcess.zbw.datalayer.ZBWTransformer",
  "mappingListTransformationFile": "mapperResultList.xsl",
  "mappingObjectTransformationFile": "mapperObject.xsl"
}
```

The parameter partnerConnectorClass defines the JAVA-Class which will be used to call the service of the partner data store. To get data out of the partner data store it’s necessary to call the API of the partner data store. For this it’s necessary to configure an already existing PartnerConnector or implement a new one with creating a new class which implements the Interface ParnterConnectorApi.

The parameter queryGeneratorClass defines the JAVA-Class which creates the query for the service. The query generator is responsible for translating the EEXCESS user profile into a query string. The user profile contains special user information as for instance special themes the user is interested in, often used keywords in
queries. Details to the user profile are specified in work package 5 and therefore not part of the present deliverable.

For example: If a user is interested in a special theme and the user puts this into his own user profile – this is under his own responsibility in order to stick to privacy guidelines – these user profile information is added to the query request sent to the PartnerRecommender. This is a first enrichment of the result list depending on user profile information.

The parameter searchEndpoint defines the URL where the service is available. The parameter transformerClass defines the JAVA-Class which transforms the results from the format of the service of the data provider to the EEXCESS format. The parameters mappingListTransformationFile and mappingObjectTransformationFile defines the transformations into the EEXCESS format for result-lists or single objects. The transformation files from the ConfigTool must be put in the sub-folder `\src\main\resources\`. The filenames must be the same as for the parameters mappingListTransformationFile and mappingObjectTransformationFile as configured in the PartnerRecommender-ConfigFile.

### 3.5 Extend the FederatedRecommender

In the next step the FederatedRecommender must be changed to call the new PartnerRecommender. Therefore open the project modules`\recommender\federated-recommender-web-service` and extend the method “initialize” in the Class FederatedRecommenderService. Add the new PartnerRecommender via a PartnerBadge like the already existing PartnerRecommenders are added.

### 3.6 Creation and Deployment of the war-File

Build the new PartnerRecommender and the modified FederatedRecommender with maven and deploy the new war-files to the Tomcat. After deploying both web-apps, the client should also show results from the newly added data provider, if some objects matches to the information need which the client detects.
4 Metadata Extraction and Enrichment

The following chapter describes the development and implementation that is done so far based on the requirements on D1.1 and the analysis described in D4.1.

4.1 Approach

Our Approach is described in D4.1 – Metadata Enrichment. Two different cases are described:

1.) Text from Special fields, e.g. Author...
2.) Free Text from e.g. description

In the first case we will call an appropriate service, depending on the type of the field which we actually process.

In the second case we follow the following steps:

1.) Textual resources are pre-processed by removing punctuation, normalizing capitalization and removing stop words
2.) Add synonyms for the terms, using WordNet
3.) Classify if the terms are entities found in DBpedia or not. If so, we also add their category
4.) Obtain a set of connected entities for the terms from FreeBase
5.) Terms identified as for instance geographic entities, add also the connected geographic terms, obtained from using GeoNames thesaurus

4.2 Proof of Concept with imdas pro

As mentioned the external data sources should be used for metadata extraction in order to make data enrichment later on. As proof of concepts the GeoNames thesaurus was integrated and analysed within the cultural information system imdas pro. This information system is highly used by cultural institutions in mainly German speaking Europe. The usage and integration of GeoNames thesaurus within imdas pro allows good analysis of the outcome of this usage.

4.2.1 Geonames Web Service

Geonames is a free geographical database. It offers a wide range of query options. At the moment the database contains over 8 million geographical names in many languages. There is additional information as coordinates, inhabitants, postcodes an so on.

Geonames has many different web services. To use these web services you have to create a user account, which is free of charge. The amount of requests is limited to 2000 credits per hour. There are also commercial accounts available (http://www.geonames.org/commercial-webservices.html).

An overview can be found at: http://www.geonames.org/export/ws-overview.html

For our prototype we used the web services “search” with the parameter “name_startsWith” and “Hierarchy” which returns all higher up locations to a given location.

The search can be limited to a specific language or you can search language independently. We decided to use the language independent search. For bigger cities there are translations into many languages.

The results of the search for Vienna (and the German name "Wien") are listed in the following two figures:
In some cases the result contains terms, which are not obvious e.g. “Vienville” for Vienna. We did not use the fuzzy search, which is also possible.

One issue we realized is that if you search for all subterms of a given term the sorting is done by the English terms. If you want to order by German terms, only these places will be shown, that have a German translation. So the tree view will not be complete for others languages than English. In the example below, “Österreich” is sorted to the English name “Austria”.

Figure 2: Search for “Vienna”

Figure 3: Search for “Wien”
The response time for queries is very fast even if you only enter some letters.
The documentation of the web services is very good. For every web service you get a short explanation and a sample. The result can be retrieved in XML or JSON format.

4.3 Implementation in the EEXCESS PartnerRecommender

In the actual deployed version of the enrichment we have already implemented some of the target services mentioned in D4.1. In the actual implementation we handle only the bullet “2) Free text” of the above list. Therefore we have implemented a subset of the mentioned steps. So we implemented 1, 3, 4 and 5 defined in the approach. But we actually don’t add the category from DBpedia.

As input for the recommend method of the PartnerRecommender an EEXCESS user profile is needed. For testing the PartnerRecommender a minimalistic user profile as shown in the following is enough.

```json
{
    "contextKeywords": [
        {
            "text": "graz",
            "weight": 0.1,
            "reason": "manual"
        }
    ]
}
```

This EEXCESS user profile must be sent with a HTTP POST to the ServiceEndPoint of the PartnerRecommender.

4.3.1 Implementation Europeana

ServiceEndPoint:
http://eexcess.joanneum.at/eexcess-partner-europeana-1.0-SNAPSHOT/partner/recommend/
4.3.2 Implementation KIM.collect
ServiceEndPoint:
http://eexcess.joanneum.at/eexcess-partner-kim-collect-1.0-SNAPSHOT/partner/recommend/

4.3.3 Implementation Mendeley
ServiceEndPoint:
http://eexcess.joanneum.at/eexcess-partner-mendeley-1.0-SNAPSHOT/partner/recommend/

4.3.4 Implementation wissenmedia
ServiceEndPoint:
http://eexcess.joanneum.at/eexcess-partner-wissenmedia-1.0-SNAPSHOT/partner/recommend/

4.3.5 Implementation ZBW
ServiceEndPoint:
http://eexcess.joanneum.at/eexcess-partner-zbw-1.0-SNAPSHOT/partner/recommend/

4.4 EEXCESS Data Enrichment

Within EEXCESS we differentiate between two types of data enrichment. One is the enrichment depending on social media which is covered in chapter 5 of this present document.
The second one is data enrichment by metadata extraction as explained above. The following chapter gives an overview on the data enrichment by metadata extraction.

4.4.1 Examples of Enriched Data

In the following section we show examples of results from two data providers. Therefore we show
- The response from the service of the data provider
- Result of the transformation: Result of executing the transformation-file from the ConfigTool on the original response
- Same record after the enrichment services

The first part of the examples show a XML file as it is sent back from the local system via the EEXCESS service. These files are used for the transformation. Results of the transformation are given under “Results after transformation”. To show – as example – one special transformation a section is highlighted in both examples in italic and marked green.

How the file structure of the result looks like after enrichment is shown under “Results after enrichment”. Enriched parts are marked bold and blue in both file parts the “original” coming from the service and in the enriched one.

Examples from ZBW
Response from Service:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<doc>
<hits>
<hit>
<book>
ViewItem>A<score>3.8187141</score>
```
Results after transformation:
Results after enrichment:

<?xml version="1.0" encoding="UTF-8"?>
<rdf:RDF xmlns:dc="http://purl.org/dc/elements/1.1/"
    xmlns:dcterms="http://purl.org/dc/terms/
    xmlns:edm="http://www.europeana.eu/schemas/edm/
    xmlns:eexcess="http://eexcess.eu/schema/
    xmlns:foaf="http://xmlns.com/foaf/0.1/
    xmlns:ore="http://www.openarchives.org/ore/terms/
    xmlns:owl="http://www.w3.org/2002/07/owl#"
    xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
    xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#
    xmlns:xsd="http://www.w3.org/2001/XMLSchema#">
    <owl:Ontology rdf:about="https://api.econbiz.de//edm/">
        <owl:imports rdf:resource="http://www.w3.org/ns/ma-ont"/>
        <owl:imports rdf:resource="http://www.openarchives.org/ore/1.0/terms"/>
    </owl:Ontology>
    <eexcess:Proxy rdf:about="http://www.econbiz.de/Record/10010341805/proxy/">
        <ore:proxyIn>
            <ore:Aggregation rdf:about="http://www.econbiz.de/Record/10010341805/aggregation"/>
            <edm:aggregatedCHO>
                <eexcess:Object rdf:about="http://www.econbiz.de/Record/10010341805"/>
            </edm:aggregatedCHO>
            <edm:preview rdf:resource="https://api.econbiz.de//edm"/>
            <edm:isShownAt rdf:resource="https://api.econbiz.de//edm"/>
            <edm:isShownBy>
                <edm:WebResource rdf:about="http://www.econbiz.de/Record/10010341805/webresource"/>
            </edm:isShownBy>
            <edm:rights rdf:resource="http://creativecommons.org/publicdomain/mark/1.0"/>
        </ore:Aggregation>
    </eexcess:Proxy>
</rdf:RDF>
Example from KIM.Collect

Response from Service:

```xml
<?xml version="1.0" encoding="UTF-8"?>
<object countHits="4" countObjects="4" resultMaskId="1086" resultMaskName="Web Ergebnisliste KIM Export">
  <id>33874</id>
  <imdasiD>048D389F7DA64D799F187D86F00E6AA4</imdasiD>
  <guid>048D389F-7DA6-4D79-9F18-7D86F00E6AA4</guid>
  <uri>http://digv525.joanneum.at/imdas-liestal/rest/record/048D389F-7DA6-4D79-9F18-7D86F00E6AA4</uri>
  <IMDASID fieldId="143">048D389F7DA64D799F187D86F00E6AA4</IMDASID>
  <AlternativeObjektbezeichnung fieldId="18927">Stange (Glas)</AlternativeObjektbezeichnung>
  <Ausgestellt fieldId="17166"/>
  <Autor fieldId="20589"/>
  <Beschreibung fieldId="129">Bierglas Ziegelhofbier, Stange, hoch, konisch mit Scheibenfuss, 3 dl, geeicht; Glas farblos; 6-farbige Einbrand-Etikette mit altem Logo und Schriftzug. Keine Nähte sichtbar; Einbrand-Etikette 6-farbig, Eichzeichen geätzt oder gesandstrahlt.</Beschreibung>
  <BiologAlter fieldId="311"/>
  <Breite fieldId="10437"/>
  <Darstellung fieldId="10289"/>
  <Datentrager fieldId="18928"/>
  <Durchmesser fieldId="10464">6.7cm</Durchmesser>
  <Eingangsart fieldId="358">Schenkung</Eingangsart>
  <Eingangsdatum fieldId="359"/>
  <Entstehungszeit fieldId="431"/>
  <Entstehungszeitvon fieldId="432"/>
  <Entstehungszeitbis fieldId="433"/>
  <Finder fieldId="10063"/>
  <Flurname fieldId="190"/>
  <Fotograf fieldId="19060"/>
  <Funddatum fieldId="10341"/>
  <Fundort fieldId="10281"/>
  <Geschlecht fieldId="345"/>
  <Gewicht fieldId="10473"/>
  <Hersteller fieldId="10094"/>
</object>
</rdf:RDF>
```
<Herstellungsort fieldId="10285"/>
<Hoehe fieldId="10428">21.6cm</Hoehe>
<Ikonographie fieldId="10290"/>
<Inschriften fieldId="13379">3 dl</Inschriften>
<InvNr fieldId="132">AMBL_00006868</InvNr>
>InventarnummerQuellsystem fieldId="23011">H 06868</InventarnummerQuellsystem>
<KoordinateX fieldId="219"/>
<KoordinateY fieldId="220"/>
<Katalogtext fieldId="131">Bierglas Ziegelhofbier, Stange, hoch, konisch mit Scheibenfuss, 3 dl, geeicht; Glas farblos; 6-farbige Einbrand-Etikette mit altem Logo und Schriftzug. Keine Nähte sichtbar; Einbrand-Etikette 6-farbig, Eichzeichen geätzt oder gesandstrahlt.</Katalogtext>
<Kuenstler fieldId="14570"/>
<Laenge fieldId="10455"/>
<Literatur fieldId="13381"/>
<Material URI="http://digv525.joanneum.at/imdas-liestal/rest/masterdata/thesaurusTerm/16824" fieldId="10287" thesaurusTermId="16824">Glas</Material>
<Objectbezeichnung fieldId="142">Bierglas Ziegelhofbier</Objectbezeichnung>
<Objectgeschichte fieldId="135"/>
<Objecttyp fieldId="366">Ethnologie</Objecttyp>
<Sachgruppe URI="http://digv525.joanneum.at/imdas-liestal/rest/masterdata/thesaurusTerm/14292" fieldId="12233" thesaurusTermId="14292">Bierkultur, Trinkbecher aus Glas oder Kristall (Trink-, Tafelgläser, geätzt, bemalt usw.)</Sachgruppe>
<Sammlung URI="http://digv525.joanneum.at/imdas-liestal/rest/masterdata/collection/3" collectionId="3" fieldId="367">
<collection>
  <id>3</id>
  <name>AMBL Europäische Ethnologie</name>
  <instituition>
    <id>1</id>
    <name>Hauptabteilung Archäologie und Museum</name>
    <isil/>
    <isca/>
  </instituition>
</collection>
</Sammlung>
<RechteamObjekt fieldId="11790">Archäologie und Museum Baselland</RechteamObjekt>
<Standort fieldId="610">BZM, Kriegackerstr. 30, Muttenz /Raum U 101</Standort>
<Stuck fieldId="450">1.0</Stuck>
<Taxon fieldId="207"/>
<Technik URI="http://digv525.joanneum.at/imdas-liestal/rest/masterdata/thesaurusTerm/16289" fieldId="10288" thesaurusTermId="16289">geblasen</Technik>
<Teilsammlung fieldId="13380">Brauerei Ziegelhof</Teilsammlung>
<Titel fieldId="216"/>
<Tiefe fieldId="10446"/>
<Zustand fieldId="218">gut</Zustand>
Results after transformation:

```xml
<?xml version="1.0" encoding="UTF-8"?>
  <owl:Ontology rdf:about="" xmlns:owl="http://www.w3.org/2002/07/owl#">
    <owl:imports rdf:resource="http://www.openarchives.org/ore/1.0/terms"/>
    <owl:imports rdf:resource="http://www.w3.org/ns/ma-ont"/>
  </owl:Ontology>
    <edm:provider xmlns:edm="http://www.europeana.eu/schemas/edm/">
      <edm:Agent rdf:about="https://kgdb.bl.ch/kim-kgs/">
        <foaf:name xmlns:foaf="http://xmlns.com/foaf/0.1/">KIM.Collect</foaf:name>
      </edm:Agent>
    </edm:provider>
    <edm:dataProvider xmlns:edm="http://www.europeana.eu/schemas/edm/">
      <edm:Agent rdf:about="https://kgdb.bl.ch/kim-kgs/">
        <foaf:name xmlns:foaf="http://xmlns.com/foaf/0.1/">KIM.Collect</foaf:name>
      </edm:Agent>
    </edm:dataProvider>
  </edm:Aggregation>
</rdf:RDF>
```
Results after enrichment:

<?xml version="1.0" encoding="UTF-8"?>
  <owl:Ontology rdf:about="https://kgdb.bl.ch/kim-kgs/edm/">
    <owl:imports rdf:resource="http://www.w3.org/ns/mansard"/>
    <owl:imports rdf:resource="http://www.openarchives.org/ore/1.0/terms"/>
  </owl:Ontology>
  <eexcess:Proxy rdf:about="http://digv525.joanneum.at/imdas-liestal/rest/record/048d389f-7da6-4d79-9f18-7d86f00e6aa4/enrichedProxy/">
    <ore:proxyFor>
      <eexcess:Object rdf:about="http://digv525.joanneum.at/imdas-liestal/rest/record/048d389f-7da6-4d79-9f18-7d86f00e6aa4"/>
    </ore:proxyFor>
    <dc:subject>TV Network</dc:subject>
    <dc:subject>City/Town/Village</dc:subject>
    <dc:subject>Logo</dc:subject>
    <ore:proxyIn>
      <ore:Aggregation rdf:about="http://digv525.joanneum.at/imdas-liestal/rest/record/048d389f-7da6-4d79-9f18-7d86f00e6aa4/enrichedAggregation/">
        <edm:dataProvider>EEXCESS data enrichment</edm:dataProvider>
        <edm:aggregatedCHO rdf:resource="http://digv525.joanneum.at/imdas-liestal/rest/record/048d389f-7da6-4d79-9f18-7d86f00e6aa4"/>
      </ore:Aggregation>
    </ore:proxyIn>
    <dc:subject>Stange</dc:subject>
    <dc:subject>Business Operation</dc:subject>
    <dc:subject>Hannah Höch</dc:subject>
    <dc:subject>Collage Artist</dc:subject>
    <dc:subject>Glas</dc:subject>
  </eexcess:Proxy>
  <dc:subject>Bierkultur, Trinkbecher aus Glas oder Kristall (Trink-, Tafelgläser, geätzt, bemalt usw.)</dc:subject>
  <dc:description>Bierglas Ziegelhofbier, Stange, hoch, konisch mit Scheibenfuss, 3 dl, geeicht; Glas farblos; 6-farbige Einbrand-Etikette mit altem Logo und Schriftzug. Keine Nähte sichtbar; Einbrand-Etikette 6-farbig, Eichzeichen geätzt oder gesandstrahlt.</dc:description>
  <dc:title>Bierglas Ziegelhofbier</dc:title>
  <dc:identifier>AMBL_000006868</dc:identifier>
  <edm:language>de</edm:language>
  <edm:europeanaProxy>false</edm:europeanaProxy>
  <ore:proxyIn>
</rdf:RDF>
5 EEXCESS Social Data Enrichment

In this paragraph we will first briefly introduce the research question that led to the concrete tasks we are currently working on. Moreover we will summarize the work that has been done so far and give a short outlook on future work.

5.1 Current status

The initial question was: Can we establish backlinks from manually selected economic blogs to EconBiz resources (either directly or via the STW – Standard Thesaurus for Economics)?

Originally, we took Twitter posts as a source for potential back linking to EconBiz resources, but we had to drop this approach for two reasons. First, there were only a few mentions of EconBiz resources in Tweets, and secondly, because of their brevity Tweets do not provide enough content to determine their relevance for Economics. Therefore, we switched to scientific blogs in Economics as our primary data source, for they provide larger text corpora to perform semantic analyses upon.

The first step to tackle the research question was to identify economists’ blogs with the highest impact. As it turned out, there were some sources that already dealt with that issue. We’ve picked a list by onalytica.com as a source for influential blogs. Afterwards we have developed a web crawler based on Scrapy that captured all available content from the 10 most influential blogs (based on the onalytica.com-list) and stored them into an elastic search database. As result we retrieved 80.000 blog posts (composed of text body, headline, included links, author, release date and some other components).

An intermediate result we discovered that there were no direct links to econbiz.eu, which could be regarded as the main EEXCESS content provider in this field.

Hence we decided to try a more sophisticated approach. The idea was to use natural language processing (NLP) algorithms and extract “relevant” information in order to figure out if a mentioned paper can be found in EconBiz. The most promising approach was, to look for URLs that point to pdf-files (~2.000 links could be found), download them and extract their plaintext. Then the plaintext was sent through the NLP-pipeline, where two processes were executed:

1. Named entities (like persons, location etc.) were filtered
2. Nouns were filtered and counted

In order to narrow down the results of the prior steps to terms from Economics, we are currently working on the utilization of ZBW’s Standard Thesaurus for Economics (STW).

The following chapter describes how the software is going to be used and where the necessary components can be downloaded from.

5.1.1 EconCrawler Documentation

This software is in a very early development state and hence experimental. Therefore the occurrence of errors is rather the rule than the exception.

1. What is the EconCrawler?

The EconCrawler is a tool that collects data from blogentries and stores them into an elasticsearch index. It is not a universal tool, since there is a finite amount of blogs that can be crawled.

1 http://www.onalytica.com/blog/posts/top-200-influential-economics-blogs-aug-2013
2 http://scrapy.org
3 http://www.elasticsearch.org/
2. Installation/Usage

This description is based on Ubuntu 14.04.

The crawler expects an Elasticsearch\(^4\) daemon to listen at http://localhost:9200. This can either be achieved by installing it as described on the elasticsearch website or by using a docker container.

An appropriate image can be found at:

http://tmfw.de/user/elasticsearchContainer.tar

Assuming you have a working docker installation, you can import the image by using this command:

docker load < elasticsearchContainer.tar

Afterwards you can start the elasticsearch-container by typing:

docker run --privileged=true -p 0.0.0.0:9200:9200 f2308d228f2c /bin/bash

In case docker prompts a message, saying that the image couldn’t be found, you should type docker images and look for the correct ID. When you visit http://localhost:9200 and no error is prompted, elasticsearch works properly.

Now we can tackle the crawler itself.

The prerequisites are:

- git
- python 2.7.x
- scrapy >= 0.22
- the python libraries python-dateutil and elasticsearch

Use your preferred package-manager (apt-get, aptitude, yum) to install python and scrapy and a python package-manager (link pip or easyinstall) for the python libraries.

A git repository containing the source code is hosted at:

http://tmfw.de/user/EconCrawler.tgz

Extract the file to a separate folder. To check out the latest code type git checkout --.

You should see several files including "crawlall.sh". When you invoke this file the crawling process starts.

The script contains the line

scrapy crawlall -a startDate=2013-06 -a endDate=2014-06.

You can modify the parameter to fit your needs.

After the first entries have been stored into elasticsearch you can start querying it. To do so, visit http://localhost:9200/_plugin/marvel/sense/index.html.

On the left hand side you can assemble your queries which you can send by hitting <ctrl>-<Enter>.

The file exampleQueries contains some example queries.

5.2 Outlook

This paragraph will give an overview of the next steps we will take.

---

\(^4\) http://www.elasticsearch.org
Until now we have tried to implement a fuzzy and therefore broad search mechanism. In the next step we will try to achieve exact matches by using techniques to extract publication title and author name to examine whether EconBiz contains that publication or not. We expect a lower hit rate, but more accurate hits.

Depending on the results of the prior described steps we will decide if we expand our effort onto other parts of the raw dataset.

Optionally we could bring the previous work (particularly the source code) into a shape that makes it useful to other people.

Subsequent research questions (RQs) to tackle:

- In addition to the analysis of topics in popular Economics' blogs, can we identify corresponding signals or indicators for those trends in other social media channels (twitter, mailing lists)?
- Can we reconstruct the social network (of sub-disciplines) through analyzing authors, comments and author mentions in blog posts and other social media channels?
- Can RQ1-RQ3 also be solved for cultural domains?
6 Data quality assurance

The task on data quality assurance aims at developing methods for assessing data and information quality of the extracted and integrated metadata from tasks 4.1, 4.2 and 4.3. The constraints of the EEXCESS system will require a high degree of automation for metadata quality assessment. Quality assessment is applied both to the input and output side, i.e., assessing the quality of data coming from the providers as well as assessing the result of mapping and enrichment. When metadata is retrieved or converted on the fly, there is no chance to react to quality issues, and it may be unreliable to draw conclusions based on a single sample. Assessment will thus be applied as a background process, collecting data for statistical analysis. The gathered data then provides feedback to metadata providers, on the definition of mappings and the performance of enrichment tools.

6.1 State of the Art

The Dublin Core Metadata Initiative has defined application profiles\(^3\) in order to specify the set Dublin Core elements supported by a specific organisation or system. Description Set Profiles have been defined in order to validate documents against application profiles [DC-DSP, 2008]. The syntax is similar to XML Schema, thus the same limitations of which constraints can and cannot be expressed, apply.

A recent work [Dröge, 2012] explores criteria for assessing the quality of vocabularies, particularly those provided as linked open data. Clearly, these criteria are not directly applicable to assessing the quality of metadata documents. However, for documents using external vocabularies, the assessment of the quality and reliability of those vocabularies is needed in order to give an assessment about the elements using these vocabularies, as well as predicting potential issues with mapping. A few of these criteria may also be relevant for assessing instances of metadata documents.

The report describes eight classes of criteria:

1. **About**: The topic, information about the creator(s) and the license.
2. **Actuality**: Creation data and recent date/frequency of updates.
3. **Semantic Representation**: Languages used, and relations described.
4. **Syntactic Representation**: Data format/model used.
5. **Size**: Topics covered, number of classes, properties, individuals and triples.
6. **Quality**: Coherence as determined by reasoned, correctness as determined by linked data validator, observing linked data principles.
7. **Linking**: Links to other vocabularies, top-level, domain and pivot ontologies, links to ontologies already used in Europeana, and number of incoming links.
8. **Context**: Type of relation and position of a specific term to be used.

One potential option, given the Europeana and EEXCESS data models, is to perform metadata validation in the RDF or OWL domain. A core issue is that in contrast to most XML-based formats, the schema may not be explicitly defined. RDF validation is thus a hard and unsolved problem [RDFValEx, 2012]. Schema representations such as RDF schema can be used to provide additional information on interpreting the RDF document. In addition, validation of RDF documents often includes closing the world and assuming the document to be self-contained. Option to implement RDF/OWL validation include OWL DL reasoning (including tools performing closed-world reasoning), or SPARQL queries checking constraints (or a series of SPARQL operations). For SPARQL, extensions such as the SPIN modelling vocabulary [Knublauch, 2013], which provides a vocabulary for checking rules and constraints using SPARQL, can be used. However, some issues using SPARQL for validation, such as that built-in functions not standardised, or the complexity of handling series of collections, have been reported in literature. An alternative approach for representing constraints, called

\(^3\) [http://dublincore.org/documents/profile-guidelines/]
bounds, has been proposed in [Skjæveland, 2013]. More details on issues and approaches around RDF validation can be found in [RDFValEx, 2012]. Alternative approaches include converting RDF documents to another representation, where validation is more straightforward and well defined.

As validation of RDF/OWL document (beyond syntax checking) is still an unsolved problem a workshop recently organised by the W3C explored requirements and potential solutions [RDFValWS, 2013]. The discussed applications include validation for Europeana and Open Government Data applications, which share many properties and requirements with EEXCESS.

The main issues identified are the following:

- Variations in syntax and semantics of RDF documents,
- Defining rules for validation (e.g. pre-/postconditions)
- Issue of completeness (required elements, cardinalities)
- Data types
  - Complex values (e.g., value and unit)
  - Checking URI patterns
- Dynamic external vocabularies, validation needs to include external sources

Solutions will include among others the following elements:

- Defining vocabulary profiles specifying closed worlds
- Validate that minimum constraints are met but do not hinder interoperability, partial validation needed
- Perform validation before and after DL reasoning, a reasoning may fulfil a constraint or cause a violation

A large share of the literature on metadata quality assessment is concerned with the definition of criteria for metadata quality. However, most of those require manual verification by an expert. [Bruce, 2004] define the following measures for quality:

- Completeness: presence of elements describing an object and use of the same element set across the collection.
- Accuracy: correctness of values, correct spelling
- Provenance: source and creation process of the metadata
- Conformance to expectations: appropriateness for target use case, controlled vocabularies matching the intended consumers
- Logical consistency and coherence: semantic correctness and absence of contradictions, but also searchability, i.e., similar objects are expected to have similar metadata in order to be returned to related queries
- Timeliness: metadata is available with the object and up-to-date
- Accessibility: readable and understandable to intended users

They then define three tiers of quality verifications. The first consists of syntactic validation and correct identifications, the second of use of appropriate controlled vocabularies, appropriate elements for specific and general target communities and provenance information, and finally conformance to an application profile and a complete provenance record including historic metadata. Clearly, some of these assumptions (e.g., on completeness across the collection) are rather applicable to homogeneous collections of a single organisation rather than heterogeneous metadata harvested from different organisations describe a range of different cultural heritage objects.

The authors of [Stvilia, 2007] have proposed a taxonomy of 22 measures for information quality, grouped into three categories: Intrinsic information quality contains measures that can determined from a single document wrt. a specification. Relational/contextual information quality concerns measures on the relation between the metadata and the content being described as well as to other metadata items. Reputational information quality is mainly determined by the origin and provenance of the information. The proposed taxonomy and its relations to activities and potential information quality problems are shown in Figure 5.
The authors of [Troncy, 2010] proposed an automatic approach for semantic validation of metadata documents wrt. to metadata standards or (application) profiles. Standard tools for syntactic level checking are widely available (e.g., XML schema validators) but the semantic expressivity of XML schema is limited. Thus validation on higher level can be only done with specific application logic. The approach transforms the XML metadata document to an RDF/OWL document, and (i) performs inference to check for inconsistencies (ii) uses rules to model additional constraints otherwise only available in textual form in the specification. The approach is deployed as a web service with both a RESTful interface and a web-based UI called VAMP\(^6\).

A recent approach to metadata assessment for cultural heritage repositories, which is very relevant for EEXCESS, is presented in [Bellini, 2013]. The authors start from viewing metadata quality as the fitness for use for a specific purpose. They propose to define quality profiles, which represent the contextualised requirements on the metadata. In their work, they asked professionals about the relevance of different elements of the Dublin Core set. Then, three metrics are proposed: completeness (the element is filled), accuracy (no syntactic and spelling errors) and consistency (correct semantics and no logical errors). Completeness checks can be automated, while accuracy can only be checked for a small set of fields (e.g., date format, well-formed URIs) and consistency can only be checked at a very limited extent (e.g., check if links can be resolved, check if MIME type of linked file is correct). The paper presents a case study using data from three Italian university repositories, on which the implemented automatic analysis framework has been tested.

\(^6\) http://vamp.joanneum.at
6.2 Approach and Planned Work

EEXCESS will target three aspects of metadata quality assessment, as detailed below. This section describes the approach and planned work for each of the aspects. For all the aspects, the approach will be chosen in order to maximise the amount of automation that can be done, and to gather information that can be fed back to improve data provision, mapping and enrichment services.

6.2.1 Assessing Input Data Quality

This will include the approach proposed by [Bellini, 2013] as a first set of quality checks. In EEXCESS we cannot assume a single application profile, against which completeness and accuracy can be checked, but the profile to be applied will depend on the type of content and the data provider (using the knowledge about the respective native data model).

In addition, the accuracy check will be extended as follows:

- A measure about the structuredness of values will be determined, assessing for example the structuredness of date, name. This is not only based on the fact whether they are structured, but also whether the format of a string is defined (e.g., using XML Schema regular expressions).
- An assessment on the type of vocabulary used, including whether it is accessible.

For the consistency check, semantic validation against constraints of the data model will be added.

6.2.2 Assessment of Enrichment Results

Enrichment results will first be checked for accuracy and consistency. In particular, the semantic consistency between existing metadata and the newly added metadata will be checked. This is expected to be a probabilistic process, yielding rather a confidence score than a binary judgment.

It is expected, that this will in particular be an issue for social enrichment. One known limitation of an automatic approach is that the consistency of information that is entirely orthogonal to what is contained in the existing metadata (and might thus be a very valuable addition) cannot be checked.

Where available, provenance metadata will be taken into account for this assessment.

6.2.3 Assessment of Mapping Results

The results of mapping outputs will be performed in order to provide feedback to the creator of the mapping. It will thus be integrated in the metadata mapping configuration tool. Three types of assessment will be investigated:

- Comparison against an expert created reference. This is a very reliable approach, but has the drawback that the ground truth mapping for each instance needs to be created. Depending on complexity of the tool used to determine the deviation between the result and the ground truth it may also not be possible to recognise alternative, but equally correct mapping outputs.
- Round trip mapping via intermediate format. If a bidirectional mapping is completely defined, mapping from any format to the intermediate format and back to the input format should be lossless. Although not directly corresponding to a practical use case, assessment can be performed automatically without creating a ground truth.
- Round trip mapping via target format. The more practical example is mapping from format A to format B and back to format A. The issue is that depending on the formats involved, this mapping may not be lossless, but due to limitations of one of the formats, some loss of information or inprecision in mapping may be expected. This needs to be specified by an expert (once per format, not per instance) and this information must be provided in machine readable form in order to be accessible by the tool checking deviations between input and result after round trip mapping.
6.2.4 Next Steps

The work will initially target assessment of input metadata, and then address assessment of mapping results. In the meantime, semantic and social enrichment tools will be further developed, and their output will be addressed.
7 Conclusions

Using the ConfigTool as basis for the mapping from the data provider input schemas to the EEXCESS output schema lead to the realisation that the mapping can be done easily if the user knows about the input structure of the data and about the target schema as well. The tools can be used by non-experts in the field of information technology as well but domain knowledge in the cultural fields regarding field semantic will be necessary as mapping of cultural data from local input schemas to formatted target schemas needs background knowledge of semantic of the input data. The chosen approach of having not only direct transformation into the EEXCESS format is very good and flexible so that further data provider can be integrated quite easily.

The transformation routine is already integrated in the PartnerRecommender. The actual Chrome Browser plugin prototype shows that the approach of having the transformation on the PartnerRecommender side that was followed within EEXCESS is very good. Improvements are necessary in the extension in the mapping of more metadata fields.

Regarding Metadata Extraction and Enrichment the first test with the actual enrichment implementation shows that the quality of the enriched content has to be increased before providing this additional content to the user.
8 References


[RDFValWs, 2013] RDF Validation Workshop Report, Practical Assurances for Quality RDF Data. 10-11 September 2013, Cambridge, MA, USA


9 Glossary

Terms used within the EEXCESS project.

**Partner Acronyms**

<table>
<thead>
<tr>
<th>Acronym</th>
<th>Full Name</th>
</tr>
</thead>
<tbody>
<tr>
<td>JR-DIG</td>
<td>JOANNEUM RESEARCH Forschungsgesellschaft mbH, AT</td>
</tr>
<tr>
<td>Uni Passau</td>
<td>University of Passau, GE</td>
</tr>
<tr>
<td>Know</td>
<td>Know-Zentrum - Kompetenzentrum für Wissenschaftsbasierte Anwendungen und Systeme Forschungs- und Entwicklungs Center GmbH, AT</td>
</tr>
<tr>
<td>INSA</td>
<td>Institut National des Sciences Appliquées (INSA) de Lyon, FR</td>
</tr>
<tr>
<td>ZBW</td>
<td>German National Library of Economics, GE</td>
</tr>
<tr>
<td>BITM</td>
<td>BitMedia, AT</td>
</tr>
<tr>
<td>KBL-AMBL</td>
<td>Kanton Basel Land, CH</td>
</tr>
<tr>
<td>CT</td>
<td>Collection Trust, UK</td>
</tr>
<tr>
<td>MEN</td>
<td>Mendeley Ltd., UK</td>
</tr>
<tr>
<td>WM</td>
<td>wissenmedia, GE</td>
</tr>
</tbody>
</table>

**Abbreviations**

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>EC</td>
<td>European Commission</td>
</tr>
<tr>
<td>EEXCESS</td>
<td>Enhancing Europe’s eXchange in Cultural Educational and Scientific resource</td>
</tr>
</tbody>
</table>

Acknowledgement: The research leading to these results has received funding from the European Union’s Seventh Framework Programme (FP7/2007-2013) under grant agreement n° 600601.