POZNAŃSKIE CENTRUM SUPERKOMPUTEROWO SIECIOWE

POZNAŃSKIE
CENTRUM
SUPERKOMPUTEROWO
SIECIOWE

Wykorzystanie standardów w masowej digitalizacji i długoterminowym przechowywaniu danych źródłowych

Adam Dudczak, Tomasz Parkoła
{maneo, tparkola}@man.poznan.pl

Wprowadzenie

- Realizacja masowej digitalizacji w Polsce
- Jak w rzeczywistości wygląda wykorzystanie standardów i narzędzi w instytucjach tworzących polskie biblioteki cyfrowe?
- Dwie ankiety realizowane w ramach projektu SYNAT:
  - **Wrzesień 2010** – ankieta dotycząca wykorzystania narzędzi OCR w polskich bibliotekach cyfrowych
  - **Styczeń 2012** – ankieta dotycząca przechowywania kopii MASTER w polskich bibliotekach cyfrowych

Wprowadzenie (2)

- Do przeprowadzenia ankiet wykorzystano kwestionariusz dostępny online
- Prośbę o wypełnienie wysłano do wszystkich instytucji tworzących polskie biblioteki cyfrowe
- Do promocji ankiety wykorzystano również strony WWW i fora internetowe wykorzystywane przez polskich bibliotekarzy cyfrowych

Ankieta nt. wykorzystania narzędzi OCR

- Ankieta składała się z 29 pytań i została podzielona na trzy części:
  - ogólną (głównie informacje nt. respondenta),
  - dotyczącą zasobów
  - dotyczącą wykorzystania aplikacji OCR (najobszerniejszą).
- W części dotyczącej zasobów pytaliśmy:
  - o liczbę obiektów opublikowanych w bibliotece cyfrowej,
  - stopień wykorzystania technologii OCR przy tworzeniu cyfrowej postaci tych obiektów oraz ich charakteru.
  - ta część ankiety zawierała również pytanie o stosowane w bibliotece zalecenia odnośnie standardów digitalizacji dla różnych typów materiałów.

Ankieta nt. wykorzystania narzędzi OCR

- Celem ostatniej części ankiety było ustalenie które cechy i funkcje oprogramowania OCR są istotne z perspektywy osób praktycznie wykorzystujących je na co dzień.
- Pytaliśmy m. in. o:
  - wykorzystywanych narzędzi OCR,
  - podstawowych funkcji np. wykorzystywanych formatów wejścia/wyjścia,
  - stosowanych metod ewaluacji jakości warstwy tekstowej

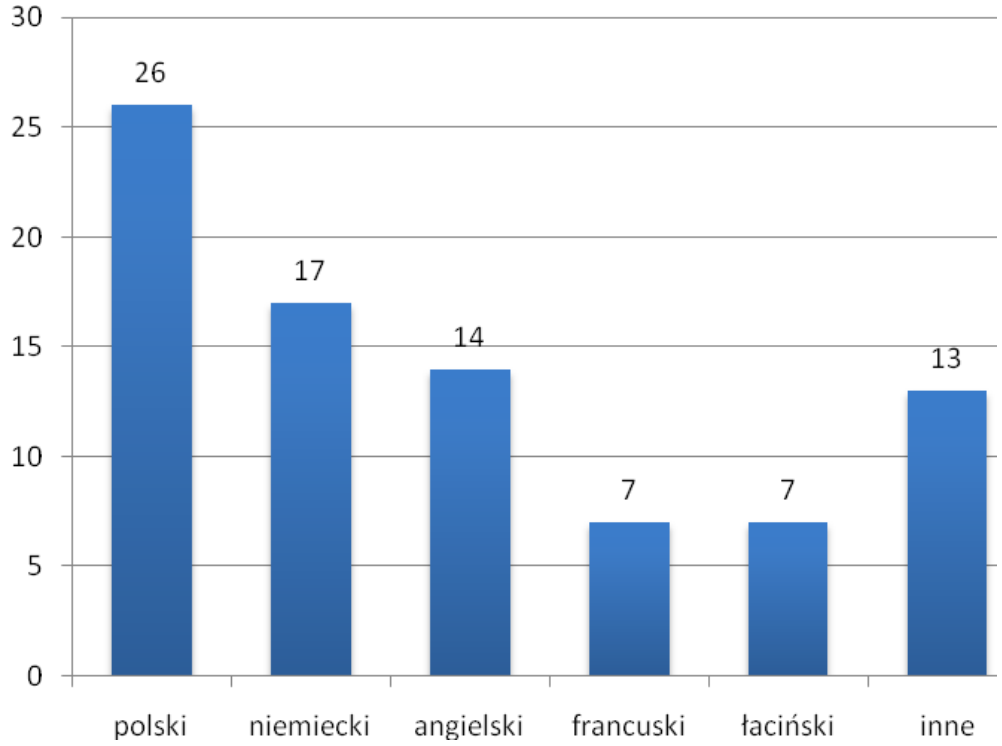
Wyniki ankiety

- W czasie (wrzesień 2010) przeprowadzania ankiety w Federacji Bibliotek Cyfrowych dostępnych było 330 000 obiektów.
- W ankiecie wzięło udział 26 instytucji tworzących takie biblioteki cyfrowej jak:
  - Wielkopolska Biblioteka Cyfrowa, Małopolska Biblioteka Cyfrowa, Cyfrowa Biblioteka Narodowa „Polona” czy też Biblioteka Cyfrowa Uniwersytetu Wrocławskiego
- Biblioteki które odpowiedziały tworzyły sumarycznie 70% zasobu dostępnego w FBC

Wykorzystanie oprogramowania OCR

- Na podstawie liczby obiektów zadeklarowanej przez respondentów obliczyliśmy procent zasobu, który został poddany przetwarzaniu OCR – 40%
- Przy założeniu że ponad 95% zasobów było opisanych jako tekstowe daje to liczbę około 130 000 obiektów
- Udostępniane są głównie czasopisma, wydawnictwa seryjne i książki do których wygasły prawa autorskie

Język przetwarzanego tekstu



Język przetwarzanego tekstu

- Blisko 70% wszystkich respondentów pracuje z zasobami, gdzie w treści dokumentu znajduje się tekst w więcej niż jednym języku.
- Wyniki wskazują również, że istotne jest wsparcie dla niestandardowych czcionek (tzw. „gotyk” został wskazany przez 8 instytucji) oraz alfabetów (wsparcie dla cyrylicy wykorzystuje również 8 instytucji)

Standardy digitalizacji na potrzeby OCR

- Rozdzielczość skanowania pomiędzy 300 a 600 DPI
- Głębia kolorów to w większości 1BPP
- Formaty wejściowy dla OCR to głównie TIFF (25 instytucji), DjVu (17 instytucji) poza tym 5 instytucji wykorzystuje format JPEG.

Wykorzystywane oprogramowania

- Document Express (16 instytucji)
- ABBY FineReader (15 instytucji)
- OmniPage (1 instytucja)
- W niektórych przypadkach biblioteki wykorzystywały zarówno FineReadera jak i Document Expressa

Zaawansowane funkcje oprogramowania OCR

- Wykorzystanie bogatej reprezentacji wyników OCR (hOCR, FineReader XML) – 5 instytucji
- Uczenie OCR dla zniszczonych, specyficznych dokumentów stosowały 3 instytucje
  - 9 ankietowanych nie było w stanie określić czy wykorzystywane przez nich oprogramowanie posiada zdolność uczenia czy nie.
- Wykorzystanie silnika OCR w środowisku serwerowym zadeklarowało 5 respondentów

Kontrola jakości wyników OCR

- 80% instytucji przegląda wybrane losowo fragmenty tekstu rozpoznanego przez OCR
- Nikt jednak nie przeprowadza korekty
- W najgorszym razie wyniki OCR nie są publikowane

Podsumowanie

- Wyniki OCR traktowane przede wszystkim jako pomoc wyszukiwawcza
  - Nie usystematyzowana kontrola jakości
  - Brak korekty wyników OCR
- Biblioteki nie mają zasobów i narzędzi, które pozwoliłyby włączyć korektę wyników OCR do procesu digitalizacji

Ankieta nt. kopii MASTER

- Ankieta składała się z **11 pytań** (wyłączając pytania identyfikujące respondenta), została podzielona na **dwie części** dotyczące:
  - przechowywania kopii MASTER (plików „matek”) w instytucji
  - formatów plików wykorzystywanych do składowania tego typu obiektów
- W ramach pierwszej części padły pytania dotyczące:
  - sposobu przechowywania kopii MASTER
  - wielkości posiadanego archiwum tego typu obiektów
  - praktyk związanych z zapewnianiem bezpieczeństwa archiwów plików
  - udostępniania kopii wzorcowych online

Ankieta nt. kopii MASTER

- Druga część to pytania dotyczące formatów plików wykorzystywanych do przechowywania materiałów:
  - ikonograficznych,
  - obiektów zawierających przeszukiwalny tekst
  - oraz plików audio-wideo
- W przypadku gdy instytucja nie przechowuje kopii MASTER prosiliśmy o krótkie uzasadnienie dla takiej praktyki

Wyniki ankiety

- W chwili przeprowadzania ankiety w portalu Federacji Bibliotek Cyfrowych dostępnych było **855 083** obiektów
- W ankiecie wzięło udział w sumie **26 instytucji**, **tworzących 24 biblioteki cyfrowe** między innymi
  - Wielkopolską Bibliotekę Cyfrową,
  - Jagiellońską Bibliotekę Cyfrową,
  - Małopolską Bibliotekę Cyfrową
  - Kujawsko-Pomorską Bibliotekę Cyfrową
  - …
- Biblioteki cyfrowe tworzone przez respondentów udostępniały w sumie **76,9%** wszystkich zasobów FBC

Wyniki ankiety

- W imieniu instytucji ankietę wypełniały osoby pełniące zarówno stanowiska kierownicze, jak i te bezpośrednio odpowiedzialne za realizacje procesu digitalizacji (bibliotekarze, administratorzy IT)
- Wśród instytucji, które brały udział w ankiecie znalazły się zarówno **biblioteki publiczne (8)**, **biblioteki akademickie (17)** oraz jedno repozytorium cyfrowe prowadzone przez organizacje pozarządową.
- Jedna instytucja odpowiedziała, że nie przechowuje kopii MASTER ponieważ w swoim repozytorium udostępnia tylko materiały natywnie cyfrowe (ang. born digital)

Średnia wielkość archiwum plików MASTER

	Biblioteki publiczne	Biblioteki akademickie
Średnia wartość [TB]	26,037	10,593
Mediana [TB]	30	5,5

Tabela 1. Wielkość archiwum kopii MASTER w polskich bibliotekach akademickich i publicznych.

- Ogólna średnia liczona z 24 wartości to: 14 656 GB, a mediana 5 660 GB
- Największe archiwum to 70TB, a najmniejsze 7.25 GB.

Średnia liczba plików w archiwum

	Biblioteki publiczne	Biblioteki akademickie
Średnia liczba plików	413 415	350 755
Mediana liczby plików	500 000	259 500

Tabela 2. Liczba plików przechowywanych w archiwach kopii MASTER w polskich bibliotekach akademickich i publicznych

- Ogólna średnia liczona z 16 wartości to: 348 469 pliki, (mediana 259 500)
- Największa wartość to 832 621 plików, a najmniejsza 881

Średnia liczba obiektów

	Biblioteki publiczne	Biblioteki akademickie
Średnia liczba obiektów	20 807	8 419
Mediana liczby obiektów	17 468	2 700

Tabela 3. Liczba obiektów przechowywanych w archiwach kopii MASTER w polskich bibliotekach akademickich i publicznych.

- Ogólna średnia liczona z 24 wartości to: 14 656 GB
  - mediana 5 660 GB
- Największe archiwum to 70TB, a najmniejsze 7.25 GB.
- Liczba obiektów zadeklarowana w ankiecie obejmowała zarówno obiekty udostępniane w bibliotece cyfrowej, ale również te, które nie zostały opublikowane.

Udostępnianie kopii MASTER

- Dokładnie połowa ankietowanych przechowuje pliki wzorcowe w więcej niż jednej kopii
- W przypadku gdy istnieje więcej niż jedna kopia tylko część (około 40%) instytucji przechowuje kopie w innych budynkach
- Większość (62%) ankietowanych nie zgadza się na publiczne (bez żadnych ograniczeń) udostępnienie kopii MASTER w sieci w sytuacji gdy istniałaby taka możliwość techniczna.
  - Nie mniej spora część ankietowanych dopuszcza udostępnianie materiału uwierzytelnionym użytkownikom

Przechowywanie dokumentów tekstowych

- Pytaliśmy o dwie rzeczy:
  - przechowywanie skanów wzorcowych dokumentów tekstowych,
  - przechowywanie wyników OCR jako części kopii matki
- W przypadku przechowywania dokumentów tekstowych 10 respondentów odpowiedziało, że nie przechowują w ogóle tekstu (np. wyników OCR) w archiwum kopii MASTER
  - Dokument składowany jest w formie plików graficznych
- Najczęściej wykorzystywanym formatem dla tego typu zasobów jest **PDF** (68% instytucji)

Przechowywanie dokumentów tekstowych (2)

- Format DjVu był w 3 przypadkach **jedyną formą przechowywania** dokumentów zawierających cyfrowy tekst.
- W pozostałych, oprócz DjVu, stosowane są również inne rozwiązania.
  - W 10 przypadkach obok plików DjVu przechowywany jest również PDF,
  - a w 6 plik tekstowy (format TXT lub XML).

Przechowywanie plików graficznych

- Formaty plików wykorzystywane do przechowywania kopii wzorcowych plików graficznych
  - JPG - 7 instytucji
  - PNG – 0
  - TIFF (z kompresją bezstratną) - 9
  - TIFF (bez kompresji) – 18
  - JPG2000 (z kompresją bezstratną) – 1
  - Inne: 2 („nie przechowujemy skanów w postaci graficznej”, „PDF”)
- Najczęściej stosowany jest TIFF (w różnych odmianach)
- W 3 przypadkach JPG to jedyna forma przechowywania!

Przechowywanie plików audio-wideo

- Formaty plików wykorzystywane do przechowywania kopii wzorcowych kolekcji audio-wizualnych (odpowiedziało 9 instytucji)
  - AVI - 2 instytucje
  - WAV - 0
  - MP3/MP4 - 3
  - FLV - 1
  - FLAC - 1
  - Inne - 3 („nie składujemy audio”, „nie dotyczy”)

Podsumowanie

- Wykorzystanie formatów stratnych do przechowywania do kopii wzorcowych
- Wykorzystanie formatów złożonych np. PDF zamiast prostych TIFF + tekst
- Lekceważące podejście do przechowywania wyników przetwarzania OCR
  - Brak późniejszej możliwości wprowadzania korekt
- Niewielki udział kolekcji audio-wizualnych
- Znaczący zasób znajduje się offline
  - Prawa autorskie
  - Brak możliwości technicznych

Projekt Succeed

Informacje ogólne

- Projekt realizowany w ramach 7PR
- Ramy czasowe projektu: 2013-2014
- Zakres projektu obejmuje aspekty digitalizacji dokumentów tekstowych
- Projekt wspiera działania Centrum Kompetencji IMPACT
- Partnerzy projektu: UA, KB, INL, Fraunhofer IAIS, PCSS, USAL, FBVMC, BNF, BL

Projekt Succeed

Informacje ogólne

- Główne zadania:
  - Stymulowanie procesu wdrażania innowacyjnych narzędzi w europejskich bibliotekach cyfrowych
  - Identyfikacja i rekomendacja standardów, formatów i licencji odnośnie narzędzi i zasobów przydatnych w digitalizacji
  - Organizacja konferencji, szkoleń, konkursów podnoszących świadomość możliwości zaawansowania procesu digitalizacji
  - Opracowanie założeń odnośnie współpracy europejskich centrów kompetencji w kontekście programu Horizon 2020

Standardy i formaty w Succeed

Założenia

- Obecnie istnieje szereg praktyk i zaleceń dotyczących procesu digitalizacji
  - Digital Curation Center, FADGI, Digital Preservation Coalition, JISC digital media guides, Biblioteka Narodowa Australii, Digital New Zeland, …
- Istniejące zalecenia w niewielkim stopniu koncentrują się na zaawansowanych aspektach procesu digitalizacji
  - OCR, zasoby lingwistyczne, interoperacyjność
- Wsparcie dla realizacji założeń zapisanych w rekomendacji UE (27.11.2011) oraz Digital Agenda for Europe

Standardy i formaty w Succeed

Wyznaczone zadania

- Stymulowanie działań digitalizacyjnych poprzez
  - Identyfikację formatów i standardów istotnych w procesie digitalizacji
  - Rekomendowanie najlepszych praktyk w kontekście dokumentów tekstowych

Standardy i formaty w Succeed

Wyznaczone zadania

- Ułatwienie wdrażania innowacyjnych rozwiązań w ramach procesu digitalizacji poprzez
  - Identyfikację narzędzi i zasobów związanych z procesem digitalizacji
  - Identyfikację istniejących narzędzi prawnych ułatwiających udostępnianie utworów
  - Rekomendowanie jednego zestawu licencji na podstawie których zalecane jest udostępnianie utworów (narzędzi i zasobów)
- Ułatwienie przygotowania tzw. one-stop shop w ramach którego dostępne są wszystkie narzędzia Succeed

Standardy i formaty w Succeed

Realizacja zadań

- Zakres:
  - Metadane (opisowe, techniczne, …)
  - Zasoby (treść, rezultaty OCR, zasoby lingwistyczne)
  - Pakiety instalacyjne do narzędzi/zasobów
- Kontekst:
  - Długoterminowe przechowywanie
  - Dostęp do zasobów online

Standardy i formaty w Succeed

Realizacja zadań

- Ramy czasowe
  - Identyfikacja i rekomendacja formatów i standardów do końca 2013
  - Rekomendacja zestawu licencji do końca 2014
- Podejście do realizacji zadania
  - Identyfikacja oraz analiza istniejących zaleceń i rekomendacji
  - Opracowanie ankiety i jej przeprowadzenie w europejskich instytucjach dziedzictwa kulturowego

Standardy i formaty w Succeed

Realizacja zadań

- Istniejące formaty i standardy – przykłady do analizy w ramach projektu
  - Reprezentacja tekstu
    - ALTO, TEI, hOCR, PAGE XML
  - Reprezentacja obrazu
    - TIFF, JPEG2000
  - Metadane
    - Dostępne online: Dublin Core, EDM
    - Długoterminowe przechowywanie: METS, PREMIS
  - Reprezentacja zasobów lingwistycznych
    - TEI, LMF

POZNAŃ SUPERCOMPUTING AND NETWORKING CENTER

Dziękuje za uwagę

kontakt:
tparkola@man.poznan.pl
maneo@man.poznan.pl

Poznań Supercomputing and Networking Center
affiliated to the Institute of Bioorganic Chemistry of the Polish Academy of Sciences,
ul. Noskowskiego 12/14, 61-704 Poznań, POLAND,
Office: phone center: (+48 61) 858-20-00,
fax: (+48 61) 852-59-54,
e-mail: office@man.poznan.pl, http://www.man.poznan.pl