Arkadiusz Pulikowski
Instytut Bibliotekoznawstwa
i Informacji Naukowej
Uniwersytet Śląski
Katowice

Widoczność publikacji naukowych w Internecie

IV Ogólnopolska Konferencja Naukowa
Zarządzanie informacją w nauce
28-29 listopada 2012, Katowice

Plan wystąpienia

- publikacje naukowe w Internecie
  - rodzaje
  - miejsca dostępu
  - korzyści
  - widoczność
- cel badań
- przebieg badań
- rezultaty
- wnioski

Publikacje naukowe w Internecie

- po latach dominacji serwisów komercyjnych coraz większą rolę w upowszechnianiu dorobku naukowego zaczynają odgrywać publikacje o dostępie otwartym
- są one umieszczane:
  - w repozytoriach
  - w bibliotekach cyfrowych
  - na stronach wydawców czasopism
  - na stronach instytucji macierzystych pracowników
  - na stronach instytucji organizujących konferencje
  - na stronach domowych autorów, blogach
  - w popularnych serwisach: Scribd, Zoho, Prezi, SlideShare, Calameo itp.
- widoczność tego typu publikacji jest przedmiotem podjętych badań

Rodzaje publikacji naukowych o dostępie otwartym

- artykuły z czasopism
- rozdziały z prac zbiorowych
- preprinty w/w
- monografie
- prace dyplomowe
- rozprawy doktorskie
- podręczniki
- prezentacje z konferencji i z wykładów
- raporty / sprawozdania z badań

Korzyści wynikające z publikowania w sieci

- ogromne zwiększenie dostępności, co przekłada się na wzrost liczby cytowań – promocja dorobku
- w przypadku preprintów skrócenie czasu prezentacji wyników badań
- ogólne usprawnienie komunikacji naukowej
- drugie „życie” dla publikacji nie mających szans na wznowienie
- dłuższe „życie” dokumentu w sieci w porównaniu do postaci tradycyjnej
- zmniejszenie kosztów publikowania

Niewidoczny Internet

- niewidoczny/ukryty/głęboki = ang. invisible/hidden/deep web - zasoby nieindeksowane przez wyszukiwarki
- do niewidocznego Internetu zalicza się:
  - bazy danych z dostępem tylko przez formularz wyszukiwawczy
  - pliki w pomijanych formatach, np. djvu, zip
  - teksty w plikach graficznych, np. skany zapisane w PDF bez OCR
  - serwisy wymagające logowania – komercyjne i niekomercyjne
  - zasoby, których wyszukiwarki nie zaindeksowały z uwagi na niedoskonałość algorytmu i ograniczenia techniczne/finansowe
- można mówić o niewidocznym Internecie w odniesieniu do:
  - zasobów niewidocznych dla wszystkich wyszukiwarek
  - zasobów niewidocznych dla każdej wyszukiwarki z osobna
- każda wyszukiwarka tworzy własny niewidoczny Internet, zależnie od przyjętego w niej algorytmu indeksującego

Publikacje naukowe a niewidoczny Internet

- korzystając z kilku serwisów wyszukiwawczych dla tych samych zapytań zwiększamy obszar widocznych zasobów
- jednocześnie udostępniając dokument w kilku różnych miejscach sieci zwiększamy jego widoczność
- zależnie od tego w jakim stopniu wyszukiwarki zaindeksują informacje o dokumencie można wyróżnić:
  - widoczność pełną (opis i pełny tekst)
  - widoczność ograniczoną (najczęściej do opisu lub jego części, rzadko do pełnego tekstu)
  - niewidoczność
- każdy powyższych rodzajów można odnieść do pojedynczych wyszukiwarek lub sumy widoczności we wszystkich branych pod uwagę

Cel badań

- w jakim stopniu (pełne teksty, opisy) są widoczne poszczególne rodzaje zasobów (repozytoria, BC itd.)?
- które serwisy wyszukiwawcze są najskuteczniejsze?
- czy warto publikować w kilku miejscach?

- uwagi:
  - najciekawsza część badań dotyczy pełnych tekstów
  - równie interesujący jest pojedynek repozytoriów i BC
  - coraz więcej publikacji naukowych trafia do bibliotek cyfrowych, stąd ocena widoczności BC jest szczególnie ważna

Charakterystyka badań

- dla każdego typu publikacji wybierano dokumenty, które były poszukiwane w kilku wyszukiwarkach po tytule lub jego fragmencie oraz po frazach pochodzących z pełnego tekstu
- wykorzystano publikacje, które pojawiły się w Internecie nie później niż na początku 2012 roku, tak by serwisy wyszukiwawcze zdążyły je zaindeksować (np. Google Scholar – ok. miesiąca)
- publikacje w większości przypadków były z zakresu informatologii i bibliologii
- język publikacji - polski
- dla każdego rodzaju zasobu wybierano przynajmniej dwa dokumenty (by ograniczyć przypadkowość wyników)

Źródła publikacji (1)

- repozytoria
  - AMUR - repozytorium Uniwersytetu im. Adama Mickiewicza w Poznaniu
  - E-LIS - E-prints in Library and Information Science
  - CEON - Repozytorium Centrum Otwartej Nauki
- biblioteki cyfrowe
  - Dolnośląska Biblioteka Cyfrowa
  - Śląska Biblioteka Cyfrowa
  - Zachodnipomorska BC Pomerania
  - Kujawsko-Pomorska BC
  - Bibliologiczna BC
  - Biblioteka Cyfrowa Politechniki Warszawskiej
  - Biblioteka Cyfrowa UMCS

Źródła publikacji (2)

- strony instytucji (materiały konferencyjne)
  - PTIN – X Forum INT 2008
  - Biblioteka Politechniki Poznańskiej - Informacja dla nauki a świat zasobów cyfrowych, 2008
- czasopisma internetowe
  - EBIB
  - iNFOTEZY
- popularne serwisy
  - SlideShare
  - Scribd

Wybrane i odrzucone wyszukiwarki

- ogólne
  - Google
  - Bing
- naukowe
  - Google Scholar
  - Scirus
- repozytoria i biblioteki cyfrowe
  - BASE
- odrzucone
  - Microsoft Academic Search
  - scienceresearch.com
  - worldwidescience.org
  - CiteSeerX
  - Core (COnnecting REpositories)

Repozytoria – badane artykuły

E-LIS

1. Derfert-Wolf L.: *Odkrywanie niewidzialnych zasobów sieci*. 2007
2. Skórka S.: *Systemy nawigacji w przestrzeni słuchowej. Analiza porównawcza*. 2011
3. Cisek S.: *Nauka o informacji na świecie: badania metanaukowe*. 2008

AMUR

4. Sidor M.W.: *Elektroniczny system wspomagający zarządzanie zasobami w bibliotece Wyższej Szkoły Biznesu – National-Louis University w Nowym Sączu*. 2010
5. Chachlikowska A.: *Badania wykorzystania przez polskie biblioteki naukowe środków europejskich, grantów ministerialnych i samorządowych oraz dotacji sponsorów w latach 2000-2008*. 2009

CEON

6. Giętkowski T.: *Zmiany lesistości Borów Tucholskich w latach 1938 – 2000*. 2009
7. Klimczuk A.: *Żyjemy razem, czyli metody budowania kapitału społecznego*. 2009

Repozytoria - wyniki

Artykuły PDF	E-LIS						AMUR				CEON			
	1.		2.		3.		4.		5.		6.		7.	
	T	F	T	F	T	F	T	F	T	F	T	F	T	F
Google	+	+	+	+	+	+	+	+	+	+	+	+	+	+
Bing	+	-	-	-	+	-	-	-	+	-	-	-	+	-
Google Scholar	+	-	-	-	-	-	+	+	+	+	+	+	-	-
Scirus	+	-	+	-	+	+	-	-	+	-	-	-	-	-
BASE	+	-	+	-	+	-	+	-	+	-	+	-	+	-

- E-LIS, AMUR i CEON wykorzystują to samo otwarte oprogramowanie – DSpace
- E-LIS – dziedzinowe, AMUR – instytucjonalne CEON - adresowany do całego polskiego środowiska naukowego

Biblioteki cyfrowe – badane artykuły

Zachodnipomorska BC Pomerania: Bibliotekarz Zachodnio-Pomorski

1. 2008, nr 1-2
2. 2010, nr 1

Śląska BC: Bibliotheca Nostra. Śląski Kwartalnik Naukowy

3. 2009, nr 2
4. 2011, nr 1

Kujawsko-Pomorska BC: Poradnik Bibliotekarza

5. 2007, nr 5
6. 2009, nr 3

Bibliologiczna BC: Zagadnienia Informacji Naukowej

7. 2006, nr 2
8. 2008, nr 2

BC – artykuły – wyniki

Artykuły	PDF				DJVU			
	ZBC		ŚBC		KPBC		BBC	
	1.	2.	3.	4.	5.	6.	7.	8.
Google	+	+	+	+	+	+	+	+
Bing	-	-	+	-	+	+	+	+
Google Scholar	-	-	-	-	-	-	-	-
Scirus	-	-	-	-	-	-	-	-

- ZBC, BBC – dLibra 4.0 ŚBC, KPBC – dLibra 5.0
- BASE – pominięty z uwagi na niemożność zastosowania
- pytania formułowane z tytułów artykułów ze spisu treści
- 5.-8. znajdowano jako odsyłacze do „czystego” OCRu generowanego przez dLibrę dla wyszukiwarek

Biblioteki cyfrowe – badane książki

Śląska BC

1. Tomaszczyk J.: *Angielsko-polski słownik informacji naukowej i bibliotekoznawstwa*. 2009
2. Roszkowski M.: *Język informacyjno-wyszukiwawczy jako narzędzie organizacji informacji w dziedzinowych systemach hipertekstowych*. 2009

Biblioteka Cyfrowa Politechniki Warszawskiej

3. Płoszajski G.: *Standardy w procesie digitalizacji obiektów dziedzictwa kulturowego*. 2008

Biblioteka Cyfrowa UMCS

4. Osiński Z.: *Biblioteka, książka, informacja i Internet.* 2010

Dolnośląska BC

5. Leśniewski D.: *Digitalizacja zasobów bibliotecznych*. 2002

BC – książki – wyniki

Książki	PDF						DJVU		HTML	
	ŚBC				BCPW		UMCS		DBC	
	1.		2.		3.		4.		5.	
	T	F	T	F	T	F	T	F	T	F
Google	+	-	+	+	+	+	+	+	+	+
Bing	+	-	-	-	+	-	-	+	+	+
Google Scholar	-	-	-	-	-	-	-	-	-	-
Scirus	-	-	-	-	-	-	-	-	-	-
BASE	+	-	+	-	+	-	+	-	+	-

- choć wyszukiwanie pełnotekstowe w dLibrze wciąż kuleje można skutecznie korzystać z Google
- słownik dr Tomaszczyka nie został zaindeksowany pełnotekstowo przez Google, ale jego kopia na Pomorskim Uniwersytecie Medycznym już tak (Bing również widzi kopię)
- Google Scholar jest zorientowany na artykuły, i to widać

Strony instytucji – badane referaty

PTIN: X Krajowe Forum Informacji Naukowej i Technicznej, 2009

1. Gajos M.: *Innowacja geoinformacyjna*
2. Sapa R.: *Warsztat naukowca a problem formatu informacji bibliograficznej generowanej przez systemy informacyjne*

Biblioteka Politechniki Poznańskiej (BPP): Informacja dla nauki a świat zasobów cyfrowych, 2008

3. Wiewiórowski W.: *Zagrożenia związane z zarządzaniem informacją prawną i prawniczą w środowisku elektronicznym*
4. Gaziński R.: *Świat informacji na nośnikach elektronicznych a humanista na przykładzie warsztatu historyka*

Strony instytucji (mat. konf.) - wyniki

PDF	PTIN				BPP			
	1.		2.		3.		4.	
	T	F	T	F	T	F	T	F
Google	+	+	+	+	+	+	+	+
Bing	-	-	+	+	+	-	-	-
Google Scholar	-	-	-	-	-	-	+	+
Scirus	-	-	-	-	-	-	-	-

- Materiały konferencyjne PTIN stanowiły prezentacje PDF
- BPP udostępniło artykuły w pełnym tekście również w formacie PDF

Strony wydawców – badane referaty

Biuletyn EBIB, nr 1/2012

1. Derfert-Wolf L.: *Archiwizacja Internetu – wprowadzenie i przegląd wybranych inicjatyw*
2. Nalewajska L.: *Archiwizowanie stron internetowych w krajach nordyckich*

iNFOTEZY, nr 1/2011

3. Uchańska A.: *Strategie przedsiębiorstwa prasowego w XXI wieku*
4. Zapała M.: *Boom komiksowy. Polski rynek historii obrazkowych w latach 2000-2003*

Strony wydawców - wyniki

	Biuletyn EBIB				iNFOTEZY			
	1.		2.		3.		4.	
	T	F	T	F	T	F	T	F
Google	+	-	+	-	+	+	+	+
Bing	+	-	+	-	+	+	+	-
Google Scholar	-	-	-	-	+	+	+	+
Scirus	-	-	-	-	-	-	-	-
BASE	-	-	-	-	+	-	+	-

- publikacje w EBIB to PDFy zaszyte w Google Docs
- artykuł 1. autorka umieściła równolegle w E-LIS, co zwiększyło jego widoczność - T i F (niezaznaczone w tabeli – inny adres)
- iNFOTEZY obsługują standardy OAI-PMH i DC oraz udostępniają pełne teksty w trzech formatach: html, epub, prc, pdf

SlideShare i Scribd – badane dokumenty

Scribd

1. *Między Regałami. Pismo Studentów Informacji Naukowej i Bibliotekoznawstwa. 2009*
2. *Jaskowska B.: Broker informacji – zawód prz(e)(y)szłości. 2011 (prezentacja)*

SlideShare

3. *Cisek S.: Dzielenie się wiedzą. 2009*
4. *Skórka S.: Architektura informacji. Dziedzina wiedzy czy rzemiosło? 2007*

SlideShare i Scribd – wyniki

	Scribd		SlideShare	
	1.	2.	3.	4.
Google	+	+	+	+
Bing	-	-	+	+
Google Scholar	-	-	-	-
Scirus	-	-	-	-

- zastosowano tylko wyszukiwanie pełnotekstowe, gdyż w obu serwisach tytuł pojawia się w obecności całego dokumentu w jedynym dostępnym widoku

Wnioski – widoczność publikacji

- nie jest obojętne, gdzie publikujemy w sieci
- pod względem widoczności najlepiej wypadły repozytoria
- dLibra wymaga dalszych prac rozwojowych – nie jest źle, ale doskonale też nie
- im więcej standardów obsługuje serwis przechowujący publikacje tym większa jego widoczność (dobry przykład – iNFOTEZY)
- jeśli publikacja znajduje się w mało widocznym miejscu, to warto dodatkowo umieścić ją w repozytorium

Wnioski - wyszukiwarki

- choć klasyczny Google jest bardzo skuteczny, to trzeba pamiętać o szumie informacyjnym jaki wprowadza do listy trafień
- serwisy dedykowane nauce są bardziej przejrzyste, ale niestety, dużo mniej skuteczne
- rozczarowały:
  - Bing
  - Scirus
  - Google Scholar
- BASE potwierdził wysoką skuteczność w odniesieniu do poszukiwań w opisach dokumentów
- zbieżność z wynikami Agnieszki Łakomy (praca magisterska z 2012 w IBiIN)

KONIEC

Dziękuję za uwagę