Cultural Heritage in CLEF (CHiC) 2012

Pilot Lab Overview

Vivien Petras
Humboldt-Universität zu Berlin
Roma, 17. September 2012

Contents

Cultural Heritage Information Systems

"Cultural heritage, as distinguished from natural heritage, consists of objects created by, or given meaning by, human activity."

(Bearman & Trant, 2002)

→ multilingual & multimedia

- general users (interested in culture, the "informed citizen"),
- cultural heritage professionals (content producers, collection managers),
- educational users (researchers, teachers, students), and
- tourist users (travelers, tourist agencies, information centers)
- the "information tourist" / casual user

CHiC Tasks (1)

- Ad-hoc
  - default IR task
  - Predetermined information need, expected outcome
  - Query → ad-hoc results
  - Binary relevance assessments / standard IR measures

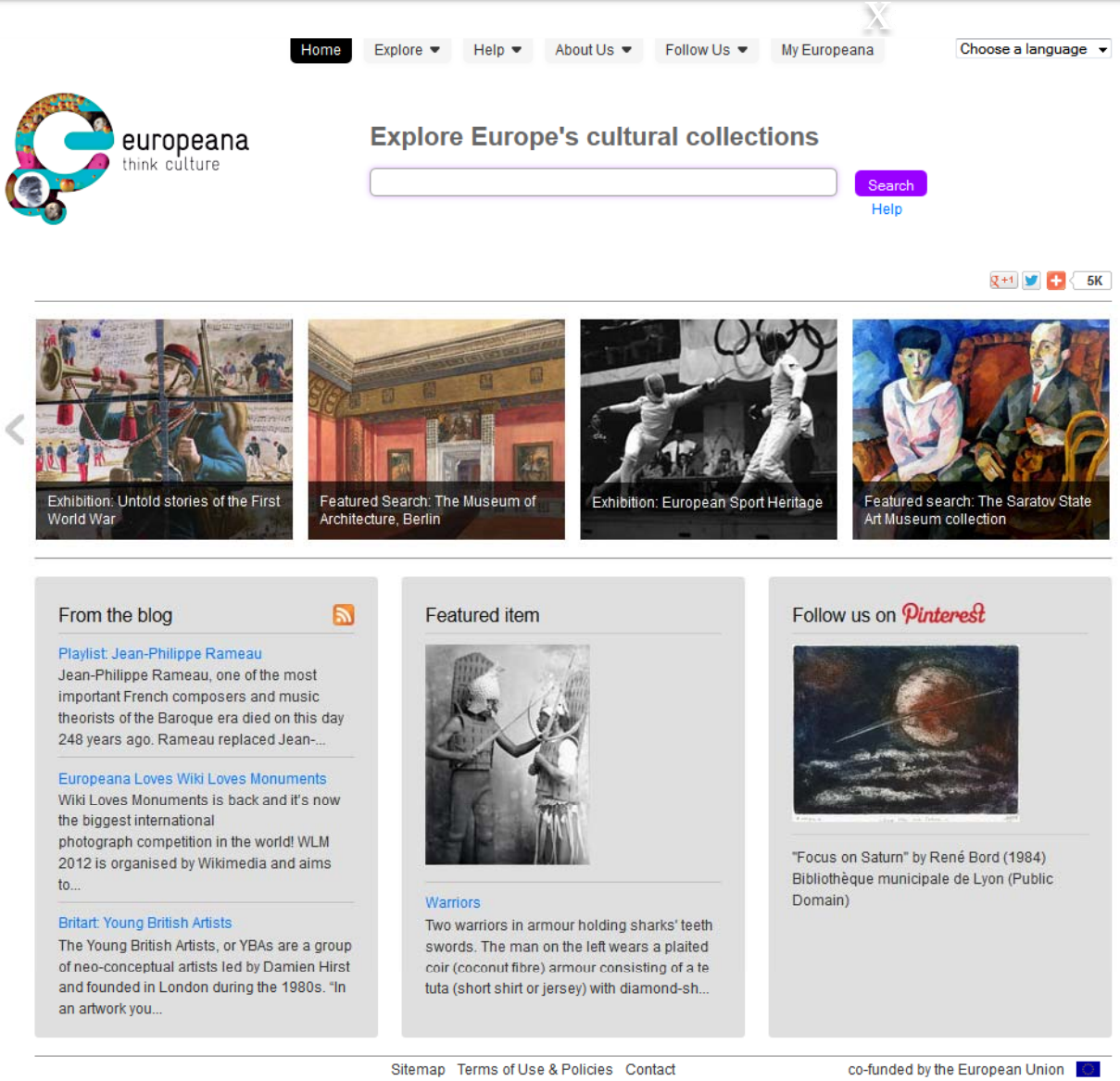
- Variability / Diversity
  - For the casual information tourist „probing“ the system
  - ad-hoc query, unexpected outcome
  - 1 result page → as diverse as possible
  - Diversity: media type, content provider, content category, …?
  - Binary relevance assessment + diversity measure (cluster recall)

CHiC Tasks (2)

- Semantic Enrichment
  - Improve semantic ambiguity of query process („Did you mean?“)
  - Ad-hoc query → 10 query suggestions
  - Internal and external resources for recommendations
  - (a) Binary relevance assessments of query suggestions
  - (b) Binary relevance assessments of IR runs using query suggestions for query expansion / standard IR measures

- Languages: English, French, German & Multilingual

CHiC Collection(s)



- Complete Europeana index (03/2012)
- 23,300,932 documents
- Metadata only + automatically added tags (content enrichment) for 30% of documents
- 62% images, 35% text, 2% audio, 1% video

CHiC Collection(s) - Documents

Free Access

View item at
British Broadcasting Corporation

Rights: All rights, including copyright, in the content of the pages submitted by the BBC are owned or controlled for these purposes by the BBC. In accessing the pages

See more

Identifier:
EUS_9E1AE3297AEB448C9896ABC

Format: Colour

Language: English

Publisher: BBC |

Publication date: 01/07/1967

Agatha Christie

Contributor: Agatha Christie (WRITER)

Date: [1967]

Geographic coverage: United Kingdom

Type: Still | Drama/Fiction |

Subject: Performed drama/TV play | Drama

Description: -

Agatha Christie wrote more than 60 novels, pla
have been adapted for television.

Data provider: British Broadcasting Corporatio

Provider: EUscreen Project | United Kingd

Explore further!

Similar content



Auto-generated tags

When

Where

Place Label: velká británie ; spojené království ; großbritannien ; suurbritannia ; großbritannien und nordirland ; inggris raya ; велика британия ; gb ; u.k. ; združeno kraljestvo velike britanije in severne irske ; marea britanie ; jungtinė karalystė ; det forenede kongerige ; ühendkuningriik ; det forenede kongerige storbritannien og nordirland ; regnum unitum ; regno unito ; обединено кралство великобритания и северна ирландия ; великобритания ; velika britanija ; britanniarum regnum ; united kingdom of great britain and northern ireland ; združeno kraljestvo (v. britanija in s. irska) ; regatul unit ; grande-bretagne ; обединено кралство ; det forente kongerike storbritannia og nord-irland ; великобритания ; wielka brytania ; reino unido ; united kingdom ; zjednoczone królestwo wielkiej brytanii ; iso-britannia ; england ;

0 0 0

Like +1 Tweet

Translate details

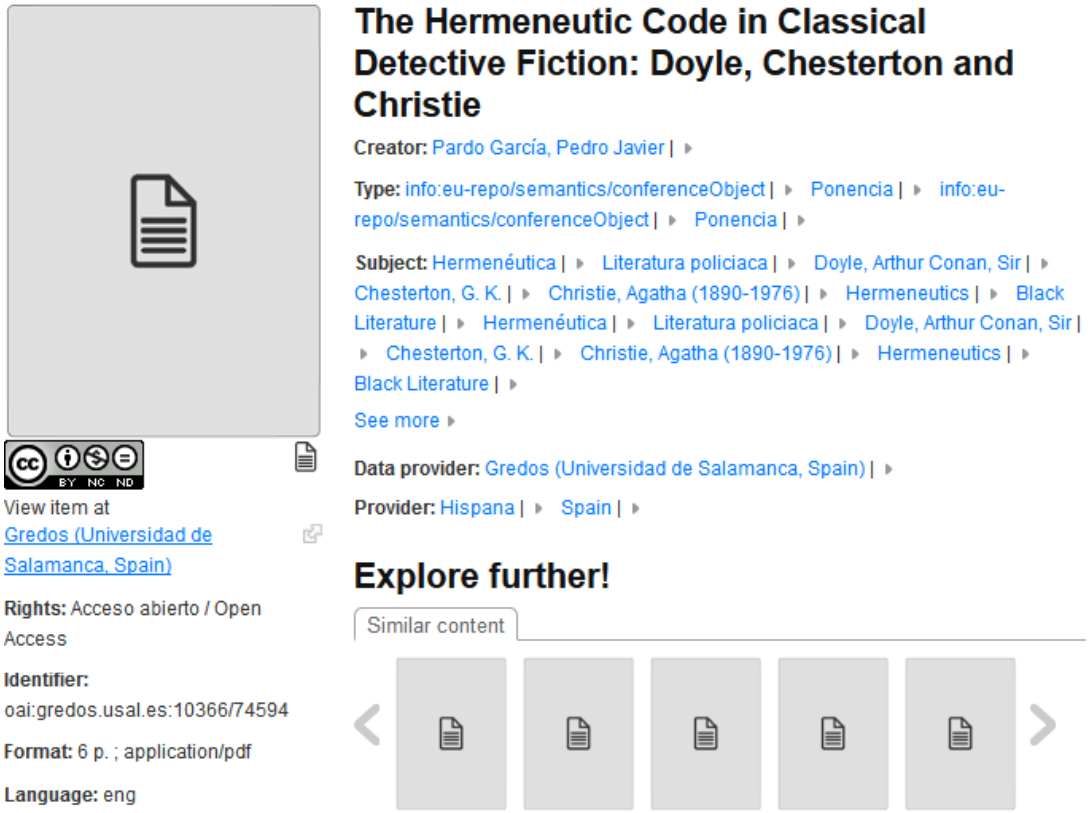
Select language

Powered by Microsoft® Translator

Cite on Wikipedia

Auto-generated tags

CHiC Collection(s) – By Language



The Hermeneutic Code in Classical Detective Fiction: Doyle, Chesterton and Christie

Creator: Pardo García, Pedro Javier | ▸

Type: info:eu-repo/semantics/conferenceObject | ▸ Ponencia | ▸ info:eu-repo/semantics/conferenceObject | ▸ Ponencia | ▸

Subject: Hermenéutica | ▸ Literatura policiaca | ▸ Doyle, Arthur Conan, Sir | ▸ Chesterton, G. K. | ▸ Christie, Agatha (1890-1976) | ▸ Hermeneutics | ▸ Black Literature | ▸ Hermenéutica | ▸ Literatura policiaca | ▸ Doyle, Arthur Conan, Sir | ▸ Chesterton, G. K. | ▸ Christie, Agatha (1890-1976) | ▸ Hermeneutics | ▸ Black Literature | ▸

See more ▸

Data provider: Gredos (Universidad de Salamanca, Spain) | ▸

Provider: Hispana | ▸ Spain | ▸

Explore further!

Similar content

View item at Gredos (Universidad de Salamanca, Spain)

Rights: Acceso abierto / Open Access

Identifier: oai:gredos.usal.es:10366/74594

Format: 6 p. ; application/pdf

Language: eng

Source: Pardo García, P. J. (1994). The Hermeneutic Code in Classical Detective Fiction: Doyle, Chesterton and Christie. En Actas del XVI Congreso de la Asociación Española de Estudios Anglo-Norteamericanos (pp. 335-343). Valladolid: Universidad de Valladolid. | ▸

Publisher: Universidad de Valladolid (España) | ▸

Publication date: 1994

- by language of content provider
- 13 of 30 with >100,000 documents
- English: 1.11 mio.
- French: 3.64 mio.
- German: 3.87 mio.
- Multilingual: all

CHiC Queries

- 50 sampled queries from Europeana query logs
- Query had to result in at least 1 full result view
- many named entities typical for cultural heritage

→ Annotated by query category: person, location, work title, topical, other

→ Translated from English to French & German

→ „information need“ added for disambiguation & relevance assessments

CHiC Queries - Disambiguation

Red kite (EN)

Cerf-volant rouge (FR-1)
Milan royal (FR-2)

Roter Drache (DE-1)
Rotmilan (DE-2)

CHiC Participation

Chemnitz University of Technology, Dept. of Computer Science	Germany
GESIS – Leibniz Institute for the Social Sciences	Germany
Unit for Natural Language Processing, Digital Enterprise Research Institute, National University of Ireland	Ireland
University of the Basque Country, UPV/EHU & University of Sheffield	Spain / UK
School of Information, University of California, Berkeley	USA
Computer Science Department, University of Neuchatel	Switzerland

- 131 runs
- all language combinations
- EN monolingual in all tasks most popular
- ad-hoc & semantic enrichment equally popular
- 2 multilingual baseline runs from Europeana

CHiC Relevance Assessments

- pools: 35,000 (EN), 22,000 (FR + DE)
- broad distribution of number of relevant documents
- topics without relevant documents:
  - EN = 14
  - FR = 11
  - DE = 2
  - Multilingual = 1
- 45 runs for semantic enrichment:
  - Semantic correctness of query suggestions
  - 45 new runs as query expansion (Lucene index)
- 32 runs for variability
  - Media types + content providers
  - Content category of document…

CHiC Relevance Assessments - Categories

View

View item at Saxon State and University Library, Dresden / Deutsche Fotothek

Identifier: http://www.deutschefotothek.de/obj70601973.html

Format: image/jpeg

Language: de-DE

Source: SLUB/Deutsche Fotothek |

Dresden, Pionierpalast "Walter Ulbricht" (Schloss Albrechtsberg). Kundgebung mit Prof. Herbert Gute, Oberbürgermeister von Dresden und Johanna Andersen-Nexö, amlässlich des 16. Jahrestages des Abwurfs der Atombombe über Hiroshima, 6. August 1961

Creator: Höhne, Erich & Pohl, Erich (Fotograf) |

Contributor: Gute, Herbert ; Andersen-Nexö, Johanna

Coverage: Dresden | Hiroshima | Japan |

Date: 1961

Type: image |

Subject: Kontaktbogen | Politik | Kundgebung | Fotos | Pressearchiv Höhne/Pohl |

Description: Dresden, Pionierpalast "Walter Ulbricht" (Schloss Albrechtsberg). Kundgebung mit Prof. Herbert Gute, Oberbürgermeister von Dresden und Johanna Andersen-Nexö, amlässlich des 16. Jahrestages des Abwurfs der Atombombe über Hiroshima, 6. August 1961 Dresden& Hiroshima & Japan

Provider: Saxon State and University Library, Dresden / Deutsche Fotothek | Germany |

0 | 0 | 0

Like | +1 | Tweet

Translate details

Select language

Powered by Microsoft® Translator

Cite on Wikipedia

Auto-generated tags

Explore further!

Similar content





CHiC Results

- Ad-hoc: best monolingual MAP

EN	52%	UPV
FR	38%	Neuchatel
DE	60%	Chemnitz

- Variability: best P@12 / # queries without relevant docs

EN	36%	UPV (SimFacets)	2
FR	15%	Chemnitz (DBPedia_Subjects)	8
DE	29%	Chemnitz (NO)	2

- Variability: avg. relative cluster recall

EN	86%	Chemnitz (BO2_3D_10T)
FR	69%	Chemnitz (NO)
DE	92%	Chemnitz (BO2_3D_10T)

CHiC Results

- Semantic Enrichment: best P@10 (semantic correctness)

EN	75%	UPV
FR	57%	Chemnitz
DE	74%	Gesis

- Semantic Enrichment: best MAP (query expansion)

EN	34%	Original
	30%	DERI
FR	32%	Original
	15%	Chemnitz
DE	57%	Original
	32%	Gesis

Approaches

- Systems: Cheshire, Indri, Lucene (Chemnitz Xtrieval), Solr
- Ranking: vector space, language modeling, DFR, Okapi
- Translation: Google Translate, Wikipedia entries, Microsoft
- Variability:
  - Chemnitz: least recently used (LRU) algorithm to prioritize documents with different media types & providers
  - UPV: maximal-marginal relevance (MMR) to cluster results & cosine similarity to select the most dissimilar documents
- Semantic enrichment:
  - Wikipedia at different levels of detail (article titles, first paragraph, full text)
  - Wordnet, DBpedia
  - co-occurrence from Europeana collection

CHiC Outlook

- Fine-tune & adjust (collections, queries)
- Ad-hoc for baselines
- Interesting experiments in realistic scenarios → but complicated to evaluate!
- More user interaction?
- More languages?

CHiC 2012 Workshop: Thursday

Organizers:

Humboldt-Universität zu Berlin / University of Padova / Europeana / University of Sheffield / Royal School of Library and Information Science Copenhagen

Thank you to:

Anthi Agoropoulou, Toine Bogers, Nicola Ferro, Maria Gäde, Antoine Isaac, Michael Kleineberg, Ivano Masiero, Mattia Nicchio, Christophe Onambélé, Oliver Pohl, Juliane Stiller, Elaine Toms, Astrid Winkelmann