„Systemy organizacji wiedzy i ich rola w integracji zasobów europejskich bibliotek cyfrowych”

Adam Dudczak
Poznańskie Centrum Superkomputerowo-Sieciowe
(maneo@man.poznan.pl)

I Konferencja „Polskie Biblioteki Cyfrowe”

Plan prezentacji

- Przedstawienie podstawowych wymagań związanych z tworzeniem wyszukiwarki zasobów wielojęzycznych
- Systemy organizacji wiedzy i obszary ich wykorzystania w europejskich bibliotekach cyfrowych
- Praktyczne spojrzenie na problemy związane z udostępnianiem treści zgromadzonych w polskich bibliotekach cyfrowych dla **Europeany**

Podstawowe wymagania i najważniejsze pytania

- Dlaczego wielojęzyczność jest problemem?
- Czy dotyczy to również polskich bibliotek cyfrowych?
- Jakie wymagania funkcyjne powinien spełniać system udostępniający treści w wielu językach?

Dlaczego wielojęzyczność jest problemem?

Dlaczego wielojęzyczność jest problemem?

flickr

Home You Orga

Search

The Commons

We couldn't find any

Did you mean pray?

Search

Photos Groups People

The Commons

paris

SEARC

Full text Tags only

We found **346 results** matching **paris**, in **The Commons**[x].

View: Most Relevant - Most Recent - Most Interesting Show: Details - Thumbnails

Paris Exposition: Pont d'Jena toward Chateau of Water, view from the, Paris, France, 1900 [correction: Pont d'lena] by **Brooklyn Museum**

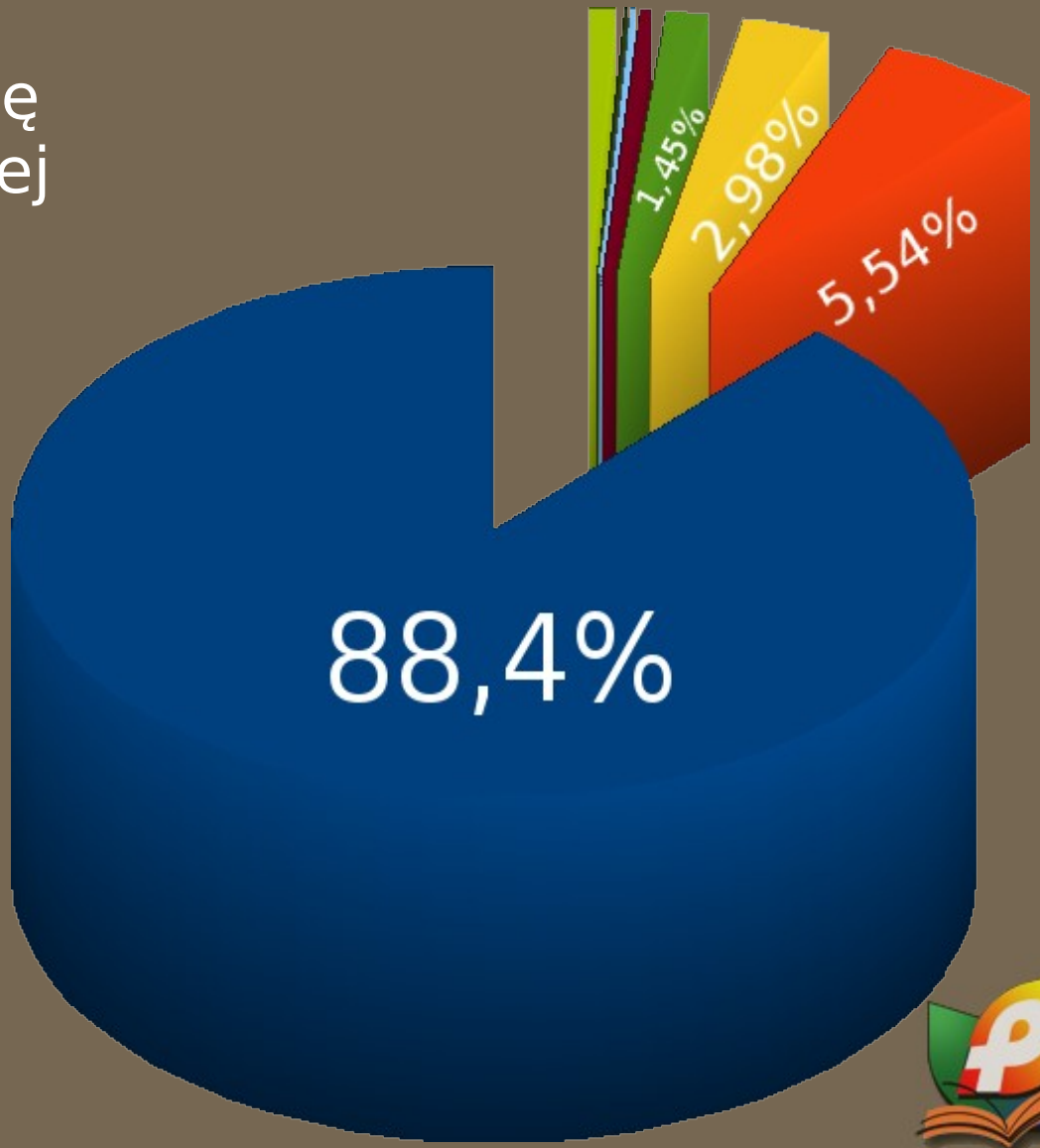
28 comments 96 faves 4 notes

Tagged with mars, paris, france, tower ...

Taken some time in 1900, uploaded May 12, 2008

Czy dotyczy to również polskich bibliotek cyfrowych?

- Języki jakimi posługują się użytkownicy Wielkopolskiej Biblioteki Cyfrowej
  - Polski – 88,4%
  - Angielski – 5,54%
  - Niemiecki – 2,98%
  - Rosyjski – 1,45%
  - Francuski – 0,4%
  - Czeski – 0,17%
  - Ukraiński – 0,13%
  - Pozostałe – 0,93%
- W sumie 77 612 wizyt
- Dane z ostatniego miesiąca (23.10 - 22.11)



Czy dotyczy to również polskich bibliotek cyfrowych?

- Materiały w wielojęzyczne w polskich bibliotekach cyfrowych
  - WBC - 21,
  - ŚBC – 14 różnych języków
- Czy użytkownik nie posługujący się językiem polskim jest w stanie odnaleźć zasoby w swoim języku?

Jakie wymagania funkcyjne powinien spełniać system udostępniający treści w wielu językach?

- Interfejs użytkownika we wszystkich wspieranych językach [1]
- Mechanizmy indeksacji uwzględniające specyfikę wspieranych języków [1]
  - Przede wszystkim analiza morfologiczna
  - Hasłowanie, lematyzacja, analiza syntaktyczna
  - Listy wyrazów pospolitych (ang.*stop words*)

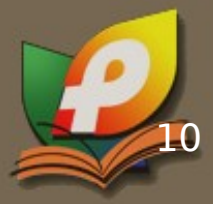
[1] EDLnet - D2.5 - Europeana Outline Functional Specification, wersja 1.0

Jakie wymagania funkcyjne powinien spełniać system udostępniający treści w wielu językach?

- Wyszukiwanie [1]
  - Zapytanie w określonym języku, w wyniku tylko obiekty w tym języku
  - Zapytanie w określonym języku, zwraca pasujące obiekty niezależnie od języka
  - Zapytanie w określonym języku jest tłumaczone na wszystkie wspierane języki i zwraca obiekty we wszystkich (wspieranych) językach

Jakie wymagania funkcyjne powinien spełniać system udostępniający treści w wielu językach?

- Wydawanie zapytań w dowolnym języku [1]
  - Wsparcie dla znaków diakrytycznych
  - Unicode
- Wyniki w języku użytkownika [1]
  - Tłumaczenie metadanych/obiektów na język w którym użytkownik wydał zapytanie



Systemy organizacji wiedzy i obszary ich wykorzystania w europejskich bibliotekach cyfrowych

- Systemy organizacji wiedzy (ang. *Knowledge organisation system*, KOS)
- Podstawowe definicje
  - Słownictwo kontrolowane
  - Tezaurus
- SKOS
- Potencjalne obszary zastosowań

Systemy organizacji wiedzy - definicje

- Słownictwo kontrolowane (SK) – zbiór unikalnych słów, które posiadają jednoznaczną definicję [2]
  - Jeżeli dany wyraz posiada wiele znaczeń uzupełnia się go o dodatkową informację
    - np. Łódź, Łódź (powiat poznański), Łódź (jednostka pływająca) – hasła z Wikipedii
  - Przykłady:
    - KABA [3],
    - Medical Subject Headings (MeSH)

[2] Wikipedia : Controlled vocabulary

[3] „Zastosowanie ontologii do wykrywania wiedzy”, Dariusz Daćko

Systemy organizacji wiedzy - definicje

- Tezaurus – to słownictwo kontrolowane zawierające informację o terminach powiązanych
  - Synonimy, pojęcia nadrzędne (hipernimy) i podrzędne (hiponimy), kolokacja i inne
  - Przykłady:
    - Library of Congres Subject Headings (LCSH)
    - WordNet,
    - synonimy.ux.pl

Simple Knowledge Organisation System

- SKOS - rodzina języków formalnych służących do zapisu struktur takich jak :
  - słowniki kontrolowane, tezaurusy, klasyfikacje
  - SKOS jest zapisywany w postaci dokumentu zgodnego z RDF i RDF Schema
  - Standard ten jest rozwijany przez konsorcjum WWW (W3C)
  - SKOS jest/ma być ważnym elementem sieci semantycznej

Jak wykorzystać SOW?

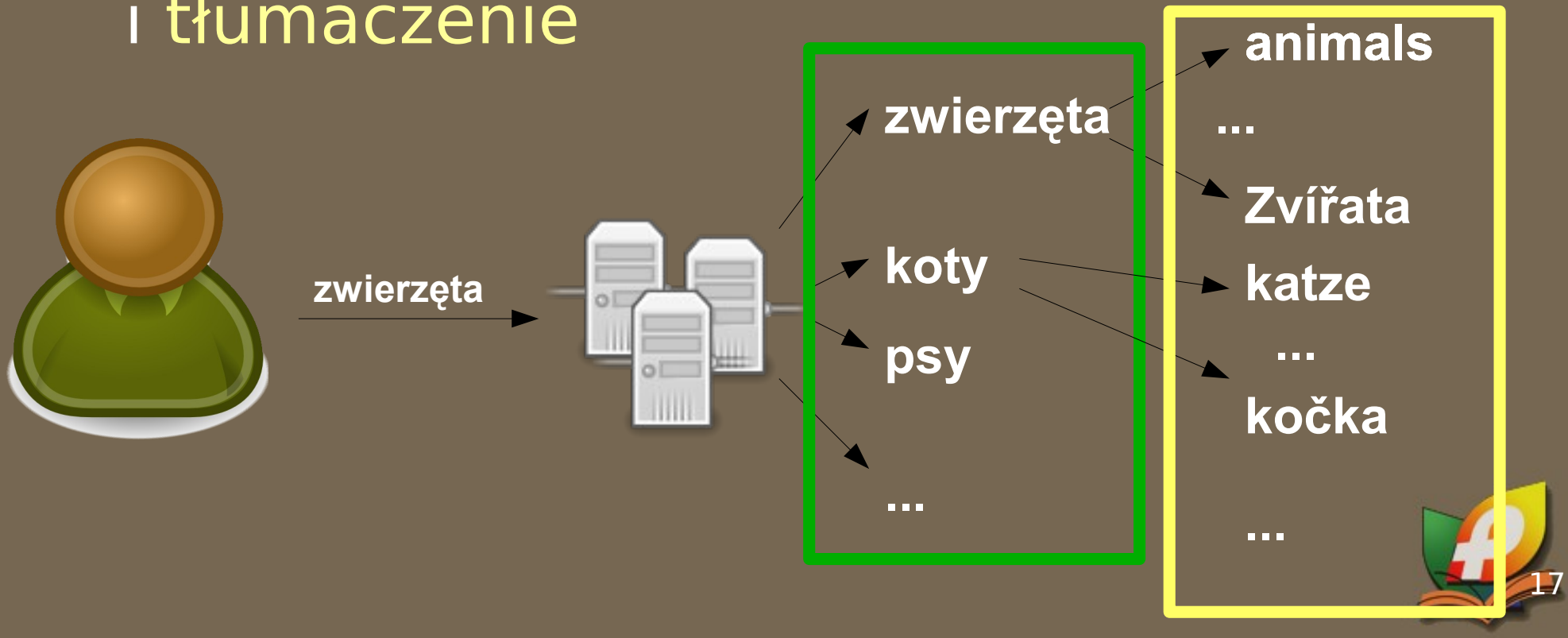
- Integracja różnych kolekcji dokmentów:
  - Tworzenie odwzorowań między dwoma zupełnie odmiennymi :
    - Słownictwami, klasyfikacjami
- Przetwarzanie zapytań
  - Tłumaczenie, rozszerzanie
- Automatyczne tłumaczenie opisów bibliograficznych
  - Indeksowanie, wyświetlanie wyników

Tworzenie odwzorowań – potencjalne problemy

- Dopasowywanie dwóch słowników kontrolowanych
  - Niezgodność znaczeniowa np. vegetables i Gemüse
  - Różne poziomy szczegółowości i dziedziny
    - Słownik A: pojazd mechaniczny, czołg
    - Słownik B: car
  - Brak odpowiadających terminów
- MACS (Multilingual Access to Subject)
  - SWD (niemiecki), RAMEAU (francuski) i LCSH (angielski)

Przetwarzanie zapytań

- Dopasowywanie zapytań użytkowników do tego co jest w KOSie
- Rozwijanie zapytania o pojęcia podrzędne i tłumaczenie



Automatyczne tłumaczenie opisów bibliograficznych

- [1] zakłada użycie :
  - odwzorowań między słownictwami lub
  - tłumaczenia maszynowego
- W przypadku opisu bibliograficznego tłumaczenie maszynowe może dać stosunkowo dobre wyniki
  - Rzeczowniki, niewiele zwartego tekstu
- Tłumaczenie całych obiektów o wiele trudniejsze (manuskrypty, OCR)

Wnioski końcowe

- Jak dotąd znamy tylko schemat metadanych który będzie wykorzystywała Europeana
  - Brak informacji o słownikach wartości
- Uspójnienie przechowywanej w polskich bibliotekach cyfrowych informacji bibliograficznej bardzo ułatwiłoby pracę nad włączeniem polskich zbiorów do Europeany

Wnioski końcowe

- Preferowanym sposobem włączania zasobów do Europeany są krajowe agregatory metadanych
- Na tym poziomie możliwe jest częściowe dokonanie przekształceń i czyszczenia przekazywanych dalej metadanych
- Rolę takiego agregatora dla Polski pełni FBC

Dziękuje za uwagę

Adam Dudczak
Poznańskie Centrum Superkomputerowo-Sieciowe
(maneo@man.poznan.pl)

I Konferencja „Polskie Biblioteki Cyfrowe”