Bringing Your Content to the User, not the User to Your Content – A lightweight approach towards integrating external content via the EEXCESS framework

Martin Höffernig, Werner Bailer
JOANNEUM RESEARCH
SWIB 2015, Hamburg, 2015-11-23
Outline (1)

• Introduction to EEXCESS
• Tools for content injection
  – Install & try Chrome plugin
• Integrating a new data provider
  – Introduction to the data model
  – PartnerWizard
  – Integrate data provider with a web-based tool
Outline (2)

• Refining data mapping
  – Introduction to mapping tool
  – Review and update mappings
  – Test and check mappings

• Metadata quality assessment
  – Checking input and mapping quality
Logistics

- Wifi
  - SSID: SWIB*
  - Password: berners-lee
- Coffee break 15.30-16.00
- Short breaks in each of the blocks before & after (flexible timing)
Materials

Links, examples etc.

http://eexcess-dev.joanneum.at/swib15.html

Accounts: see handout

Slides: will be made available on EEXCESS website
EEXCESS - Enhancing Europe’s eXchange in Cultural Educational and Scientific resources

- EU FP7 project (Feb. 2013-Jul. 2016)
- 10 partners
  - technical partners
  - scientific partners
  - cultural institutions
Overview
Motivation

- Vast amounts of digital cultural and scientific resources available
- Still memory organisations (i.e. library, museums, archives) face challenges in disseminating their content
- Two reasons, addressed by EEXCESS:
  - Today's content dissemination processes are optimised for mainstream content
  - Long tail content needs contextualisation
Motivation

• Content provider strategies
  – Dedicated portals
  – Search engine optimisation
  – Social network marketing

• User strategies
  – Use major search engines
  – Use Wikipedia
The Long Tail Content

- Few sites get a large share of visits
- Large number of sites get a low share of visits
- A big, short “head”, but a (very) long tail

Challenges of the Long Tail

- High specialisation
- Low contextualisation
- Most items are unrelated
- Not easy to consume
- Low # of users per item
The value of long tail content

Value of Long Tail Content
- Discover new knowledge
- Verify information
- Enrich other content

Cultural Heritage content
- Multimedia Artefacts
- Original Material
- Explanations

Scholarly content
- Discourse
- Validated facts
- Additional explanations
Long Tail content dissemination
Challenges of today’s methods

Challenges
- Competition with mainstream content
- Highly commercialised
- Unawareness of existing portals
- Content is not contextualised
- User triggered

Search Engine Optimization
Social Media Marketing etc.
EEXCESS Vision

Unfold the treasure of cultural heritage and scholarly long-tail content for

• discovering new knowledge,
• triggering serendipitous effects,
• verifying consumed information,
• enriching new content

by “bringing the content to the user, not the user to the content”
Approach

Idea

„Bring the content to the user, not the user to the content“

- Inject cultural and scientific content into existing web channels
  - Websites (Wikipedia, etc.)
  - CMS/LMS
  - Social media channels (Twitter, etc.)
  - Support “head-channels” as well as tail-channels

- Contextualise Long Tail content
  - Context of the web channel
  - User Context
  - User Task

- Gather user and usage feedback such that memory organisations can optimise their resource distribution
Approach
Overview

Content Consumption (e.g. Browsing, SNA)

Involved in

Content Creation (e.g. Writing Blogs, Editors)

Involved in

Recommendation

Content

context

content

content

context

ZBW Content
Mendeley Content
AMBL Content
Europeana
CT Content
Open Access
Approach
Test Beds

3 User Groups as Test Beds

• Educational Support
  - Cultural/scientific resources injected to LMS
  - Pupils, teachers

• Scholarly Communication
  - Interconnecting cultural and scientific resources
  - Students, lecturers, researchers

• General Public Education
  – Disseminate cultural/scientific content to the general public
  – Regionally interested users, culturally interested users, media consumers
Objectives

• Adaptive Augmentation User Interfaces
• Personalized Recommendation
• Integration and Enrichment
• User and Usage Mining
• Privacy Preservation
Architecture

• Distributed data storage
  – Data remains with data providers
  – No central index

• Partner Recommender
  – Interface between data provider’s API and EEXCESS system

• Federated Recommender
  – Aggregates and ranks results
Architecture
Recommendation flow

14
Recommendation flow

• Implications from architecture
  – transformation and enrichment must work on the fly
  – configuration can be checked and revised manually, but transformation results cannot
  – no issues due to enrichment with resources that are no longer available
Querying partner sites

• Two step process
  – Speed up retrieving initial results
  – Reduce load on partner sites
• Initial query
  – Get basic metadata of entries
• Detail query
  – Additional metadata
  – Images
Metadata Enrichment

- Enriching textual information with named entities
- Type of metadata field is used to constrain entity type (e.g. persons) – search for entities with appropriate type
- Classify if words are entities in DBpedia
- Add synonyms using WordNet
- Add connected geographic terms using GeoNames
Content Injection – Chrome Browser Extension

Content Consumption

- A sidebar for recommending cultural/scientific content while browsing
Content Injection – Content Management Plugin (Wordpress)

Content Creation

• Inject cultural heritage and scholarly content into social media creation process
• Multiplier effect in the Blogging Community by providing a Wordpress Plugin
Content Injection – Google Docs App

Content Creation

• Inject cultural heritage and scholarly content into collaborative word processing
• Support writing reports, grant requests, homeworks
• Google Apps Market for Google Documents as high-potential dissemination platform

The vision of EEXCESS is to reveal and interconnect the treasuries of European culture, science and education.

European museums, archives, libraries and other cultural aggregators hold vast resources of cultural and educational content—scientific research, historical sound recordings, images of sculptures, films, sheet music and much more. This highly specialised, carefully curated content is still largely unexplored and invisible to the general public. In the Web context it is therefore often referred to as the long tail (i.e. a huge body of specialised knowledge existing in the Wide Wide Web, but hidden from most users).

The vision of EEXCESS is to connect those valuable resources with the mainstream content available via Internet giants such as Google, Facebook and Wikipedia. Our research is based on the simple principle of taking the content to the user, not the user to the content. We want to bring the information directly to the users’ working environment, on their favourite platforms (Facebook, Twitter etc) and their preferred devices (tablets, smartphones etc.). Instead of navigating through a multitude of libraries, repositories and databases, users will find relevant and specialised information in their habitual environment. By this means, we will bring the long-tail content to the surface, where it can be put to new and different uses.

As these improvements depend on a large degree on persons providing information about their usage behaviour (i.e. click-rates, technical details of the devices used etc.), privacy protection issues play a large role in the research goals of EEXCESS. Data will only be collected with prior and express permission of users and anonymised.

Thus unfolding the treasures of culture, science and education for the benefit of all users in both their professional and private lives, EEXCESS will contribute to cultural, scientific and economic prosperity throughout the European Union.

As a simple example, if you’d write an article about Ada Lovelace, you could include interesting resources found with EEXCESS using the Google Docs Plugin.
Content Injection – Collection Management System
Content Injection – Collection Management System
Content Injection – Learn Management Systems

Content Creation for Educational Support

• Inject cultural heritage content into Learn Management Systems
• Moodle and BitMedia‘s SITOS LMS
Privacy vs. Personalisation trade-off?
Privacy vs. Personalisation trade-off?

Privacy

Personalisation/Quality
Privacy vs. Personalisation trade-off?

User Awareness (and Transparency)

User Empowerment

User Privacy Protection (Privacy Proxy)
PEAS: Unlinkability Protocol

- PEAS: Private, Efficient, and Accurate web Search
- Hypothesis
  - only the user’s device is trusted
- Split the Privacy Proxy into two pieces
  - Receiver: knows the user, but not the content of the query
  - Issuer: knows the content of the query, but not the user
  - Both are supposed “honest but curious” and do not collude
PEAS: Unlinkability Protocol (simplified)

**Privacy Proxy**

- **$u$:** User
- **Receiver**
  - $b = \text{generateKey}()$
  - $q' = \text{encrypt}_a(q + b)$
- **Issuer**
  - $q' = \text{decrypt}_a(q')$
- **FR**
  - $R' = \text{encrypt}_b(R)$
  - $R = \text{decrypt}_b(R')$
  - $q + b = \text{decrypt}_a(q')$
PEAS: Indistinguishability Protocol (simplified)

• Protocol divided into two parts
  – Obfuscation (done at the user’s side): add fake queries
    • to mislead attackers, fake queries have the same structure as the original one, are built other users’ queries, but are semantically different from the original query
  – Filtering: remove irrelevant results
PEAS: Indistinguishability Protocol (simplified)

\[ q^+ = \text{obfuscation}(q) \]

\[ R^+ = \text{filtering}(R^+) \]
PEAS: Combination of Protocols

User

$q^+ = \text{obfuscation}(q)$

$R^+ = \text{unlinkability}(q^+)$

$R = \text{filtering}(R^+)$
Privacy Settings

- Transparent to user
- Choice which information to expose
- Choice to switch on/off different privacy features
Data Model
Data model

• Need to combine search results from different providers
• Perform duplicate removal, ranking
• Perform semantic enrichment
• Provide metadata in unified format to the client applications
EEXCESS Ontology

- Based on existing data models (EDM/PROV)
- Analysed data providers’ formats
  - data providers investigated their data formats
  - identified overlaps and core metadata elements
- Defined EEXCESS Ontology
- Validated ontology by mapping data providers’ formats
EEXCESS Ontology

• Europeana Data Model - EDM
  – Represents metadata of cultural heritage objects (CHO)
  – **CHO**: real world resource
  – **Proxy**: representation CHO from one source
  – **Agent**: data provider
  – **Aggregation**: puts CHO, Agent and Proxy in relation

• EDM and EEXCESS
  – Objects are modeled as EDM CHO
  – Annotations are modeled using EDM Proxies
  – Data providers are modeled as EDM Agents
  – Aggregation is used as in EDM
EDM – Main entities
EDM – Proxy example

Context-specific “view” on object
EEXCESS Ontology

• W3C PROV
  – describes how things are created or delivered
  – **Entity**: physical, digital, conceptual, or other kinds of things
  – **Activity**: how entities are created or changed
  – **Agent**: takes a role in performing an activity

• PROV and EEXCESS
  – Objects and Proxies are modeled as PROV entities
  – Metadata creation is modeled as PROV activity
  – Creator of metadata is modeled as PROV agent
EEXCESS Ontology

- **eexcess:Object**
  - Single item curated by a data provider

- **eexcess:Agent**
  - Data provider
  - Annotator of existing content

- **eexcess:Proxy**
  - Groups metadata from one source
EEXCESS Ontology, EDM and W3C PROV

- `ore:Proxy` subClassOf `eexcess:Proxy`
- `ore:Aggregation`
- `edm:ProvidedCHO` subClassOf `eexcess:Object`
- `edm:Agent` subClassOf `eexcess:Agent`
- `prov:Agent` subClassOf `eexcess:Agent`
- `prov:Activity` subClassOf `prov:Entity`
Representation

• Serialisation
  – RDF/XML
  – JSON-LD

• Not stored, but exchanged between Partner Recommenders, Federated Recommender and clients
PartnerWizard
Motivation

• Connect more data providers to the EEXCESS system
• Make it easy to achieve basic integration
• Allow setup without the need to write code
• Jump start software development by starting from a template
Overview

Build a new PartnerRecommender
• Create a new project
• Configure QueryGeneration, API-endpoints, ...
• Implement special Classes e.g. QueryGeneration, Transformation, ...
• Configure for EEXCESS-DEV-Server
• Deployment on local PC/Server
• New PartnerRecommender register on DEV-FederatedRecommender
• Download Chrome plugin from WebStore
• Configure Chrome plugin to EEXCESS-DEV-Server

User will see their data integrated in the Chrome plugin
maven archetype

- Projects are built with maven
  - Defines dependencies incl. version of the lib
  - repositories
- maven archetype – project templating toolkit
- maven provides command to create an archetype from an existing project
maven archetype

- Existing PartnerRecommender as input
- Defining Parameters for the new archetype
- Replaced the specific code with placeholder
Parameters for maven archetype: EEXCESS archetype

package=at.joanneum
version=0.1-SNAPSHOT
groupId=eu.eexcess
artifactId=myPRTTest
partnerName=Partner Name
partnerURL=http://example.org/
dataLicense=unknown license
partnerAPIsearchEndpoint=https://kgapi.bl.ch/solr/kim-portal.objects/select/xml?q=_fulltext_:{query}&rows=${numResults}
partnerAPIsearchTerm=s
partnerAPIsearchMappingFieldsLoopXPath=/response/result/doc/
partnerAPIsearchMappingFieldsXPathID=str[@name='uuid']
partnerAPIsearchMappingFieldsXPathURI=str[@name='uuid']
partnerAPIsearchMappingFieldsXPathTitle=str[@name='_display_']
partnerAPIsearchMappingFieldsXPathDescription=str[@name='beschreibung']
partnerAPIdetailEndpoint=https://kgapi.bl.ch/solr/kim-portal.objects/select/xml?q=uuid:${detailQuery}
partnerAPIdetailTerm=s
partnerAPIdetailMappingFieldsLoopXPath=/response/result/doc/
partnerAPIdetailMappingFieldsXPathID=str[@name='uuid']
partnerAPIdetailMappingFieldsXPathURI=str[@name='uuid']
partnerAPIdetailMappingFieldsXPathTitle=str[@name='_display_']
partnerAPIdetailMappingFieldsXPathDescription=str[@name='beschreibung']
Query Optimiser

- Optimise query to partner sites
- Test different query options, e.g.
  - AND vs. OR of query terms
  - use of query expansion
- Expert selection from examples
- Automatically adjust query configuration of PartnerRecommender
Query Optimiser
Query Optimiser

EEXCESS Partner Wizard

Partner Wizard Query Generation Configuration

With this tool you can optimize the query generation strategy for your system.

Customize example queries

Keyword: Napoleon

Keyword: Frankreich

Keyword: Schweiz

Keyword: Helena

Add Keyword

Remove Query

Keyword: Roman Missal

Add Keyword

Remove Query

Add Query

Start
Query Optimiser

EEXCESS Partner Wizard

Keywords:
Roman Missal

List 1

<table>
<thead>
<tr>
<th>DVD: Naturparadies Bruderholz</th>
</tr>
</thead>
<tbody>
<tr>
<td>DVD: Naturparadies Bruderholz</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td>DVD: Naturparadies Bruderholz</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
</tbody>
</table>

List 2

<table>
<thead>
<tr>
<th>DVD: Naturparadies Bruderholz</th>
</tr>
</thead>
<tbody>
<tr>
<td>DVD: Naturparadies Bruderholz</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
<tr>
<td>DVD: Naturparadies Bruderholz</td>
</tr>
<tr>
<td>-----------------------------</td>
</tr>
</tbody>
</table>

EEXCESS Partner Wizard

Finished, here is your configuration:

```json
{"keywords":null,"queryGeneratorClass":"eu.excess.partnerrecommender.reference.OrQueryGenerator","queryExpansionEnabled":false,"querySplittingEnabled":false}
```

Keywords:

- queryGeneratorClass: eu.excess.partnerrecommender.reference.OrQueryGenerator
- queryExpansionEnabled: false
- querySplittingEnabled: false
Metadata Mapping Configuration Tool
Motivation

- Convert XML-based metadata documents between different metadata formats
  - Data providers’ formats from and to the EEXCESS data model
- Define and configure mapping instructions
  - Avoid hand-crafted 1:1 mappings
  - Infer mapping instructions
  - Mappings are easier to maintain
  - Adding new metadata formats without side effects
Metadata Mapping Configuration Approach

- Derive mapping instructions based on a mapping ontology
Metadata Mapping Configuration Approach

• **Mapping Ontology**
  – Define mappings between metadata properties from different formats
  – Formalized with respect to a conceptual representation of metadata properties serving as hub
  – Additional localization and context information

• **Structural description of the target metadata format**

• **Result: XSL template**
Metadata Mapping Configuration Workflow

- Define format-specific metadata concepts
- Define mappings of the format-specific concepts to the conceptual representation
- Adding data type, localisation, structure information to format-specific concepts
- Create/edit structural representation of target format
- Create mapping instructions
  - Retrieve mapping parameters from mapping ontology
  - Merged into output structure
Metadata Mapping Configuration Tool

• Implemented as web application
• Configuration of metadata mapping
• Define relations between metadata fields by drag and drop
• Define data type mappings
• Define the output structure
• Preview of created mappings
Metadata Mapping Configuration Tool

- Demo

Metadata Conversion Configuration Tool 0.8.3.BETA

Project: (no project selected)

Logout in 58 minutes

New Project
- Please enter a name for the new project:
  - wissenserver2eexcess

Source Schema
- Wissensserver

Target Schema (optional)
- EEXCESS

Create New Project

Recent Projects
- imdas2eexcess
- w1
Metadata Mapping Configuration Workflow

Concept Mappings

- based on meon ontology
Metadata Mapping Configuration Workflow
Datatype Representations
Metadata Mapping Configuration Workflow

Derive Mapping Parameters

• Mapping Parameters Inference
Create Mapping Instructions

Example

Output Structure:
<xsl:stylesheet>
  <xsl:element name="eexcess:Proxy">
    ...<xsl:call-template name="Main.Description"/>
    ...
  </xsl:element>
</xsl:stylesheet>

Mapping Parameters:
Template Name: Main.Description
XPath: /intro
Output Structure: dc:Description
Mapping Template: StringToString

Mapping Instructions:
<xsl:template name="Main.Description">
  <apply-templates select="intro"/>
</xsl:template>

<template match="intro">
  <element name="dc:description">
    <call-template name="StringToString"/>
  </element>
</template>
Metadata Quality
Motivation

• Metadata from many sources
• Heterogeneous formats (and thus conversions)
• Different workflows
• Context
Three subproblems

• Assessing Input Data Quality
• Assessing Enrichment Results
• Assessing Mapping Results
Input data quality – metrics

• Statistics about input data

• Completeness of records
  – fields/record (min, max, average)
  – # empty fields/record

• Structuredness of data
  – for example the structuredness of date, name fields
  – Structured element or format specification (e.g. using XML Schema regular expressions)
Input data quality – metrics

• Use of controlled vocabularies
• Availability of linked resources

• Evaluated on data collected during testbed on 6K records
Completeness

mean data fields / record / provider

- KIMCollect
- ZBW
- Europeana
- Wissenmedia
- Mendeley
- DDB
Completeness
Completeness
Structuredness

- Length of value -> histogram
- Group characters and numbers
- Infer candidate patterns – e.g. Height: 00.0aa Width: 0.0aa
- Histogram of candidate patterns
- Detect known particles (e.g. SI unit abbreviations)
URLs in record

• Counting URLs in responses
• Check if URL accessible
• Check type of response
  – XML/RDF, XML, HTML
  – determine if result is machine readable
URLs used in records

![Bar chart showing mean links per record per provider.](chart.png)
URLs used in records (resolvable)
Enriching and transforming data

• Apply the same metrics before and after transformation or enrichment
• Compare values, e.g.
  – decrease in number of empty fields
  – increase in use of controlled vocabularies
  – Increase in resolvable URLs in the data
Use of input metadata quality results

• Statistics, completeness, etc.
  – Provide feedback to data provider
  – Improve result representation returned by data providers

• Structuredness
  – More appropriate mapping
  – Detect outliers on the fly (avoid errors)
Use of input metadata quality results

• Use of controlled vocabularies
  – Need for detecting/replacing named entities
  – Detect need to map vocabulary (to a standard and/or accessible one)
Mapping Quality Assessment

• Assessment of mapping results
  – Comparison against an expert created reference
  – Round trip mapping via intermediate format
    • e.g., ZBW -> MEON -> ZBW
    • no expected loss
  – Round trip mapping via target format
    • e.g., ZBW -> EEXCESS -> ZBW
    • possibly expected loss
Mapping Quality Assessment

**Recent XML Data Files**

- k1.xml (167B)

**Mapping Quality Assessment**

<table>
<thead>
<tr>
<th>Input XPath</th>
<th>Output XPath</th>
<th>Result</th>
</tr>
</thead>
<tbody>
<tr>
<td>/objects[1]/object[1]/Fotograf[1]</td>
<td>/objects[1]/object[1]/Hersteller</td>
<td>Mapping missing</td>
</tr>
</tbody>
</table>

**Add Locator Association**

**Edit Locator Association**

- /objects/object/Hersteller
- /objects/object/Autor
- /objects/object/Fotograf
Data Quality Assessment – Result Representation

- Requirements
  - Well-defined
  - Structured
  - Machine-readable
Data Quality Assessment – Result Representation

- W3C Data Quality Vocabulary (DQV) - First Public Working Draft 25 June 2015
  http://www.w3.org/TR/2015/WD-vocab-dqv-20150625/
  - Data Catalog Vocabulary (DCAT) – Recommendation(2014)

- Dataset(DCAT)
- Distribution(DCAT)
- Metric(DQV)
- QualityMeasure(DQV)
Data Quality Assessment – Result Representation

```xml
<dataset rdf:about="#eexcessDataset">
    <title>My EEXCESS dataset</title>
    <distribution>
        <distribution rdf:about="#eexcessDatasetZBWDistribution">
            <title>My EEXCESS ZBW dataset</title>
            <prov:wasGeneratedBy rdf:resource="#ZBW"/>
        </distribution>
    </distribution>
    <distribution>
        <distribution rdf:about="#eexcessDatasetZBWTransformationDistribution">
            <title>My EEXCESS ZBW Transformation dataset</title>
            <prov:wasGeneratedBy rdf:resource="#EEXCESSTransformation"/>
            <prov:wasDerivedFrom rdf:resource="#eexcessDatasetZBWDistribution"/>
        </distribution>
    </distribution>
    <distribution>
        <distribution rdf:about="#eexcessDatasetZBWEnrichmentDistribution">
            <title>My EEXCESS ZBW Enrichment dataset</title>
            <prov:wasGeneratedBy rdf:resource="#EEXCESSEnrichment"/>
            <prov:wasDerivedFrom rdf:resource="#eexcessDatasetZBWTransformationDistribution"/>
        </distribution>
    </distribution>
</dataset>
```
Data Quality Assessment – Result Representation

<daq:Metric rdf:about="#eexcessDataQMetricNumberOfRecords">
</daq:Metric>
<daq:Metric rdf:about="#eexcessDataQMetricNumberOfFields">
</daq:Metric>

<dqv:QualityMeasure rdf:about="#measureNumberOfRecordsZBW">
  <daq:value rdf:datatype="http://www.w3.org/2001/XMLSchema#double">102</daq:value>
  <daq:computedOn rdf:resource="#eexcessDatasetZBWDistribution"/>
  <daq:metric rdf:resource="#eexcessDataQMetricNumberOfRecords"/>
</dqv:QualityMeasure>

<dqv:QualityMeasure rdf:about="#measureNumberOfFieldsZBW">
  <daq:value rdf:datatype="http://www.w3.org/2001/XMLSchema#double">10</daq:value>
  <daq:computedOn rdf:resource="#eexcessDatasetZBWDistribution"/>
  <daq:metric rdf:resource="#eexcessDataQMetricNumberOfFields"/>
</dqv:QualityMeasure>

<dqv:QualityMeasure rdf:about="#measureNumberOfFieldsZBWAfterTransformation">
  <daq:value rdf:datatype="http://www.w3.org/2001/XMLSchema#double">10</daq:value>
  <daq:computedOn rdf:resource="#eexcessDatasetZBWTTransformation"/>
  <daq:metric rdf:resource="#eexcessDataQMetricNumberOfFields"/>
</dqv:QualityMeasure>
Visualisation from DQV

- Generate diagrams using XSLT