POZNAŃSKIE CENTRUM SUPERKOMPUTEROWO-SIECIOWE

Zaawansowane usługi dla rozproszonych bibliotek cyfrowych

Marcin Werla

Poznańskie Centrum Superkomputerowo-Sieciowe

IV Warsztaty „Biblioteki cyfrowe”

Toruń, 2007

Plan prezentacji

- Federacja Bibliotek Cyfrowych
- Przechowywanie skanów wysokiej jakości
- Duplikaty w sieci bibliotek cyfrowych
- Spójność opisów bibliograficznych

Federacja Bibliotek Cyfrowych

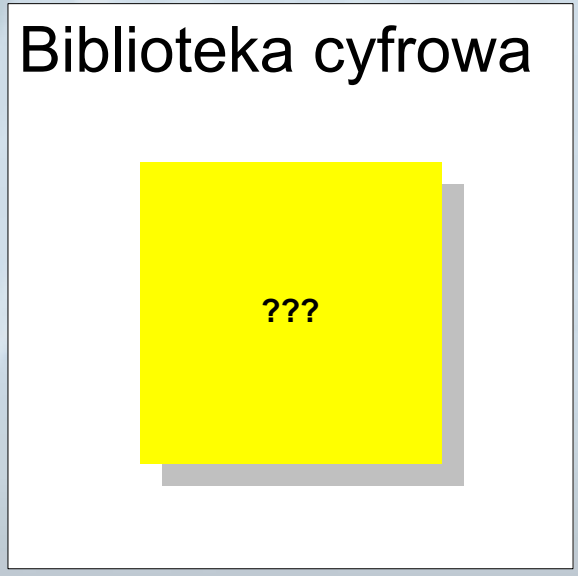
http://fbc.pionier.net.pl/

Agnieszka Lewandowska

FEDERACJA BIBLIOTEK CYFROWYCH

1996

Początek prac

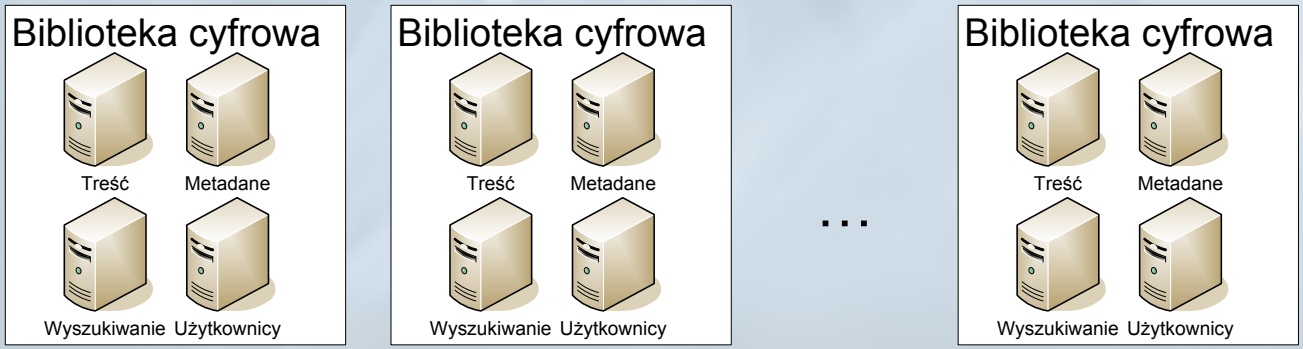


1999

Określenie architektury

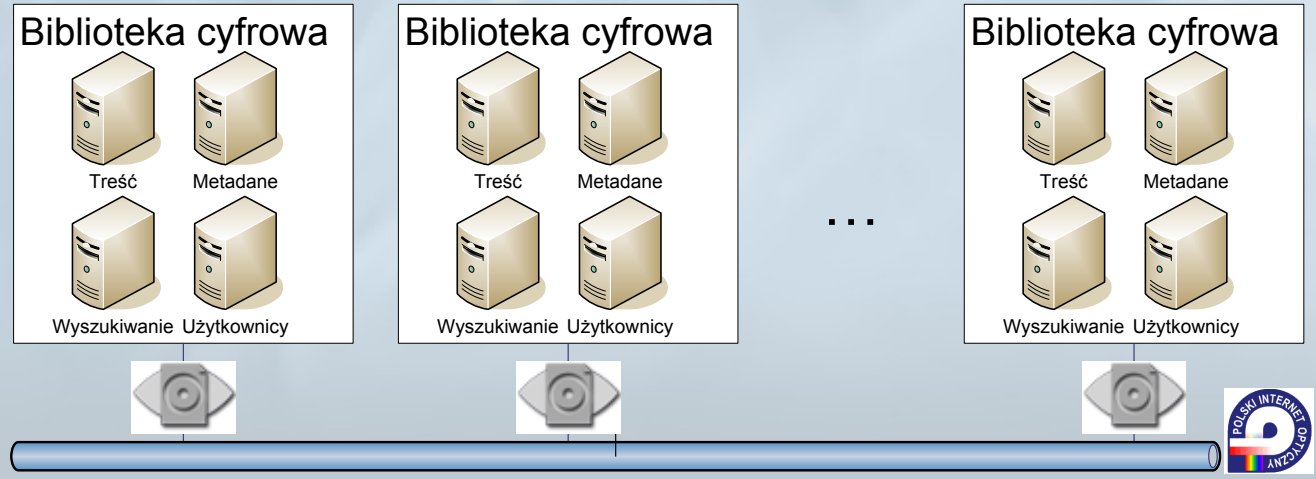
2002-...

Rozpoczynają się wdrożenia

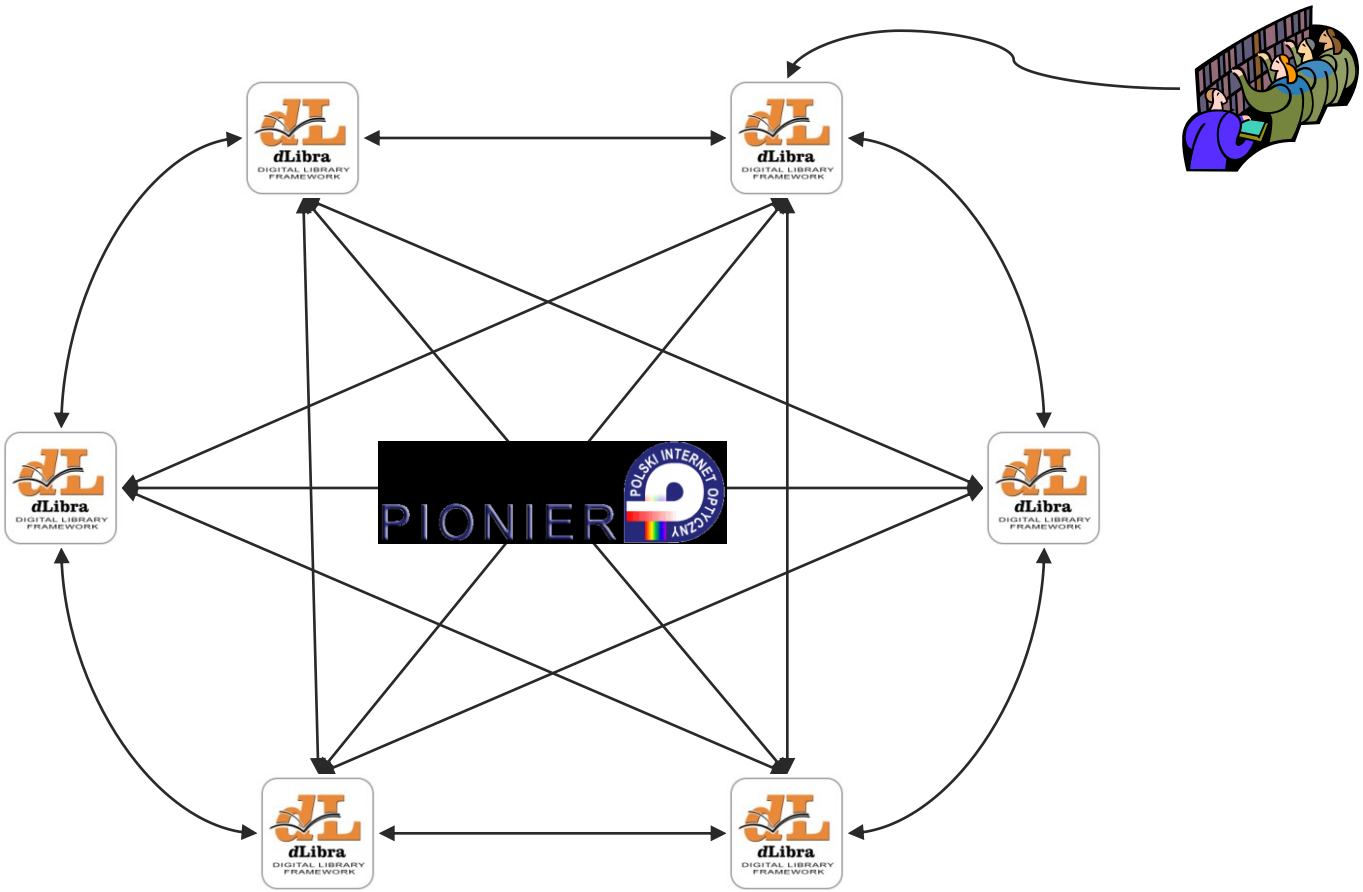


2005-...

Możliwość dostępu poprzez OAI-PMH

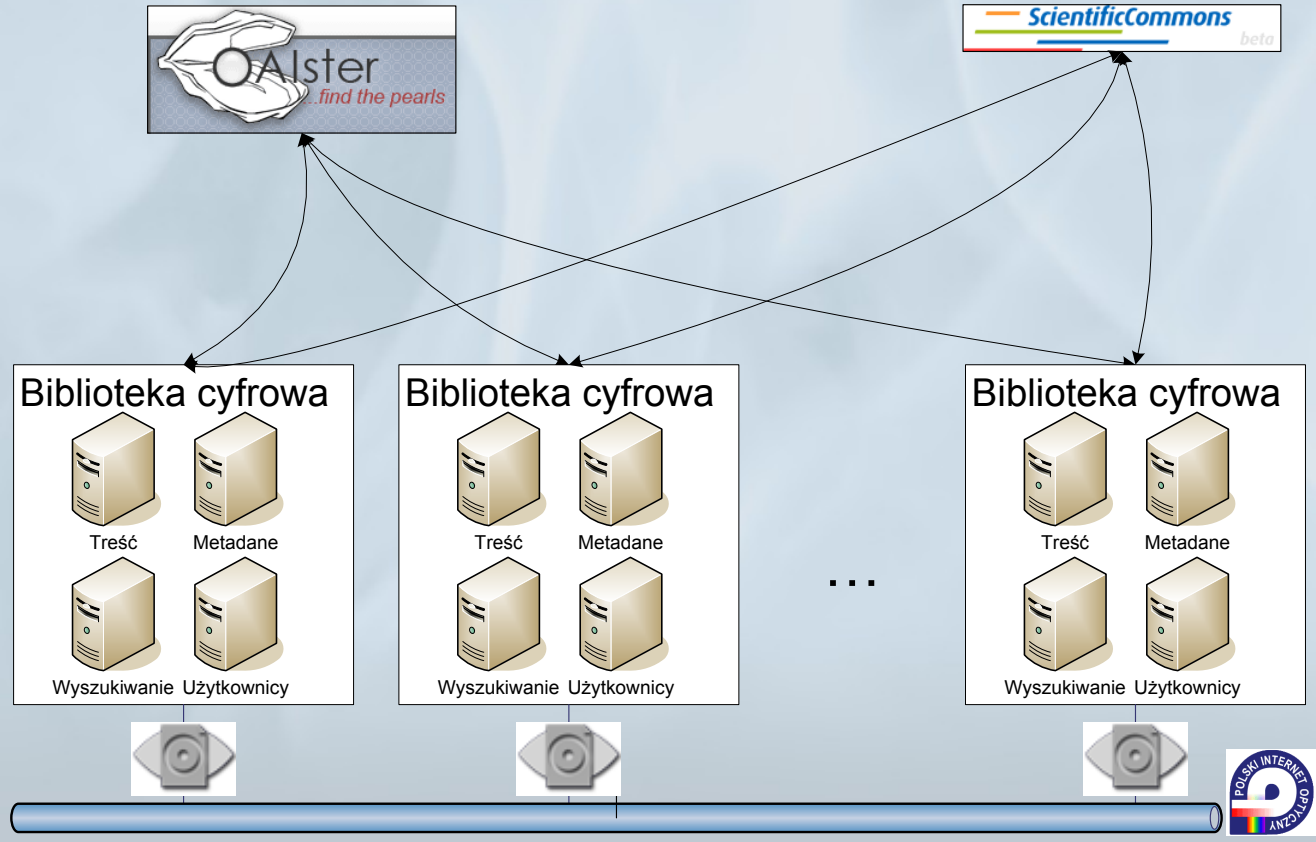


Wyszukiwanie zasobów rozproszonych w systemie dLibra



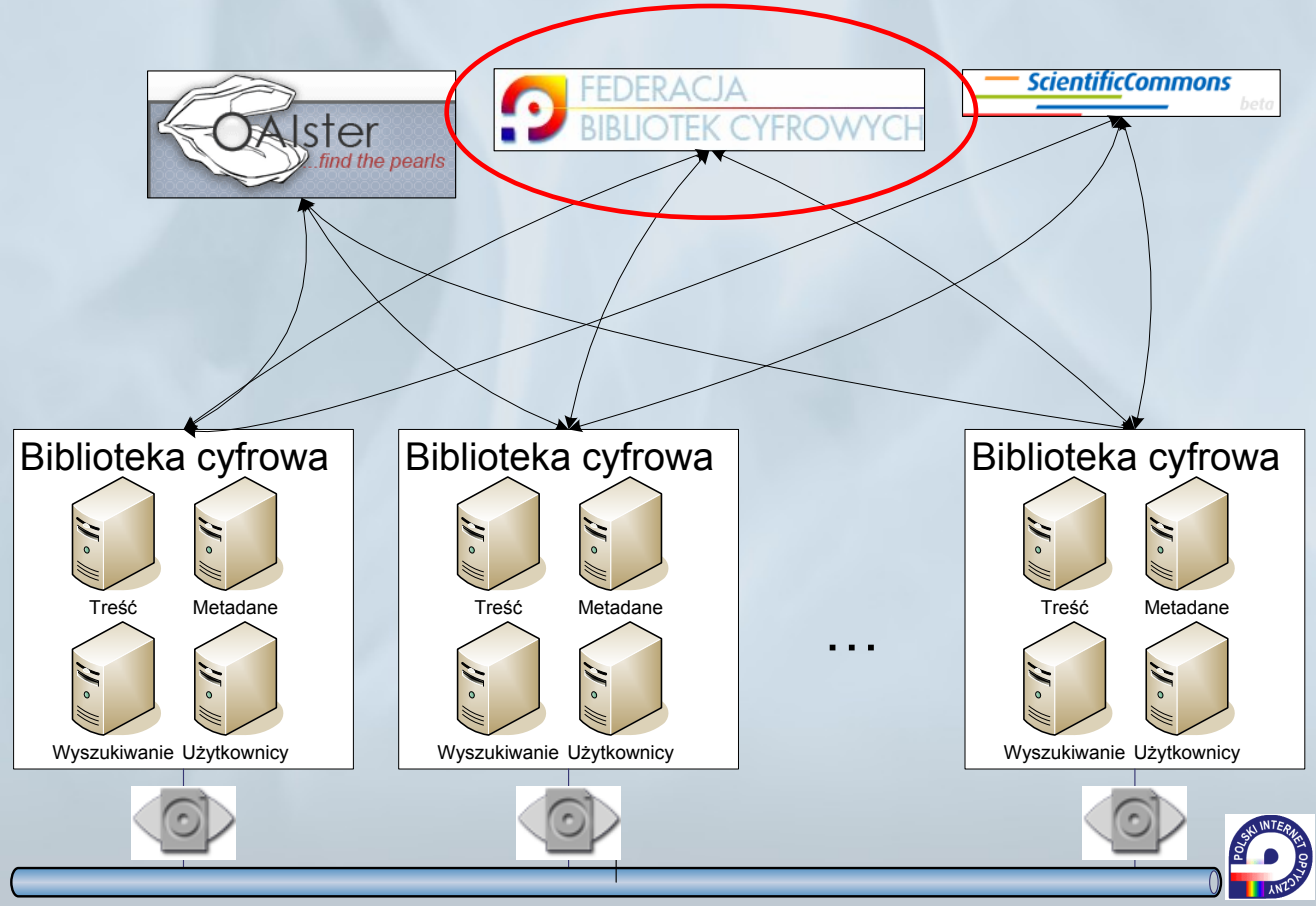
2005-...

Polskie zasoby w światowych wyszukiwarkach OAI-PMH

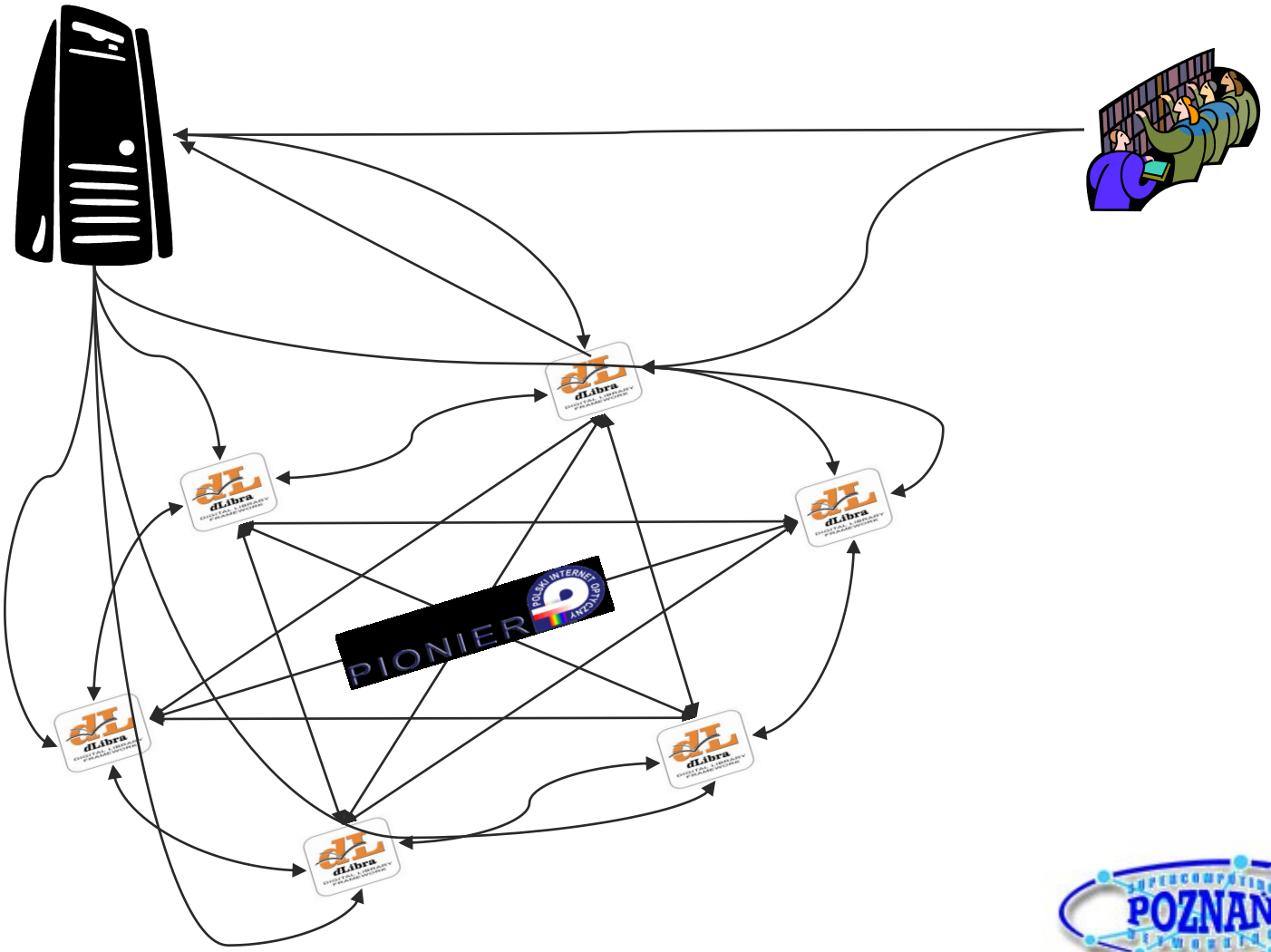


2007

Udostępnienie FBC



Wyszukiwanie zasobów rozproszonych w systemie dLibra



Federacja Bibliotek Cyfrowych

- Cel
  - Ułatwienie wykorzystania zasobów polskich bibliotek cyfrowych i repozytoriów
  - Zwiększenie widoczności zasobów polskich bibliotek cyfrowych i repozytoriów w Internecie
  - Udostępnienie użytkownikom Internetu nowych, zaawansowanych usług sieciowych opartych na zasobach polskich bibliotek cyfrowych i repozytoriów

Federacja Bibliotek Cyfrowych

- Podstawowe założenia
  - Nie ma konieczności przekazywania zasobów na rzecz FBC
  - Nie ma opłat za korzystanie z FBC
  - Podstawą działania są otwarte standardy
    - Możliwość użycia różnych rozwiązań technicznych przez poszczególne biblioteki cyfrowe

Federacja Bibliotek Cyfrowych

- Dostępne funkcje
  - Przeszukiwanie dostępnych publikacji
  - Plany digitalizacji
    - Przeszukiwanie
    - Zestawienie
  - Rozwiązywanie identyfikatora OAI
  - Wykrywanie duplikatów
    - Raport
    - Bezpośrednie wsparcie dla redaktorów bibliotek cyfrowych

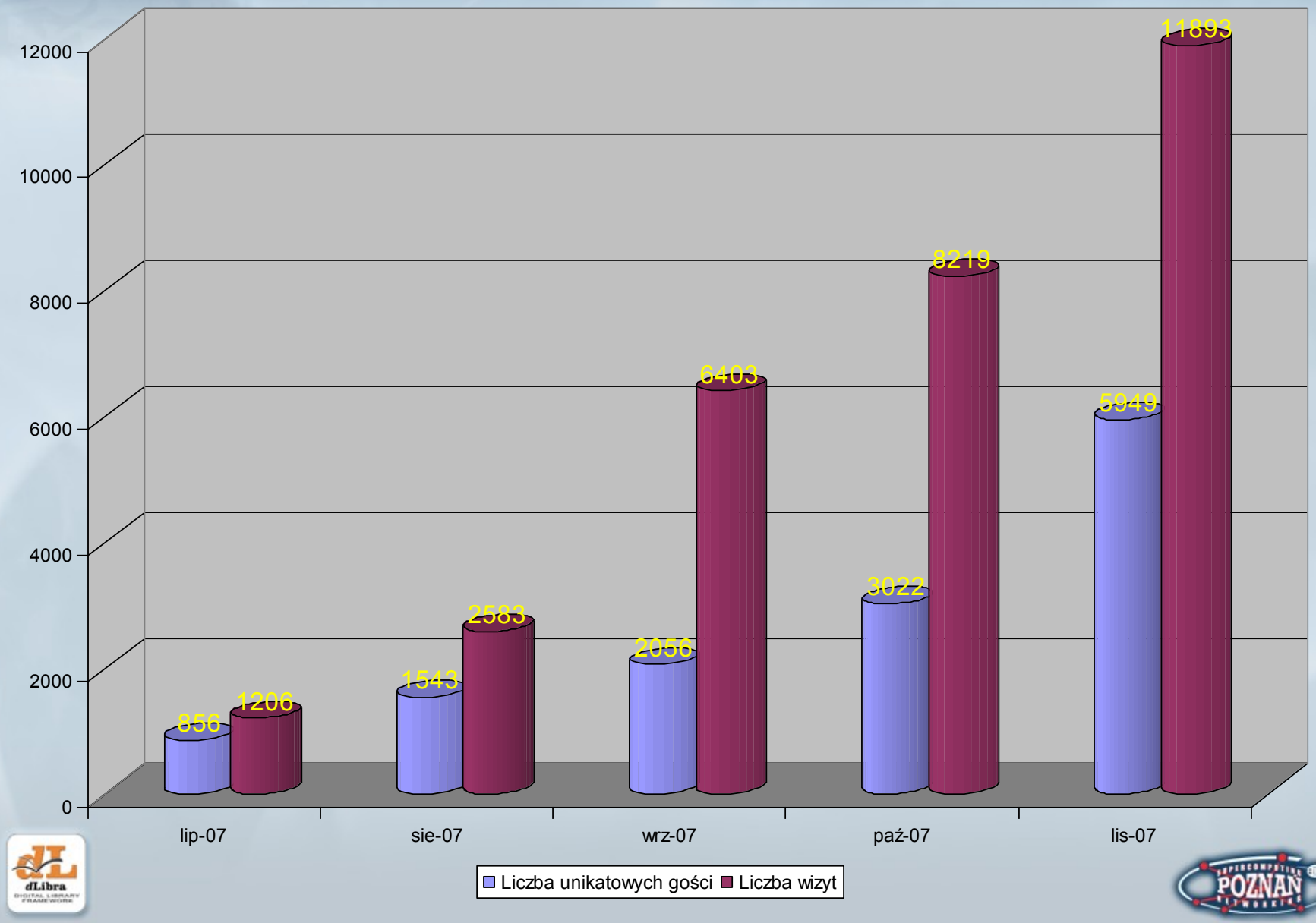
Federacja Bibliotek Cyfrowych

- Wewnętrzna struktura:
  - Dwie usługi wyszukiwania rozproszonego opracowane w ramach projektu dLibra
    1. Wydania
    2. Publikacje planowane
  - Aplikacja WWW dająca dostęp do usług wyszukiwania rozproszonego i realizująca dodatkowe funkcje
  - Baza danych PostgreSQL

Federacja Bibliotek Cyfrowych

- Platforma sprzętowa - serwer Sun Fire 440
  - 4 procesory UltraSPARC IV
    - 1 593 GHz
    - 1MB cache
  - 16 GB RAM
  - 4 dyski twarde 143 GB, 10000 RPM Ultra320 SCSI

Wykres popularności FBC



Możliwe kierunki rozwoju

- Obsługa „zamówień” czytelników
- Sieciowe konto czytelnika
- Implementacja interfejsu SRU
  - FBC w KARO
- Przeszukiwanie treści publikacji
- Nowe funkcje oparte o ujednolicone metadane
  - Wyszukiwanie
  - Przeglądanie
  - Wizualizacje

Plan prezentacji

- Federacja Bibliotek Cyfrowych
- **Przechowywanie skanów wysokiej jakości**
- Duplikaty w sieci bibliotek cyfrowych
- Spójność opisów bibliograficznych

Przechowywanie skanów…

- Obecnie większość bibliotek cyfrowych to dwie biblioteki:
  - Biblioteka postaci prezentacyjnych (DjVu/PDF/JPG/…)
  - Biblioteka postaci archiwalnych (TIFF/…)
    - Płyty CD/DVD
    - Macierze dyskowe
- Wyjątki
  - Biblioteki, które nie zabezpieczają TIFFów…

Przechowywanie skanów…

- Postać prezentacyjna
  - Główna funkcja: udostępnianie
  - Względnie małe pliki
  - Pliki widoczne w Internecie
  - Używane regularnie
  - Wymagany szybki dostęp

Przechowywanie skanów…

Postać archiwalna

- Główna funkcja: długoterminowe przechowywanie
- Duże pliki
- Dostęp tylko dla redaktorów
- Rzadko potrzebne
  - Do stworzenia postaci prezentacyjnej
  - W razie awarii
  - Na „specjalne okazje”
- Możliwe kilkusekundowe opóźnienia w dostępie

Przechowywanie skanów…

- Przechowywanie postaci archiwalnej w bibliotece cyfrowej
  - Zwiększone bezpieczeństwo przechowywania
  - Lepsze zarządzanie postaciami archiwalnymi
  - Prostszy dostęp dla redaktorów i administratorów

Przechowywanie skanów…

- Konieczność opracowania rozszerzeń biblioteki cyfrowej umożliwiających przechowywanie postaci archiwalnej w dedykowanych systemach
  - Macierze dyskowe
  - Archiwizatory
  - Outsourcing
  - Krajowy Magazyn Danych

Krajowy Magazyn Danych

KMD
krajowy magazyn danych

Krajowy Magazyn Danych

System przechowywania danych:

- **wiarygodny** i **bezpieczny**:
  - replikacja geograficzna
  - szyfrowanie
- **trwałość** składowanych danych przy niskich kosztach:
  - wewnętrzne mechanizmy migracji między technologiami składowania:
    - np. dysk -> dysk magneto-optyczny -> taśma LTO4
    - cykliczne, audyty spójności danych i meta-danych oraz stanu mediów
    - automatyczne, przeźroczyste dla użytkownika
  - możliwość wykorzystania różnego typu mediów, np. dysk vs taśma w zależności od potrzeb i możliwości finansowych użytkownika
- **rozproszony**, brak centralizacji
  - wiele fizycznych punktów dostępu
  - wiele replik – możliwość optymalizacji dostępu
- **dostępny** i **wydajny**:
  - krajowy „zasięg” – centra danych w głównych centrach KDM
  - dostęp szerokopasmowy

Krajowy Magazyn Danych

- **Metody dostępu i usługi (1):**
  - **zdalny, wirtualny, logiczny system plików:**
    - logiczny system plików:
      - jedna przestrzeń nazw (z pkt. widzenia użytkownika)
      - oddzielne przestrzenie dla użytkowników
    - fizycznie:
      - dane na macierzach dyskowych, serwerach plików, w systemach HSM
      - replikacja
    - metody dostępu:
      - standardowe protokoły do przesyłu plików: SCP, (s)FTP, HTTP(s)
      - architektura umożliwia stworzenie tzn. „wtyczek dostępowych”
    - można wykorzystać w bibliotekach cyfrowych?

Krajowy Magazyn Danych

- **Metody dostępu i usługi (2):**
  - **usługa kopii zapasowej, archiwizacji oraz odtwarzania**
    - kopie pełne, przyrostowe i różnicowe :
      - optymalizacja ilości przesyłanych danych
    - wersjonowanie
      - możliwość powrotu do dawnych wersji plikó
    - automatyzacja procesu wykonywania kopii danych i archiwizacji:
      - na podstawie polityk zdefiniowanych przez użytkownika
      - zwolnienie użytkownika z „myślenia” o kopiach zapasowych
    - możliwa automatyczna replikacja danych:
      - w obrębie centrum danych (rozłączne media lub grupy/typy mediów)
      - replikacja geograficzna
    - interesujące z pkt. widzenia bibliotek cyfrowych
      - jak biblioteki cyfrowe zabezpieczone są przed:
        - „wandalizmem”,
        - „pomyłką” administratora lub użytkownika,
        - awariami sprzętu przechowującego dane?

KMD – Infrastruktura

Infrastruktura docelowa:

- 4 główne węzły przechowywania
- 4 węzły aplikacyjne
- osadzone w sieci PIONIER

Węzły przechowywania:

- realizują udostępnianie i zarządzanie obiektami danych
- zarządzają odwzorowaniem logicznej struktury plików na fizyczne systemy przechowywania: macierze dyskowe, systemy HSM
- kontrolują elementy infrastruktury: systemy przechowywania, serwery dostępowe i aplikacyjne, sieć

Węzły dostępowe:

- świadczą usługi dostępowe do KMD
- mogą realizować dodatkowe usługi, np. zarządzanie zawartością, wyszukiwanie na podstawie meta-danych itd. (otwarte pole do działania)

KMD – Uczestnicy projektu

- Politechnika Białostocka, Centrum Komputerowych Sieci Rozległych
- Akademickie Centrum Komputerowe
- Centrum Komputerowe Politechniki Łódzkiej
- Uniwersytet Marii Curie-Skłodowskiej w Lublinie
- Poznańskie Centrum Superkomputerowo-Sieciowe
- Politechnika Częstochowska
- Politechnika Gdańska
- Politechnika Wrocławska

Plan prezentacji

- Federacja Bibliotek Cyfrowych
- Przechowywanie skanów wysokiej jakości
- **Duplikaty w sieci bibliotek cyfrowych**
- Spójność opisów bibliograficznych

Duplikaty

- FBC pozwoliło na budowanie nowych funkcji na bazie zasobów bibliotek cyfrowych dostępnych przez OAI-PMH
- Jedna z pierwszych nowych usług: mechanizm automatycznego wykrywania **potencjalnych** duplikatów

Mechanizm automatycznego wykrywania duplikatów

- Bazuje na indeksach wyszukiwawczych FBC
- Wykonuje automatyczną analizę porównawczą wszystkich opisów dostępnych w FBC
- Obecnie podstawą do analizy jest
  - Tytuł
  - Autor
  - Data wydania

Mechanizm automatycznego wykrywania duplikatów

- Wykrywanie duplikatów pomimo, że:
  - Nie wszystkie analizowane atrybuty są wypełnione
  - Wartości w analizowanych atrybutach różnią się od siebie nieznacznie

Mechanizm automatycznego wykrywania duplikatów

- System uczy się na błędach
  - Jest udoskonalany na podstawie analizy opisów faktycznych duplikatów, których sam nie wykrył
  - Wkrótce na FBC formularz do zgłaszania duplikatów

Czy duplikaty to coś złego?

- Liczba publikacji: 100 000
- Liczba faktycznych duplikatów: 50?
- Na 1 potencjalny duplikat na 2 000 publikacji
- Ale – jak zdefiniować publikację (pod względem liczby stron)?

  Theatrum Chemicum
  vs
  Telegram Kościuszkowski
- Duplikat = kopia bezpieczeństwa ☺

Dlaczego powstają duplikaty? Jak zmniejszyć ich liczbę do minimum?

- Przyczyna: brak wymiany informacji między bibliotekami
  - Rozwiązanie: mechanizm publikacji planowanych
    - Informacje o planach digitalizacji dla innych bibliotek
    - Automatyczne sprawdzenie w FBC, czy jest dostępna podobna publikacja już na etapie planowania digitalizacji
      - Uwzględnia obiekty planowane i zdigitalizowane ze wszystkich bibliotek widocznych w FBC

Dlaczego powstają duplikaty? Jak zmniejszyć ich liczbę do minimum?

- Przyczyna: spójność i kompletność posiadanych kolekcji
  - Rozwiązanie: publikacje/kolekcje wirtualne
    - Umieszczanie w bibliotece cyfrowej obiektów cyfrowych pochodzących z:
      - innych bibliotek cyfrowych (identyfikator OAI)
      - innych systemów sieciowych (adres URL)
    - Co z prawami?

Dlaczego powstają duplikaty? Jak zmniejszyć ich liczbę do minimum?

- Przyczyna: Zła jakość obiektów już zdigitalizowanych
  - Rozwiązanie: standardy digitalizacji?

Czy duplikaty to coś złego?

Jeżeli tak, to dlaczego powstają i jak zmniejszyć ich liczbę do minimum?

Dyskusja

Plan prezentacji

- Federacja Bibliotek Cyfrowych
- Przechowywanie skanów wysokiej jakości
- Duplikaty w sieci bibliotek cyfrowych
- **Spójność opisów bibliograficznych**

Spójność opisów bibliograficznych

- FBC pozwoliło na budowanie nowych funkcji na bazie zasobów bibliotek cyfrowych dostępnych przez OAI-PMH
- Pierwszym poważnym problemem jaki się pojawia są różnice w opisach pochodzących z różnych bibliotek cyfrowych

Spójność opisów bibliograficznych

- Przykład 1: typ

Wartość atrybutu	Liczba wystąpień	Udział %
gazeta	26782	27%
czasopismo	16281	16%
Czasopismo	10913	11%
Gazeta	7877	8%
gazety	5415	5%
książka	4960	5%
fotografia	2149	2%
grafika	1860	2%
artykuł z czasopisma	1333	1%
pocztówka	1094	1%
Czasopisma	1040	1%
starodruk	1033	1%
czasopisma	927	1%
Książka	858	1%
mapa	749	1%

Spójność opisów bibliograficznych

- Przykład 2: język

Wartość atrybutu	Liczba wystąpień	Udział %
pol	82181	82%
ger	10813	11%
lat	2528	3%
und	941	1%
fre	889	1%
eng	847	1%
polski	151	0%
ita	124	0%
mul	114	0%
pl	82	0%
cze	80	0%
pol/ger	80	0%
lat/pol	79	0%
rus	64	0%
pol ; ger	61	0%

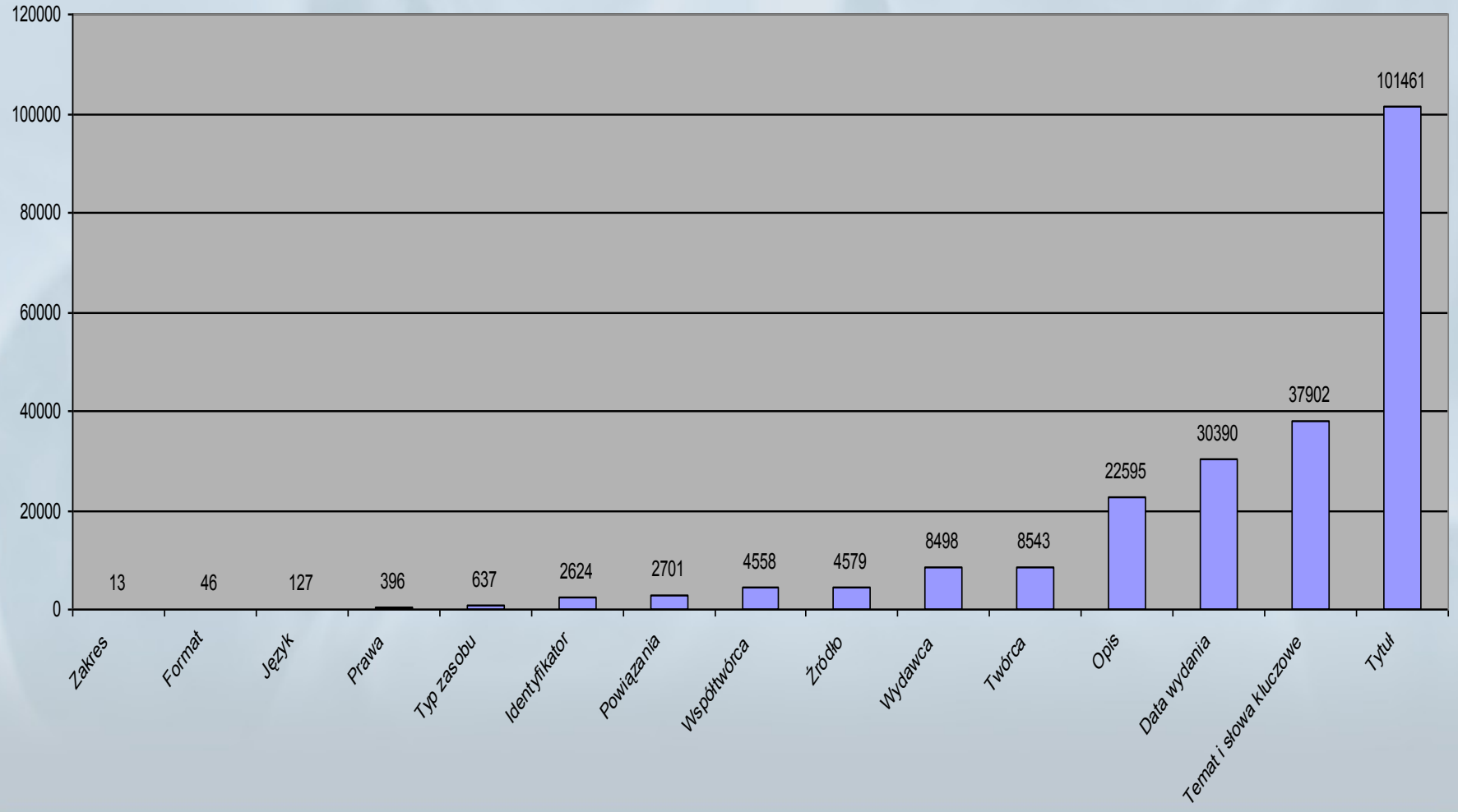
Spójność opisów bibliograficznych

- Konsekwencje niespójności opisów
  - Utrudnione wyszukiwanie
    - Np.: zapytania pisane pod kątem specyficznego sposobu opisywania zasobów
  - Uniemożliwione automatyczne przetwarzanie opisów w celu realizacji nowych funkcji
    - Np.: wyszukiwanie po zakresie dat

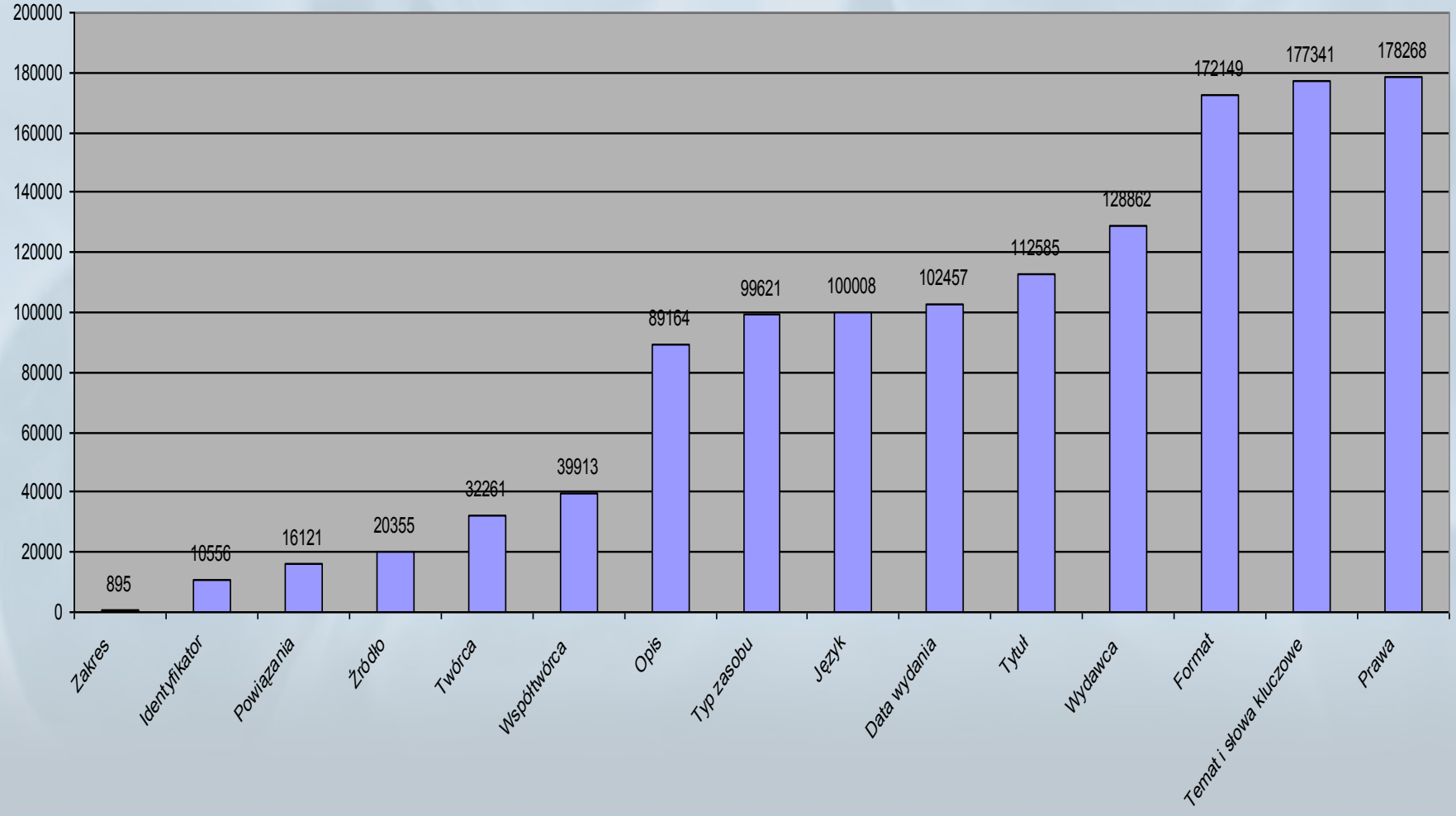
Spójność opisów bibliograficznych

- Kluczowe atrybuty
  - Te, które posiadają względnie małą liczbę różnych wartości

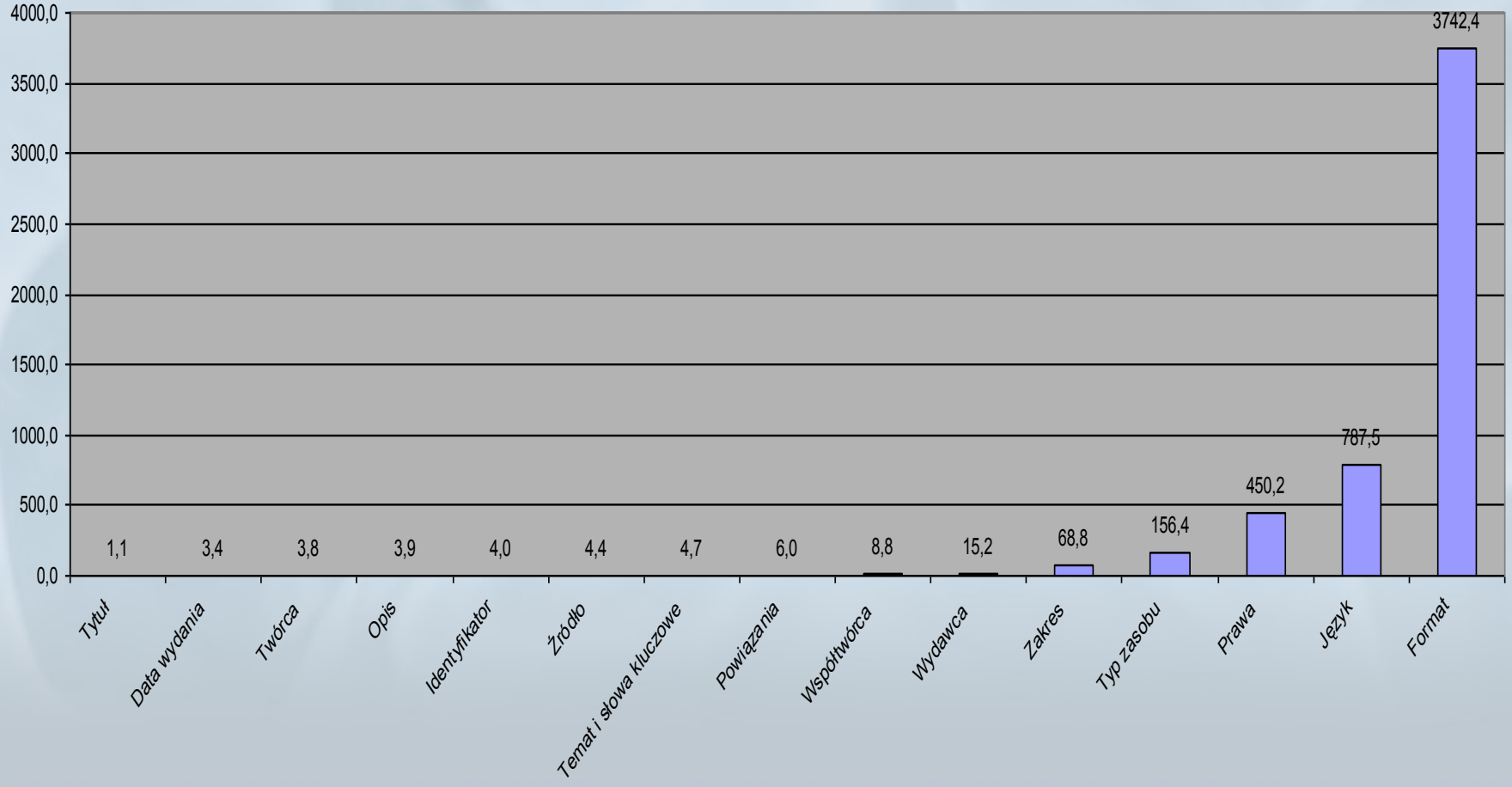
Liczba unikalnych wartości poszczególnych atrybutów



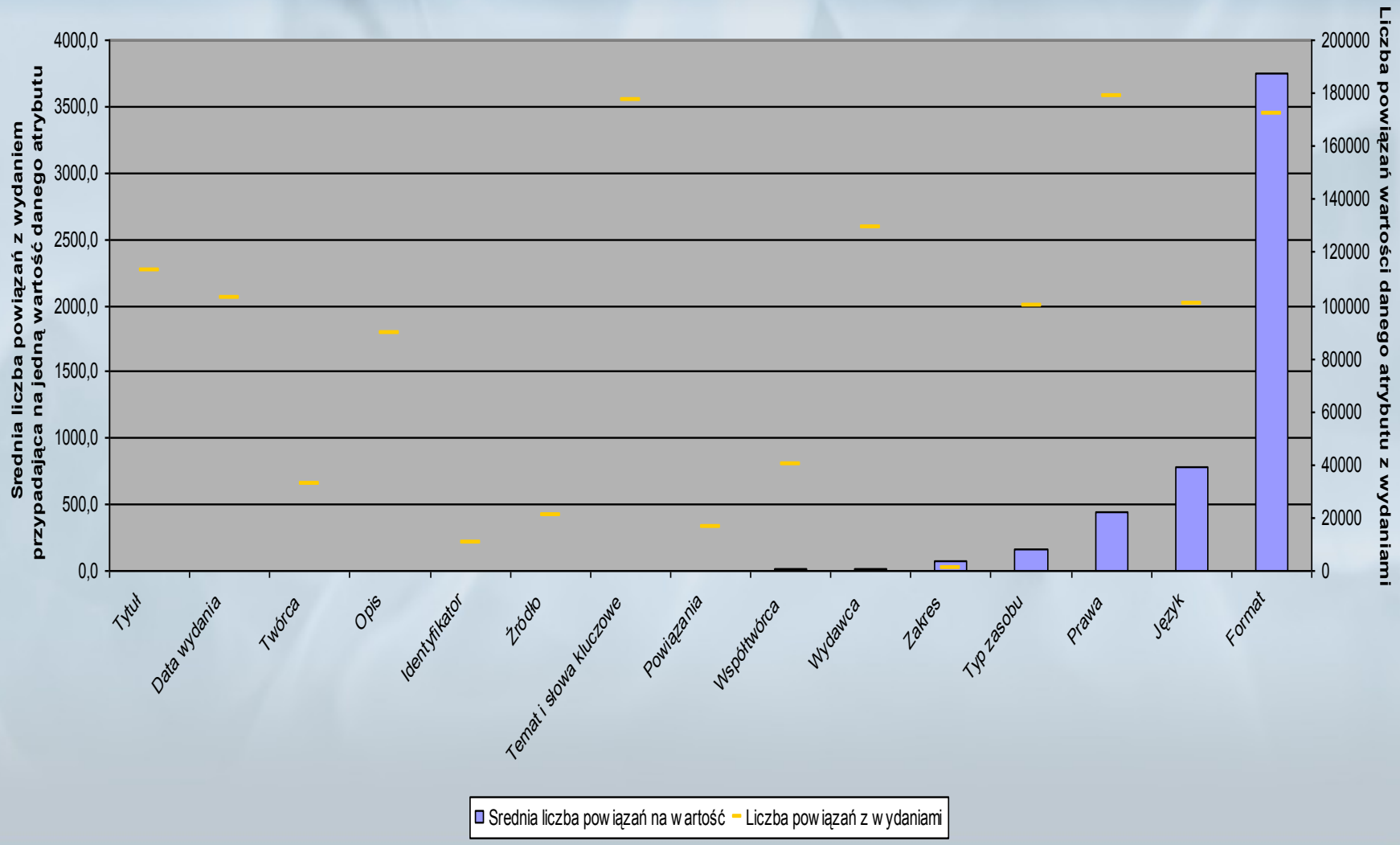
Liczba powiązań wartości danego atrybutu z wydaniami



Średnia liczba powiązań z wydaniem przypadająca na jedną wartość danego atrybutu



Wykorzystanie wartości atrybutów



Spójność opisów bibliograficznych

- Kluczowe atrybuty
  - Te, które posiadają względnie małą liczbę różnych wartości: **typ, format, język, prawa**

Spójność opisów bibliograficznych

- Kluczowe atrybuty
  - Te, które posiadają względnie małą liczbę różnych wartości: **typ, format, język, prawa**
  - Te, które mogłyby posiadać pewien podstawowy „słownik” dowolnie rozszerzany przez poszczególne biblioteki: **temat i słowa kluczowe, …**
  - Te, w których sposób zapisu ma znaczenie: **data**

Spójność opisów bibliograficznych

- Jak zapewnić/poprawić spójność opisów bibliograficznych?
  - Tak jak Google ;-)
  - Ustalając zasady opisu w ramach sieci bibliotek cyfrowych
  - Korzystając z pomocy „zewnętrznej” instytucji

Jak zapewnić/poprawić spójność opisów bibliograficznych?

- Tak jak Google ;-)
  - Podpowiadanie wartości atrybutów w Aplikacji Redaktora oparte na statystykach użycia poszczególnych wartości wg FBC
    - Dominacja WBC?
    - Co z datami?

Jak zapewnić/poprawić spójność opisów bibliograficznych?

- Ustalając zasady opisu w ramach sieci bibliotek cyfrowych
  - Wykorzystując opracowania takie jak „e-Poradnik redaktora zasobów cyfrowych” opracowany na Uniwersytecie Wrocławskim

    http://fbc.pionier.net.pl/id/oai:www.bibliotekacyfrowa.pl:17703

Jak zapewnić/poprawić spójność opisów bibliograficznych?

- Korzystając z pomocy „zewnętrznych” baz
  - NUKAT (p. Agnieszka Kasprzyk)

Jak zapewnić/poprawić spójność opisów bibliograficznych?

Dyskusja

Dziękuję za uwagę!