



Wikipedia w wyszukiwarce NetSprint.pl

Jak miksuje Mikser?

Michał Kosmulski, Jakub Kacprzak
NetSprint.pl

Prezentacja jest dostępna na licencji GNU Free Documentation License
lub Creative Commons Attribution-Share Alike.
download: www.netsprint.pl/publikacje/wikimedia2007/



- Kilka słów o NetSprint.pl
- Nieco o Mikserze
- Szczegółowo o Wikipedii i Mikserze



- Czym jest NetSprint:
 - Podmiot prawny - NetSprint.pl Sp. z o.o.
 - Silnik wyszukiwarki, dostępny na:
 - WP.pl, o2.pl, Hoga.pl, Gery.pl, Polska.pl ... (w sumie prawie 150 serwisów)
 - Serwis wyszukiwarki www.netsprint.pl



- Dynamiczny rozwój Internetu:
 - Rosnąca liczba stron internetowych:
 - wyzwanie dla silnika wyszukiwarki - jak przeszukiwać coraz większą liczbę stron i zapewnić optymalne wyniki?
 - wyzwanie dla internautów - jak w tym gąszczu wyników znaleźć szukane informacje?
- Potrzeby internautów:
 - Czego tak naprawdę szukają?? Stron, czy **informacji**?



- ...a do tego inspiracja: www.Ask.com

The screenshot shows the Ask.com search results for the query "Poland". The search bar contains "Poland" and the search button is labeled "Search". The results are displayed under the heading "Web Search" and show 1-10 of 17,100,000 results.

Poland | Save
Capital: Warsaw; **Population:** 38,635,144
Location: Central Europe, east of Germany
Chief of State: President Aleksander Kwasniewski, **Head of Government:** Prime Minister Marek Belka
Languages: Polish [More »](#)
[Encyclopedia](#) | [BBC Profile](#) | [History](#) | [Anthem](#) | [Flag](#)
[FCO Advice](#) | [Tourist Attractions](#) | [Current Weather](#) | [Local Time](#) | [Polish National Football Team](#)

News about Poland
 Skanska Starts Three Commercial Projects in **Poland** and Czech Republic Totaling A... [Business Wire UK](#) 36 hours ago
Poland animates film screenings [The Scotsman](#) 4/23 11:44 AM

Polish National Tourist Office
 Promoting travel to **Poland** from North America. In English, French, and German.
www.polandtour.org/ • [Cached](#) • [Save](#)

Ryanair
 Ryanair.com Great Britain - The Low Fares Airline, cheap flights from Europe, UK and Ireland. Cheapest flight tickets, discount airline tickets.
www.ryanair.com/ • [Save](#)

PolishWorld
 Internet guide to **Poland** and Polonia (Polish communities worldwide). In English or Polish.
www.polishworld.com/ • [Cached](#) • [Save](#)

Poland: History, Geography, Government, and Culture — ...
 Information on **Poland** — geography, history, politics, government, economy, population statistics, culture, religion, languages, largest ...
www.infoplease.com/ipa/A0107891.html • [Cached](#) • [Save](#)

The History Of Poland
 History from 960 to the end of WWII. ... [Main Index](#) | [History Of Poland](#) | [Krakow Index](#) | [Books on Poland](#) | [Links ...](#)
www.kasprzyk.demon.co.uk/www/HistoryPolska.html • [Cached](#) • [Save](#)

Narrow Your Search
 Facts about **Poland**
 Map of **Poland**
Poland Cities
Poland History
 Population of **Poland**
Poland's Culture
 Interesting Facts about **Poland**
 Clickable Map of **Poland**
Poland Economy
 Fun Facts about **Poland**
Poland Government
 Weather in **Poland**
 Holidays in **Poland**
Poland News
 What Is the Capital of **Poland**
[More »](#)

Expand Your Search
 Polish Websites
 Warsaw
 Polish Flag
 Polish Food



- Mechanizm pozwalający na wyświetlanie na szczycie listy wyników wyszukiwania **nie stron**, lecz konkretnych **informacji** pochodzących z **wiarygodnych źródeł**, dzięki czemu **skróceniu** ulega droga dotarcia do informacji oraz zagwarantowana jest ich **jakość**.



[Strona główna](#)

netsprint™

[WWW](#) | [Grafika](#) | [Wiadomości](#) | [Firmy](#) | [Encyklopedia](#) | [Słowniki](#)

Wikipedia [z wikipedii]

20 ▾

Szukaj

[Zaawansowane Preferencje](#)

[Dodaj stronę i telefon](#) | [Promocja w wyszukiwarce](#) | [Znajdź odpowiedź. Najszybciej.](#)

WWW: **Wikipedia**



Wikipedia (źródło: [wikipedia](#), licencja: [GNU FDL](#), [autorzy](#))

[Oceń ten wynik/zasób](#)

Wikipedia to wielojęzyczny projekt internetowej encyklopedii, działającej na zasadzie otwartej treści. Działa w oparciu o oprogramowanie wiki, dzięki czemu pozwala na edycję każdemu użytkownikowi odwiedzającemu stronę. Słowo *Wikipedia* to połączenie wyrazów *wiki* i *encyklopedia*. [[więcej](#)]

Inne: [galeria zdjęć](#)

Najnowsze wiadomości na temat: wikipedia z wikipedii

✉ [Codzienny mail na temat wikipedia z wikipedii »](#)



[Wikipedia 0.5 na CD!](#) - Linux.pl - 27-04-2007

[Wikipedia offline](#) - Weblinside.pl - 27-04-2007

[Wikipedia nie chce reklam](#) - Wirtualne Media - 26-04-2007

[Strona główna - Wikipedia, wolna encyklopedia](#) ★ popularna strona

Źródło: " http://pl.wikipedia.org/wiki/S_g%C5%82%...

[pl.wikipedia.org](#) ✿ aktualna strona - kopia strony - kolejne podstrony

Kanał #irc Wikipedii

Kanał **#wikipedia.pl** Aby zacząć rozmowę na kanale... ..ksywkę i wejdź. **Wikipedia-pl** Wikinews Wikibooks Wikisłownik Wikimedia-pl... ..wejsć jeszcze na jakiś kanał: /join **#wikipedia-pl** - polska **Wikipedia** /join **#wikinews-pl**...

[adamdziura.9g.pl/wikipedia/irc](#) kopia strony - kolejne podstrony



- Mikser przechowuje rekordy z różnych, wcześniej wybranych źródeł i wybiera najbardziej użyteczne
- Dla bardzo ogólnych zapytań próbujemy odgadnąć, czego szuka użytkownik i podać mu od razu gotową odpowiedź



Jakie dane prezentuje Mikser i skąd one pochodzą?

- Dane statyczne
 - **Wikipedia**
 - **WP.pl:**
 - aktorzy, filmy, wykonawcy, płyty
 - **Filmpolski.pl:**
 - aktorzy, filmy, seriale
 - **Autocentrum.pl:**
 - samochody
 - **IDG.pl**
 - programy
 - **Gry-online.pl:**
 - gry
 - **Firmy.NetSprint.pl/PF.pl:**
 - Dane teleadresowe firm
 - ...
- Dane dynamiczne
 - **News.NetSprint.pl:**
 - aktualności
 - **WP.pl:**
 - kursy walut
 - notowania giełdowe
 - prognoza pogody
 - wyniki lotto
 - repertuary kin
 - ...



- www.Przewodnik.NetSprint.pl:
 - Rozwiązywanie problemów użytkowników:
 - Piszesz wypracowanie? Masz klasówkę?
 - Nie wiesz, na co iść do kina?
 - Szukasz pożyczki lub spłacasz kredyt?
 - Planujesz wypoczynek za granicą?
 - Interesujesz się sportem?
 - Chcesz szybko sprawdzić prognozę pogody na najbliższe dni?
 - ...




- www.Przewodnik.NetSprint.pl:

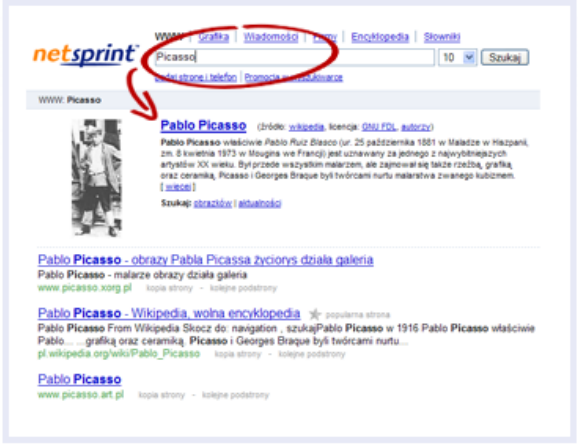
Masz pytanie? Zobacz, jak znaleźć odpowiedź. Najszybciej.
 Ponad 500 000 odpowiedzi w wynikach wyszukiwania NetSprint.pl!

w serwisie w internecie

wszystkie następny >>



Piszesz wypracowanie? Masz klasówkę? Pomożemy Ci. Dzięki NetSprintowi uzyskujesz dostęp do encyklopedycznych informacji z wielu dziedzin naukowych.



W NetSprint.pl znajdziesz wiele informacji historycznych. Zaczynając od postaci historycznych - [Bolesław Chrobry](#) - czy wielkich bitew - [Bitwa pod Warną](#). Idąc dalej przez sztukę, malarstwo czy rzeźbę: [Picasso](#), [Krzyk](#), [Michał Anioł](#). A kończąc na geografii, astronomii, matematyce czy chemii: [kwas pruski](#), [kometa](#), [twierdzenie sinusów](#), [sublimacja](#), [kenozoik](#).

Inne w kategorii Edukacja

- > Jak wyznaczyć długość i szerokość geograficzną?
- > Wielcy twórcy polskiego renesansu.
- > Problem z matmy?
- > Czym się różni komórka od tkanki?
- > Znasz układ okresowy Mendelejewa?

[więcej ...](#)



Wikipedia w Mikserze



- **Wikipedia** jest źródłem gotowych odpowiedzi dla haseł encyklopedycznych: nazw miast, państw, nazwisk znanych ludzi i in.
- Co prezentujemy w wynikach:
 - tytuł
 - skrócony opis
 - rysunek
 - linki do strony domowej, galerii zdjęć
 - pola tabelaryczne (ustrukturyzowane dane statystyczne i inne): ludność miast, powierzchnia państw, nazwa łacińska dla roślin i zwierząt, ...



[Strona główna](#)

netsprint™

WWW | [Grafika](#) | [Wiadomości](#) | [Firmy](#) | [Encyklopedia](#) | [Słowniki](#)

Białowieża [wieś]

20 ▾

Szukaj

[Zaawansowane](#)
[Preferencje](#)

[Dodaj stronę i telefon](#) | [Promocja w wyszukiwarce](#) | **Znajdź odpowiedź. Najszybciej.**

WWW: **Białowieża**



Białowieża (źródło: [wikipedia](#), licencja: [GNU FDL](#), [autorzy](#))

[Oceń ten wynik/zasób](#)

Ludność: 2670 (dane na rok 2002)

Białowieża — duża wieś w Polsce położona w województwie podlaskim, w powiecie hajnowskim, w gminie Białowieża o charakterze małomiasteczkowym. Wieś jest siedzibą gminy Białowieża oraz Białowieżskiego Parku Narodowego, obejmującego 17% polskiej części Puszczy Białowieżskiej. Białowieża jest także siedzibą sołectwa obejmującego miejscowości: Podolany, Krzyże i Zastawa. [[więcej](#)]

Inne: [strona domowa](#)

Szukaj: [obrazków](#) | [aktualności](#)

Inne: [galeria zdjęć](#)

Czy chodziło Ci o: [Białowieża, województwo mazowieckie \[wieś\]](#) [Białowieża, województwo opolskie \[wieś\]](#) [Białowieża, województwo świętokrzyskie \[wieś\]](#) [Białowieża, powiat nakielski \[wieś\]](#) [Białowieża, powiat tucholski \[wieś\]](#)

[Hotel Białowiecki w Białowieży](#)

Hotel Białowiecki - Białowieża. Wypoczynek i rekreacja w trzygwiazdkowym luksusowym obiekcie za przystępną cenę. Gwarantujemy doskonałą regionalną kuchnię, spokój i ciszę w sąsiedztwie prastarej Puszczy Białowieżskiej

[www.hotel.bialowieza.pl](#) - [Link Sponsorowany](#)

[Hotel Żubrówka w Białowieży](#)

Wyjątkowy hotel, wyjątkowe miejsce. Luksusowy obiekt czterogwiazdkowy z niepowtarzalną atmosferą, położony w sercu Puszczy Białowieżskiej. Tradycyjna kuchnia regionalna i myśliwska "Białowieżskie Jadło"

[www.hotel-zubrowka.pl](#) - [Link Sponsorowany](#)

[Białowieża Turystyka - Biuro Turystyki Przyrodniczej Białowieża](#)

Białowieża, Puszcza Białowieża, noclegi, leśniczówki, wycieczki, ogniska, safari, integracja, informacja. Nie zwlekaj, **Białowieża** czeka! lang pl Białowieża, Białowieża Primaeval Forest, Wilderness Wetlands - Białowieża, Biebrza, Narew, National Parks. Don't wait! Białowieża waits.



- Wyszukiwanie: ścisłe dopasowanie nazwy rekordu + dodatkowe „synonimy”
 - Mikołaj Kopernik → Kopernik (algorytm)
 - Uchatka kalifornijska → Uchatka (algorytm)
 - McDonnell Douglas F-15 Eagle → F-15 (algorytm)
 - Muhammad Ali → Cassius Clay (redirect)
- Nazwy typów i podświetlanie pól tabelarycznych
 - Typy rekordów: „Warszawa [samochód]”
 - Dostajemy gotową odpowiedź na zapytania typu „powierzchnia Gruzji”
 - Rozpoznajemy dopełniacze tam gdzie to możliwe



[Strona główna](#)

netsprint™

[WWW](#) | [Grafika](#) | [Wiadomości](#) | [Firmy](#) | [Encyklopedia](#) | [Słowniki](#)

Wenecja

20

Szukaj

[Zaawansowane
Preferencje](#)

[Dodaj stronę i telefon](#) | [Promocja w wyszukiwarce](#) | [Znajdź odpowiedź. Najszybciej.](#)

WWW: **Wenecja**



Wenecja (źródło: [wikipedia](#), licencja: [GNU FDL](#), autorzy)

[Oceń ten wynik/zasób](#)

Ludność: 266 181 (dane na rok 2004)

Miasto **Wenecja** (wł. *Venezia*), to miejscowość i gmina na północy Włoch nad Adriatykiem, stolica regionu Wenecja Euganejska. Ludność: 271 tys. mieszk. (2001). [[więcej](#)]

Inne: [strona domowa](#)

Szukaj: [obrazków](#) | [aktualności](#) | [w encyklopedii](#)

Inne: [galeria zdjęć](#)

Czy chodziło Ci o: [Wenecja \[utwór literacki\]](#) [Wenecja \[z wikipedii\]](#) [Wenecja \[wieś\]](#)

[Wenecja, województwo warmińsko-mazurskie \[wieś\]](#) [Wenecja i Veneto \[utwór literacki\]](#)

[Wenecja. Przewodnik kieszonkowy \[utwór literacki\]](#) [Galerie dell' Accademia. Wenecja \[utwór literacki\]](#)

[Wenecja.Art - serwis o Wenecji Bydgoskiej i...](#)

Wenecja Bydgoska to najbardziej malowniczy zakątek naszego miasta, charakteryzują go wznoszące się nad wodą XIX-wieczne kamieniczki. Nie musisz jechać do Włoch, bo wystarczy odwiedzić Bydgoszcz! Wejdziesz, aby ocalić to miejsce od zapomnienia!

[www.wenecja.art.pl](#) [kopia strony](#) - [kolejne podstrony](#)

[Wenecja - restauracja](#)

Restauracja **Wenecja**

[www.wenecja.bydgoszcz.pl](#)  [adres i telefon](#) [kopia strony](#) - [kolejne podstrony](#)

[WENECJA - AGENCJA REKLAMY KRAKÓW, REKLAMA, PLAKATOWANIE, BTL, ...](#)

WENECJA - AGENCJA REKLAMY KRAKÓW, REKLAMA, PLAKATOWANIE, BTL, PUBLIC RELATIONS, MEDIA, KSERO

[www.wenecja.com.pl](#)  [adres i telefon](#) [kopia strony](#) - [kolejne podstrony](#)



[Strona główna](#)

netsprint™

[WWW](#) | [Grafika](#) | [Wiadomości](#) | [Firmy](#) | [Encyklopedia](#) | [Słowniki](#)

Wenecja [wieś]

20

Szukaj

[Zaawansowane
Preferencje](#)

[Dodaj stronę i telefon](#) | [Promocja w wyszukiwarce](#) | [Znajdź odpowiedź. Najszybciej.](#)

WWW: **Wenecja**



Wenecja (źródło: [wikipedia](#), licencja: [GNU FDL](#), [autorzy](#))

[Oceń ten wynik/zasób](#)

Wenecja – wieś w Polsce położona w województwie kujawsko-pomorskim, w powiecie żnińskim, w gminie Żnin. W latach 1975-1998 miejscowość należała administracyjnie do województwa bydgoskiego. [[więcej](#)]

Szukaj: [obrazków](#) | [aktualności](#)

Czy chodziło Ci o: [Wenecja, województwo warmińsko-mazurskie \[wieś\]](#)

[Wenecja.Art - serwis o Wenecji Bydgoskiej i...](#)

Wenecja Bydgoska to najbardziej malowniczy zakątek naszego miasta, charakteryzują go wznoszące się nad wodą XIX-wieczne kamieniczki. Nie musisz jechać do Włoch, bo wystarczy odwiedzić Bydgoszcz! Wejdziesz, aby ocalić to miejsce od zapomnienia!

[www.wenecja.art.pl](#) kopia strony - kolejne podstrony

[Wenecja - restauracja](#)

Restauracja **Wenecja**

[www.wenecja.bydgoszcz.pl](#)  adres i telefon kopia strony - kolejne podstrony

[WENECJA - AGENCJA REKLAMY KRAKÓW, REKLAMA, PLAKATOWANIE, BTL,...](#)

WENECJA - AGENCJA REKLAMY KRAKÓW, REKLAMA, PLAKATOWANIE, BTL, PUBLIC RELATIONS, MEDIA, KSERO

[www.wenecja.com.pl](#)  adres i telefon kopia strony - kolejne podstrony

<http://www.petersburg.ovh.org/>

[www.petersburg.ovh.org](#) - kolejne podstrony



- Surowy wikitekst jest przeznaczony do prezentacji, dane nie posiadają struktury

Logowanie / rejestracja

artykuł dyskusja edytuj historia i autorzy

Edytujesz "Żubr"

Nie jesteś zalogowany. Twój adres IP będzie zapisany w historii edycji strony.

[[disambigR|''dużego ssaka''|[[Żubr (ujednoznacznienie)|inne znaczenia słowa "Żubr"]]]

```

{{Zwierzę infobox
|Nazwa zwyczajowa=Żubr
|Lacinska="Bison bonasus"
|Zoolog={{Karol Linneusz|Linnaeus}}. 1758)
|Obrazek=Żubr_Bison_bonasus.jpg
|Opis_obrazka=Żubr kaukaski w poznańskim [[Ogród zoologiczny|200]]
|Status_IUCN=EM
|Gromada=[[ssaki]]
|Podgromada=[[ssaki żyworodne]]
|Szczep=[[łożyskowce]]
|Rząd=[[parzystokopytne]]
|Rodzina=[[krętorogie]]
|Rodzaj=[[Bison (zwierzęta)|Bison]]
|Gatunek=żubr
|Podgatunki="B. bonasus bonasus"<br>"B. bonasus caucasicus"&agger;
|Commons=Category:Bison bonasus
|Wikispecies=Bison bonasus
}}
{{gatunek pod ochroną}}
'''Żubr''' (''Bison bonasus'') - [[łożyskowce|ssak łożyskowy]] [[parzystokopytne|parzystokopytny]]. Aktualnie populację w [[Białowieża|Puszczy Białowieżskiej]], [[Puszcza Borecka|Puszczy Boreckiej]] oraz w [[Bieszczady|Bieszczadach]] (dalej w [[Knyżyńska|Puszczy Knyżyńskiej]])

```

Zmiany są widoczne natychmiast po zapisaniu. Eksperymentuj!

Strona główna

WWW | Grafika | Wiadomości | Firmy | Encyklopedia | Słowniki

Żubr | 20 | Szukaj

[Zaawansowane Preferencje](#)

[Dodaj stronę i telefon](#) | [Promocja w wyszukiwarce](#) | [Znajdź odpowiedź. Najszybciej.](#)

WWW: Żubr

Żubr (źródło: [wikipedia](#), licencja: [GNU FDL](#), autorzy) [Oceń ten wynik/zasób](#)



Nazwa łacińska: *Bison bonasus* **Gromada:** ssaki **Rząd:** parzystokopytne **Rodzina:** krętorogie **Rodzaj:** *Bison*

Żubr (*Bison bonasus*) - ssak łożyskowy z rodziny krętorogich, rzędu parzystokopytnych. Aktualnie populacja żubrów na świecie liczy ok. 3400 osobników, z czego 630 żyje w Polsce na terenie Puszczy Białowieżskiej, Puszczy Boreckiej, Puszczy Pilskiej, Puszczy Knyżyńskiej oraz w Bieszczadach (dalszych ok. 150 żubrów znajduje się w zamkniętych hodowlach). [[więcej](#)]

Szukaj: [obrazków](#) | [aktualności](#)

Czy chodziło Ci o: [Żubr \[samochód\]](#) [Żubr \[z wikipedii\]](#) [Żubr pierwotny \[zwierzę\]](#) [LWS-4 Żubr \[samolot\]](#) [Joanna Żubr \[wojskowy\]](#) [Żubr, piwo \[z wikipedii\]](#)



- Problemy

- MediaWiki zamienia wikitekst na stronę HTML, my potrzebujemy tylko początkowej części artykułu oraz danych z tabel
- MediaWiki jest napisana w PHP, nasz konwerter w Javie (do tego dochodzą kwestie licencyjne)
- Wydajność
- Aktualizacje danych



- Pierwotny parser: wyrażenia regularne
 - znaczna liczba błędów parsowania
 - powolny
- Obecny parser: JFlex
 - redukcja liczby błędów formatowania
 - kilkukrotny wzrost wydajności
 - na podstawie wikitekstu tworzy streszczenie artykułu (uproszczony HTML bez tabel itp.) oraz zbiór informacji o wykorzystanych szablonach, odnośnikach itd.



- Domyślnie bierzemy pierwszy rysunek w artykule
 - Akceptujemy rysunek tylko gdy rozmiary są „rozsądne”
 - Potrzebujemy listy ikon i innych rysunków pomocniczych, które należy ignorować
- Dla wybranych typów staramy się pobrać rysunek wskazany przez odpowiednie pole szablonu
 - herby dla miast
 - flagi dla państw
 - rysunek umieszczony w infoboksie dla niektórych innych typów



- Pola tabelaryczne pobieramy z szablonów
 - głównie szablony typu „infobox”
 - linki do wikiźródeł i wikicytatów (szablon „siostrzane projekty” i podobne)
- Przykłady
 - powierzchnia, ludność (państwa)
 - nazwa łacińska (rośliny, zwierzęta)
 - wysokość (szczyty górskie)
 - rozpiętość skrzydeł, pułap, prędkość maksymalna (samoloty)
 - przynależność partyjna i funkcja (politycy)



- Czasami można określić na podstawie szablonu (tak zwykle jest dla miejscowości)
- Następnie próbujemy wydobyć link z sekcji „Linki zewnętrzne”
 - Heurystyka określa, które linki wydają się wskazywać na stronę domową
 - Pewną część linków do stron domowych ignorujemy ze względu na ograniczenia algorytmu



- Na podstawie szablonów występujących w treści artykułu (zwykle infoboksy)
- Na podstawie kategorii w Wikipedii
- Na podstawie nazwy - np. „Potop (powieść)”



- Wydobycie wartości określonych w szablonie (pola tabelaryczne, strona domowa)
 - Zalety
 - Dane są ustrukturyzowane
 - Względna łatwość parsowania
 - Ograniczenia
 - Różne nazwy pól w różnych szablonach
 - Różna składnia pól w różnych szablonach (np. strona WWW z prefiksem http:// lub bez, czasem link a czasem zwykły tekst)
- Wykorzystanie przekierowań do tworzenia synonimów nazw rekordów



- Prawie wszystko!
- Wydobycie streszczenia artykułu
 - Wikitekst jest przygotowany pod kątem prezentacji, a nie struktury tekstu
 - Trzeba pomijać tabele, jeżeli nie są one używane do formatowania strony
 - Jeśli użyto zwykłej tabeli zamiast szablonu, to z artykułu nie wyciągniemy wartości pól tabelarycznych



- W znacznym stopniu nastawiony na prezentację, a nie na strukturę
- Język „dla człowieka” a nie „dla komputera”: znaczenie symboli zależy od kontekstu
- Nieudokumentowane zachowania, które trzeba odtworzyć, aby poprawnie przetwarzać większość artykułów
- Składnia złożonych tabel jest trudna do parsowania (gdy nas interesuje struktura, a nie prosta zamiana na odpowiedniki HTML)



- Chcieliśmy je wykorzystać tak jak przekierowania do znajdowania synonimów, ale:
 - Brakuje ustalonej składni
 - Większość heurystyk zawodziła dla pewnych stron ujednoznaczniających, produkując niekiedy zabawne synonimy (Francja → troglodyta)
- Obecnie na podstawie ujednoznaczeń rozpoznajemy tylko skróty (łatwo sprawdzić czy synonim pasuje do hasła, np. Portable Document Format → PDF)



- Służą do rozróżniania rekordów o identycznej nazwie
- Określają relację typu „X jest Y”, np. „Lublin jest miastem”
- Nie tworzą hierarchii
- Jest kilkadziesiąt rozłącznych typów, w tym typ „inne”



- Służą do porządkowania rekordów przy przeglądaniu
- Określają relację typu „X kojarzy się z Y”, np. gra go kojarzy się z innymi grami oraz z Japonią
- Tworzą graf skierowany
- Tysiące kategorii od bardzo ogólnych („Geografia”) do bardzo szczegółowych („1995 w grach komputerowych”)
- Każdy artykuł może należeć do wielu kategorii, jednych całkiem oczywistych, innych mniej



- Zamiana **kategorii** Wikipedii na **typy** Miksera
 - Wyrażenia regularne dopasowywane do kategorii wiki
 - Niektóre **kategorie** można przypisywać do **typów** Miksera wraz z **podkategoriami**, innych lepiej nie:
 - „Dzieci” obejmuje m.in. „Aktorzy dziecięcy” ale też „Lektury szkolne”
 - „Miasta wojewódzkie” → „Lublin” → „Ludzie urodzeni w Lublinie”
- Nazwy kategorii często się zmieniają, trudno nad tym zapanować. Często potrzebne są zmiany w konfiguracji konwertera.



- Wikipedia jako źródło
 - Wartościowe dane z wielu dziedzin, gotowe wyniki dla konkretnych zapytań
 - Przyjazna licencja
 - Musimy popracować nad automatycznym przetwarzaniu wikitekstu
 - Tempo zmian czasami przyprawia o zawrót głowy



Dziękujemy 😊