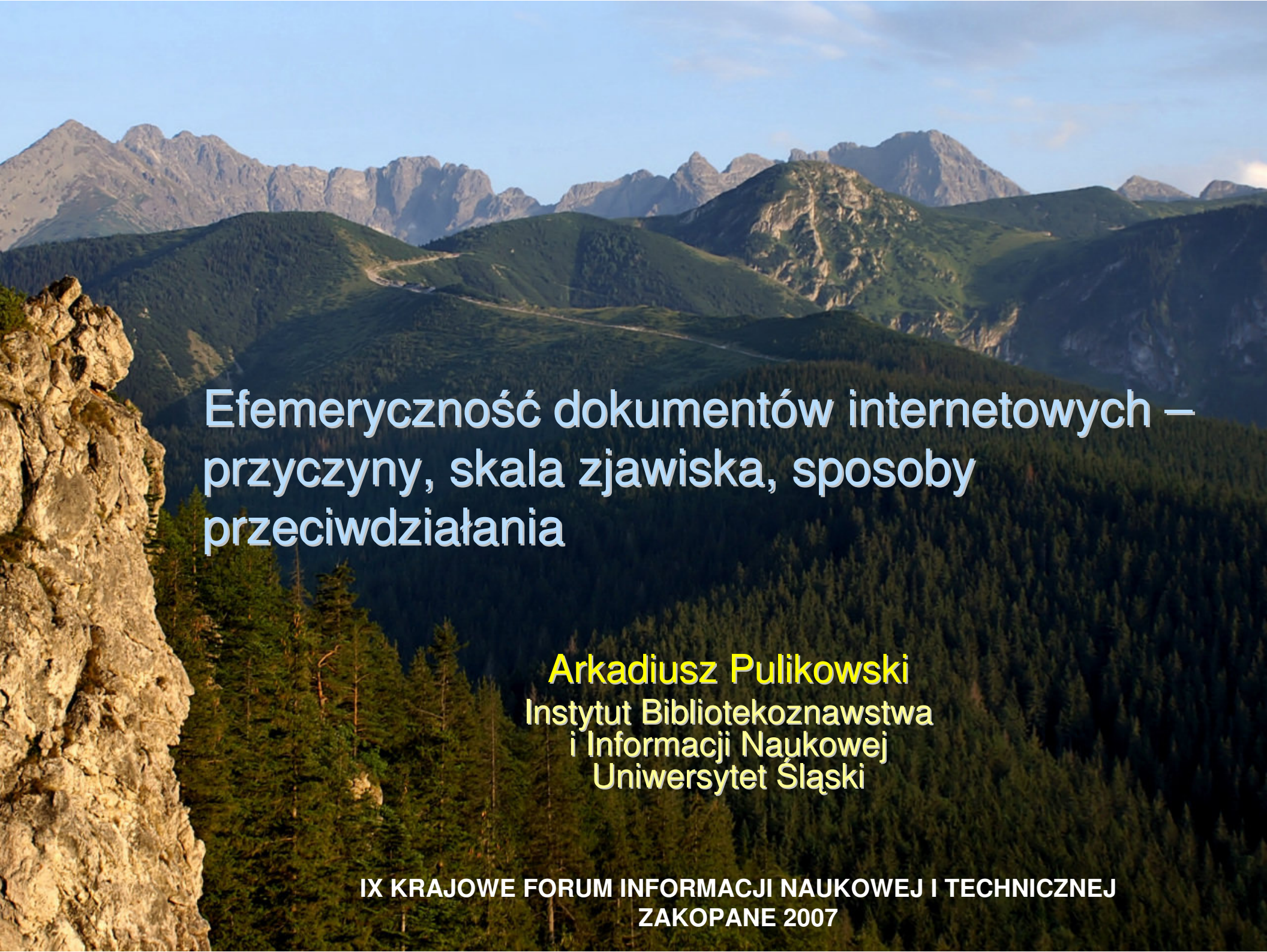


Efemeryczność dokumentów internetowych – przyczyny, skala zjawiska, sposoby przeciwdziałania

Arkadiusz Pulikowski
Instytut Bibliotekoznawstwa
i Informacji Naukowej
Uniwersytet Śląski

IX KRAJOWE FORUM INFORMACJI NAUKOWEJ I TECHNICZNEJ
ZAKOPANE 2007

Plan wystąpienia

- określenie przyczyn „znikania” dokumentów internetowych,
- zbadanie skali zjawiska, na podstawie analizy odsyłaczy hipertekstowych z biuletynu EBIB i kwartalnika PTINT,
- wskazanie obecnie dostępnych metod odzyskiwania nieaktywnych dokumentów,
- przedstawienie koncepcji systemu, który mógłby w przyszłości definitywnie rozwiązać omawiany problem.

Najczęstsze przyczyny braku dostępu do cytowanego dokumentu internetowego

Permanentne

- zaprzestanie działalności serwera,
- usunięcie dokumentu z serwera,
- zmiana ścieżki dostępu do pliku,
- błędne przepisanie adresu dokumentu.

Przejściowe

- czasowe wyłączenie serwera (aktualizacja oprogramowania, awaria sprzętowa),
- awaria sieci lokalnej, w której funkcjonuje serwer przechowujący interesujące nas dokumenty.

Założenia przyjęte dla badania odsyłaczy hipertekstowych z biuletynu EBIB i kwartalnika PTINT

- przedział czasu: 2001- 2006,
- analizowane odsyłacze pochodziły z treści artykułu, bibliografii i przypisów,
- dla danego artykułu każdy odnośnik liczony był raz, niezależnie od tego ile razy pojawił się w dokumencie,
- przekierowania wykonywane przez oprogramowanie serwera prowadzące do właściwego dokumentu były traktowane jako poprawnie działający odsyłacz.

Założenia przyjęte dla badania odsyłaczy hipertekstowych z biuletynu EBIB i kwartalnika PTINT – c.d.

- nie były rozróżniane dokumenty chwilowo niedziałajace od tych trwale niedostępnych,
- dla EBIB-u artykuły pochodziły z działów: Artykuły, Badania, teorie wizje, oraz rzadko pojawiających się Opinii, Polemik i Technologii,
- dla PTINT-u były to wszystkie działy z wyjątkiem: Wydarzeń krajowych i zagranicznych oraz recenzji,
- łączna liczba artykułów objętych badaniem dla EBIB-u wyniosła 455, a dla PTINT-u 131.

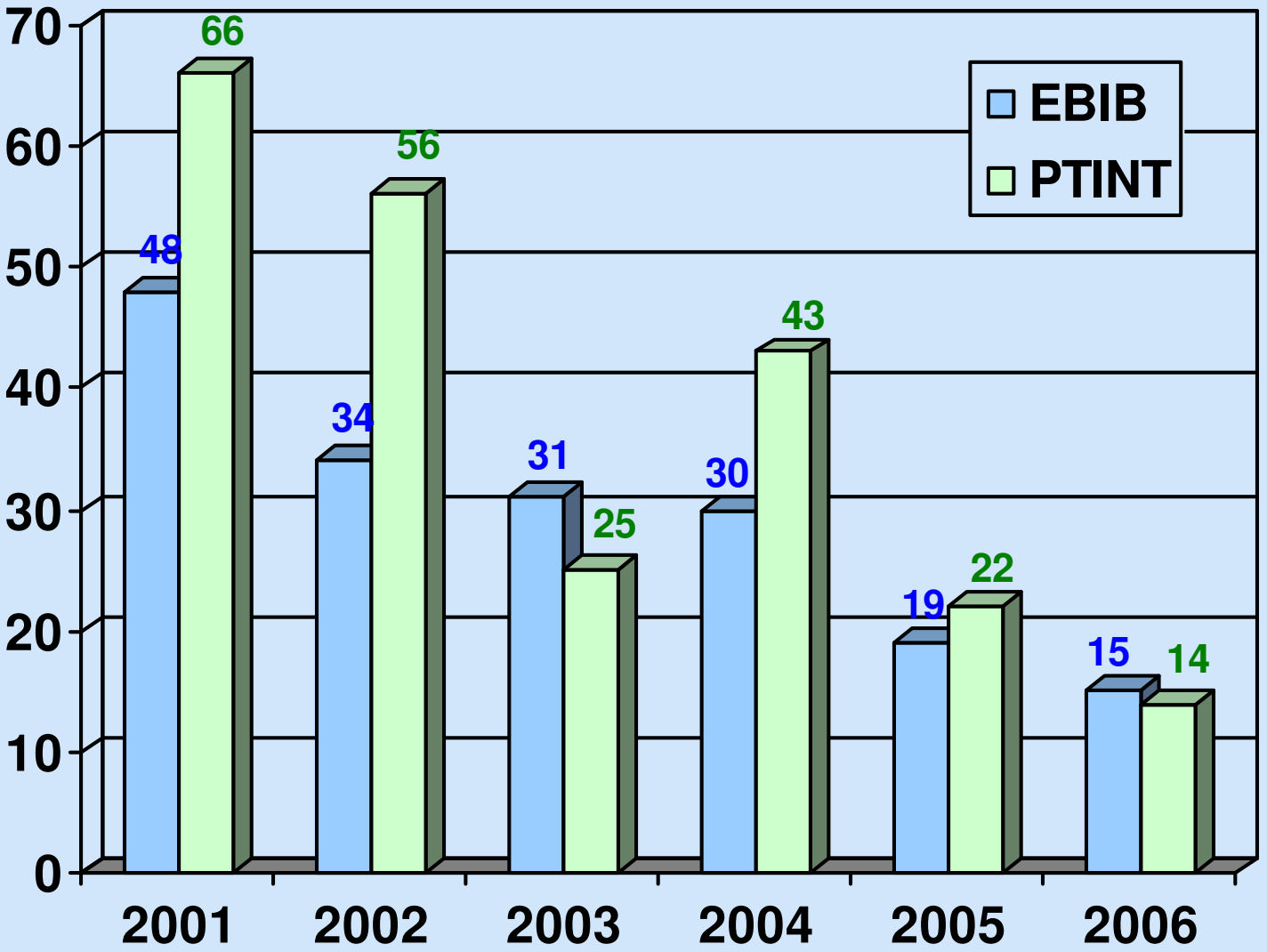
Wyniki badania czasopisma EBIB w ujęciu tabelarycznym

Rok	liczba numerów rocznika	Liczba badanych artykułów	łączna liczba odsyłaczy	śr. liczba odsyłaczy w art.	Liczba odsyłaczy niedział.	procent odsyłaczy niedział.
2001	11	46	147	3,2	70	48
2002	11	71	279	3,9	95	34
2003	11	78	302	3,9	95	31
2004	10	94	344	3,7	103	30
2005	9	81	522	6,4	99	19
2006	11	85	590	6,9	90	15

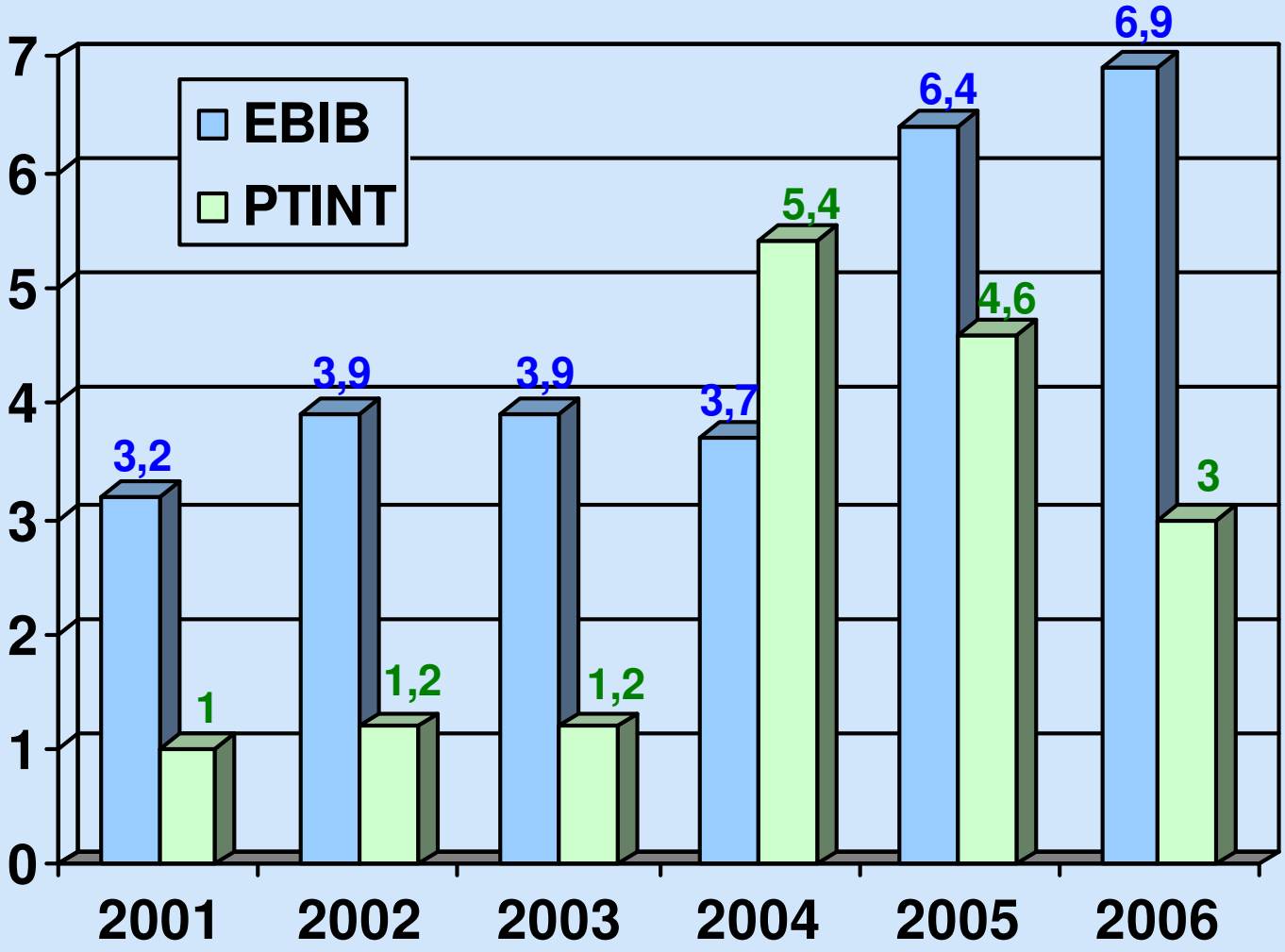
Wyniki badania czasopisma PTINT w ujęciu tabelarycznym

Rok	liczba numerów rocznika	Liczba badanych artykułów	łączna liczba odsyłaczy	śr. liczba odsyłaczy w art.	Liczba odsyłaczy niedział.	procent odsyłaczy niedział.
2001	3	23	24	1,0	16	66
2002	3	20	25	1,2	14	56
2003	3	24	28	1,2	7	25
2004	3	20	108	5,4	47	43
2005	4	21	91	4,6	20	22
2006	4	23	70	3,0	10	14

Odsyłacze niedziałające w procentach



Średnia liczba odsyłaczy w artykule



Metody odzyskiwania dokumentów internetowych

- analiza adresu URL,
- wykorzystanie kopii przechowywanej przez wyszukiwarki,
- wyszukiwanie w serwisie Wayback Machine - web.archive.org.

Analiza adresu URL

- sprawdzenie poprawności zapisu adresu URL dokumentu,
- skracanie adresu URL dokumentu do domeny serwera, a następnie przeglądanie zawartości witryny (o ile adres ten działa),
- skracanie adresu URL dokumentu do domeny i wyszukiwanie z wykorzystaniem wyszukiwarki dostępnej na tym serwerze lub użycie polecenia site: google'a, host: Szukacza itp.
- wyszukiwanie w adresach URL za pomocą Google'a lub innej wyszukiwarki z wykorzystaniem operatora allinurl: lub formularza zaawansowanego.

Wykorzystanie kopii przechowywanej przez wyszukiwarki

- większość wyszukiwarek w trakcie indeksowania tworzy kopię dokumentu,
- każda wyszukiwarka robi to w innym czasie, dzięki czemu jeśli w jednej nie znajdziemy kopii poszukiwanego dokumentu można próbować w kolejnej,
- metoda daje najlepsze rezultaty dla poszukiwań dokumentów zaginionych w ostatnich miesiącach – znajdujących się jeszcze w indeksach wyszukiwarek.

Kopie dokumentów na listach trafień wybranych wyszukiwarek

- Google,

Kolekcja matematyczno-fizyczna
Prosimy o przesyłanie uwag na adres bwm@icm.edu.pl · International Journal of Applied Mathematics and Computer Science. Aktualnie do Państwa dyspozycji: ...
matwbn.icm.edu.pl/spis.php?wyd=11 - 3k - Kopia - Podobne strony

- Szukacz,

16. Biblioteka Główna Politechniki Gdańskiej
[...] Bazy pełnotekstowe • Bazy abstraktowe • Bazy bibliograficzne • **Open Access** • Bazy prawnicze • Czasopisma online • Lista A *[...]*
URL: www.bg.pg.gda.pl/ – mod. 2007-06-22 – arch. 2007-06-22 – 24 kB – Kopia z archiwum
Więcej dokumentów z www.bg.pg.gda.pl (jeszcze 1)

- Live.

Open access - Wikipedia, the free encyclopedia
Open access (OA) is the free online availability of digital content. It is best-known and most feasible for peer-reviewed scientific and scholarly journal articles, which scholars publish without ...
en.wikipedia.org/wiki/Open_access · Buforowana strona

Wyszukiwanie w archiwum web.archive.org

- Wayback Machine działa od 1996 r.,
- gromadzi w chwili obecnej około 85 miliardów dokumentów internetowych,
- pliki są dodawane z kilkumiesięcznym opóźnieniem (nowsze można znaleźć korzystając z kopii wyszukiwarek),
- odnośniki wewnątrz kopii odsyłają do kolejnych kopii utworzonych w podobnym czasie,
- do dokumentów przechowywanych w archiwum można tworzyć odsyłacze np.: http://web.archive.org/web/20031225004248/http://ibin.us.edu.pl/

Wayback Machine

- na 70 niedziałających odsyłaczy z rocznika 2001 EBIB-u Wayback Machine znalazł aż 51 – 73%

Przykład wyszukiwania z Wayback Machine

INTERNET ARCHIVE
WayBackMachine

Enter Web Address: http:// All Take Me Back Adv. Search Compare Archive Pages

Searched for http://sunsite.berkeley.edu/SICI/sici.pdf **10** Results

* denotes when site was updated.

Search Results for Jan 01, 1996 - Sep 25, 2007

1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007
0 pages	0 pages	0 pages	0 pages	1 pages	3 pages	0 pages	3 pages	1 pages	2 pages	0 pages	0 pages
				Aug 24, 2000 *	Nov 18, 2001 * Nov 23, 2001 Nov 27, 2001		Mar 08, 2003 * Apr 08, 2003 Oct 03, 2003	Apr 22, 2004	Feb 17, 2005 May 22, 2005		

Home | Help

Internet Archive | Terms of Use | Privacy Policy

- oprócz kopii poszukiwanego dokumentu możemy również zobaczyć czy i jak zmieniała się jego zawartość w czasie oraz kiedy mniej więcej został utworzony i do kiedy był dostępny online.

Wady wymienionych metod walki ze „znikającymi” dokumentami

- są czasochłonne,
- nie do końca rozwiązują problem, ponieważ wielu dokumentów i tak nie odnajdziemy,
- nawet gdy odnośnik działa to nie możemy mieć pewności, że dokument jest w takiej samej postaci w jakiej widział go autor odsyłający do niego.

Koncepcja nowego systemu

- powinien tworzyć kopie „na życzenie” użytkownika (osoby tworzącej opis bibliograficzny),
- każda kopia otrzymywałaby unikalny identyfikator, na kształt obecnie funkcjonującego DOI – Digital Object Identifier,
- identyfikator byłby dodawany do opisu bibliograficznego obok adresu URL,
- na jego podstawie ze strony WWW projektu można by szybko dotrzeć do kopii dokumentu, niezależnie od tego czy jest on jeszcze dostępny,
- przed utworzeniem identyfikatora dla dokumentu, system sprawdzałby, czy zgłaszany adres URL nie ma już przypisanego jednego lub kilku numerów,
- zgłaszający mógłby skorzystać z istniejącego już identyfikatora do tworzenia własnego opisu gdyby uznał, że kopia do której odsyła jest taka sama lub prawie identyczna ze zgłaszanym dokumentem.


Dziękuję za uwagę