
The
European
Library
Connecting knowledge

Linked Open Data in Aggregation Scenarios: The Case of The European Library

Nuno Freire
The European Library

SWIB14 Semantic Web in Libraries Conference
Bonn, December 2014

Outline

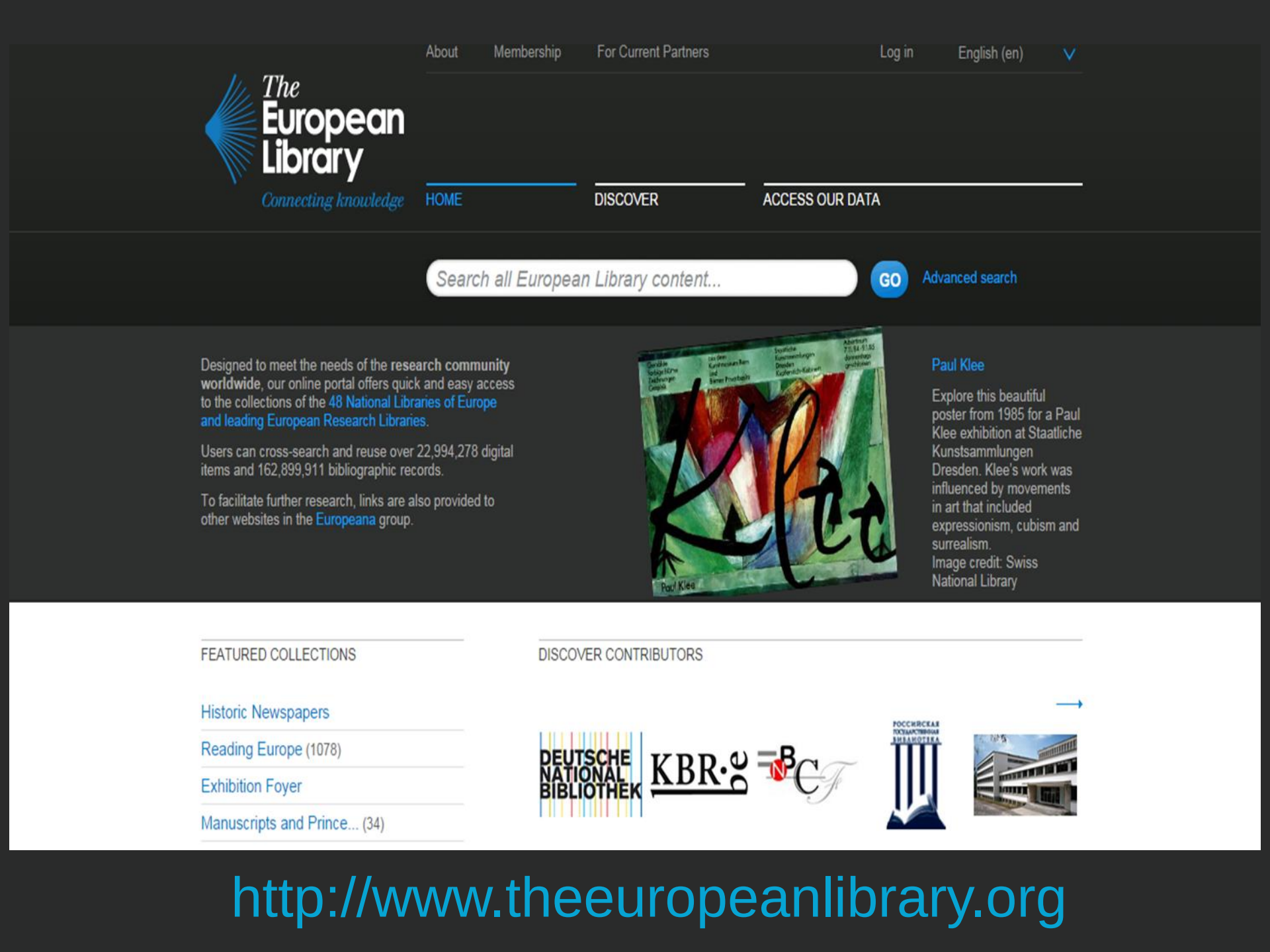
- *Introduction to The European Library*
- *The European Library Open Dataset*
  - *What data is included*
  - *The data model*
  - *How is it made available*
- *Linking Data*
  - *Managing and linking person names*
  - *Managing and linking place names*
  - *Managing and linking concepts*

Introduction to The European Library

www.theeuropeanlibrary.org

What is The European Library?

- Project started 1996, full operational service from 2005
- European hub of metadata, collections and increasing amount of full text
- Membership of national and research libraries of 47 Council of Europe states
- Non-profit, owned and managed by member libraries



http://www.theeuropeanlibrary.org

What does The European Library offer?

Large-scale aggregation Infrastructure

Data and digital content of Europe's libraries

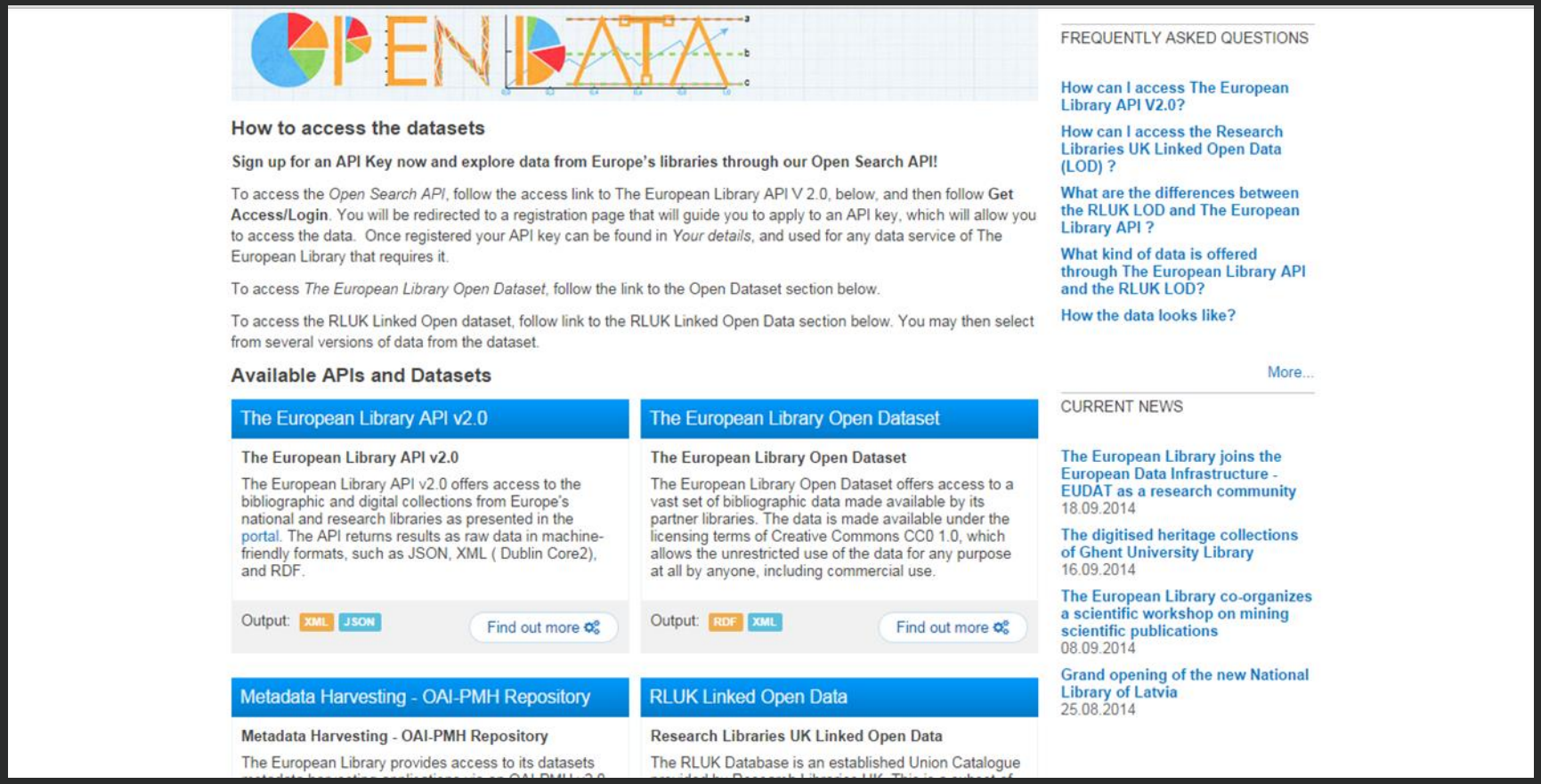
Data enrichment Linked open data

Experienced European project partner

Data distribution

Open data distribution

http://www.theeuropeanlibrary.org/tel4/access



OPEN DATA

How to access the datasets

Sign up for an API Key now and explore data from Europe's libraries through our Open Search API!

To access the *Open Search API*, follow the access link to The European Library API V 2.0, below, and then follow **Get Access/Login**. You will be redirected to a registration page that will guide you to apply to an API key, which will allow you to access the data. Once registered your API key can be found in *Your details*, and used for any data service of The European Library that requires it.

To access *The European Library Open Dataset*, follow the link to the Open Dataset section below.

To access the RLUK Linked Open dataset, follow link to the RLUK Linked Open Data section below. You may then select from several versions of data from the dataset.

Available APIs and Datasets

The European Library API v2.0

The European Library API v2.0

The European Library API v2.0 offers access to the bibliographic and digital collections from Europe's national and research libraries as presented in the portal. The API returns results as raw data in machine-friendly formats, such as JSON, XML (Dublin Core2), and RDF.

Output: XML JSON

Find out more

The European Library Open Dataset

The European Library Open Dataset

The European Library Open Dataset offers access to a vast set of bibliographic data made available by its partner libraries. The data is made available under the licensing terms of Creative Commons CC0 1.0, which allows the unrestricted use of the data for any purpose at all by anyone, including commercial use.

Output: RDF XML

Find out more

Metadata Harvesting - OAI-PMH Repository

Metadata Harvesting - OAI-PMH Repository

The European Library provides access to its datasets

RLUK Linked Open Data

Research Libraries UK Linked Open Data

The RLUK Database is an established Union Catalogue

FREQUENTLY ASKED QUESTIONS

- How can I access The European Library API V2.0?
- How can I access the Research Libraries UK Linked Open Data (LOD) ?
- What are the differences between the RLUK LOD and The European Library API ?
- What kind of data is offered through The European Library API and the RLUK LOD?
- How the data looks like?

More...

CURRENT NEWS

- The European Library joins the European Data Infrastructure - EUDAT as a research community
  18.09.2014
- The digitised heritage collections of Ghent University Library
  16.09.2014
- The European Library co-organizes a scientific workshop on mining scientific publications
  08.09.2014
- Grand opening of the new National Library of Latvia
  25.08.2014

The European Library
Open Dataset

www.theeuropeanlibrary.org

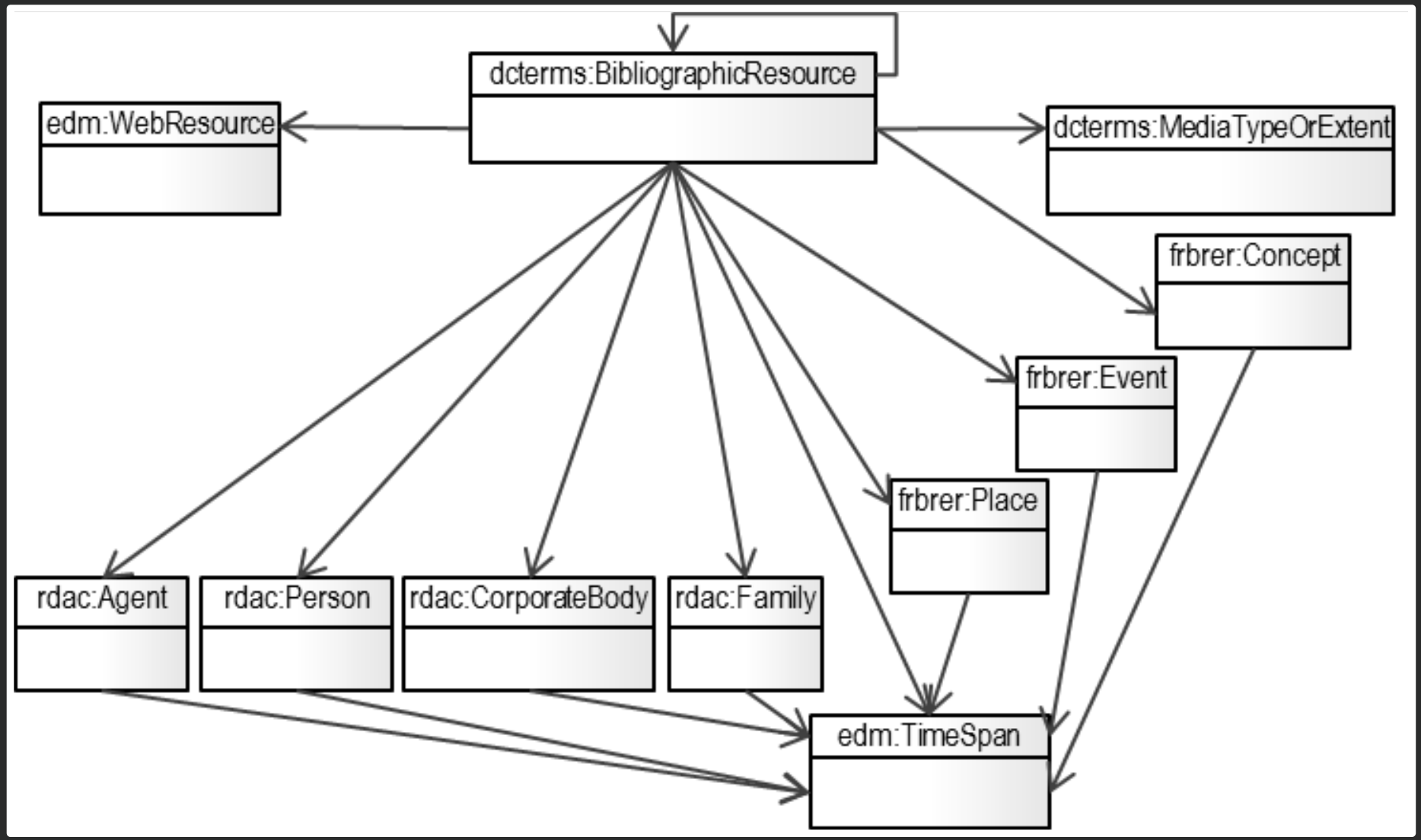
Library LOD
Leveraging on aggregation networks

- Aggregation networks provide:
  - An existing information and communication technology infrastructure
  - Technical expertise may be focused on the aggregating organizations
  - Centralized data, enabling for more linking to be established
    - Linking bibliographic within aggregated data is easier than across distributed datasets
    - Each library benefit from the linking done for other libraries
    - Each external dataset liked to, benefits all libraries' data

Library LOD
Leveraging on aggregation networks

- The European Library also leverages on other aggregators of library data
- Its first major release of LOD was focused on the Research Libraries UK consortium
  - The dataset was the focus of the RLUK Hack Day in May 2014
  - It was a subset of the RLUK database comprising nearly 20 million bibliographic records from 34 libraries

The Data Model



The Data Model

- RDA Element Vocabularies
  - The most extensivelly used vocabularies
  - Used entensivelly in the properties of the Bibliographic Resources
- FRBRer model
  - Used for context
  - Not used for Item, Manifestation, Expression, Work
    - The LOD data is derived from non-FRBR MARC data
- Europeana Data Model
  - Used for Web Resources
- OWL 2 Web Ontology Language
  - Used for linking to external datasets
  - For linking duplicate Bibliographic Resources within libraries
- Dublin Core Terms
  - Used where more general semantics could/should be applied
- WGS84 Geo Positioning

Resulting usage o classes (from MARC data)

- Statistics from the RLUK dataset

Class	URI	Number of RDF triples
Concept	http://iflastandards.info/ns/fr/frbr/frbrer/C1007	61492571
Person	http://rdaregistry.info/Elements/c/C10004	28938624
Corporate Body	http://rdaregistry.info/Elements/c/C10005	28876282
Place	http://iflastandards.info/ns/fr/frbr/frbrer/C1010	24243355
Bibliographic Resource	http://purl.org/dc/terms/BibliographicResource	20022485
Media Type or Extent	http://purl.org/dc/terms/MediaTypeOrExtent	6748632
Event	http://iflastandards.info/ns/fr/frbr/frbrer/C1009	4614086
Time Span	http://www.europeana.eu/schemas/edm/TimeSpan	1926618
Web Resource	http://www.europeana.eu/schemas/edm/WebResource	1125852
Aggregation	http://www.openarchives.org/ore/terms/Aggregation	1040182
Agent	http://rdaregistry.info/Elements/c/C10002	396280
Family	http://rdaregistry.info/Elements/c/C10008	119701

Resulting properties usage (from MARC data)

- Statistics from the RLUK dataset

Property	URI	Number of RDF triples
label	http://www.w3.org/2000/01/rdf-schema#label	72832213
subject	http://purl.org/dc/terms/subject	66500103
Has Part	http://purl.org/dc/terms/hasPart	52300808
extent	http://purl.org/dc/terms/extent	42884658
identifierForTheManifestation	http://rdaregistry.info/Elements/m/P30004	39593358
placeOfPublication	http://rdaregistry.info/Elements/u/P60163	37805533
nameOfThePlace	http://rdaregistry.info/Elements/u/P60366	36545591
sameAs	http://www.w3.org/2002/07/owl#sameAs	33277509
nameOfThePerson	http://rdaregistry.info/Elements/a/P50111	28938624
nameOfTheCorporateBody	http://rdaregistry.info/Elements/a/P50032	27099104
dateOfPublication	http://rdaregistry.info/Elements/u/P60073	25195496
latitude	http://www.w3.org/2003/01/geo/wgs84_pos#lat	24244006
longitude	http://www.w3.org/2003/01/geo/wgs84_pos#long	24244006
languageOfTheContent	http://rdaregistry.info/Elements/u/P60099	20344741
titleProper	http://rdaregistry.info/Elements/m/P30156	19765445
modeOfIssuance	http://rdaregistry.info/Elements/m/P30003	19609578
publisher	http://rdaregistry.info/Elements/u/P60444	19534818
contentType	http://rdaregistry.info/Elements/u/P60049	19526623
intendedAudience	http://rdaregistry.info/Elements/u/P60520	18401952
statementOfResponsibility	http://rdaregistry.info/Elements/u/P60339	15964881
contributor	http://rdaregistry.info/Elements/u/P60398	15960077

External LOD Datasets Linked To

- Links to external datasets linked are available for the following:
  - VIAF Virtual Union Authority File
  - Geonames
  - Library of Congress Subject Headings
  - Library of Congress Children's Subject Headings
  - Library of Congress Classification
  - data.bnf.fr
  - Gemeinsame Normdatei
  - Dewey Decimal Classification
  - ISO639-2 Languages
  - MARC Countries

External LOD Datasets Linked To

- Availability of links

Subject Heading Systems

Subject Heading System	Language	Number of RDF triples
Library of Congress Subject Headings	English	2388129
Rameau (data.bnf.fr)	French	746067
Gemeinsame Normdatei	German	424388
Library of Congress Children's Subject Headings	English	699

Classification Systems

Classification System	Number of RDF triples
Library of Congress Classification	5371905
Dewey Decimal Classification	556600

External LOD Datasets Linked To

- Availability of links

Other datasets and ontologies

Dataset/Ontology	Number of RDF triples
Geonames	24243355
MARC Countries	21997155
ISO639-2 Languages	20344741
VIAF Virtual Union Authority File	283429

The European Library Open Dataset

Current Status

Datasets statistics:

Data Providers	Countries	Collections	Records	RDF Triples
25	22	157	82,800,212	3,454,724,561

Download combined dataset from all providers:

Format	Download files
Dublin Core	datasets.dc.tar (13.1 GB)
Rdf	datasets.rdf.tar (23.4 GB)
Turtle	datasets.ttl.tar (20.6 GB)

Downloadable data files (organized by provider):

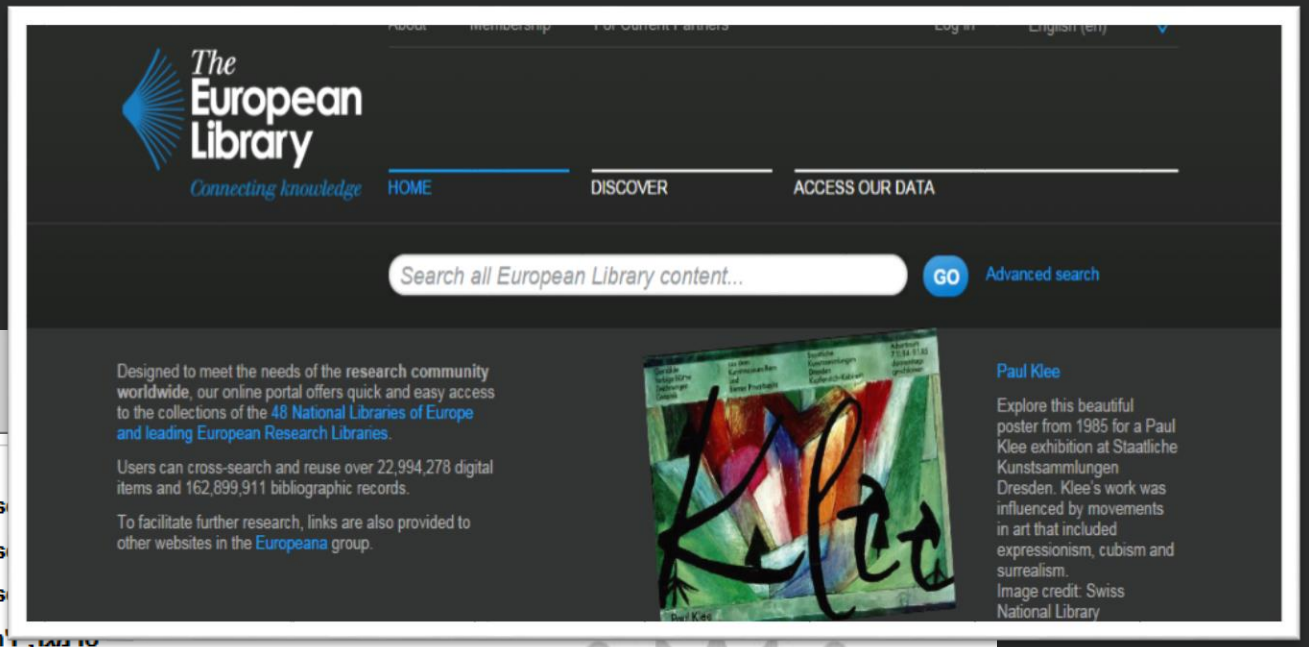
Data Provider/Collections	Records	RDF Triples	Download files
"Lucian Blaga" University of Sibiu	508	20,590	
Sibiu / Hermannstadt in old Postcards	110	3,687	RDF: rdf+xml (18.9 KB), RDF: turtle (16.5 KB), Dublin Core: xml (9.6 KB), RDF data statistics
History of Romanian People in Old Books	6	318	RDF: rdf+xml (8.5 KB), RDF: turtle (7.8 KB), Dublin Core: xml (6.8 KB), RDF data statistics
Sibiu - European Capital of Culture 2007	152	6,964	RDF: rdf+xml (123.7 KB), RDF: turtle (116.6 KB), Dublin Core: xml (98.3 KB), RDF data statistics

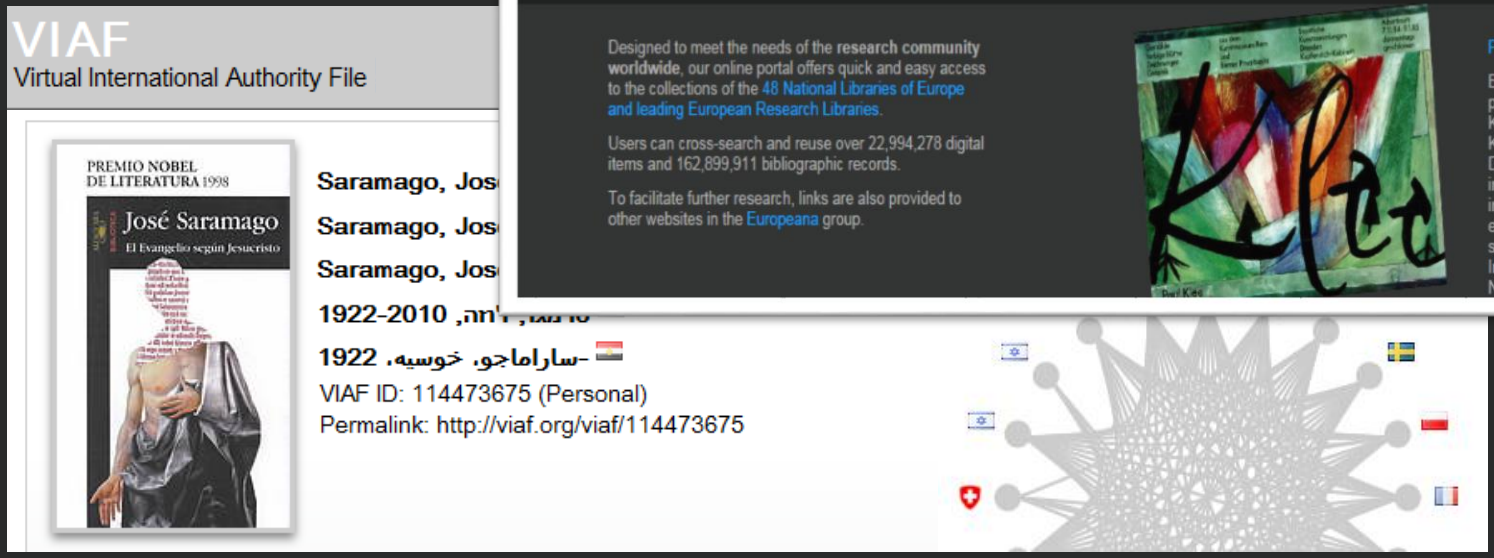
Linking Data

www.theeuropeanlibrary.org

Linked Data at The European Library

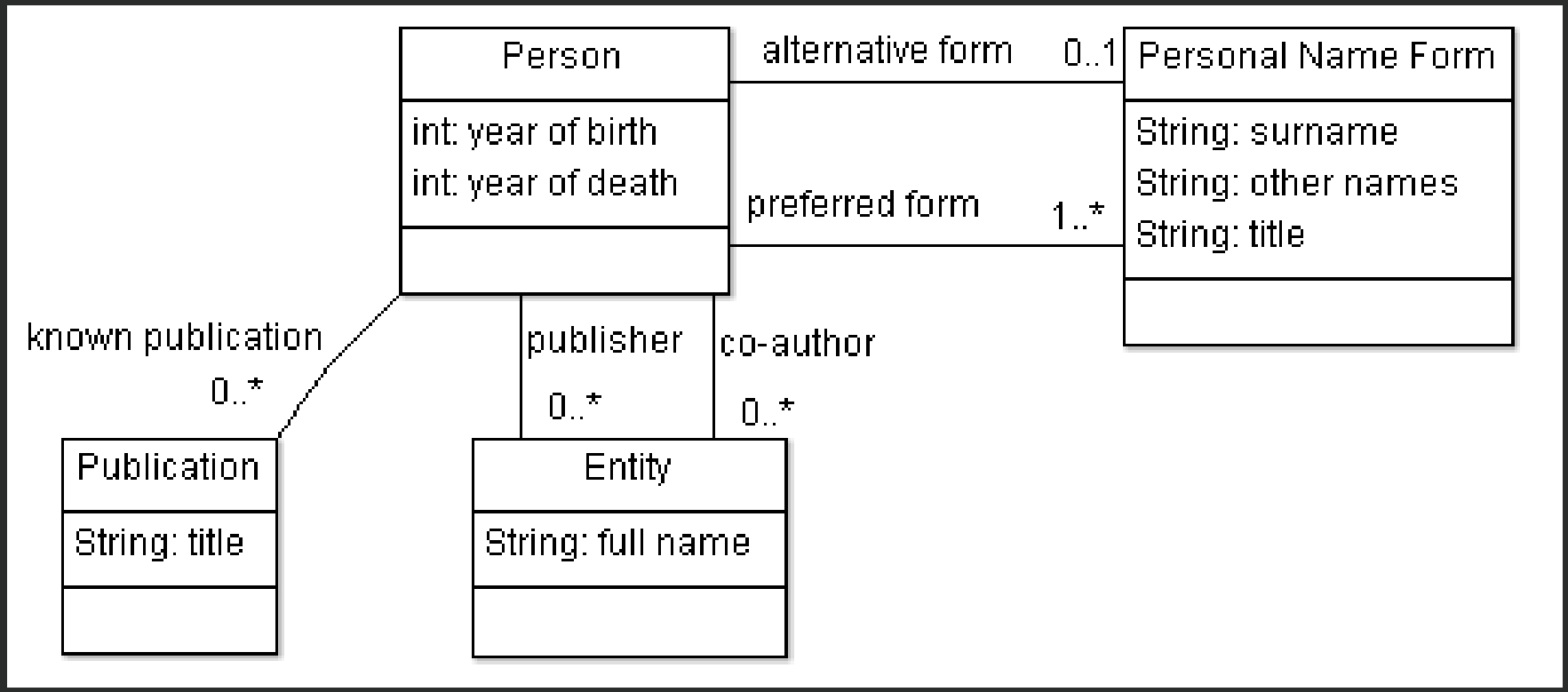
Managing and linking person names





The matching process

- VIAF data used for matching, disambiguation, and match probability



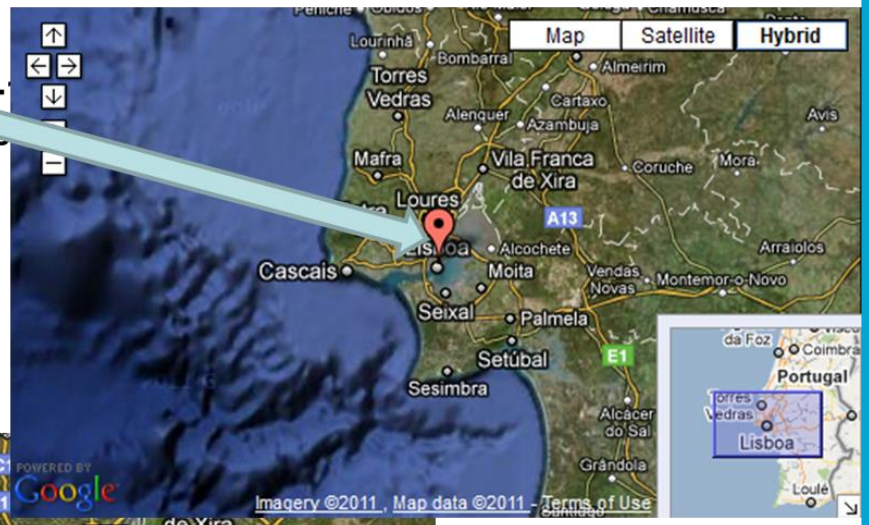
Matching work contributors with VIAF

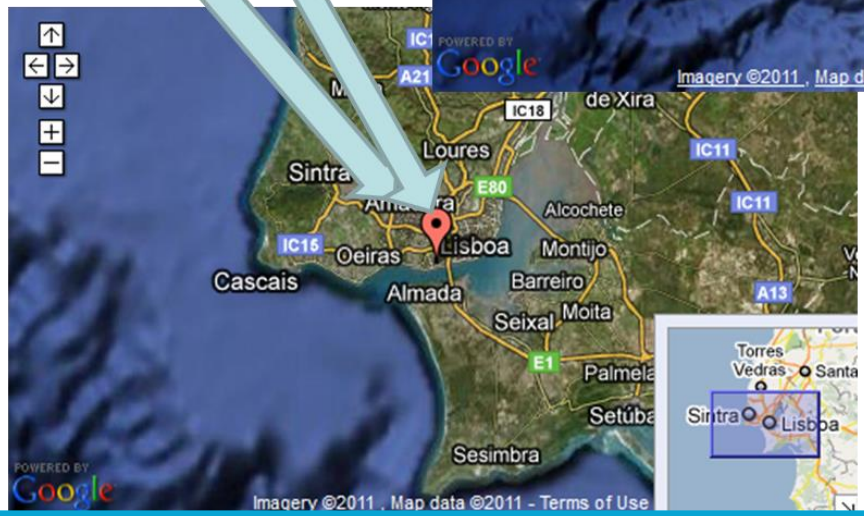
- Names are matched by similarity
- Confirmation of the correctness of a name match is taken from other matching data
  - The dates of birth and death
  - The title of the work is compared against the list of titles available in VIAF
  - All the contributors of the work are matched against the list of known co-authors in VIAF
  - The publisher(s) of the work are matched against the list of known publishers in VIAF
- A match is only chosen if enough supporting evidence is found

Linked Data at The European Library

Managing and linking place names

identifier: BN-E. 4639 P.
title: Belem Castle, Lisbon
creator: Sta[illegible]field, William Clar[illegible]1793-
description: V[illegible]ta da Torre de Belém rod[illegible]
pescadores a p[illegible]ar as re[illegible]s
publisher: J.Murra[illegible]
publisher: London
contributor: Finden, E[illegible]ard, [illegible]91-1857
date: 1832





The approach for place name linking

- The alignment is performed with Geonames
  - Using the RDF dump of Geonames
- A generic approach not using any language specific information
  - The words themselves are not used as evidence
    - We use only characteristics of the words (capitalization, size, etc)
  - Wordnets, part-of-speech analysis, morphological analysis, etc., are not used.
  - … in order to allow the use of this approach in a language independent manner

Resolution of the place names

- This task aims to find a single entity in the geographic ontology for linking to the place name
- The first step of this task is to find all possible candidates for the resolution in Geonames
- Uses a heuristic based predictive model:
  - Assigns a probability for each resolution candidate as a *match*
  - A link is established if a minimum probability threshold for a match is achieved.

Which information supports the place name resolution

Feature	Description
Number of words	The number of words in the place name.
Name match	If the recognized place name matched: the main name of the place, an alternate name, etc.
Exact name match	If the recognized place name matched exactly the place name.
Relative population	Relative population of the candidate in comparison with other candidates.
Geographic feature type	The type of geographic feature: continent, country, city, etc.
Related places found	The number of other place names found in the administrative hierarchy.
Relative related places	The relative number of administrative divisions found in the subject heading
In source country	If it is located in one of the source countries of the subject heading system.

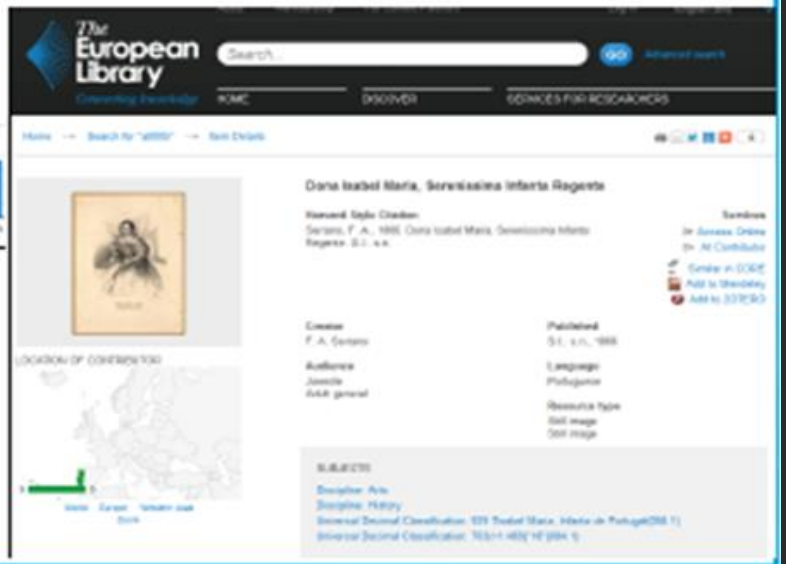
Linked Data at The European Library

Managing and linking concepts

Cross-Language Subject Browsing

Subject normalization centered in the Common European Research Classification Scheme

- A research and development oriented knowledge organization system
- It is part of the Common European Research project Information Format



Subject browsing is based on normalization of bibliographic data regarding the subject of the resources

Linking Subject Indexing and Classification Data

- The context
  - The centralization of bibliographic metadata enables resource access under a unified knowledge organization system
- The challenges
  - Diversity of languages
  - Diversity of knowledge organization systems in use across European libraries
  - Heterogeneous levels of details in subject information
- Current status at The European Library
  - Use of alignments between ontologies:
  - Alignments were created manually or semi-automatically
  - Alignments in use include: CERIF, MACS (LCSH, RAMEAU, SWD), UDC and DDC

The European Library

References

Further details may be consulted in the following publications:

- Freire, N, 2014, 'Word Occurrence Based Extraction of Work Contributors from Statements of Responsibility'. International Journal on Digital Libraries: Volume 14, Issue 3 (2014), Page 141-148. DOI: 10.1007/s00799-014-0113-3.
- Charles, V., Freire, N, Antoine, I., 2014, 'Links, languages and semantics: linked data approaches in The European Library and Europeana', in 'Linked Data in Libraries: Let's make it happen!' IFLA 2014 Satellite Meeting on Linked Data in Libraries.
- Freire, N, Muhr, M, 2013, 'Use of Authorities Open Data in the ARROW Rights Infrastructure' in proceeding of the DC-2013 Linking to the Future Conference, 2013.
- Freire, N, 2013, 'Visualization and navigation of knowledge in pan-European resources: the case of The European Library' in proceedings of International UDC Seminar on Classification & Visualization: interfaces to knowledge.
- N. Freire, et al., "Author Consolidation across European National Bibliographies and Academic Digital Repositories", 11th International Conference on Current Research Information Systems, 2012.
- N. Freire, J. Borbinha, P. Calado, "A Language Independent Approach for Aligning Subject Heading Systems with Geographic Ontologies", International Conference on Dublin Core and Metadata Applications 2011, 2011.
- N. Freire, J. Borbinha, P. Calado, B. Martins, "A Metadata Geoparsing System for Place Name Recognition and Resolution in Metadata Records", ACM/IEEE Joint Conference on Digital Libraries, 2011.

Thank you

Nuno Freire
nuno.freire@theeuropeanlibrary.org

The European Library