# The Europer 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



Users can cross-search and reuse over 22,994,278 digital items and 162,899,911 bibliographic records.

To facilitate further research, links are also provided to other websites in the Europeana group.



Explore this beautiful poster from 1985 for a Paul Klee exhibition at Staatliche Kunstsammlungen Dresden. Klee's work was influenced by movements in art that included expressionism, cubism and surrealism. Image credit: Swiss National Library

#### FEATURED COLLECTIONS

DISCOVER CONTRIBUTORS

Historic Newspapers

Reading Europe (1078)

Exhibition Foyer

Manuscripts and Prince... (34)





### http://www.theeuropeanlibrary.org

## What does The European Library offer?

#### <sup>•</sup>Large-scale aggregation Infrastructure



OUT-D-LIB | CAREENTISSUE | ARCHIVE | INDEXES | CALENDAR | ALTHOREGAREELINES | SUB



rahip fee has been expanded, with the structure simplifie

#### Data and digital content of Europe's libraries

#### Data enrichment Linked open data



#### Experienced European project partner

Aline: Usability Testing Report of the The European Website Alpha Version (March 2012) wy Porg on E Mach 26, 2012 # 342 pm

#### Report's Executive Summary Ma Europeana Web Site)

In other to measure the quality of the new European Library ports? The European Library offse conducted a series of solatility tests in January and Fibrary 2012 to the apta resion of the welders' network whether the solary in the test of the solar test conducted the testory, as well as interfuent guinessty of Appel Discrete and the European Library start from phene searchers' the Caspitor of the User Case and The Telephene Library start and the solar start in the Search and the solar start National Library starts along 60 to guine search of the solar guine the Search and Searc

summary. Hot text users spontaelevely identified the man purpose of The European tables potcht to dire science and early access to the collections of European and anoth beners. As a whole, they expected the offinite of a catalogies assert hancion with near beners, and the functions that ensure immediate access to the indicated a Additionally, they supercised a quick enseme of the collections.

Data distribution



More Conferences and Meetings
(77) If used for Uniformity Comp Topology

News and Announcements 3/28/12

In today's means: a Bowker ebook study: project Europeane Libraries nears in end; Ex Libris robuste DX Bot Articles; OCLC launches ArchiveGoti; Istelaliton acquires Mohile Context; Gougle instructors third party commenting instrum to rival Facebook; and Publishing Technology joins GSE Research to true online partal.

#### eleases results of global ebook research

VALUE on a poer mean that reports according to Boyner Accelet only Child eldoch Albane "Initial, Annual, the U.K. and the US and the or of a books, with more than 300 promet of respondents reporting books in the six months piers to the acover. Respondent in Planies in the set Billing's taken guidence and a solow, the group errors and a Utility purchase behavior works by county, presentence While purchase behavior works by county, presentence

## **Open data distribution**

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



#### 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.

#### 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?

More

How the data looks like?

#### Available APIs and Datasets

The European Library API v2.0	The European Library Open Dataset	CURRENT NEWS		
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: X011 JSON Find out more S	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 CCO 1.0, which allows the unrestricted use of the data for any purpose at all by anyone, including commercial use.	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		
Metadata Harvesting - OAI-PMH Repository	RLUK Linked Open Data	Grand opening of the new National Library of Latvia 25.08.2014		
Metadata Harvesting - OAI-PMH Repository The European Library provides access to its datasets	Research Libraries UK Linked Open Data The RLUK Database is an established Union Catalogue			

## 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	LIBI	Number of	
Class		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	http://purl.org/dc/terms/MediaTypeOrExtent	6748632	
Extent			
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	<b>RDF Triples</b>	Download files
"Lucian Blaga" University of Sibiu	508	20,590	
Sibiu / Hermannstadt in old Postcards	110	3,687	RDF: <u>rdf+xml</u> (18.9 KB), RDF: <u>turtle</u> (16.5 KB), Dublin Core: <u>xml</u> (9.6 KB), <u>RDF data statistics</u>
History of Romanian People in Old Books	6	318	RDF: <u>rdf+xml</u> (8.5 KB), RDF: <u>turtle</u> (7.8 KB), Dublin Core: <u>xml</u> (6.8 KB), <u>RDF data statistics</u>
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

## www.theeuropeanlibrary.org



## **Linking Data**



## 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



# 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							
The European Search	releasing for Cannot Plathau Decomen	LINE CONTRACTOR FOR FRAME ENGLISHING FOR FRAME ENGLISHING FOR FRAME	g th Unglish (or) Advanced search watchests	Subject norma mon Europe Scheme • A research and organization system	alization cent an Researc d development tem	ered in the h Classif oriented kn	e Com- ication
Pursuenties (H12012) Pursuenties (H12012)	Natural Actions and reduced as (1.24.4.524)	English CH, 307(411) Barrian (H.2018) French (H.2017) 100 English (H.2017) 100 Egenical (J.2018) Decision (J.2018) Decision (J.2018) Decision (J.2018) Public (J.2017) Public	Later (VIBI JUM) Kalano (NET-A28) Postoparene (JBI, JBI) Botzki Konengun (JEC-29) Brownson (NO-20) Brownson (NO-20) Brownson (NO-20) Brown (JEC-20) Brown (JEC-20) Brown (JEC-20) Kalanie (JEC-20) Kalanie (JEC-20) Kalanie (JEC-20)	It is part of the Information Form European Library	Common Europe		
Subject brows of bibliograph of the resource	ing is based ic data rega	on norma rding the	lization subject	Here - Band to 'allor' - Band And	Dana kashel Maria, Berwaian Hawad Indo Chadan Sarana, P.A., HHI Cara Lubol Ma Engene B.J. v.k. Control F.A. (Serges Rolleren Jacob Rolleren Jacob Rolleren Banghe An Jacob Physical Checkloster B Banghe An	ina Infanta Ragante III, Somissing Mints St. 11, 1988 Langunge Philipine Remota Type Still mage Still mage Still mage	Contraction of the second seco

## 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



#### 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.



#### Nuno Freire nuno.freire@theeuropeanlibrary.org

