

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.

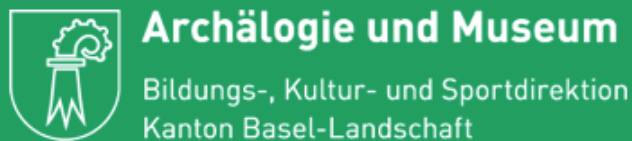
<http://eexcess-dev.joanneum.at/swib15.html>

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





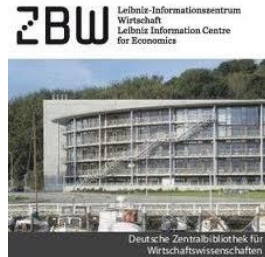


Overview



Motivation

- Vast amounts of digital cultural and scientific resources available



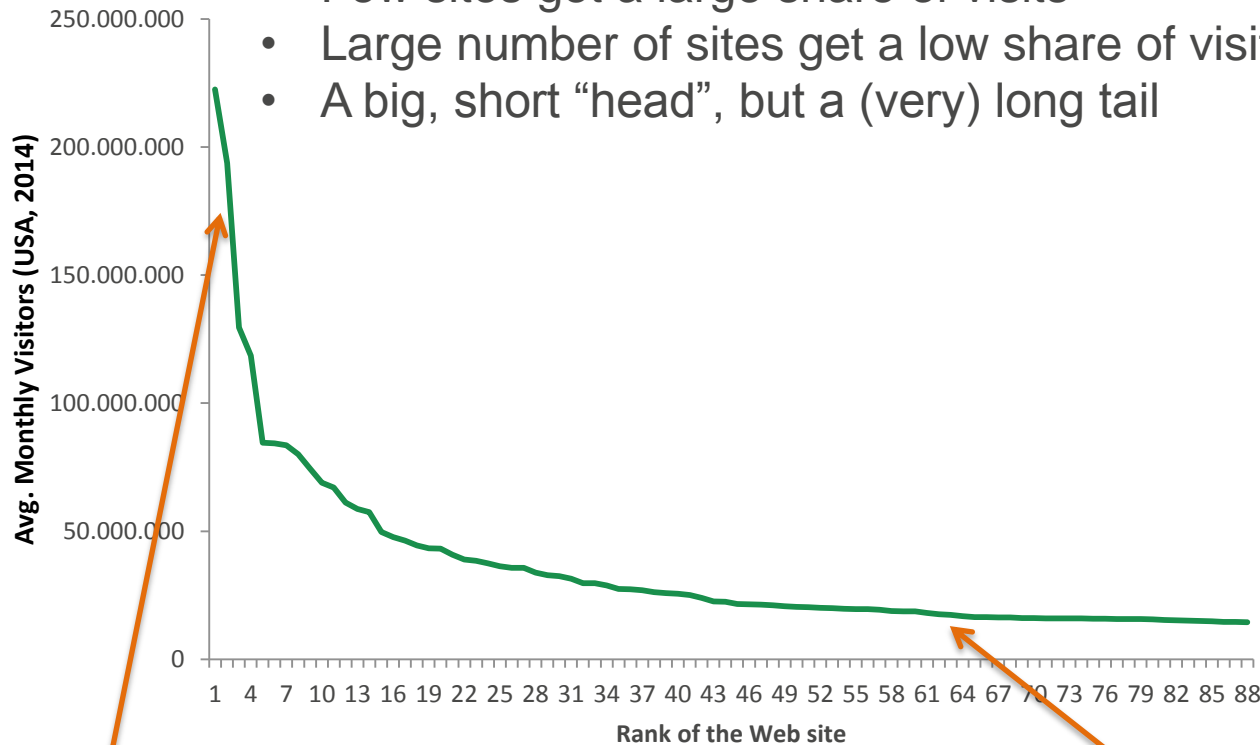
- Still memory organisations (i.e. library, museums, archives) face challenges in disseminating their content
- Two reasons, addressed by EEXCESS:
 - Today's content dissemination processes are optimised for mainstream content
 - Long tail content needs contextualisation

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

Google



You
Tube



ZBW

MENDELEY

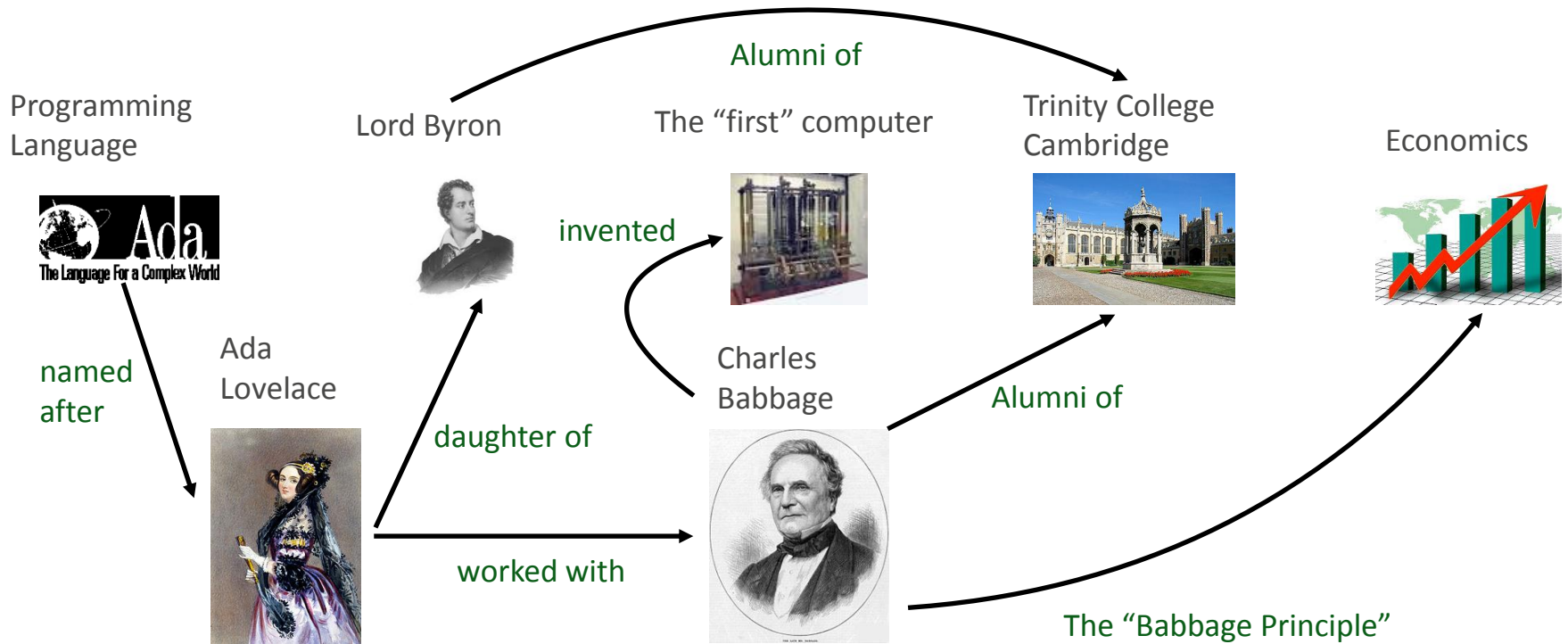
Archäologie und Museum
Bildungs-, Kultur- und Sportdirektion
Kanton Basel-Landschaft

wissenmedia

Collections
Trust

bit
best in training

The value of long tail content



- Cultural Heritage content
- Multimedia Artefacts
 - Original Material
 - Explanations



Value of Long Tail Content

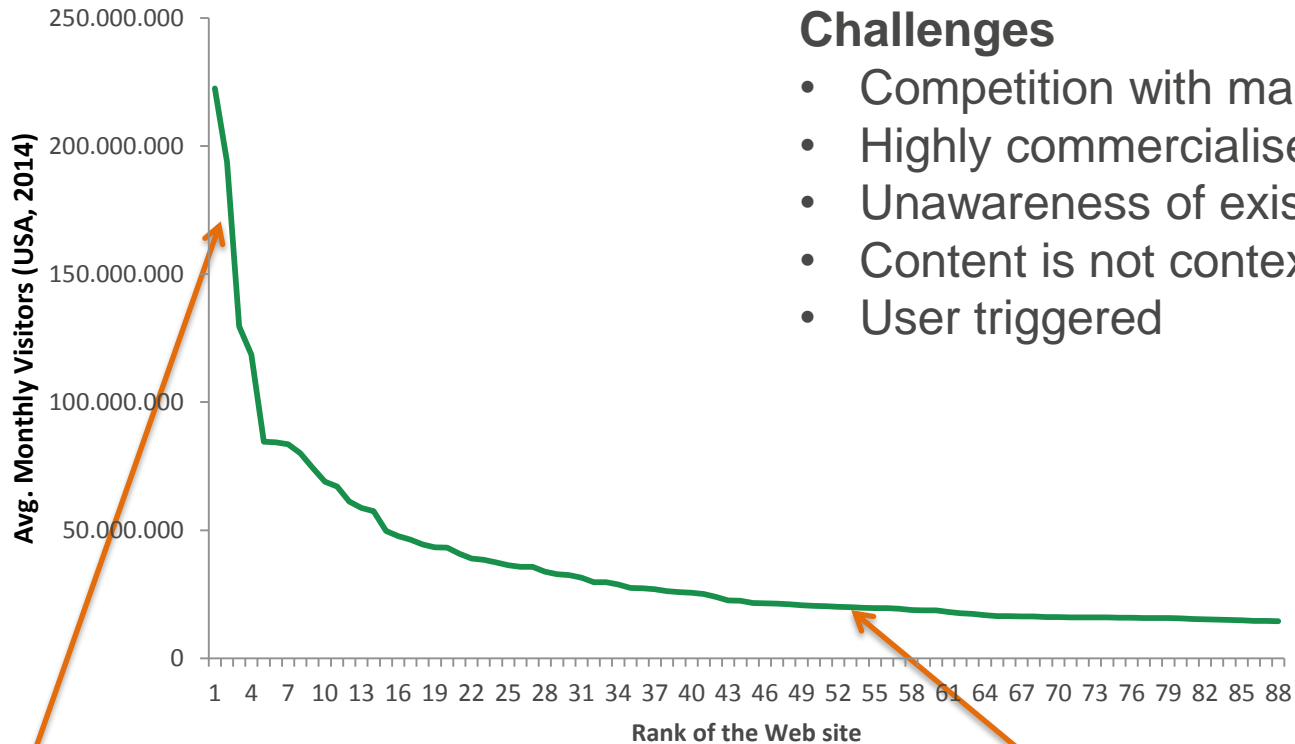
- Discover new knowledge
- Verify information
- Enrich other content

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

Google



Search Engine Optimization
Social Media Marketing etc.

YouTube



ZBW

MENDELEY

Archäologie und Museum
Bildungs-, Kultur- und Sportdirektion
Kanton Basel-Landschaft

wissensmedia

Collections Trust

LEXbit
best in training



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

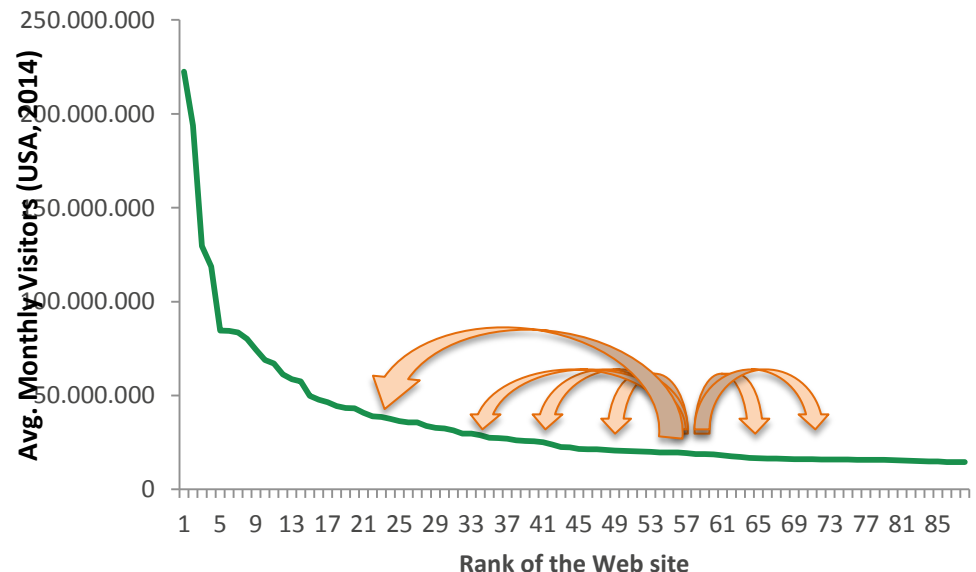
by “bringing the content to the user, not the user to the content”

Approach

Idea

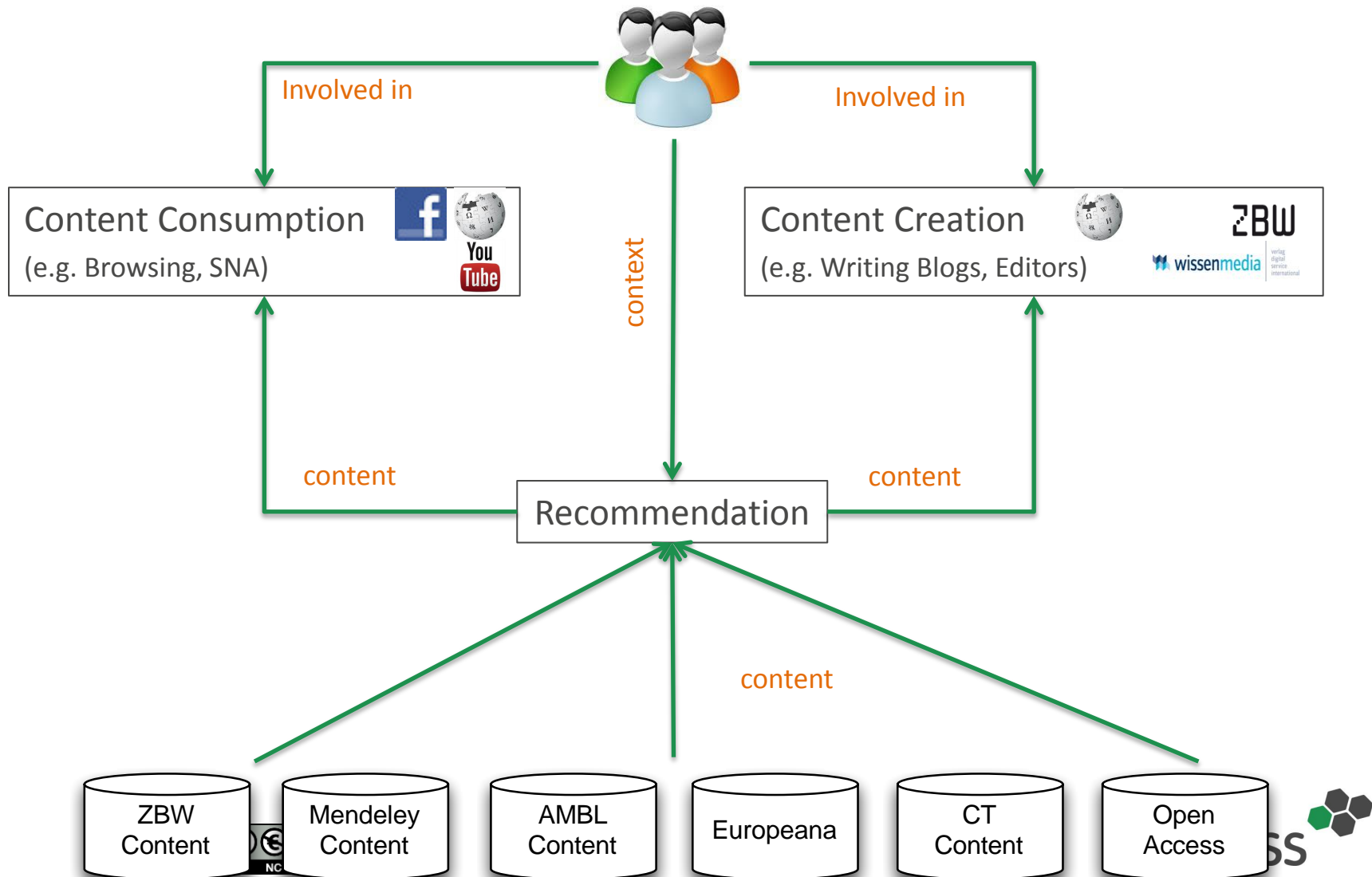
„Bring the content to the user, not the user to the content“

- 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



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



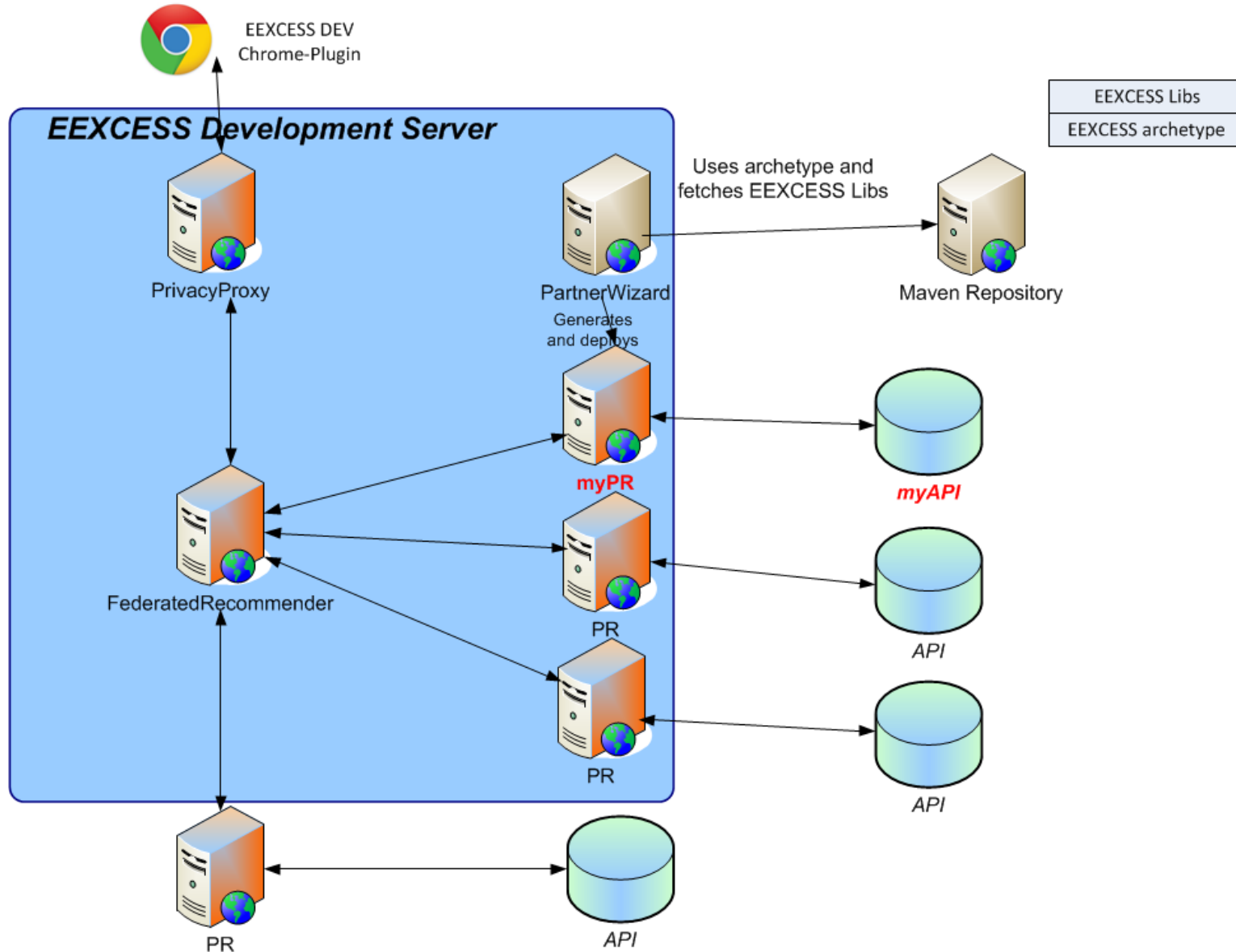
Objectives

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

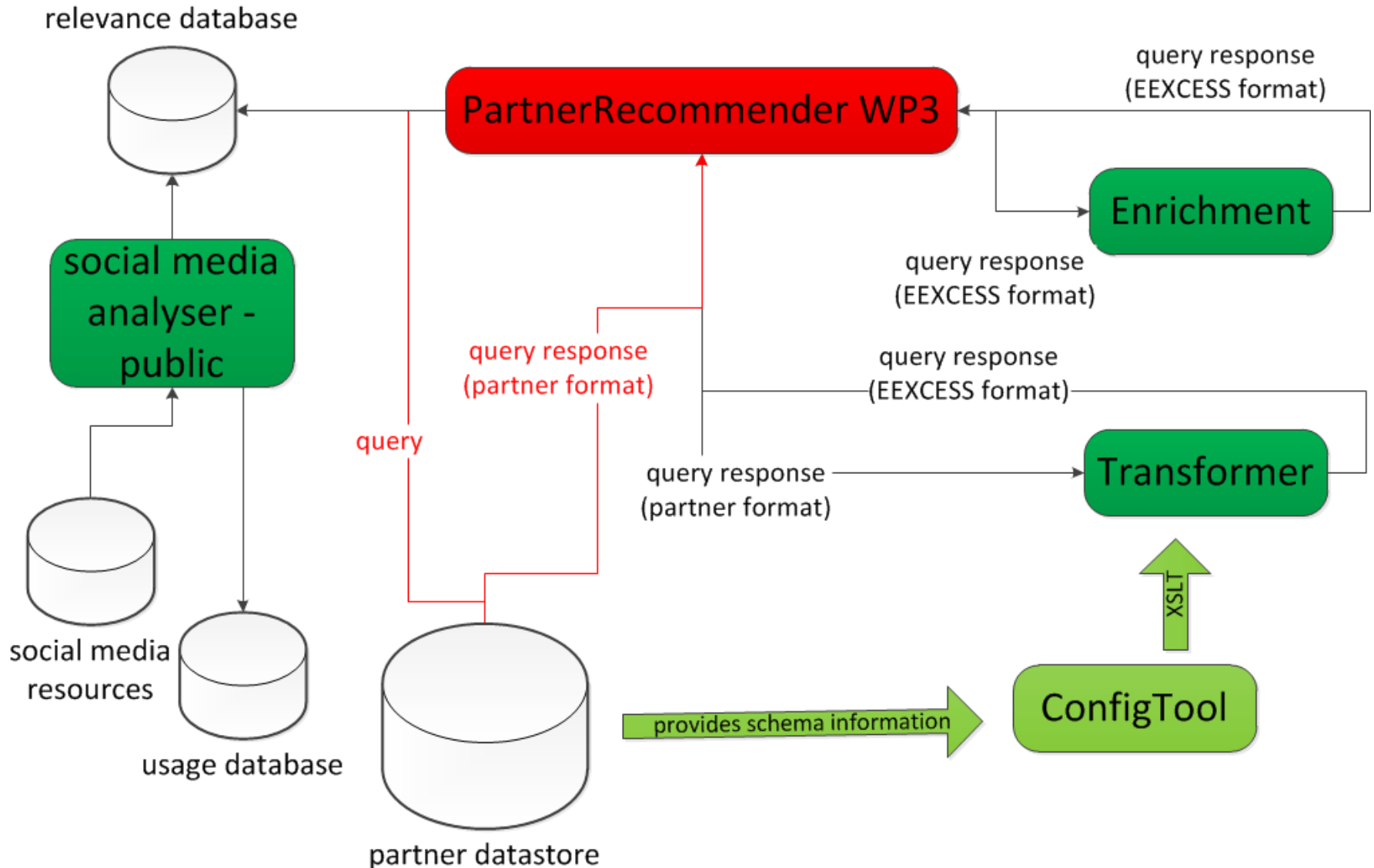
Architecture

- 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



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

The screenshot shows a Wikipedia article for Ada Lovelace. The article text is partially visible, discussing her work on Charles Babbage's Analytical Engine. A sidebar on the right, titled "Ada, Countess of Lovelace", features a portrait of her. Below the article, a grid of recommended content is displayed, including various articles and images related to mathematics, science, and history. The interface includes a search bar, navigation links, and a footer with page information.

WIKIPEDIA The Free Encyclopedia

Article Talk

Read Edit View history Search

European Science Photo Competition 2015 – participate now!

Ada Lovelace

From Wikipedia, the free encyclopedia

Augusta Ada King, Countess of Lovelace (née **Byron**; 10 December 1815 – 27 November 1852) was a British mathematician and writer, chiefly known for her work on Charles Babbage's early mechanical general-purpose computer, the *Analytical Engine*. Her notes on the engine include what is recognised as the first *algorithm* intended to be carried out by a machine. Because of this, she is often regarded as the first computer programmer.^{[1][2][3]}

Lovelace was born 10 December 1815 as the only legitimate child of the poet *George Gordon, Lord Byron* and his wife *Anne Isabella Milbanke, Baroness Wentworth*.^[4] All *Byron's other children* were born out of wedlock to other women.^[5] Byron separated from his wife a month after Ada was born and left England forever four months later, eventually dying of disease in the *Greek War of Independence* when Ada was eight years old. Ada's mother remained bitter towards Lord Byron and promoted Ada's interest in mathematics and logic in an effort to prevent her from developing what she saw as the insanity seen in her father, but Ada remained interested in him despite this (and was, upon her eventual death, buried next to him at her request).

Ada described her approach as "poetical science"^[6] and herself as an "Analyst (& Metaphysician)"^[7] As a teenager, her mathematical talents led her to an ongoing working relationship and friendship with fellow British mathematician Charles Babbage also known as 'the father of computers', and in particular, Babbage's work on the Analytical Engine. Between 1842 and 1843, she translated an article by Italian military engineer *Luigi Menabrea* on the engine, which she supplemented with an elaborate set of notes of her own, simply called *Notes*.

These notes contain what many consider to be the first computer program—that is, an algorithm designed to be carried out by a machine. Lovelace's notes are important in the early *history of computers*. She also developed a vision of the capability of computers to go beyond mere calculating or number-crunching, while others, including Babbage himself, focused only on those capabilities.^[8] Her mind-set of "poetical science" led her to ask questions about the Analytical Engine (as shown in her notes) examining how individuals and society relate to technology as a collaborative tool.^[5]

Ada, Countess of Lovelace

Landschap met koeien

Elementary logic for philosophy of science and economic methodology

"Un atelier de sculpture"

Een minnezanger, naar Couture

The core can be accessed with a bounded number of blocks.

Charles Babbage

All but one: How pioneers of linear economics overlooked Perron-Frobenius mathematics

Programmer la recherche : facile mais... impossible

Mathematics

Charles Babbage: an inadvertent development economist

Elemek:

Interest rates, debt and intertemporal allocation: evidence from notched mortgage contracts in the United Kingdom

Une sieste en Italie

Sectoral shocks and monetary policy in the United Kingdom

"Un torrent"

Algorithmic Aspects of k-tuple Domination in Graphs

Sectoral shocks and monetary policy in the United Kingdom

L'amour de l'or

On the Mathematical Discovery of new Physical Phenomena

On the Mathematical Discovery of new Physical Phenomena

Conceptions of mathematics among diploma in actuarial science students

"La messe pendant la moisson dans la campagne de Rome"

De droom van Byron

Rechttoeige hanger aan collier samengesteld uit diverse materialen

MATHEMATICS KINDERGARTEN TO GRADE 6

Folded copper sheet mathematical model.

Landschap in de manier van Rembrandt

Matematikai kincsek Bolyai János kéziratás hagyatékából.

Het nieuw vermakelijk ganzenspel / voor de lieve jeugd aan onze hoogeschool

show all images (3) text (27) original order title date

show all Mathematics Charles Babbage x Lord Byron x Algorithm x Mathematician x Science x United Kingdom x Society x Logic x Ada Lovelace x Anne Isabella Byron, Baroness Byron x England x Programmer x Insanity x Military engineering x

30 results

EEXCESS

Seite 18

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

The screenshot displays the WordPress 'Add New Post' editor. The main content area contains a paragraph of text about Augusta Ada Byron King, Countess of Lovelace, which is a snippet from the EEXCESS plugin. The text reads: "Augusta Ada Byron King, [Countess of Lovelace](#), allgemein als [Ada Lovelace](#) bekannt (geborene [Augusta Ada Byron](#)) * 10. Dezember 1815 in London; † 27. November 1852 ebenda), war eine britische Mathematikerin. Für einen nie fertiggestellten mechanischen Computer, die Analytical Engine, schrieb sie das erste veröffentlichte Programm und nahm wesentliche Aspekte späterer Programmiersprachen wie etwa das Unterprogramm oder die Verzweigung vorweg. Aus diesem Grund wird sie heute als erste Programmiererin der Welt bezeichnet [2][3] - fast einhundert Jahre vor den modernen Pionieren der Programmierung wie Grace Hopper, Jean Bartik oder Howard Aiken. Die Programmiersprache Ada[4] und die [Lovelace Medal](#) wurden nach ihr benannt."

Below the text, the EEXCESS search results are visible, showing a list of items related to the search term "Countess of Lovelace". The results include:

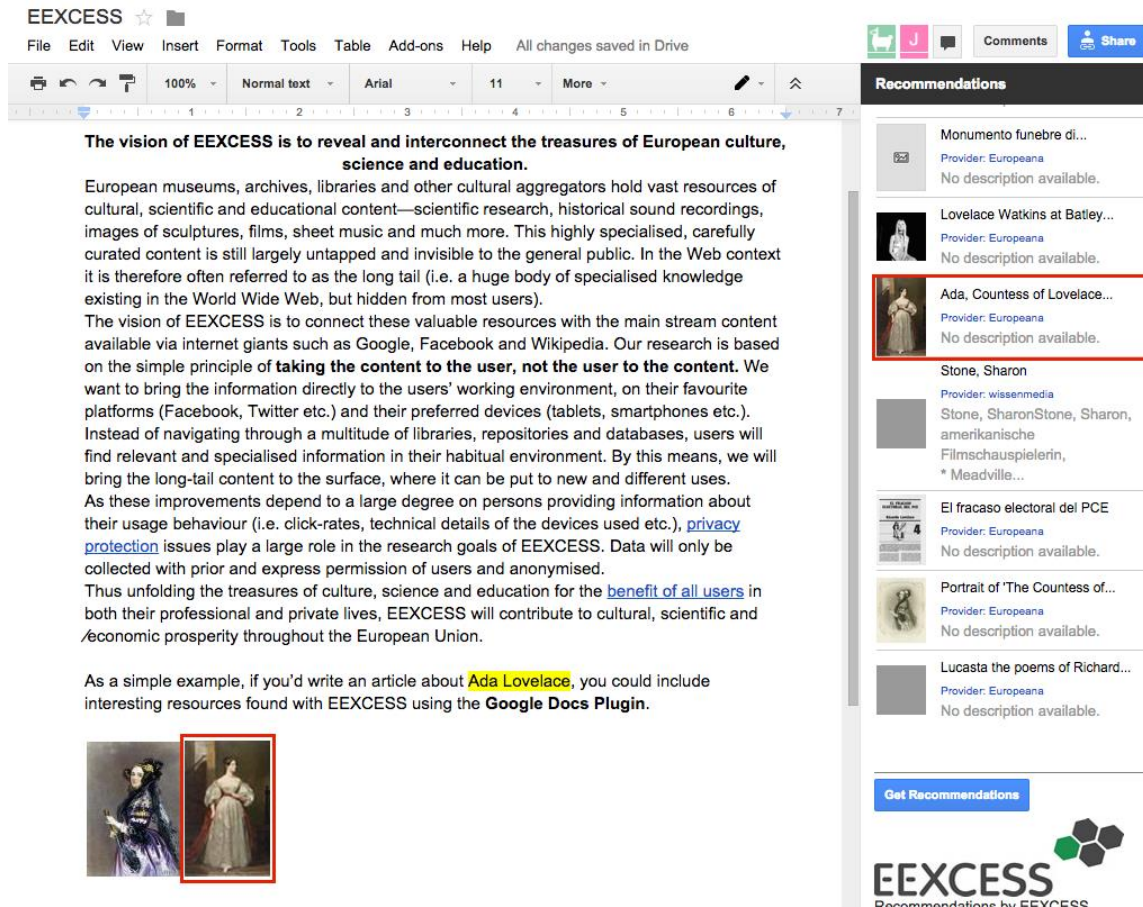
- Parish Church, c 1790 view from South West.
- General Pichegru, Bear Courtesis De Part
- Portrait of Ada, Countess of Lovelace
- L.A. Henry de Bourbon - Duc d'Angoulême
- L.A. Henry de Bourbon - Duc d'Angoulême - Count de Calv, John Munné (Litgewer)
- Diabetes and managed care: the Lovelace Health System's Episode of Care Program
- Management of diabetes mellitus in the Lovelace Health System: EPISODES OF CARE program.
- Vision and kinesthesia in accuracy of hand movement
- Right Stuff, Wrong Sex: The Lovelace Woman in Space Program (1960-1962).
- St Mary Magdalene's Church, Hucknall - from the south
- Ada, Countess of Lovelace (1815-1852) Mathematician; Daughter of Lord Byron

The interface also shows the WordPress dashboard sidebar on the left, the 'Publish' and 'Format' panels on the right, and the 'EEXCESS' search results panel at the bottom.

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 screenshot shows a Google Docs interface with a document titled "EEXCESS". The document content includes a mission statement and a paragraph about cultural heritage. A sidebar on the right displays a "Recommendations" list with several items, one of which is highlighted with a red box. At the bottom of the document, two portrait images are shown, also highlighted with a red box.

EEXCESS ☆

File Edit View Insert Format Tools Table Add-ons Help All changes saved in Drive

100% Normal text Arial 11 More

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

European museums, archives, libraries and other cultural aggregators hold vast resources of cultural, scientific 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 untapped 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 World Wide Web, but hidden from most users).

The vision of EEXCESS is to connect these valuable resources with the main stream 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 to 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**.

Recommendations

- Monumento funebre di...
Provider: Europeana
No description available.
- Lovelace Watkins at Batley...
Provider: Europeana
No description available.
- Ada, Countess of Lovelace...**
Provider: Europeana
No description available.
- Stone, Sharon
Provider: wissenmedia
Stone, SharonStone, Sharon,
amerikanische
Filmschauspielerin,
* Meadville...
- El fracaso electoral del PCE
Provider: Europeana
No description available.
- Portrait of 'The Countess of...
Provider: Europeana
No description available.
- Lucasta the poems of Richard...
Provider: Europeana
No description available.

Get Recommendations

EEXCESS
Recommendations by EEXCESS

Content Injection – Collection Management System

imdas pro 6.1.1 Benutzer: SA DB-Version: 06.00.0002 Datenbank: IMDASKIRLING - [Objektbearbeitung - (3 / 5610 Objekte)]

Datei Bearbeiten Recherche Objekt Module Navigation Ansicht Stammdaten Thesaurus Fenster ?

Betrachten Museumsobjekt: Objektbeschreibung für das Web freigegeben

Museumsobjekt

Museumsobjekt

- Objektbearbeitung
 - Objektbeschreibung
 - Objektverwaltung
 - Gesamtansicht (Standard)
 - Registrierung- Eingangsbuch (Standard)
 - Registrierung- Eingangsbuch (Erweitert)
 - Inventarisierung
- Detailinformationen
 - Sacherschließung (2)
 - Literaturhinweise (Liste)
 - Literaturhinweise
 - Wiss. Dokumentation
- Objektbeziehungen
 - Materialien zum Objekt (1)
 - Materialien zum Objekt (Bild) (1)
 - Materialien zum Objekt (Dokumente)
 - Handbuch
- Geschäftsprozesse
 - Leihverkehr
 - Ausstellungen
 - Deakzession
 - Inventur
 - Erhaltung / Präparation
- Ausdrucke
 - Registrierbeleg (Museum)
 - Objektinformation (Museum)
 - Restaurierbericht (Museum)
 - Leihschein (Museum)
 - als neue Merkliste erstellen

Objektbezeichnung: Ansichtskarte Eing.Nr.: 10003

Inv.Nr.: AK000003 Sammlung: Ansichtskarten

Standort: Stück: 1

Objektbeschreibung

Eingangsort: Eing.dat: 30.06.2006 Text Altbestand

Objektgeschichte: Entstehungszeit:

Schlagworte:

Hersteller: Herstellungsort: Länge: cm Breite: cm

Material: Kurzbeschreibung: Höhe: cm Tiefe: cm

Technik: Durchmesser: cm Gewicht: g

Darstellung: Zustandsmerkmale:

Ereignisse	Notiz	Dateinamen
		1
		\\avm-medialStorageWithBackupKIERLING(EHS)\Bi

Beschreibung:

Suche in EEXCESS: (2.98 Sekunden)

ERMANN FLEISCHMARKT SOMMERFRISCHE GUGGING SEEHÖHE

Treffer: 99

- KIMPortal: 100
- Kierling: 0
- RijksMuseum: 0
- Landesarchiv Kärnten: 0
- CORE.ac.uk: Waited too long for partner system 'CORE.ac.uk'

- Camelina microcarpa DC. Datum: unknown
- Fotografie (Dia), Alte Postkarte Datum: unknown
- Malerei, Zeichnung, Panorama vom Wisenberg Datum: unknown
- Illustrierter Brief von Karl Jauslin an Mutter und Geschwister Wien den 17ten Juni 1875 Datum: unknown
- Dianthus deltoides L. Datum: unknown
- Fotografie, Denkmal Bauernkriegdenkmal (Obelisk) auf Postkarte an Edgar Seiler von seiner Tante und seiner Cousine Bertha. Datum: unknown
- Aquarellmalerei, S2010-48 Datum: unknown
- Skizzen, Studien, Panorama Datum: unknown
- Fotografie, Blick über Mitteldorf, Breite, Rössligasse Ansichtskarte an Fraulein Marie Grollmund z.Z. in Brugg. Links Schulhaus und Turnhalle Breite. Datum: unknown
- Brief von Karl Jauslin an Mutter und Geschwister Wien den 24 August 1875 Datum: unknown

Powered by EEXCESS

BY NC SA

imdas pro 6.1.1 B... DE 21:32 20.11.2015

Content Injection – Collection Management System

imdas pro 6.1.1. Benutzer: SA DB-Version: 06.00.0002 Datenbank: IMDASKIRLING - [Listen]

Datei Bearbeiten Recherche Objekt Module Navigation Ansicht Stammdaten Thesaurus Fenster ?

Aufgabenbereich
Objekt-/Medientypen la...


- Akte
- Archivalie
- Bibliographisches Objekt
- Bild
- Dokument
- Film
- Konvolut
- Medienobjekt
- Museumsobjekt
- Ton

Objektauswahl laden
Erste Schritte
Administration
Makros
Objekt-/Medientypen laden
Recherche

Startseite Museumsobjekt GIS

Google Maps Zoom: 3 Markers: 62

Karte Satellit











Google Kartendaten © 2011

EEXCESS

Suche in EEXCESS: (4.45 Sekunden)

turm

-  Blick hinter die Kulissen: was ist eigentlich Kultur?: neues Forum für kulturwissenschaftliche Forschung in Frankfurt
Datum: unknown 92
-  Fotografie, Ziefen, Glockenweihe & Glockenaufzug
Datum: unknown
-  Bild: Bleistiftzeichnung, Palazzo Vecchio Florenz oberer Teil Fassade und Turm
Datum: unknown 4
-  Scherenschnitt, Kaiser Karl Gedächtnis-Jahrbuch 1930/04
1) Li: Palme, darunter sitzender bärtiger Mann, weibliche Heilige überreicht ihm Becher, re Brunnen mit Seilzug und Schöpffäß; 2) Li: Baum, darunter Jesus und 2 Jünger, re Kirche mit Stiegenaufgang; 3) Höhle, li: Baum, 2 knieende Ordensmänner, re: segnender Heiliger, Andachtskruzifix 4) Dom mit 6-stöckigem Turm, Häuser, Wolken 5) Kirche mit 3 Zwiebeltürmen
Datum: unknown
-  Fotografie, Liestal, Obertor (Törl) im Winter, Abendstimmung
Datum: unknown 96
-  Scherenschnitt, Kinderschnitte 97
Vorderseite: Kinderschnitte: 33 Schnitte: Herzen, Anker, Kreuz, Laterne, Pflanze, Leuchter, Vase, Krone
Rückseite: 3 3-teilige Schnitte: 2 Pflanzen, in der Mitte Bischof (Nikolaus?); über ihm ein 3-teiliges Blatt (beschnitten); 2 Pflanzen, Zaun, Kirche mit 3 Türmen (mittig); 2 Pflanzen, Zaun Haus (mittig)
Datum: unknown
-  Scherenschnitt, Kirche 98
Kirche mit runder Apsis und 2 Türmen
Datum: unknown
-  Fotografie, Muttenz, Wartenberg, hintere Ruine 99
Datum: unknown

Frankfurt im Spätmittelalter und die Dominanz der Patrizier : wie urbane Planung und Architektur das soziale Gefüge bestimmten
Powered by EEXCESS

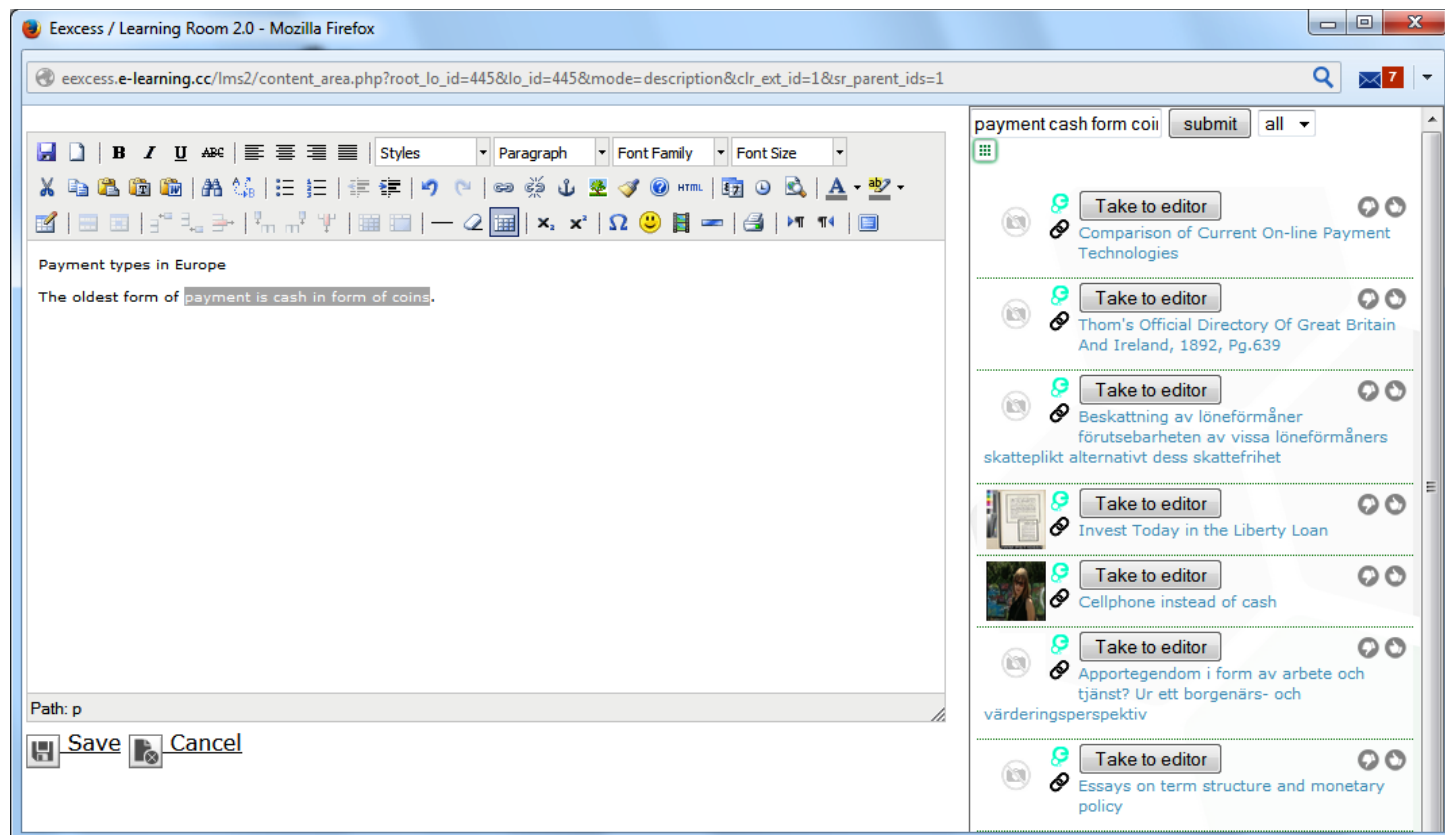
DE 21:38 20.11.2015

BY NC SA

Content Injection – Learn Management Systems

Content Creation for Educational Support

- Inject cultural heritage content into Learn Management Systems
- Moodle and BitMedia's SITOS LMS



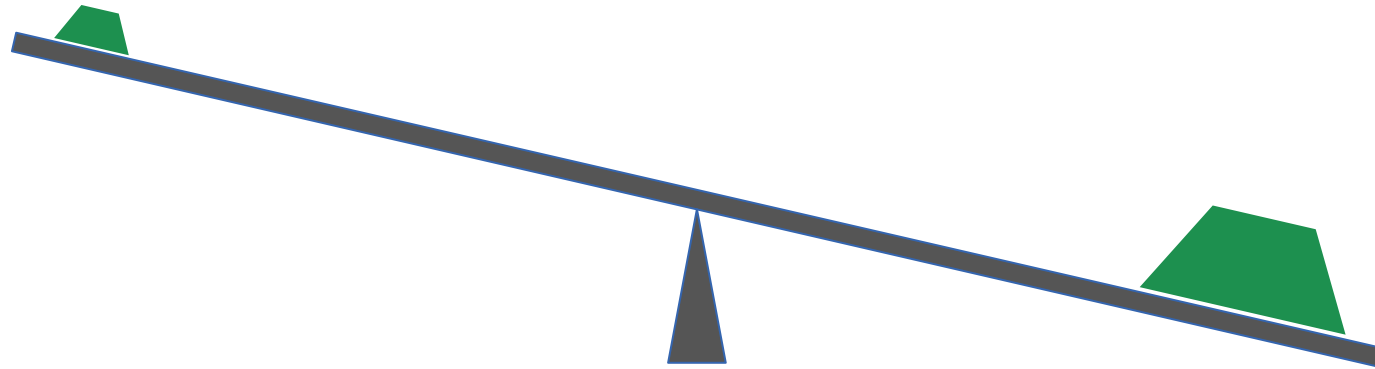
The screenshot displays a Mozilla Firefox browser window titled "Eexcess / Learning Room 2.0 - Mozilla Firefox". The address bar shows the URL: `eexcess.e-learning.cc/lms2/content_area.php?root_lo_id=445&lo_id=445&mode=description&clr_ext_id=1&sr_parent_ids=1`. The main content area features a rich text editor with a toolbar and a text box containing the sentence: "The oldest form of payment is cash in form of coins." Below the editor, a "Path: p" field and "Save" and "Cancel" buttons are visible. On the right side, a sidebar titled "payment cash form coi" contains a list of content items, each with a "Take to editor" button and a "submit" button. The items include:

- Comparison of Current On-line Payment Technologies
- Thom's Official Directory Of Great Britain And Ireland, 1892, Pg.639
- Beskattning av löneförmåner förutsebarheten av vissa löneförmånens skatteplikt alternativt dess skattefrihet
- Invest Today in the Liberty Loan
- Cellphone instead of cash
- Apportegendom i form av arbete och tjänst? Ur ett borgenärs- och värderingsperspektiv
- Essays on term structure and monetary policy

Privacy vs. Personalisation trade-off?

Privacy

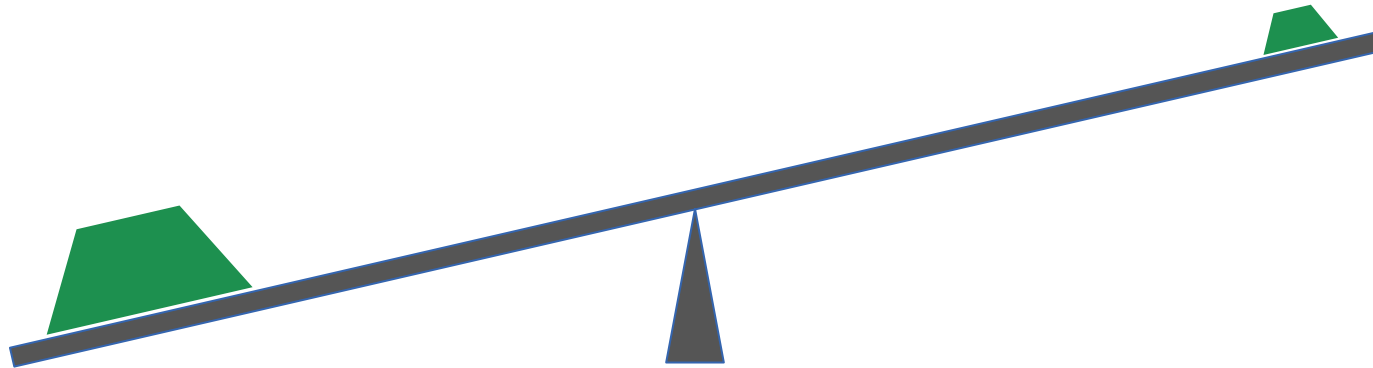
Personalisation/Quality



Privacy vs. Personalisation trade-off?

Privacy

Personalisation/Quality



Privacy vs. Personalisation trade-off?

User Awareness (and Transparency)

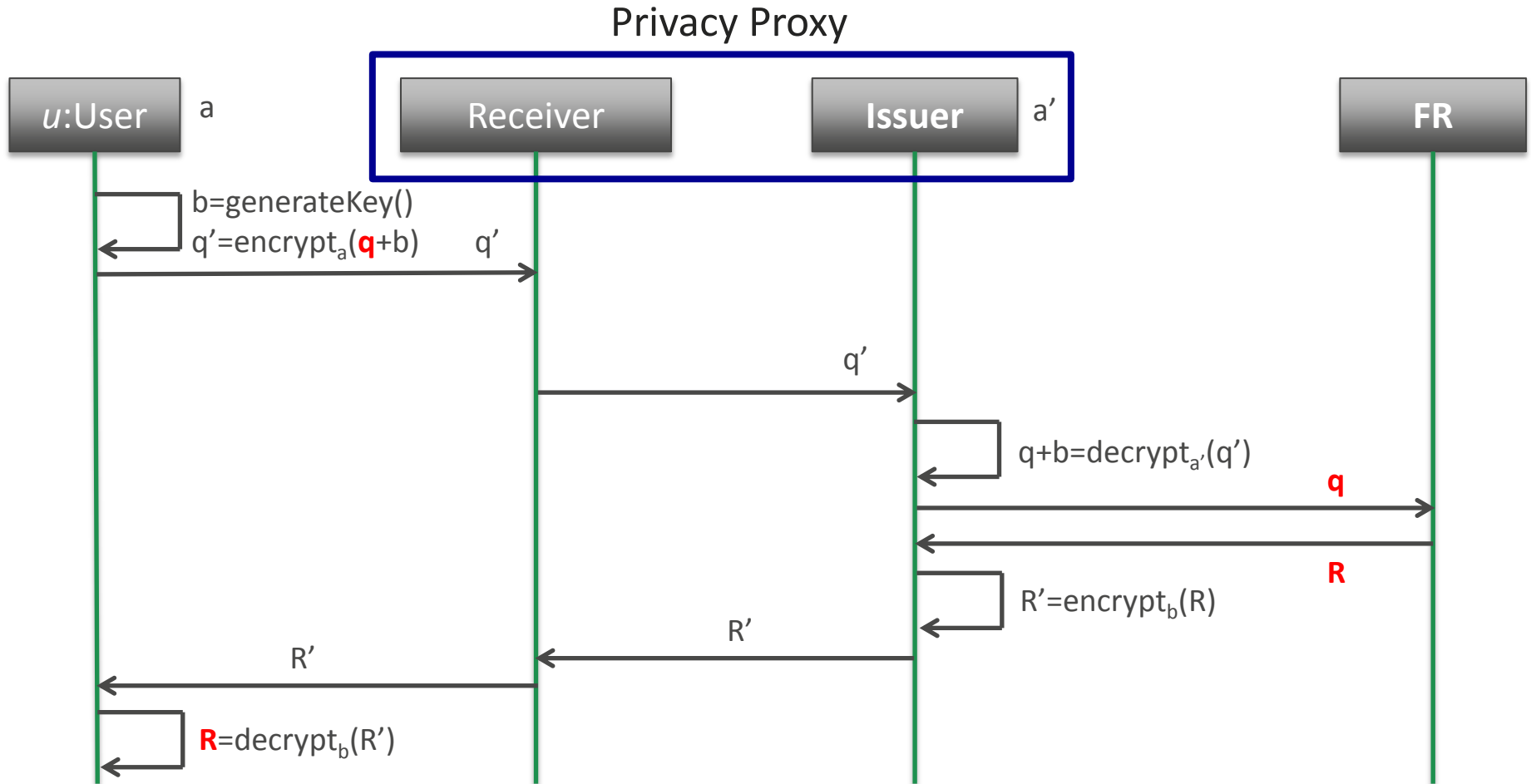
User Empowerment

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

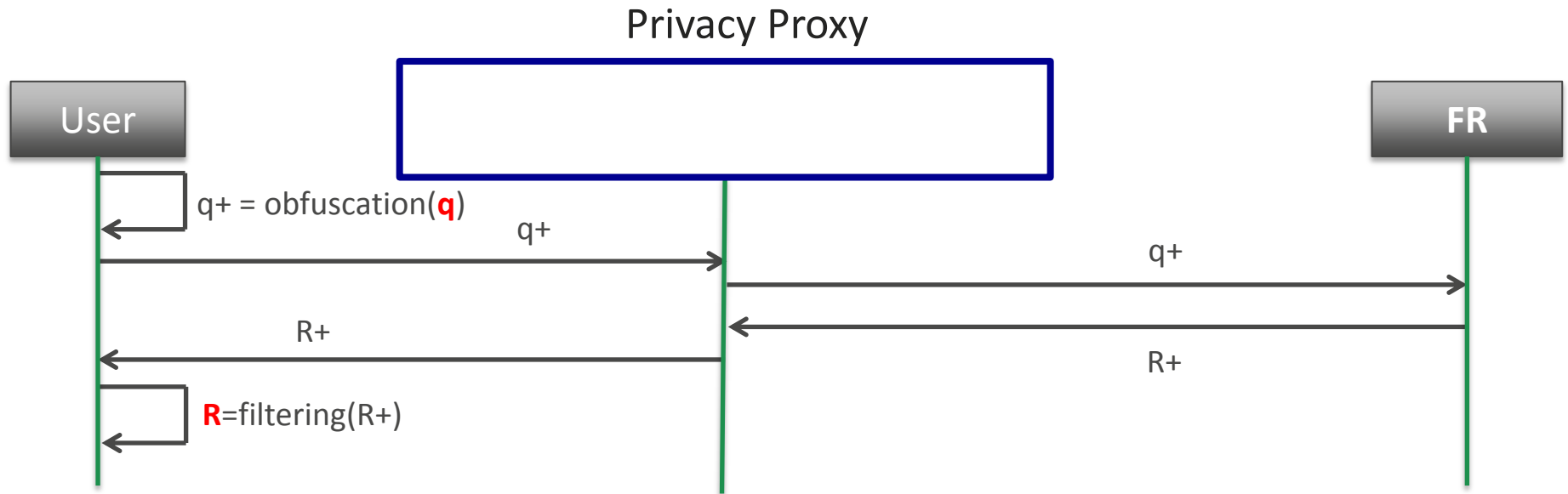
PEAS: Unlinkability Protocol (simplified)



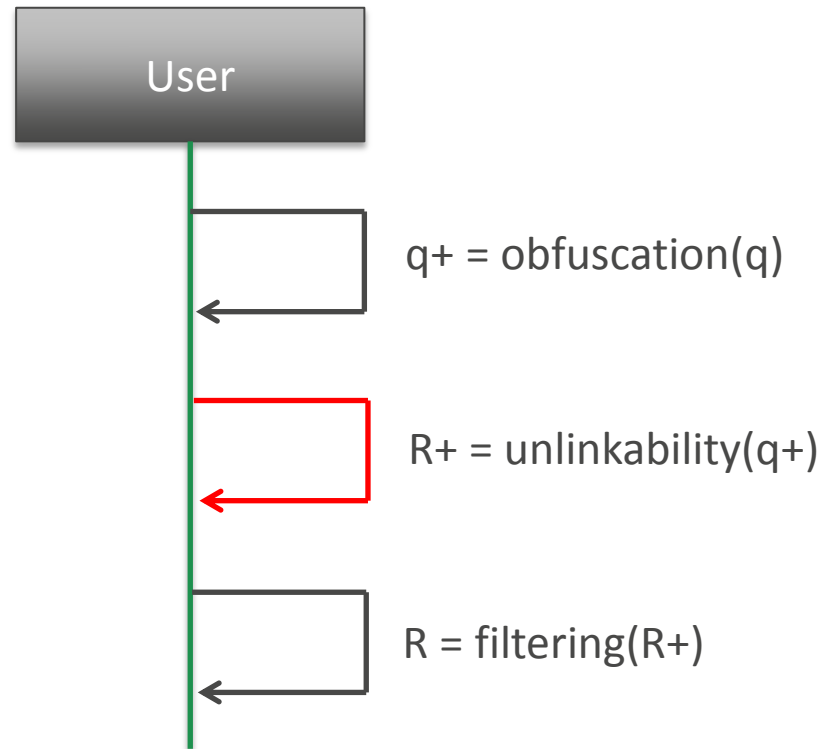
PEAS: Indistinguishability Protocol (simplified)

- 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)



PEAS: Combination of Protocols



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

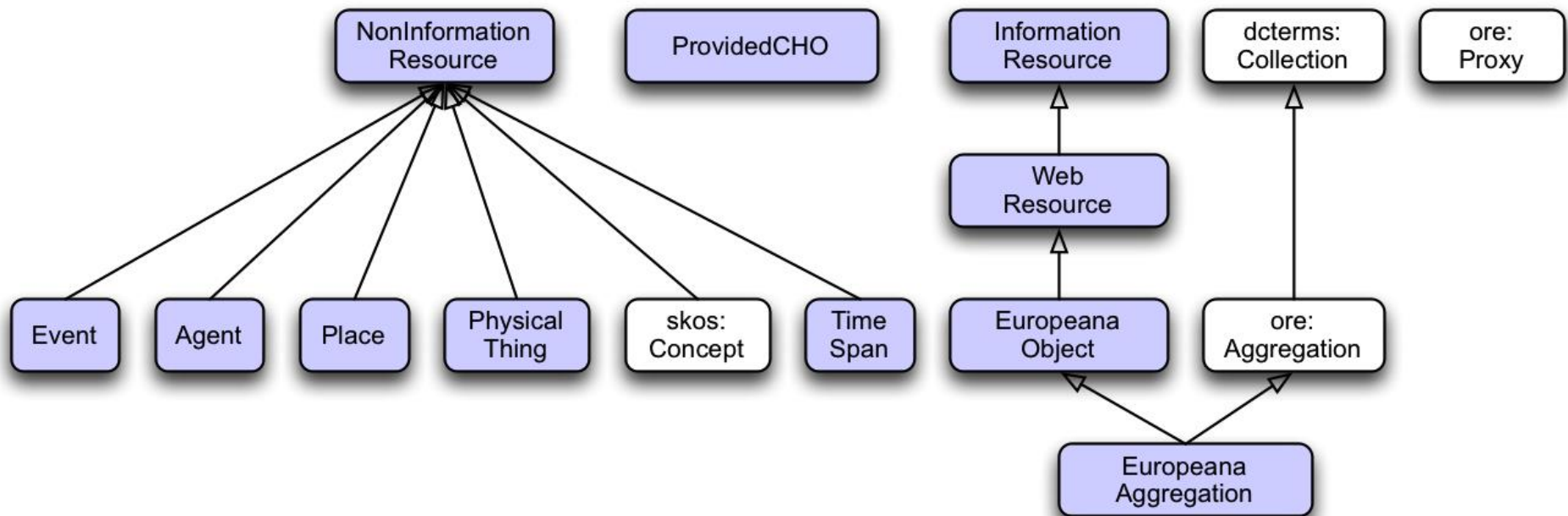
EEXCESS Ontology

- 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

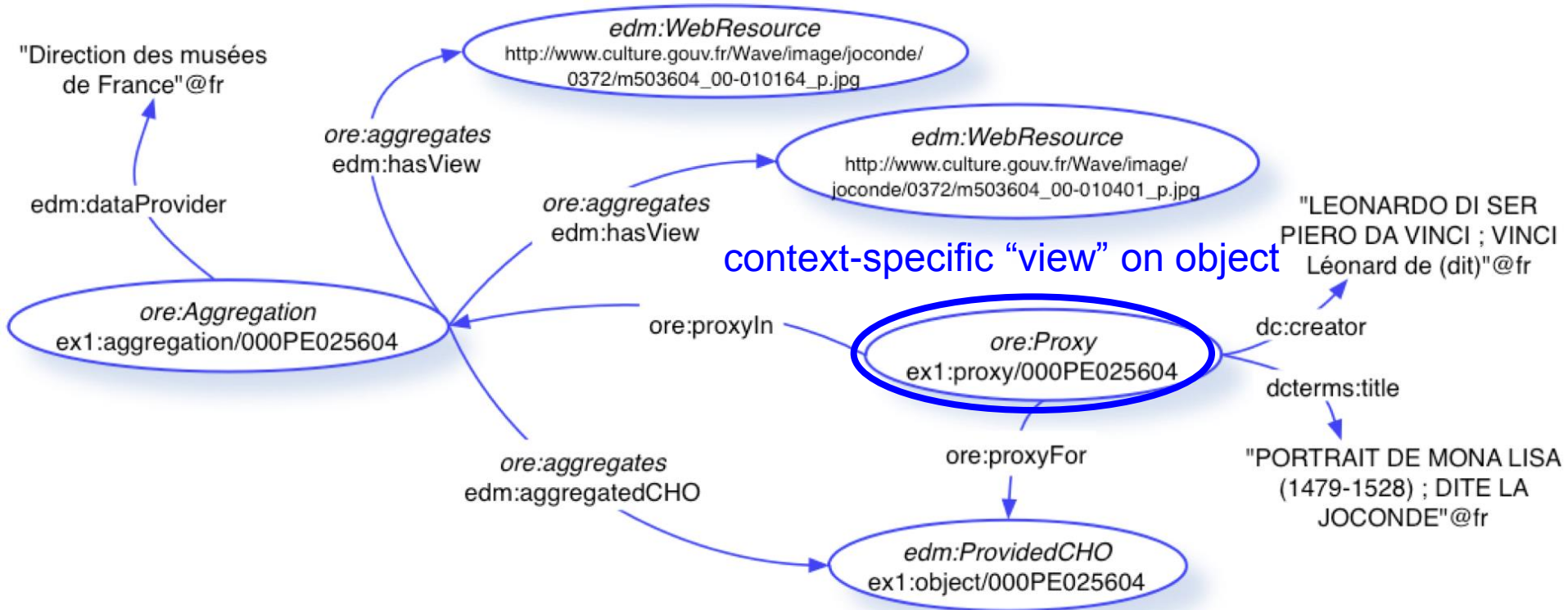
EEXCESS Ontology

- 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 CHOs
 - Annotations are modeled using EDM Proxies
 - Data providers are modeled as EDM Agents
 - Aggregation is used as in EDM

EDM – Main entities



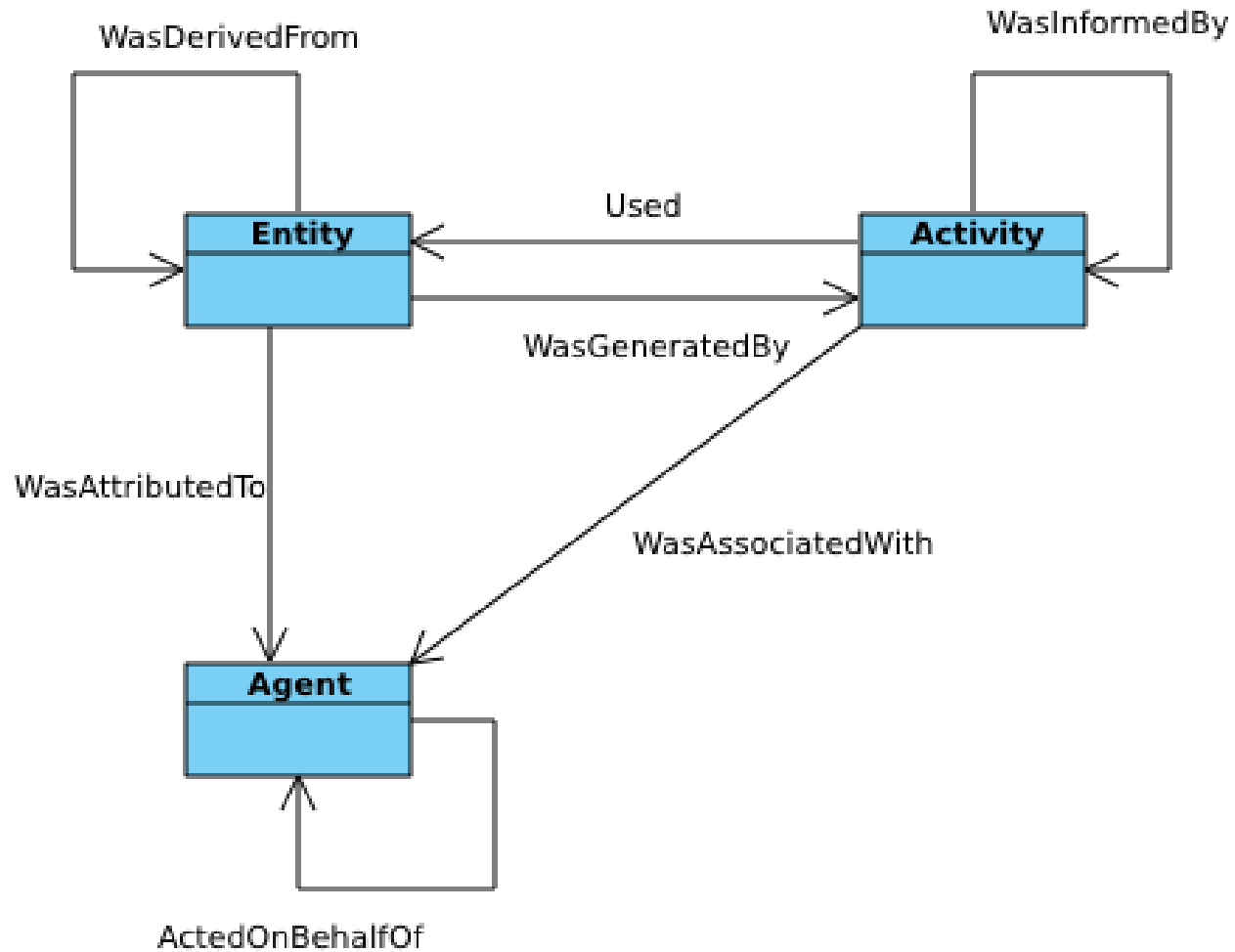
EDM – Proxy example



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

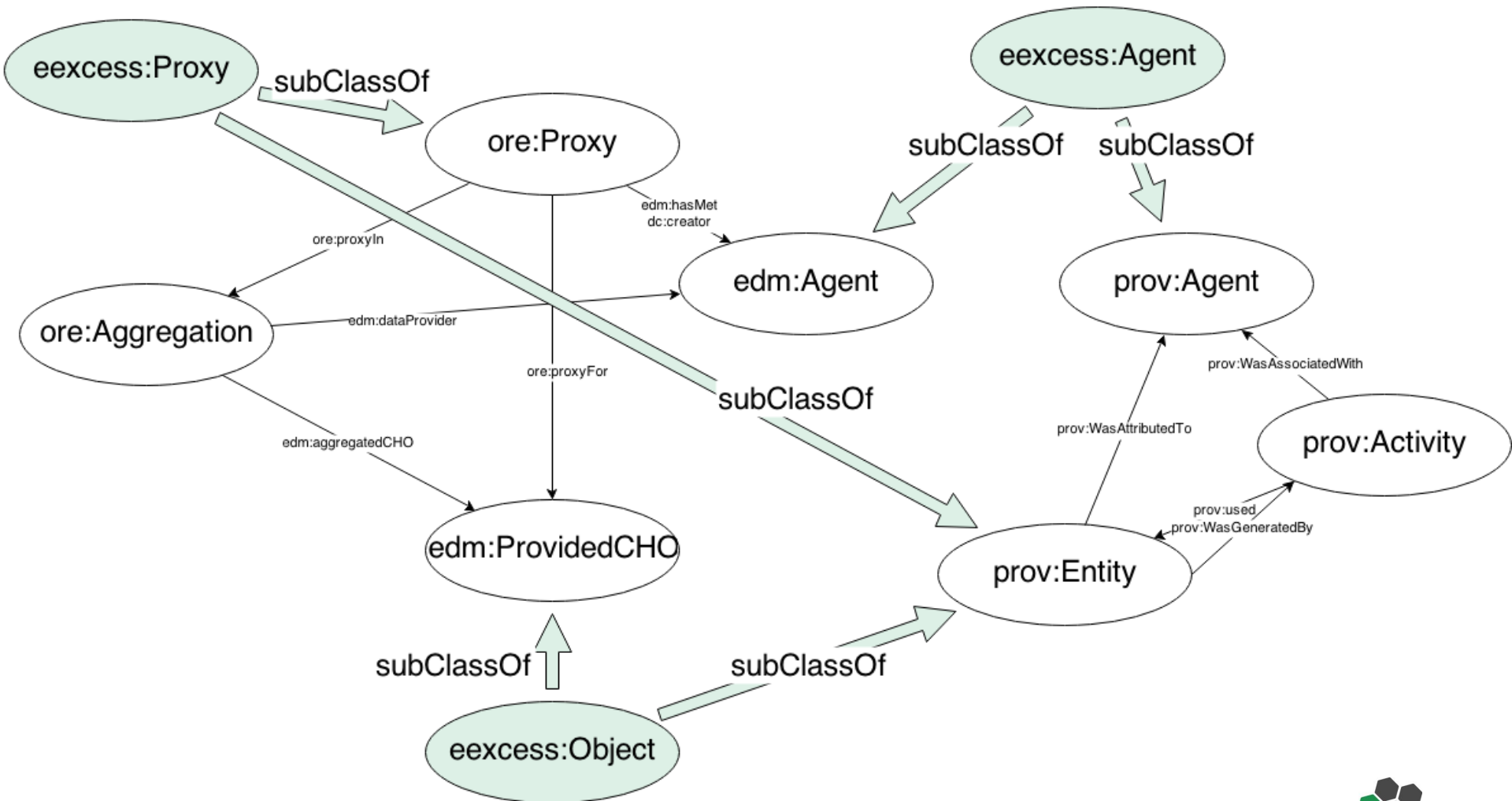
W3C PROV



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



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

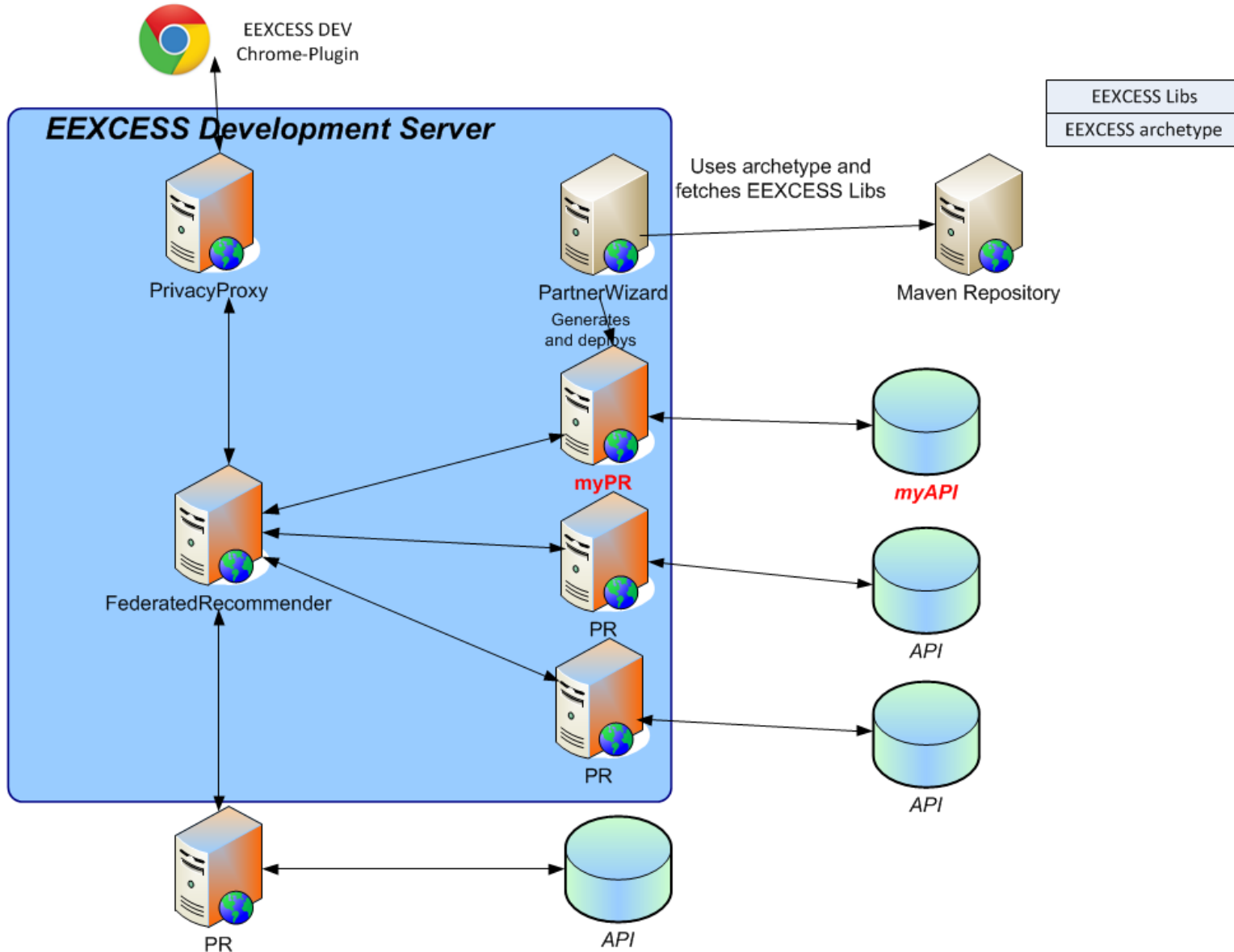
Overview

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

Architecture

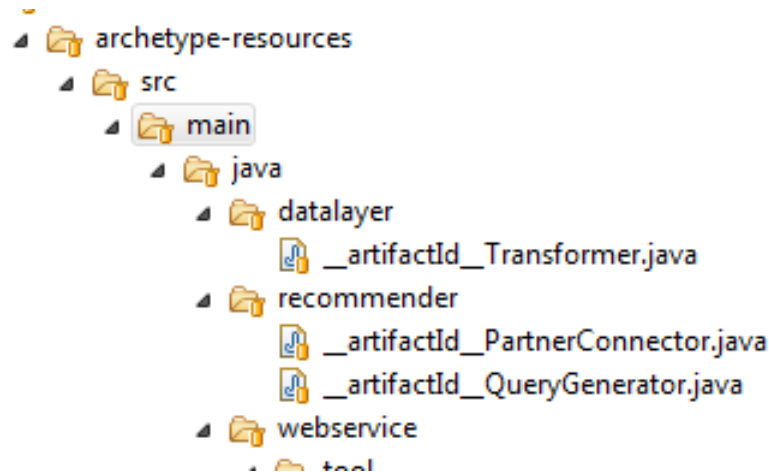


maven archetype

- 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



```
17 package ${package}.datalayer;
18
19+ import javax.xml.xpath.XPath;
36
37 public class ${artifactId}Transformer extends Transformer{
38
```

maven archetype

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']

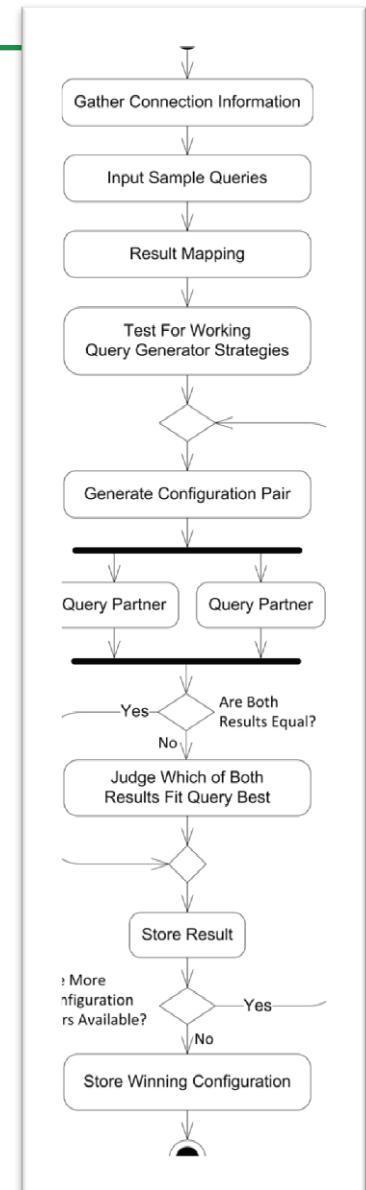
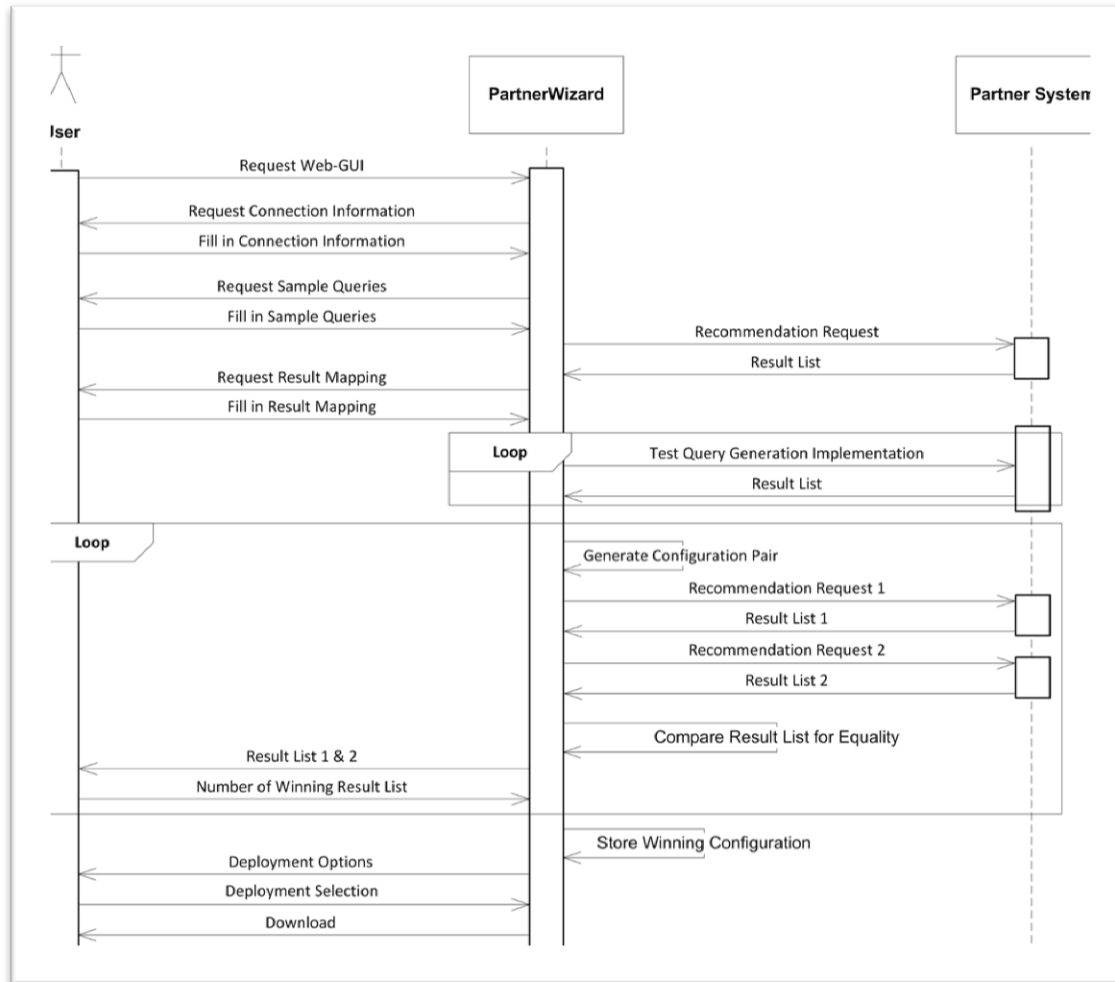
partnerAPIdetailMappingFieldsXPathTitle=str[@name='_display_']

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 strategie for your system.

← Customize example queries

Keyword	<input type="text" value="Napoleon"/>	is Main Topic	<input type="checkbox"/>	<input type="checkbox"/>
Keyword	<input type="text" value="Frankreich"/>	is Main Topic	<input type="checkbox"/>	<input type="checkbox"/>
Keyword	<input type="text" value="Schweiz"/>	is Main Topic	<input type="checkbox"/>	<input type="checkbox"/>
Keyword	<input type="text" value="Helena"/>	is Main Topic	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="button" value="+ Add Keyword"/>				

Remove Query

Keyword	<input type="text" value="Roman Missal"/>	is Main Topic	<input checked="" type="checkbox"/>	<input type="checkbox"/>
<input type="button" value="+ Add Keyword"/>				

Remove Query

Query Optimiser

KNOW
Center

EEXCESS Partner Wizard | Imprint | Help

EEXCESS Partner Wizard

Keywords:
Roman Missal

List 1

DVD, Naturparadies Bruderholz
Bildlegende: „Er rang sich empor und raffte einen Stuhl auf und zerschlug ihn auf dem Kopfe des ersten“ Darüber Titel und Autor: „Mit dem Leben belohnt von Gotthold Roman“, darunter: „für National-Kalender 1904 Aarau“

DVD, Naturparadies Bruderholz
Bildlegende: „Er rang sich empor und raffte einen Stuhl auf und zerschlug ihn auf dem Kopfe des ersten“ Darüber Titel und Autor: „Mit dem Leben belohnt von Gotthold Roman“, darunter: „für National-Kalender 1904 Aarau“

DVD, Naturparadies Bruderholz
Bildlegende: „Er rang sich empor und raffte einen Stuhl auf und zerschlug ihn auf dem Kopfe des ersten“ Darüber Titel und Autor: „Mit dem Leben belohnt von Gotthold Roman“, darunter: „für National-Kalender 1904 Aarau“

DVD, Naturparadies Bruderholz
Bildlegende: „Wer sind Sie?“ links daneben Titel und Autor: „Mit dem Leben belohnt v. Gotthold Roman“, rechts: „für National-Kalender 1904 Aarau“

DVD, Naturparadies Bruderholz
Bildlegende: „Wer sind Sie?“ links daneben Titel und Autor: „Mit dem Leben belohnt v. Gotthold Roman“, rechts: „für National-Kalender 1904 Aarau“

[Choose List 1](#)

List 2

DVD, Naturparadies Bruderholz
Bildlegende: „Wer sind Sie?“ links daneben Titel und Autor: „Mit dem Leben belohnt v. Gotthold Roman“, rechts: „für National-Kalender 1904 Aarau“

DVD, Naturparadies Bruderholz
Titel 1: Naturtagebuch 2004 15. Mai - 15. August, Bottminger Bruderholz Titel 2: Naturnah (2006)

DVD, Naturparadies Bruderholz
Titel 1: Naturtagebuch 2004 15. Mai - 15. August, Bottminger Bruderholz Titel 2: Naturnah (2006)

DVD, Naturparadies Bruderholz
Titel 1: Naturtagebuch 2004 15. Mai - 15. August, Bottminger Bruderholz Titel 2: Naturnah (2006)

DVD, Naturparadies Bruderholz
„Befestigte gotische Dorfkirche St. Arbogast in Muttenz. Die Kirche selbst unter Benützung älterer roman(t)ischer Bauteile nach dem Erdbeben unter Konrad Münch von Münchenstein im 3. Viertel des 14. Jahrhunderts, die Befestigungen wahrscheinlich unter des vorigen Sohn, Hans Münch, zwischen 1378 - 1399, der Kirchturm wohl unter dem Enkel Hans Thüring Münch zwischen 1390 - 1449 erbaut.“

[Choose List 2](#)

KNOW
Center

KNOW
Center

EEXCESS Partner Wizard | Imprint | Help

EEXCESS Partner Wizard

Finished, here is your configuration:

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

keywords:
queryGeneratorClass: eu.eexcess.partnerrecommender.reference.OrQueryGenerator
queryExpansionEnabled: false
querySplittingEnabled: false

KNOW
Center

@ 2015 Know-Center

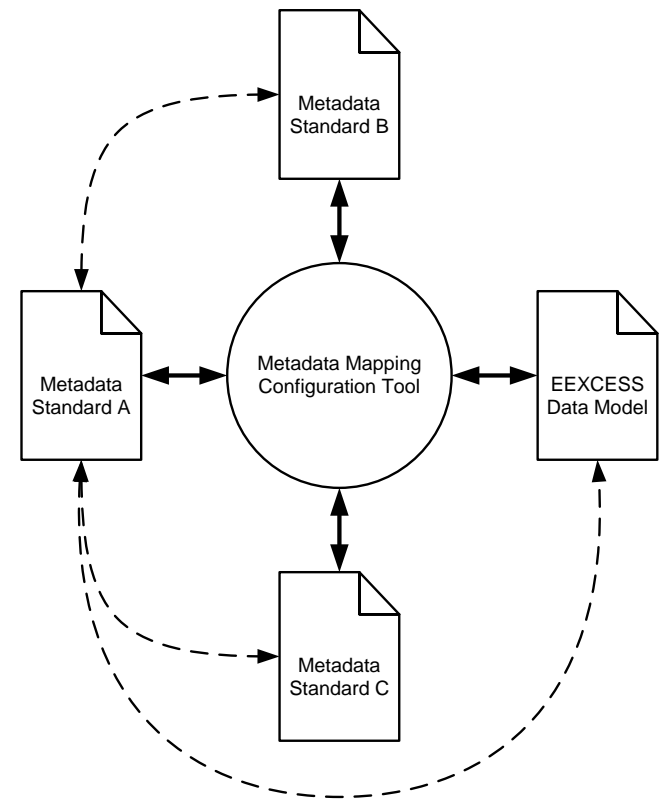


Metadata Mapping Configuration Tool



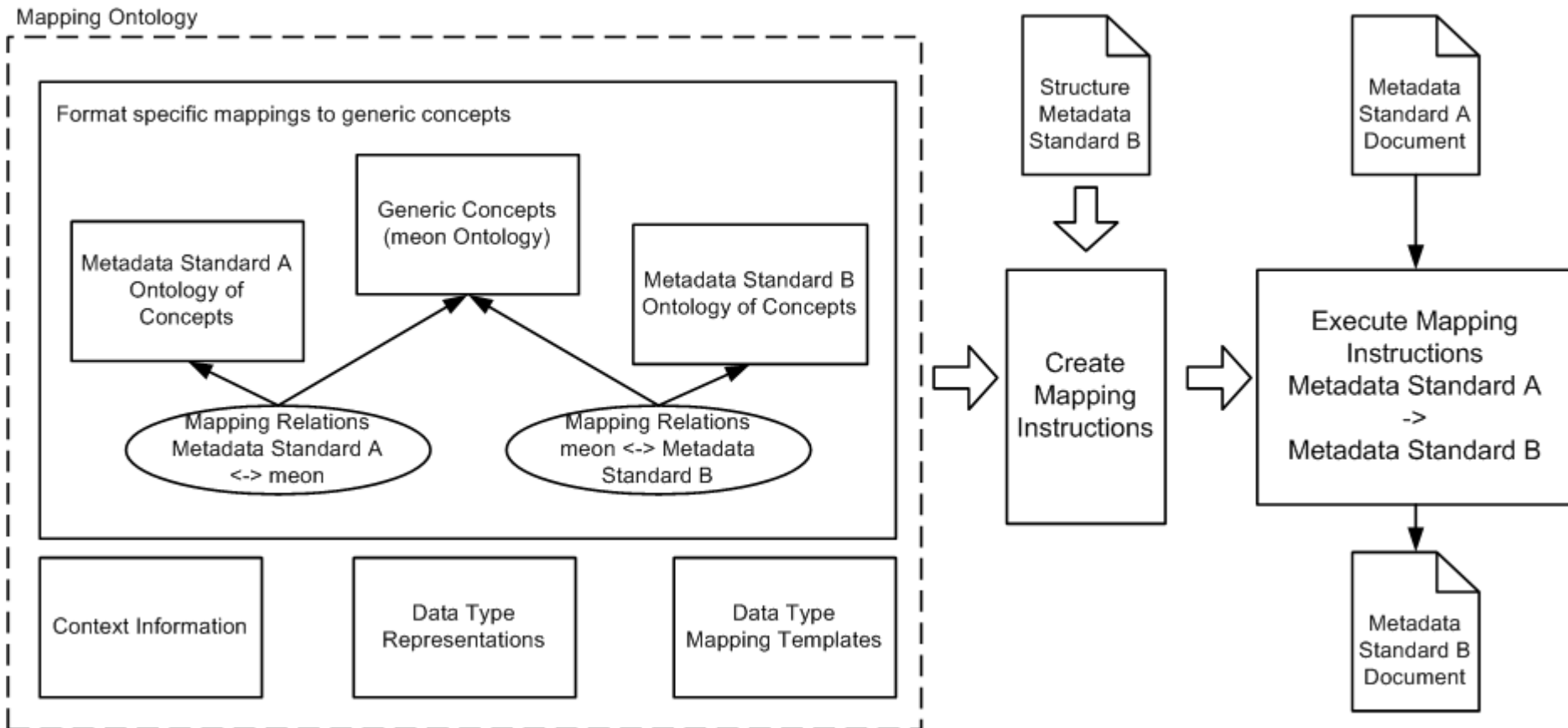
Motivation

- 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 on 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



 Logout User: hoi

New Project

Please enter a name for the new project:

Source Schema

Wissensserver

Target Schema (optional)

EEXCESS

Create New Project

Recent Projects

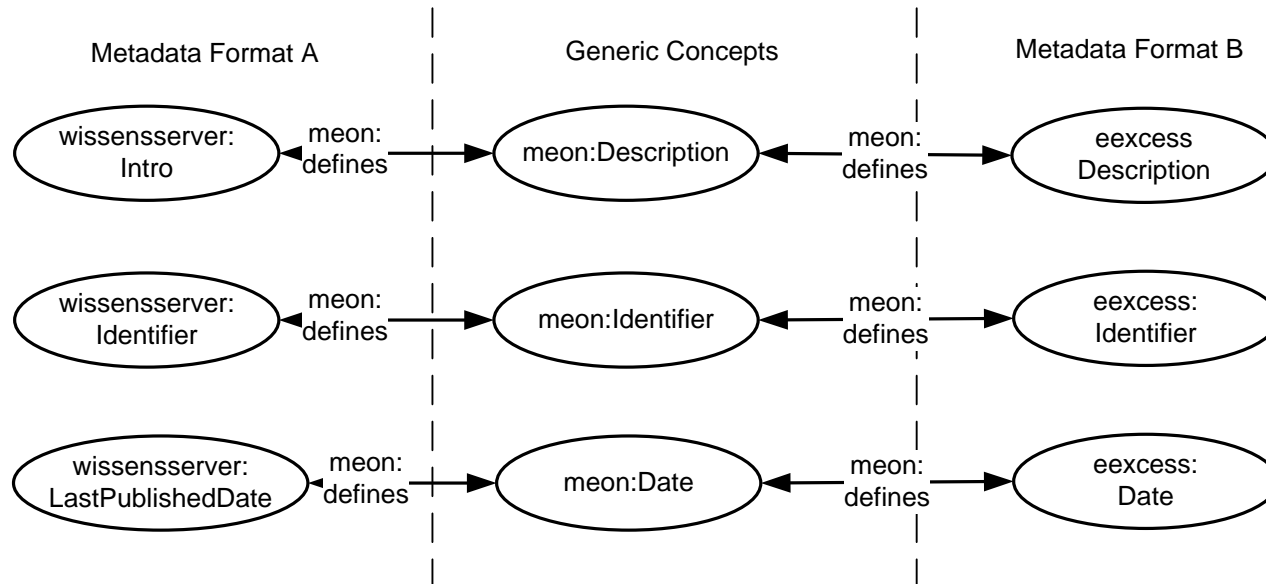
 [imdass2eexcess](#) 

 [w1](#) 

Metadata Mapping Configuration Workflow

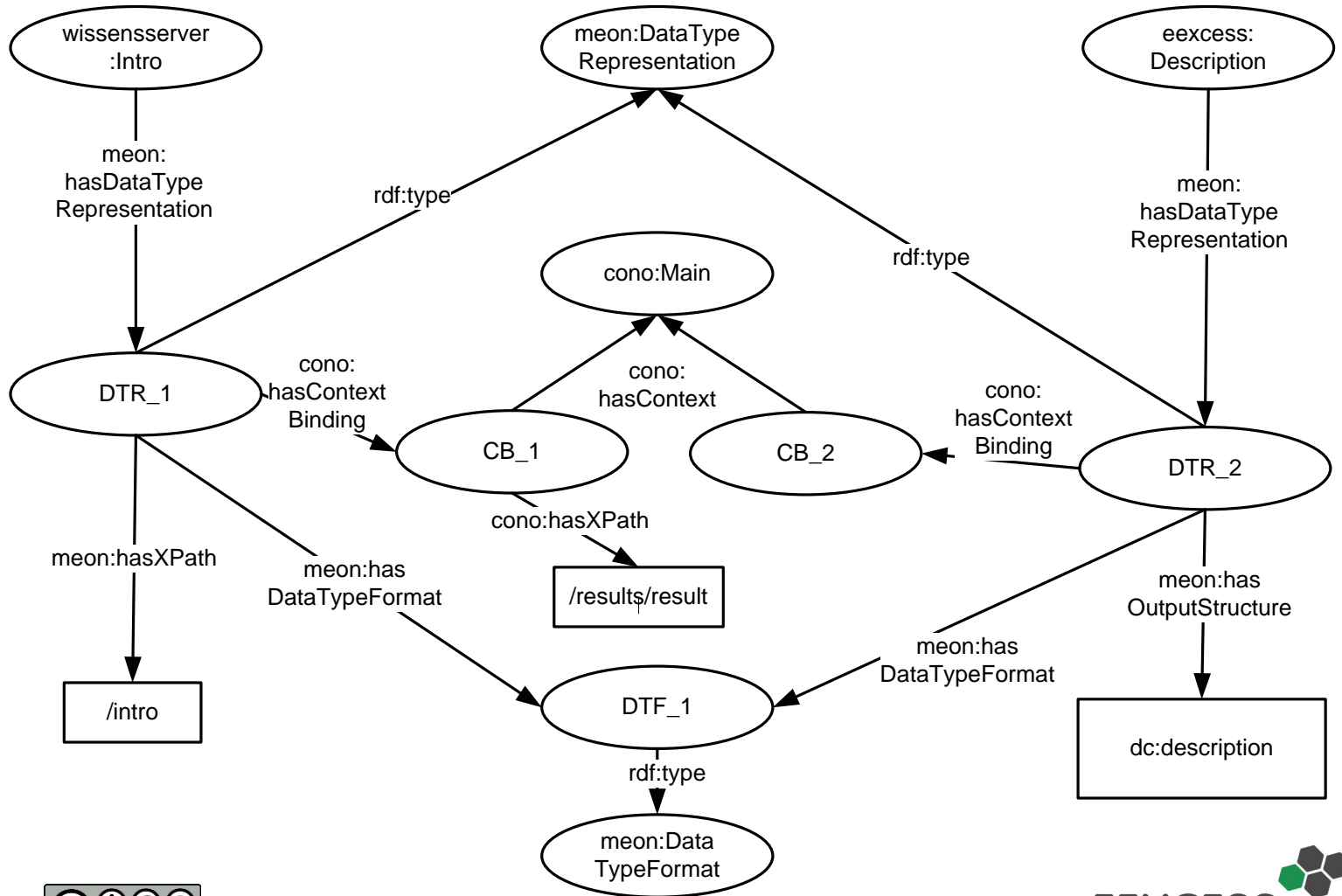
Concept Mappings

- based on meon ontology



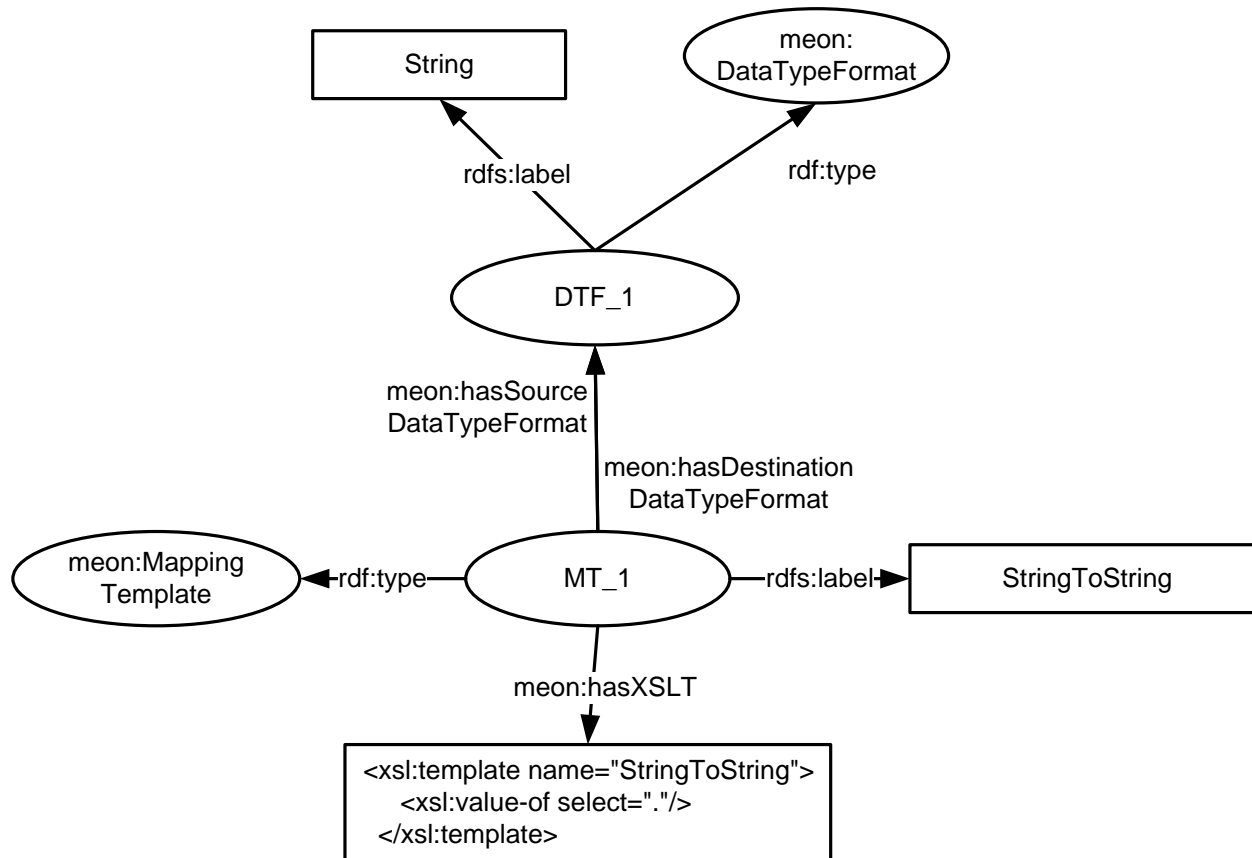
Metadata Mapping Configuration Workflow

Datatype Representations



Metadata Mapping Configuration Workflow

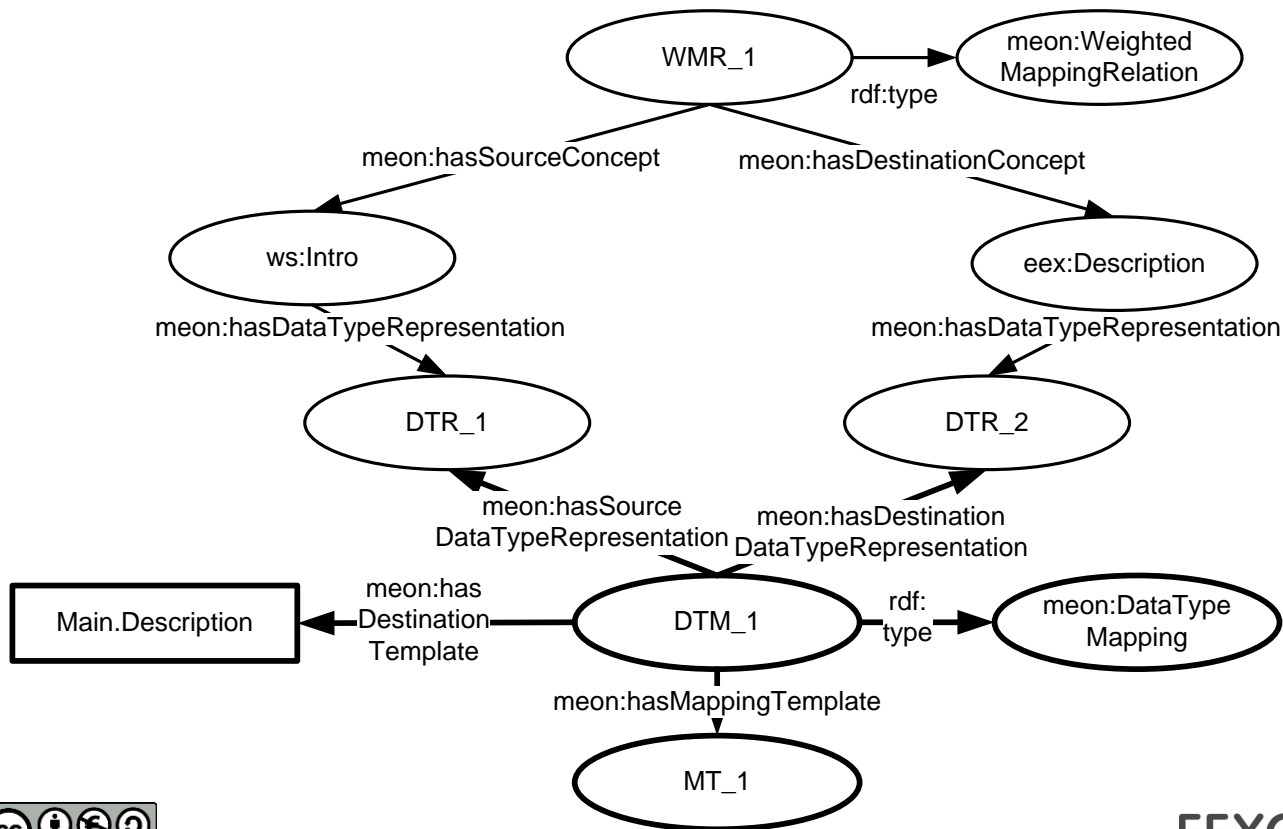
Mapping Template



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:stylesheet>
```

Mapping Parameters:

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

Mapping Instructions:

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

```
<template match="intro">
  <element name="dc:description">
    <call-template name="ToString"/>
  </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

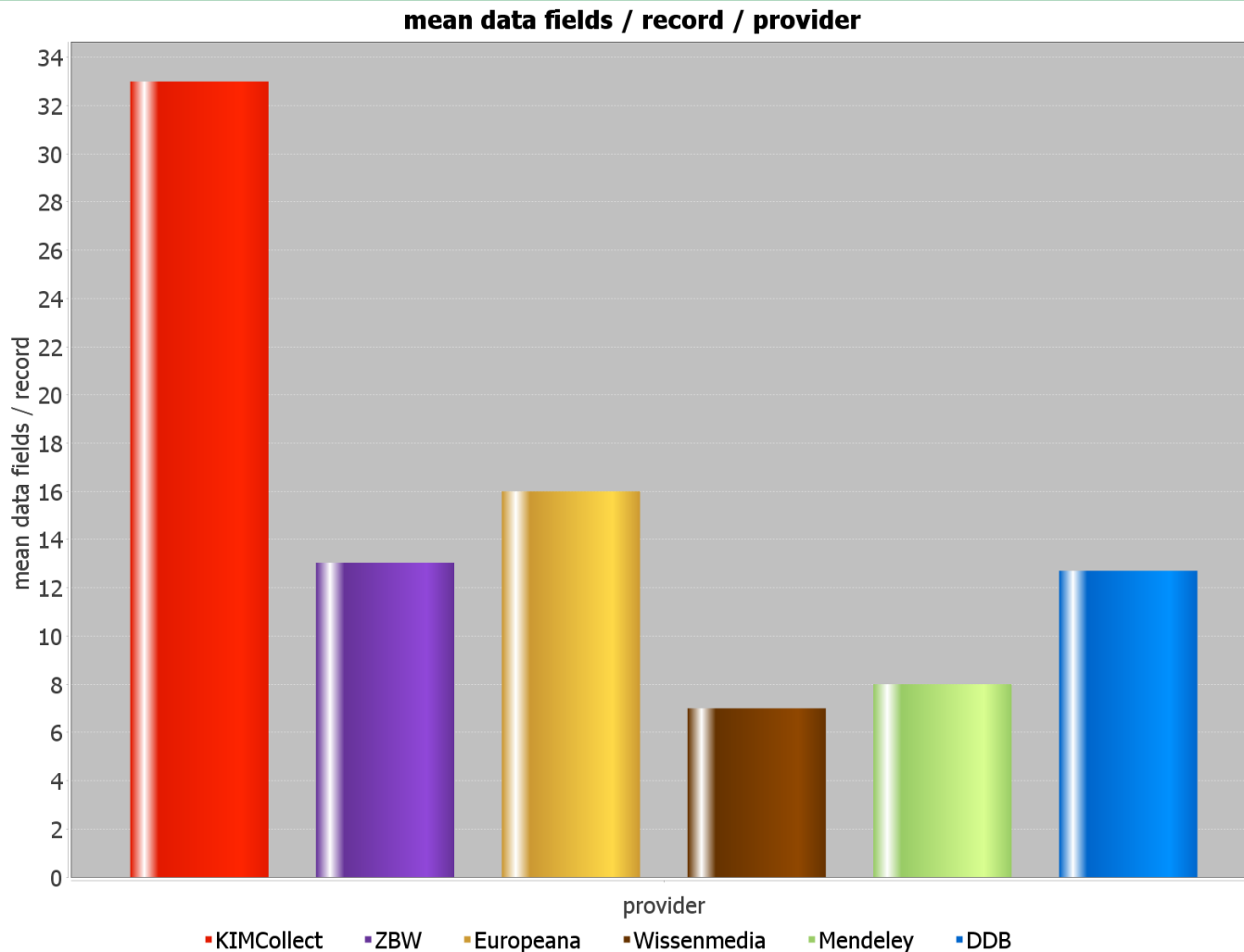
Input data quality – metrics

- 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)

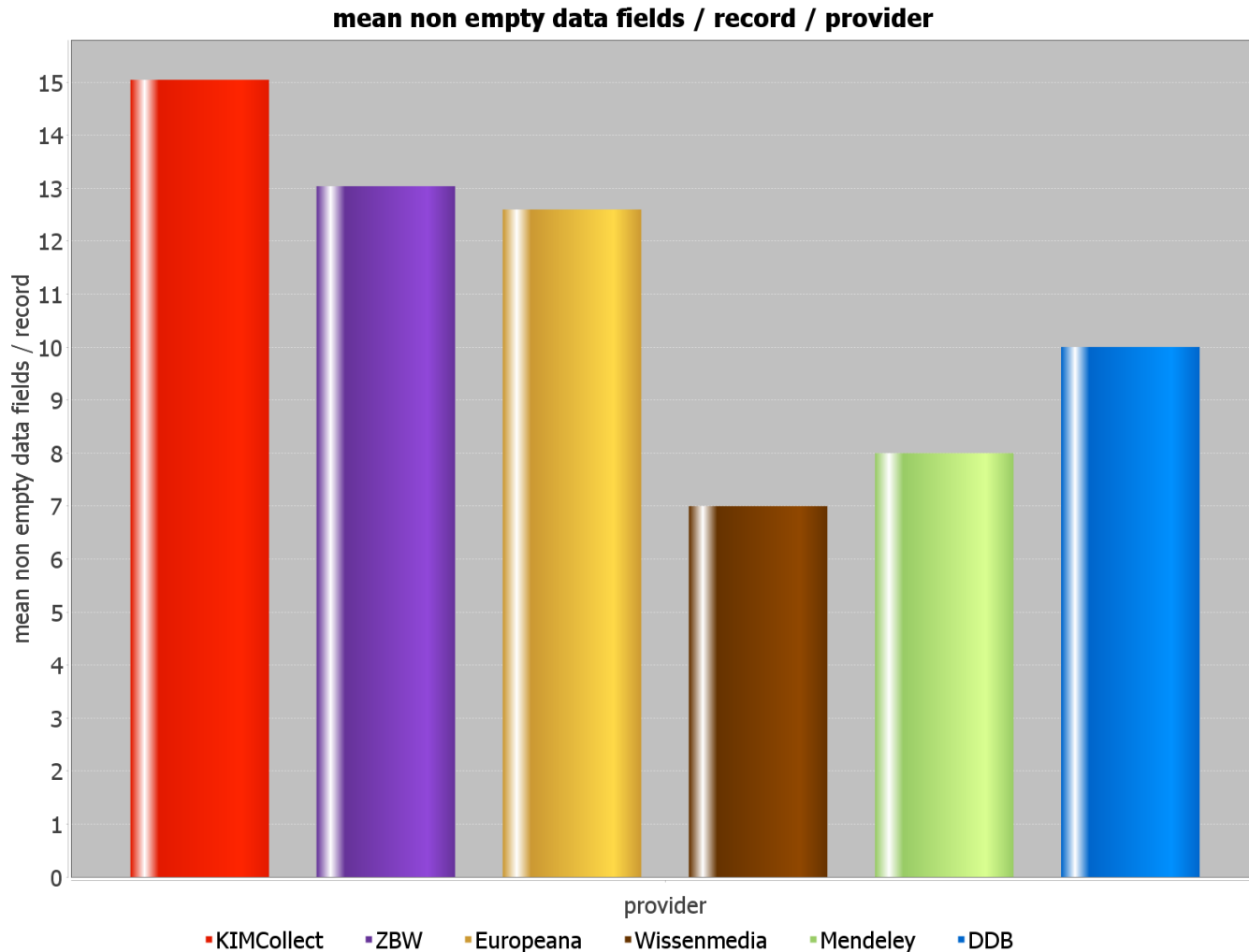
Input data quality – metrics

- Use of controlled vocabularies
- Availability of linked resources
- Evaluated on data collected during testbed on 6K records

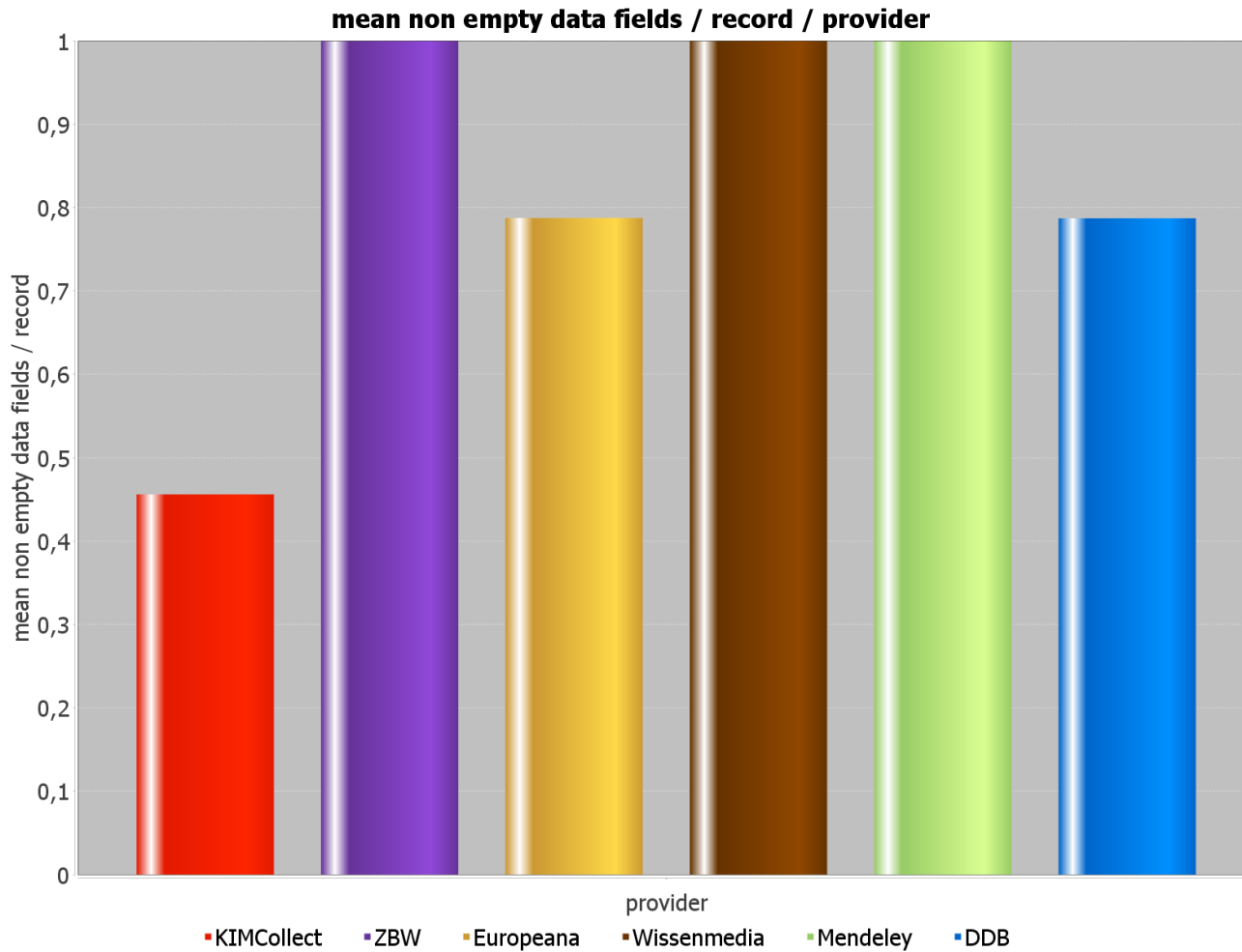
Completeness



Completeness



Completeness



Structuredness

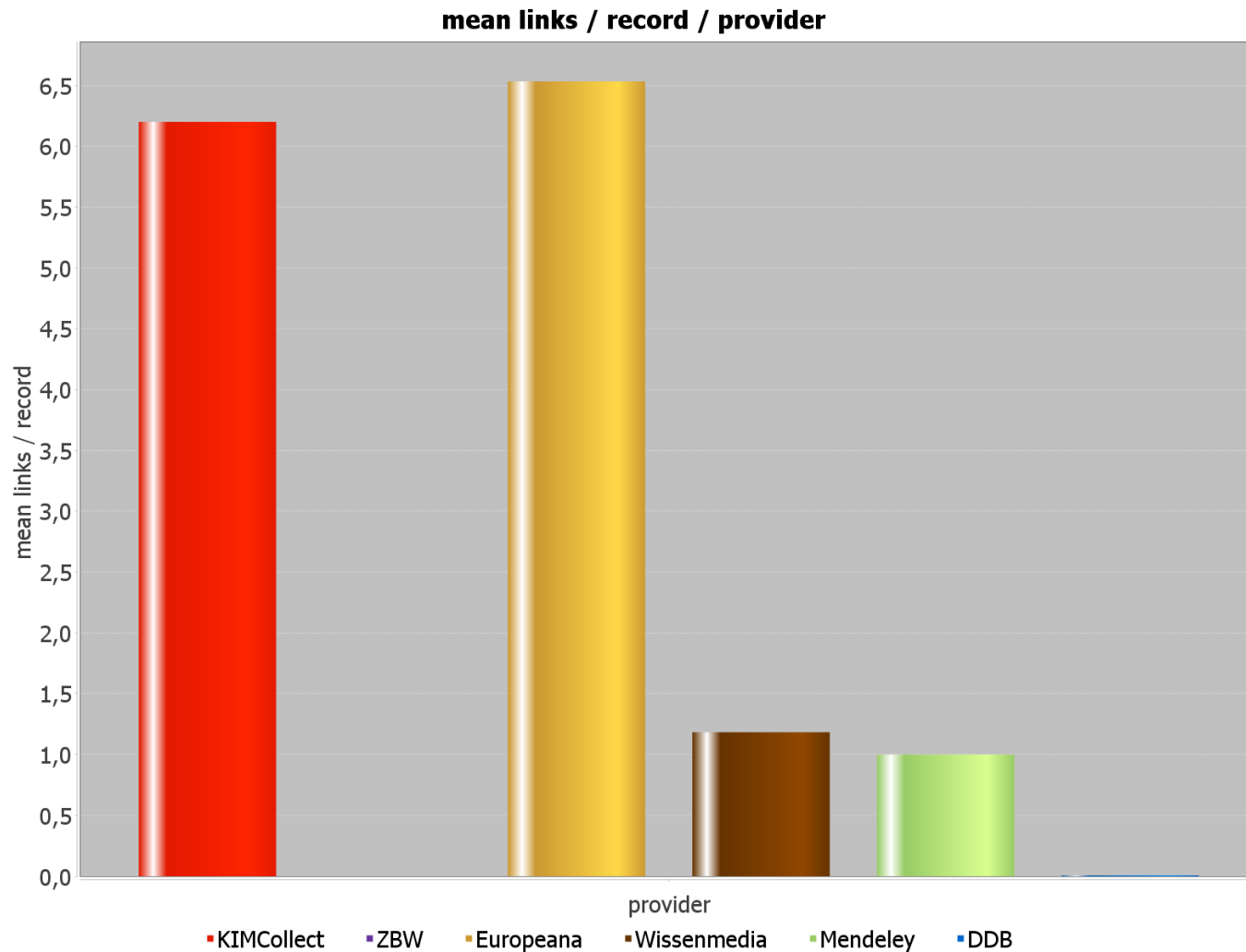
- Length of value
-> histogram
- Group characters and numbers
- Infer candidate patterns
 - e.g. Height: 00.0aa
Width: 0.0aa
- Histogram of candidate patterns
- Detect known particles (e.g. SI unit abbreviations)

Time of origin	Start time of origin	End time of origin	Height	Width
1902	1902.0000	1902.0000	43.0cm	2.5cm
1868	1868.0000	1868.0000	35.0cm	1.7cm
2002			21.0cm	0.5cm
1904	1904.0000	1904.0000	47.0cm	2.7cm
1869	1869.0000	1869.0000	35.0cm	1.7cm
1870 - 1871	1870.0000	1871.0000	34.5cm	3.0cm
1872 - 1873	1872.0000	1873.0000	40.0cm	4.0cm
1874 - 1875	1874.0000	1875.0000	40.5cm	5.0cm
1876 - 1877	1876.0000	1877.0000	40.5cm	5.6cm
1878 - 1879	1878.0000	1879.0000	42.0cm	5.5cm
1880 - 1881	1880.0000	1881.0000	40.5cm	4.8cm
1882 - 1883	1882.0000	1883.0000	41.0cm	4.5cm
1884 - 1885	1884.0000	1885.0000	40.5cm	5.5cm
1886 - 1887	1886.0000	1887.0000	41.0cm	5.0cm
1888 - 1889	1888.0000	1889.0000	41.5cm	5.0cm
1890 - 1891	1890.0000	1891.0000	44.0cm	6.0cm
1892	1892.0000	1892.0000	44.3cm	2.5cm
1893	1893.0000	1893.0000	43.8cm	2.5cm

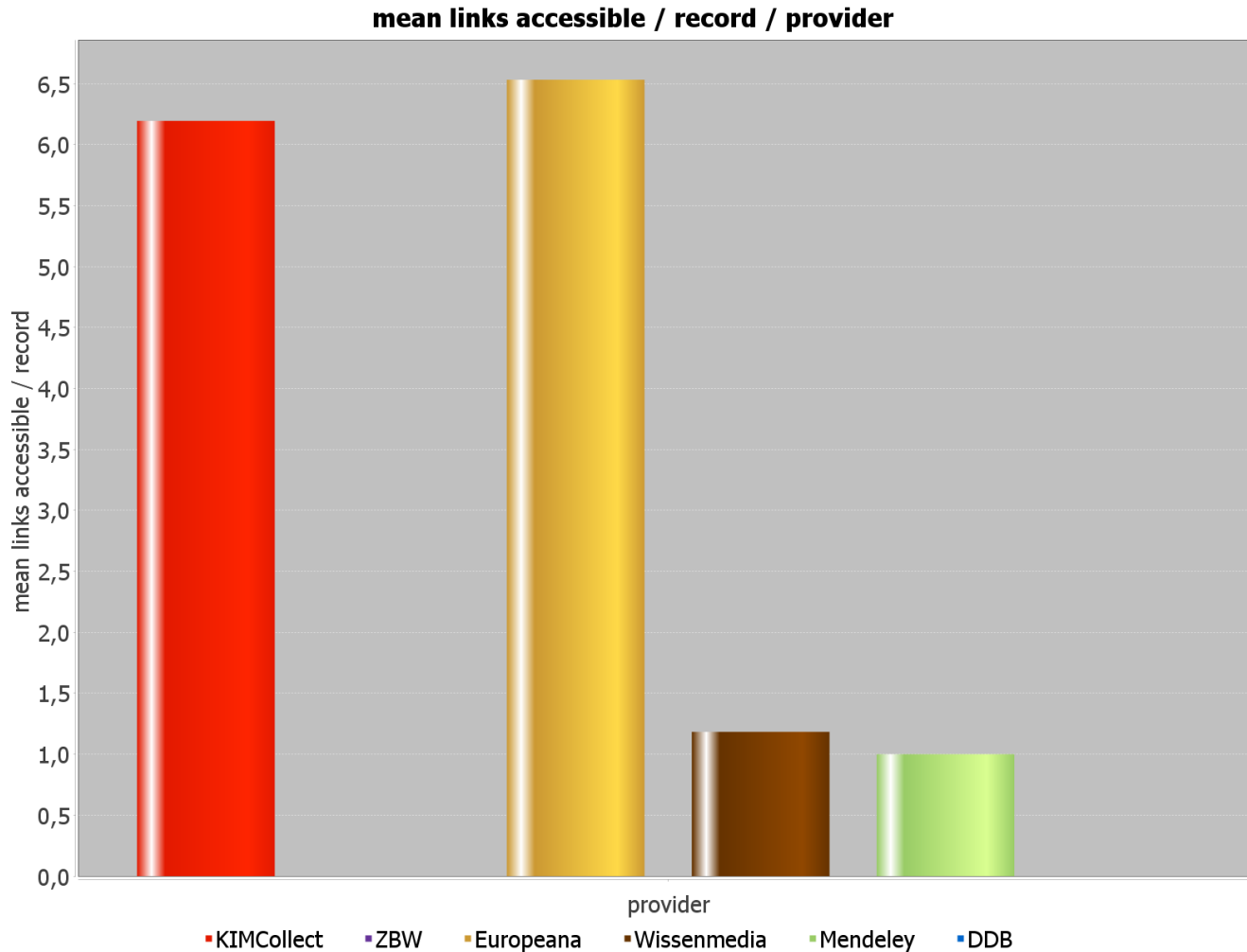
URLs in record

- 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



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

Add Locator Association

Edit Locator Association

	Input XPath	Output XPath
Edit Delete	/objects/object/Hersteller	/objects/object/Hersteller
Edit Delete	/objects/object/Autor	/objects/object/Hersteller
Edit Delete	/objects/object/Fotograf	/objects/object/Hersteller

Recent XML Data Files

k1.xml (167B)



Mapping Quality Assessment

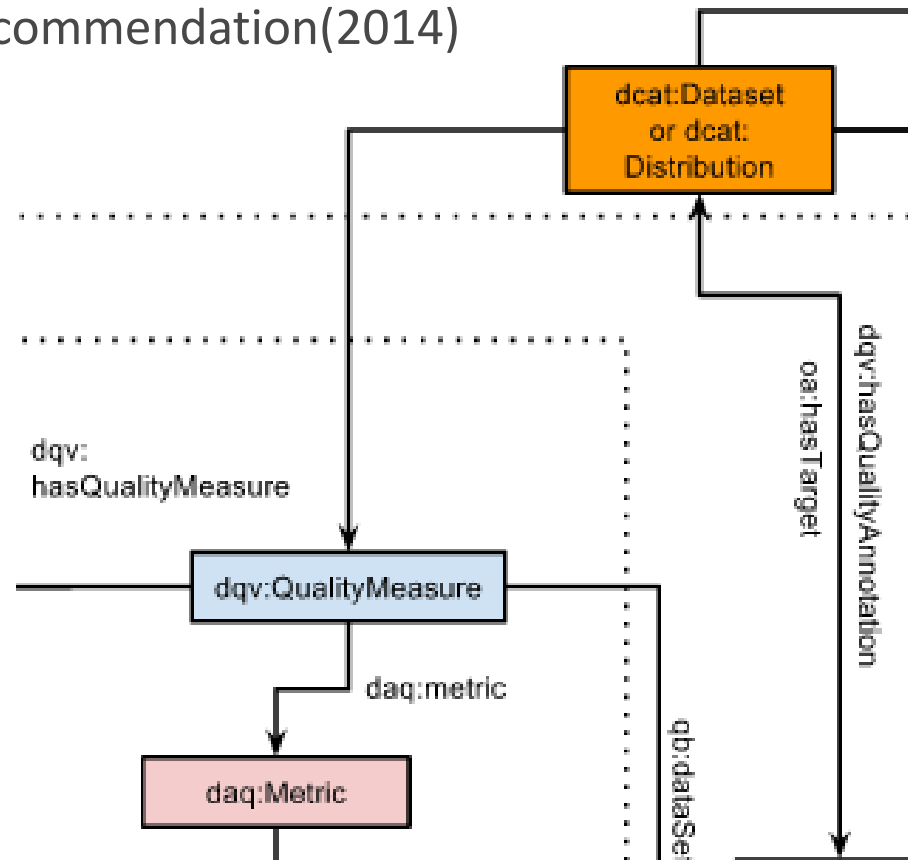
Input XPath	Output XPath	Result
/objects[1]/object[1]/Hersteller[1]	/objects[1]/object[1]/Hersteller[1]	Mapping ok
/objects[1]/object[1]/Autor[1]	/objects[1]/object[1]/Hersteller[2]	Mapping ok
/objects[1]/object[1]/Fotograf[1]	/objects[1]/object[1]/Hersteller	Mapping missing

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

```
<dcats:Dataset rdf:about="#eexcessDataset">
  <dcats:title>My EEXCESS dataset</dcats:title>
  <dcats:distribution>
    <dcats:Distribution rdf:about="#eexcessDatasetZBWDistribution">
      <dcats:title>My EEXCESS ZBW dataset</dcats:title>
      <prov:wasGeneratedBy rdf:resource="#ZBW"/>
    </dcats:Distribution>
  </dcats:distribution>
  <dcats:distribution>
    <dcats:Distribution rdf:about="#eexcessDatasetZBWTransformationDistribution">
      <dcats:title>My EEXCESS ZBW Transformation dataset</dcats:title>
      <prov:wasGeneratedBy rdf:resource="#EEXCESSTransformation"/>
      <prov:wasDerivedFrom rdf:resource="#eexcessDatasetZBWDistribution"/>
    </dcats:Distribution>
  </dcats:distribution>
  <dcats:distribution>
    <dcats:Distribution rdf:about="#eexcessDatasetZBWEnrichmentDistribution">
      <dcats:title>My EEXCESS ZBW Enrichment dataset</dcats:title>
      <prov:wasGeneratedBy rdf:resource="#EEXCESSEnrichment"/>
      <prov:wasDerivedFrom rdf:resource="#eexcessDatasetZBWTransformationDistribution"/>
    </dcats:Distribution>
  </dcats:distribution>
</dcats: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/XMLSchemaDouble">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/XMLSchemaDouble">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/XMLSchemaDouble">10</daq:value>
```

```
  <daq:computedOn rdf:resource="#eexcessDatasetZBWTransformation"/>
```

```
  <daq:metric rdf:resource="#eexcessDataQMetricNumberOfFields"/>
```

```
</dqv:QualityMeasure>
```


Visualisation from DQV

- Generate diagrams using XSLT

Data Quality Measurements

