

# XML i nowoczesne metody zarządzania treścią

Wykład 13: Wyszukiwanie informacji

Maciej Ogrodniczuk

MIMUW, 9 stycznia 2011

Tematem wykładu nie będzie wyszukiwanie w Internecie (w sensie „składni Google’a” itp.) ani w bazie danych, lecz tzw. *enterprise search*, czyli indeksowanie, przeszukiwanie i prezentacja dokumentów firmowych uprawnionym użytkownikom.

Idea: wyszukiwarka jest najważniejszym komponentem systemu zarządzania dokumentami.

Specyfika (w porównaniu z wyszukiwarką internetową):

- ograniczony zakres (zwykle intranet, znane prawa dostępu, ...)
- możliwość wykorzystania firmowych standardów (np. rodzaju używanych metadanych) – ale niekoniecznie,
- ranking dokumentów nie aż tak istotny,
- konieczność przeszukiwania wielu źródeł, w wielu formatach,
- lepsza aktualność listy wyników,
- brak śmieci!

Z punktu widzenia użytkownika:

- cel: „znajdź mi najlepszą odpowiedź na zadane pytanie”,
- wprowadź zapytanie,
- oczekuj na listę wyników.

W systemie:

- zanalizuj zapytanie,
- uzyskaj dostęp do danych,
- zanalizuj je,
- utwórz listę wyników zgodnych z zapytaniem, uwzględniając uprawnienia użytkownika,
- uporządkuj listę wyników,
- zaprezentuj ją użytkownikowi.

# Jak naprawdę działa wyszukiwanie?

Naiwne wyobrażenie o wyszukiwaniu:

- 1 przeglądamy dokumenty jeden po drugim,
- 2 jeśli w danym dokumencie znajduje się poszukiwany element (metadana, wyraz, fraza, wzorzec) – powiadom użytkownika,
- 3 jeśli wyniki mają być uporządkowane, zbieraj je w jakiejś strukturze danych i wyświetl po przejrzaniu wszystkich.

Działa dla 100 dokumentów. A dla 100.000?

Dużych ilości danych nie da się szybko przeszukiwać bez wstępnego **zindeksowania** treści.

**Indeks** to struktura danych porównywalna z wykazem rzeczowym w książce, zoptymalizowana pod kątem jej przeszukiwania, typowo zawierająca informację:

- o wystąpieniu słowa w dokumencie,
- zwykle także o miejscu jego wystąpienia.

## Ważne:

- indeks musi być aktualny – to na jego podstawie użytkownik dostanie wyniki,
- częstość i sposób synchronizacji indeksu ze stanem faktycznym zależy od zastosowania i możliwości technicznych:
  - często nie da się przyrostowo, tylko trzeba całościowo,
  - często nie da się na bieżąco, tylko trzeba cyklicznie,
- proces indeksowania może generować dodatkowe dane przydatne podczas wyświetlania wyników (np. streszczenia dokumentów).

# Czy indeks powinien zawierać wszystkie słowa?

**Słowa nieznaczące** (ang. *stopwords*) to w kontekście wyszukiwania często występujące słowa języka nie niosące samodzielnie żadnego znaczenia (przedimki, spójniki, przyimki, zaimki...)

Czy wyszukiwać słowa nieznaczące?

- samodzielnie – na pewno nie, bo wystąpią w (prawie) każdym dokumencie,
- + to dlaczego Google znajduje?

Czy indeksować słowa nieznaczące?

- bez nich indeks będzie mniejszy, a wyszukiwanie prostsze (po usunięciu ich z zapytania),
- + bez nich nie wyszukamy fraz takich jak „Take That” czy „The The”.

Dobra rada: dobrze, by podsystem wyszukiujący pozwalał na definiowanie listy słów nieznaczących.

# Co to znaczy „słowa”?

Podział tekstu na jednostki podstawowe (np. dla celów umieszczenia w indeksie) to **segmentacja tekstu** (tokenizacja).

Segmenty tekstu mogą być w niektórych przypadkach krótsze niż słowa ortograficzne („od spacji do spacji”).

Problemy tokenizacji:

- ważna znajomość języka tekstu (brak spacji w językach azjatyckich...),
- ważna znajomość kodowania znaków (dla oddzielenia znaków białych od zawartości – vide U+00A0),
- jaki znak dzieli segmenty? (→ *can't*, *długośmy*, *przyszedłby*, *doń*, *polsko-niemiecki*, *pięćdziesięciotowy*),
- czy spacja zawsze dzieli segmenty? (*dwadzieścia pięć*),
- co z datami? liczbami? walutami?
- co ze skrótami (prof. = profesor?)
- ...

# Co począć z odmianą wyrazów?

Problem: chcemy się dowiedzieć „*czym karmić kota*” – wpisujemy w wyszukiwarkę i...

Nie znajduje się, bo w dokumencie było „*Czym karmić koty*”.

Placebo: użyć symboli wieloznacznych (ang. *wildcards*) w zapytaniu: „*Czym karmić kot\**”.

Panaceum: wyszukiwanie z fleksją (stemming, lematyzacja).

W indeksie zapisywane są gramatyczne formy podstawowe kolejnych słów (np. mianownik rzeczownika, bezokolicznik czasownika). W ten sam sposób normalizowane jest zapytanie.

Problem nie jest trywialny, zwłaszcza po polsku:

„*Dudek, obciąć pensję*” vs. „*Real obetnie pensję Dudkowi*”.

Uwaga: zmiany w indeksie mogą sprawić, że niektóre słowa staną się nieodróżnialne (ang. *bore* = nudziarz vs. *bear* = niedźwiedź, Google: *zęby* vs. *żeby*)!

Najlepszy analizator morfologiczny dla zainteresowanych: <http://sgjp.pl/morfeusz/>

Uwaga oczywista: potrzebujemy modułów dla każdego języka:

- stemujemy dokumenty,
- stemujemy zapytanie,
- dopasowujemy zapytanie do dokumentów.

Uwaga mniej oczywista: za pomocą którego modułu stemować?

Możemy wykryć język dokumentu (za pomocą metod statystycznych – działa niemal idealnie):

- badając występowanie słów charakterystycznych dla języka,
- badając występowanie n-gramów (zbitek n-literowych),
- porównując kompresowalność tekstu ze wskaźnikiem dla tekstu porównawczego,
- ...

I musimy jeszcze tylko wykryć (ustawić ręcznie, przypisać domyślny) język zapytania, co może być trudne dla krótkich tekstów (vide *Stare forty*).

Jak to działa?

- użytkownik wprowadza zapytanie,
- system „zgaduje”, o co mogło chodzić użytkownikowi – na bieżąco lub po wyświetleniu wyników (gdy ich brak lub nie może znaleźć wyników spełniających podane kryteria).

Dwa sposoby korekty zapytania:

- automatyczny, poprzez zamianę błędnie napisanego słowa na jego poprawną wersję,
- półautomatyczny, poprzez wyświetlenie użytkownikowi poprawnej pisowni i umożliwienie mu ponownego wykonania wyszukiwania z użyciem poprawnej wersji zapytania.

Zwykle sposób kontroli pisowni można konfigurować regulując:

- 1 stopień podobieństwa dwóch słów, który musi być zachowany, by można było przyjąć, że jedno jest złą pisownią drugiego,
- 2 stopień podobieństwa, przy którym zła pisownia winna być automatycznie poprawiana.

Jak konstruuje się podpowiedzi?

- na bazie listy słów z tekstów indeksowanych,
- ewentualnie wpisów z oddzielnego słownika ortograficznego.

Żelazna zasada: nie podpowiadaj niczego, czego nie da się znaleźć.

Źródła tekstu to nie tylko pliki, ale także:

- zawartość dostępna online,
- pola CLOB w bazie danych (pobierz tekst, użyj nazw kolumn do ekstrakcji metadanych),
- maile,
- ...

dlatego oprócz samego silnika wyszukiwarki ważny jest też interfejs dostępu do danych.

Podobnie z ekstrakcją tekstu:

- prosta dla plików czysto tekstowych,
- bardziej złożona dla innych formatów tekstowych (np. usuń znaczniki HTML-owe, weź pod uwagę atrybut alt i metadane ze znaczników <meta>; nie popsuj polskich liter w PDF-ach),
- wymagająca użycia dedykowanych narzędzi dla pozostałych formatów.

Co to znaczy, że dokument „pasuje” do zapytania?

Pozycja wyniku na liście odzwierciedla ważoną średnią wielu parametrów – często nieoczywistych:

- wystąpienia słów w dokumencie,
- wystąpienia słów w metadanych, tekstach linków (*PageRank*),
- kto da więcej?
- ...

Najpopularniejszy sposób reprezentacji dokumentów i zapytań: *model przestrzeni wielowymiarowej* (ang. *vector space model*), którą tworzą wszystkie słowa zawarte w indeksowanych dokumentach (każde słowo jest osobnym wymiarem tej przestrzeni).

Dokument (zbiór słów) i zapytanie mogą zostać przedstawione jako wielowymiarowe wektory, których współrzędne przechowują informację o występowaniu poszczególnych słów w treści dokumentu/zapytania (wystąpienie oznacza niezerową wartość). Przy takiej reprezentacji jakość wyniku jest funkcją odległości między *wektorem dokumentu* a *wektorem zapytania* – zgodnie z określoną metryką.

Dwie najpopularniejsze metryki:

- binarna – wektor zawiera wartość 1 dla każdego słowa zawartego w dokumencie i 0 dla wszystkich innych słów,
- TF/IDF – stosunek częstości wystąpień słów w dokumencie (ang. *Term Frequency*) do ich „istotności” – tym niższej, im więcej różnych dokumentów zawiera dane słowo (ang. *Inverse Document Frequency*).

Podobnie możemy obliczać odległość między wektorami dokumentów, tworząc listę *dokumentów podobnych* do danego.

# Kiedy wyszukiwanie pełnotekstowe to za mało

Wyszukiwanie pełnotekstowe działa najlepiej, gdy użytkownik jest w stanie „przewidzieć”, jakich słów użyto w przeszukiwanych dokumentach, a dokumenty opisane są dokładnie (→ ręcznie).

Niestety, rzadko kiedy jest tak pięknie:

- w zapytaniach używane są synonimy pojęć z dokumentów („błękitny sweterek”),
- dokumenty zawierają pojęcia szczegółowe bez wskazania szerszej klasy („KDL-19S5730E śnieży”), podczas gdy zapytania mogą być ogólne („metro na Węgrzech”),
- użytkownik nie do końca wie, czego szuka („TV nie działa”).

Rozwiązania:

- wyszukiwanie ze słownikiem synonimów,
- wyszukiwanie sterowane dialogiem,
- drzewa decyzyjne,

czyli ogólnie rzecz biorąc – wyszukiwanie z modelem (wiedzy).

Metadane tekstowe mogą zawierać dowolny tekst (np. tytuł artykułu) lub wartość z ograniczonego zestawu (dział gazety, państwo itp.) – słownika.

Mogą być wypełniane na dwa sposoby:

- „ręcznie”, tj. na podstawie pól dokumentu, pól formularza, ...
- automatycznie – w wyniku analizy zawartości dokumentu.

Zestaw metadanych wraz z opisem ich zawartości tworzy prosty model danych dla wyszukiwania, który może zostać użyty podczas:

- wyświetlania interfejsu użytkownika (wyszukiwarki zaawansowanej),
- dialogu z użytkownikiem w celu doprecyzowania kryteriów zapytania,
- filtrowania wyników,
- ...

Gdy użytkownik nie uwzględnił w swoim zapytaniu wszystkich kryteriów nakładanych na istotne metadane (**atrybuty** modelu), a wyników jest na tyle dużo, że warto doprecyzować zapytanie, system może przed wykonaniem wyszukiwania automatycznie wygenerować pytania uszczegóławiające.

Przykład modelu – oferty sprzedaży samochodów:

- *producent* – Audi, Fiat, ...
- *model* – zależny od producenta: A4, A6, A8, TT, ...
- *rok produkcji, cena, przebieg, kolor, ...*

Zapytanie: *Audi za mniej niż 30.000 złotych.*

Formularz umożliwiający doprecyzowanie kryteriów zawierający pytania zwrotne bazujące na zindeksowanych dokumentach:

- jaki model? A4, A6, A8? (bo nie ma TT w tej cenie),
- jaki rok produkcji?
- jaki przebieg? mniej niż 100K, 100K-200K, powyżej 200K?

Na podobnej zasadzie mogą zostać udostępnione filtry umożliwiające zawężanie listy wyników już zwróconej po wykonaniu zapytania.

Korzyść w porównaniu z wielokrotnym wypełnianiem formularza:

- dynamiczna, aktualna informacja o liczbie wyników w danej klasie,
- możliwość wyeliminowania pustych odpowiedzi – brak zgadywania.

**Drzewa decyzyjne** (ang. *decision trees*) to technika ułatwiająca interaktywną diagnozę trudnych do zdefiniowania problemów.

Drzewo jest grafem wyborów prowadzących do rozwiązania problemu:

- każdy węzeł w drzewie wyboru (miejsce rozgałęzienia) składa się z pytania i możliwych odpowiedzi,
- odpowiedzi mogą prowadzić do dalszych wyborów (czyli pytań) lub w końcu do rozwiązań.

Jak to działa?

- udzielenie odpowiedzi powoduje ustalenie wartości określonych atrybutów zapytania (jak dla wyszukiwarki zaawansowanej),
- już ustalone wartości mogą być dostępne jako parametry kolejnych zapytań,
- jeśli dialog nie doprowadzi do uzyskania rozwiązania, użytkownik może zwykle posłużyć się „standardowym” wyszukiwaniem.

Zastosowanie: szczególnie do rozwiązywania problemów wymagających diagnostyki:

- w centrach obsługi klienta (*call center, contact center*),
- serwisach,
- innych specjalistycznych zastosowaniach wymagających samoobsługi.

Dodatkowe możliwości drzew decyzyjnych:

- narzędzia graficzne do budowy drzew,
- wielojęzyczność na poziomie węzła,
- ścieżki zależne od roli/grupy użytkownika,
- odpowiedzi domyślne.

# Drzewa decyzyjne – przykład środowiska graficznego

The screenshot displays a graphical decision tree editor with a grid background. The tree starts at a root node 'power\_general' (blue box). A green question box asks 'Who is the producer of your appliance?'. From this question, several arrows lead to different paths: 'Sony', 'Samsung', 'Grundig', 'PVC', 'Panasonic', and a list box containing 'Philips', 'Sony', 'Toshiba', 'Panasonic', and 'Philips'. The 'Sony' path leads to a green question box: 'Switch off the <=>POWER SAVE MODE <=> Does it work now?'. From here, a 'Yes' path leads to a blue box 'Problem solved!', and a 'No' path leads to another green question box: 'Did you just install or move the recorder?'. From this second question, a 'No' path leads to a blue box 'Hand the recorder to your local dealer ...', and a 'Yes' path leads to a blue box 'The recorder was autom. switche...'. At the bottom, a 'Translation' window shows the German and English text for the first question.

	de	en
Question Who...		
Question	Von welchem Hersteller ist ihr Gerät?	Who is the producer of your appliance?
Process		
Role		
Comment		
Attribute	Att_Manufacturer	
Parent		

Złote rady z notatnika projektanta interfejsu:

## 1 Najważniejsza jest **łatwość użycia**:

- żaden system podpowiedzi nie zastąpi intuicji użytkownika, który chciałby „od razu móc wyszukiwać” ,
- warto zainwestować w grafika,
- warto przeprowadzić testy użyteczności ekranów (okna zapytania i listy wyników).

## 2 **Każdemu według potrzeb**:

- zdecydowana większość zapytań ma nie więcej niż 3 słowa, a tylko 5% z nich używa operatorów,
- ale: zaawansowani użytkownicy mogą potrzebować więcej, więc nie można rezygnować z „wyszukiwania zaawansowanego” .

## 3 **Metadane**:

- włącz najważniejsze z nich (np. typ/format dokumentu) nawet do najprostszego formularza wyszukiwania,
- nie przeładuj metadanymi pojedynczego wyniku.

## 4 Postać listy wyników:

- dołącz nagłówek: czego szukano, ile wyników, podpowiedzi pisowni itd.,
- ogranicz widok podstawowy do najważniejszych metadanych,
- podawaj zawsze rozmiar dokumentu oryginalnego (by ostrzec przed otwieraniem dużych dokumentów); jeśli wynikiem są multimedia, dołącz odtwarzacz,
- użyj kontekstu z podświetlonymi elementami zapytania, by zawsze było jasne, dlaczego właśnie ten wynik znalazł się na liście,
- zapewnij sortowanie, stronicowanie, grupowanie, filtrowanie wyników.

## 5 Mniej często znaczy więcej:

- pokaż różne aspekty danych już na pierwszej stronie nie zwiększając ich ilości, lecz dodatkowo je kategoryzując,
- ułatw wyszukiwanie w już zwróconych wynikach,
- używaj zakładek i filtrów do ograniczania zakresu wyszukiwania,
- utrzymanie taksonomii kosztuje, ale korzystanie z niej jest dobrze odbierane przez użytkowników.

## 6 Nie przesadzaj z wodotryskami:

- drzewo hiperboliczne jest dobre na demo, ale czy sprawdza się w codziennej pracy?
- jeden nowy komponent odkrywający nieoczywiste związki między danymi zawsze się obroni (ciekawy filtr metadanych, sensowne linki kontekstowe, ...)

## 7 Nie martw się malkontentami – używania każdego interfejsu, nawet najlepiej przetestowanego i najbardziej intuicyjnego, użytkownicy będą się musieli nauczyć.

# Wydajność i dostępność wyszukiwania (i innych)

Dwie podstawowe metody zwiększenia wydajności i dostępności:

- 1 zrównoleżenie instalacji,
- 2 jej rozdzielenie na moduły (np. modułu aktualizacji indeksu od zwracania wyników, podział indeksu na części...)

Ważne:

- zapewnienie stuprocentowej dostępności systemu jest praktycznie niemożliwe, a każde podwyższenie jej o jedno miejsce po przecinku (np. z 99% do 99,9%) generuje koszty o jeden rząd wielkości wyższe w stosunku do poniesionych poprzednio,
- przed rozpoczęciem podnoszenia wydajności zawsze rozważa się stosunek kosztów jej wprowadzenia do strat związanych z obniżoną wydajnością.

Inna ważna zasada projektowania bezpieczeństwa: nie przesadzać, podwyższając koszty rozwiązania obsługującego nierealistyczne zagrożenia.

Dwa podstawowe rozwiązania zapewniające kontrolę wyświetlania wyników:

- ① indeksowanie treści wraz z prawami dostępu (ang. *early binding*) – niedozwolone dokumenty nie trafiają w ogóle na listę wyników,
- ② wymuszanie praw na etapie dostępu do zasobu (ang. *late binding*):
  - a wyświetlanie wszystkich dokumentów na liście wyników, sprawdzenie uprawnień przy próbie dostępu,
  - b usuwanie niedozwolonych dokumentów z listy wyników jeszcze przed jej wyświetleniem.
- ③ oba naraz!

## Konsekwencje:

- ❶ szybkość (weryfikacja dostępu jest czasochłonna, warto ją robić offline), ale każda zmiana praw dostępu skutkuje zmianami w indeksie, co bywa czasochłonne! może istnieć moment, gdy użytkownik może uzyskać dostęp do dokumentu, do którego nie ma praw – bo w indeksie jeszcze się nie zmieniło...
- ❷ rzeczywiste odzwierciedlenie najświeższych praw dostępu, indeks niezależny od komponentu autoryzacyjnego, ale wcześniejsze uwzględnienie praw mogłoby zmniejszyć porcję indeksu do przejrzania...
  - a wydaje się najbezpieczniejsze, ale czy chcemy pozwolić Kowalskiemu na znalezienie dokumentu  
Wypowiedzenie-Kowalski-1-lutego.doc?
  - b świetny pomysł, ale często trzeba jeszcze czyścić listy dokumentów powiązanych, przeliczać liczby wyników...  
co z odpowiedziami?

Dlaczego nie mogę znaleźć tego, czego szukam?

Rady dla użytkownika:

- może zrobiłem literówkę?
- użyłem synonimu?
- użyłem zbyt ogólnego zapytania?
- użyłem zbyt szczegółowego zapytania?
- problem z językiem?
- nie mam uprawnień do wyświetlenia wyników?

Rady dla administratora:

- analizuj logi:
  - jakiego rodzaju zapytania pojawiały się najczęściej?
  - jakie błędy użytkownicy popełniali najczęściej?
  - jakiego rodzaju zapytania nie zwróciły wyników?
- dodaj moduły, które mogą pomóc użytkownikom.

# Jak wybrać najlepsze narzędzie wyszukiwawcze?

Trzy podstawowe typy systemów wyszukiwawczych:

- programowe,
- sprzętowe (ang. *search appliances*),
- zdalne (ang. *remote search services*).

Na czym nam właściwie zależy?

- jakie metody wyszukiwawcze?
- elastyczny interfejs, łatwo integrowalny z firmową witryną?
- tylko formularz i lista wyników, czy także API?
- dostępność wyników w formacie XML?
- jakie moduły? czy można łatwo doimplementować nowe?
- jakie źródła dokumentów?
- jaka szybkość, skalowalność, metody równoważenia obciążenia, ...

Jak je porównać?

- zindeksować reprezentatywne źródła danych,
- porównać listy zindeksowanych dokumentów,
- porównać listy wyników dla danego zestawu zapytań testowych.

- podejście ewolucyjne: na początek zindeksuj najważniejszy serwer, potem „cały intranet” ,
- monitoruj, co robią użytkownicy i rozszerzaj zakres,
- pomyśl, co indeksować – najlepiej wszystko, bo nie wiadomo, co przyda się użytkownikom, ale może dla uniknięcia śmietnika warto wybrać (stworzyć?) serwer zawierający najważniejsze informacje?
- rozsądne minimum:
  - systemy plików, serwery sieciowe, archiwa mailowe,
  - serwery wiki, systemy śledzenia błędów (bugzilla i spółka),
  - artykuły portalowe, zawartość tekstową baz danych,
  - dane programów specjalistycznych, o ile dostarczają właściwych interfejsów,
- przemyśl strategię aktualizacji indeksu – co 5 minut czy w nocy? przesyłając dane do procesów indeksujących (*push*) czy pozwalając na ich niezależne ściąganie (*pull*)?

Ważne pytania:

- czy jest możliwe usuwanie treści z indeksu? jak to zrobić?
- jak indeksowana jest nowa treść? jak aktualizuje się indeks?
- jak system radzi sobie z problemami technicznymi (np. odwołaniami do nieosiągalnych źródeł zewnętrznych)?
- jak zmiany w taksonomii wpływają na działanie i aktualność systemu?
- czy można (i jak) dodawać nowe metadane? nowe typy dokumentów? czy to wymaga przebudowy indeksu? zmiany interfejsu?

Uwaga: zmiany w modelu są operacją potencjalnie kosztowną – aby nowe atrybuty i wartości mogły zostać uwzględnione w wyszukiwaniu, trzeba przeindeksować całe repozytorium!

## Ważne pytania:

- czy temat w ogóle nas dotyczy?
- jak szybko zwiększa się liczba naszych dokumentów?
- jak szybko przyrasta liczba zapytań?
- jakie możliwości skalowania oferuje system wyszukiwawczy?
  - wyłącznie poprzez zakup wydajniejszego sprzętu?
  - poprzez rozdzielenie modułów na różne maszyny?
  - poprzez równoważenie obciążenia poszczególnych modułów przez wiele maszyn – które procesy (indeksowanie, wyszukiwanie)?

OSW to polski sztab analityków monitorujących sytuację polityczną, ekonomiczną i społeczną w państwach wschodnich.

Główne założenia dla systemu:

- ułatwienie przeszukiwania wielojęzycznej bazy dokumentów (co najmniej: PL, EN, DE, RU, UA),
- spójne udostępnienie różnorodnych źródeł danych (dokumenty własne, serwisy informacyjne, portale o tematyce wschodniej),
- możliwość uzyskiwania odpowiedzi w ciągu kilku sekund,
- użycie narzędzi lingwistycznych i zaawansowanych technik wyszukiwawczych,
- wyszukiwarka dostępna z poziomu przeglądarki WWW z portletowym GUI,
- możliwości personalizacyjne,
- system praw dostępu.

Efekt:

- integracja różnych źródeł danych w pojedyncze repozytorium,
- dwa tryby dodawania dokumentów:

### Wyszukiwanie dokumentów

Zapytanie:  Szukaj Wyczyść

Szukaj:  wszystkich wyrazów  któregośkolwiek z wyrazów

Język zapytania:

Teksty własne  Teksty własne tajne  Biogramy  Bazy działowe  Wiadomości  Subskrypcje  Agencje informacyjne

Sortowanie wg: **trafności** | daty

Liczba wyników: 1032 Znznaczonych: 0

1 2 3 4 5 6 7 8 9 10 1 · 10

**Niemcy/Merkel: Rosja solidnym partnerem**

5.3.Berlin (PAP) - Kanclerz Niemiec **Angela Merkel** uważa **Rosję** za solidnego partnera Unii Europejskiej. "Stawiamy na to, że **Rosja** jest solidnym partnerem. Dlatego popieram zasadniczo rosyjskie zaangażowanie gospodarcze w Europie" - powiedziała **Merkel** w wywiadzie opublikowanym w poniedziałek na łamach dziennika "Süddeutsche Zeitung". Dodała, że Europejczycy są z kolei zainteresowani...

98% Agencje informacyjne, PAP/Zagranica TEXT 1 KB Data mod.: 2007-03-05 10:48

**Rosja/ Merkel: problem polskiego mięsa wciąż nierozwiązany**

21.1.Moskwa (PAP/AFP) - Kanclerz Niemiec **Angela Merkel** przyznała w niedzielę, że problem zakazu importu polskiego mięsa do **Rosji** jest wciąż nierozwiązany. Wyraziła jednak nadzieję, że rozmowy w sprawie porozumienia o partnerstwie UE-**Rosja** rozpoczną się przed końcem pierwszego półrocza tego roku. (PAP) mw/ ap/ 3847 18:18 07/01/21

97% Agencje informacyjne, PAP/Zagranica TEXT 386 B Data mod.: 2007-01-21 18:23

**Niemcy/ Merkel zaniepokojona rozwojem demokracji w Rosji**

7.1.Berlin (PAP/AP) - Kanclerz Niemiec **Angela Merkel** w rozmowie z tygodnikiem "Der Spiegel" wyraziła zaniepokojenie rozwojem demokracji w **Rosji**. Wyniada okazać się w poniedziałek. "Tę możemy (...) przenieść naszej koncepcji demokracji do **Rosji**" - powiedziała **Merkel**. - Jednocześnie, przynajmniej, dochodzi (w **Rosji**) do zdarzeń, które postrzegam jako powód z troską, na przykład...

96% Agencje informacyjne, PAP/Zagranica TEXT 1 KB Data mod.: 2006-01-07 19:15

**Rosja/ Merkel o energetyce i kolejnym szczycie G-8**

17.7.Petersburg (PAP/Reuters) - Kanclerz Niemiec **Angela Merkel** powiedziała w poniedziałek na swej konferencji prasowej po szczycie G-8, że Niemcy nie obawiają się zależności energetycznej od **Rosji**. Dodała, że mimo zaufania do **Rosji** jako dostawcy, Niemcy rozważają dywersyfikację swoich źródeł energii. Kanclerz **Merkel** zapowiedziała, że wśród tematów kolejnego...

96% Agencje informacyjne, PAP/Zagranica TEXT 777 B Data mod.: 2006-07-17 17:15

**Premier rozmawiał z Angielą Merkel**

19.5.Warszawa (PAP) - O stosunkach Unii z **Rosją** rozmawiał w sobotę po południu premier Jarosław Kaczyński z kanclerz Niemiec **Angielą Merkel**. Porównował PAP rzecznik rządu Jan Dziędziczak. "Premier odbył ponad pół godzinny rozmów telefoniczny z niemiecką kanclerz **Angielą Merkel**. Przedmiotem rozmowy były najbardziej aktualne kwestie związane z relacjami UE...

96% Agencje informacyjne, PAP/Nrca TEXT 955 B Data mod.: 2007-05-19 14:19

**Rosja/ Merkel chce niezawodność dostaw rosyjskich ropy naftowej**

energi do Europy (krótka2) 21.1.Moskwa (PAP) - Kanclerz Niemiec **Angela Merkel** powiedziała w niedzielę na konferencji prasowej w Soczi po spotkaniu z prezydentem Władimirem Putinem, że rozumie dążenie **Rosji** do wprowadzenia zasad rynkowych w handlu ropy naftową i gazem ziemnym z Ukrainą i Białorusią. Jednocześnie **Merkel** podkreśliła, że Niemcy opowiadają się za niezawodnością rosyjskich dostaw ropy naftowej do Europy. Niemiecka kanclerz dodała, że miała z Putinem "dobrą rozmowę" o tym, jak Europa i **Rosja** powinny lepiej się komunikować. W jej opinii, Europa i **Rosja** są sobie wzajemnie potrzebne. (PAP) mal/ ap/ 17:29 07/01/21

96% Agencje informacyjne, PAP/Zagranica TEXT 705 B Data mod.: 2007-01-21 17:33

**Angela Merkel pojedzie do Moskwy na początku 2006 r.**

25.11.Berlin (PAP/AFP.dpa) - Kanclerz Niemiec **Angela Merkel** uda się do Moskwy na początku 2006 r., aby spotkać się z prezydentem **Rosji** Władimirem Putinem - ogłosił w piątek rząd niemiecki. W piątkowej rozmowie telefonicznej **Angela Merkel** i Władimir Putin "potwierdził, że doskonale stosunki niemiecko-rosyjskie powinny być pogłębiane i rozwijane się" - dodała kanclerka niemiecka. Prezydent Putin był jednym z pierwszych szefów państw, który przedstawił nowej kanclerz gwałtownie wyrażając "wiarę" w "pogłębianie partnerstwa strategicznego" między dwoma krajami.(PAP) kd/ ro/ 3453 15:03 05/11/25

96% Agencje informacyjne, PAP/Zagranica TEXT 695 B Data mod.: 2005-11-25 15:10

1 2 3 4 5 6 7 8 9 10 1 · 10

### Filtry

Wyszukiwanie we wszystkich dokumentach.

**Bazy**

- Bazy
- Teksty własne
- Biogramy
- Wiadomości [49]
- Bazy działowe
- Subskrypcje i agencje [983]
  - PAP [982]
  - Interfax [1]

**Regiony**

- Świat
- Azja [318]
- Azja Centralna [68]
- Kazachstan [38]
- Kirgistan [10]
- Tadżykistan [5]
- Turkmenistan [0]
- Uzbekistan [7]
- Kaukaz [113]
- Turcja [93]
- Ankara [741]

**Daty**

- Daty
- 2005 [241]
- 2006 [381]
- 2007 [410]
- 01 [96]
- 02 [39]
- 03 [56]
- 04 [26]
- 05 [82]
- 06 [02]
- 07 [741]

**Języki**

PL [1031]  EN [0]  DE [0]  RU [1]  inny [0]

10 przykazań Google'a odnośnie wyszukiwania korporacyjnego:

- 1 Poprawiaj zadowolenie użytkowników – jeśli wyszukiwarka pomoże im w pracy, nie będą pamiętać, ile kosztowała. Pytaj użytkowników, co jeszcze możesz dla nich zrobić.
- 2 Przyspieszaj wyszukiwanie – łatwość zadawania zapytań i szybkość pozyskania wyniku zachęcają do użycia wyszukiwarki.
- 3 Promuj wyszukiwanie – staraj się, by pole zapytania było wszechobecne, dostępne na każdej stronie serwisu intranetowego firmy oraz zawsze także na liście wyników, zachęcając do użycia wyszukiwarki i precyzowania kryteriów zapytania.
- 4 Odchudź strony zapytania i wyników – niech główna strona zapytania pozostanie prosta. Przenieś dodatkowe opcje na inną stronę, ogranicz do minimum niezwiązane z wyszukiwaniem komponenty nawigacyjne na stronie wyników.

Z listu prof. Marciszewskiego:

*W tym semestrze ankiety oceniające zajęcia ponownie będą przeprowadzone w formie elektronicznej.*

*Ankiety będą dostępne dla studentów w USOS-ie od 9 do 22 stycznia.*

*Pisałem o tym do wszystkich studentów, ale również proszę Państwa o zachęcanie uczestników zajęć do wzięcia udziału w ankietach.*

Egzamin odbędzie się **2 lutego** br. (czwartek) w godz. 10-13 w sali 2180 (drugi termin:  $\approx$  początek marca).

O czym mówiłem na początku:

- dopuszczenie do egzaminu wymaga zaliczenia pracowni,
- ocena z pracowni przekłada się na punkty,
- ocena z przedmiotu wynika z sumy punktów z egzaminu i pracowni.

I jeszcze:

- egzamin będzie się składać z 16 pytań testowych wielokrotnego wyboru (prawdziwa co najmniej 1 z 4) lub opisowych, nie wolno korzystać z materiałów,
- pytanie testowe jest zaliczone, gdy zaznaczone są wszystkie poprawne odpowiedzi oraz nie jest zaznaczona żadna niepoprawna,
- każde pytanie jest warte 1 pkt, w przypadku pytań opisowych możliwe są także oceny 0,5 pkt.

# Czego się spodziewać (na przykładzie poprzednich edycji)?

## Pytania, jakich nie lubię:

Które z następujących języków są zastosowaniami SGML-a?

- a) HTML (HyperText Markup Language),
- b) XML (Extensible Markup Language),
- c) CALS (Computer-Aided Acquisition and Logistic Support),
- d) DSSSL (Document Style Semantics and Specification Language).

## Pytania, jakie lubię:

Przy pomocy DTD nie można:

- a) zadeklarować, że zawartość elementu musi być liczbą,
- b) zadeklarować elementu zawierającego sekwencję określonych podelementów, występujących zawsze w określonej kolejności,
- c) zadeklarować elementu, który jest opcjonalny,
- d) zadeklarować elementu o zawartości mieszanej (ang. *mixed content*).

Co do tej pory mogło kojarzyć nam się z zarządzaniem wiedzą?

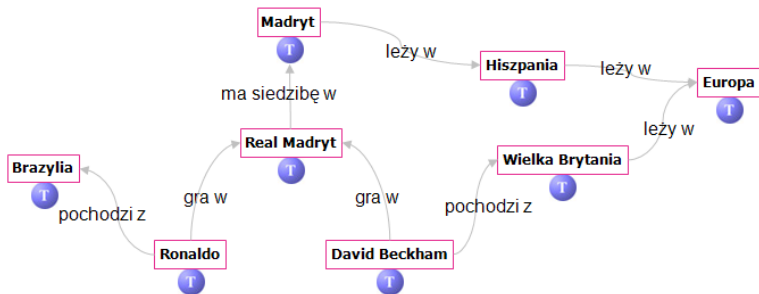
- DITA?
- moduł synonimów w wyszukiwarce (szukam „błękitnej bluzeczki”, znajduje się „niebieski sweterek”)?
- model podobieństw (szukam Toyoty Yaris, znajduje się Opel Corsa)?

Co można rozumieć jako zarządzanie wiedzą?

- dobre praktyki zarządzania przedsiębiorstwem?
- kulturę organizacyjną?
- rozwiązania technologiczne usprawniające pracę (CMS, portal korporacyjny, ...)?
- różne obszary zainteresowań sztucznej inteligencji (automatyczne wnioskowanie, uczenie maszynowe, systemy eksperckie)?

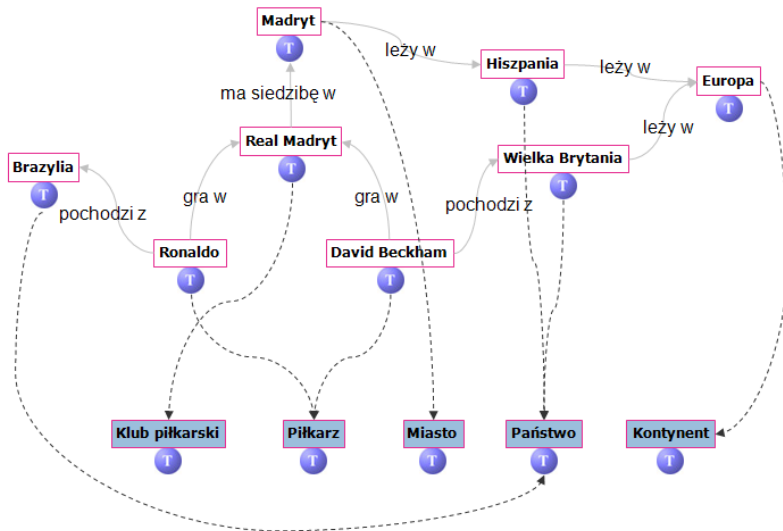
**Czym tak naprawdę jest wiedza?**

# Intuicyjny model wiedzy – siatka pojęć

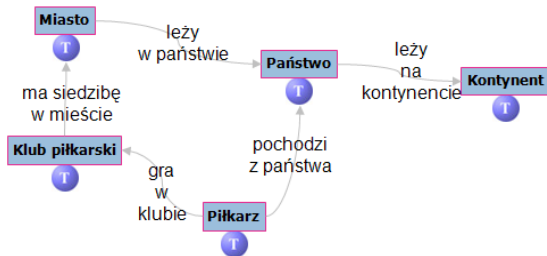


Niektóre problemy:

- płaski model: pojęcie „Europa” jest słabo odróżnialne od pojęcia „Ronaldo”,
- siatka może się rozrastać w niekontrolowany sposób!



Relacje między klasami są abstrakcją relacji pomiędzy pojęciami:



klasy + relacje = schemat mapy wiedzy (ontologia)

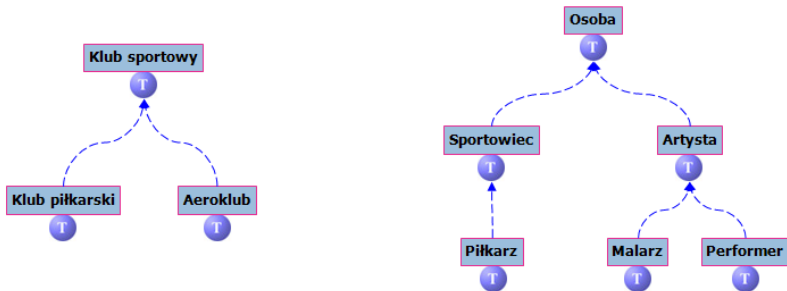
Po co tworzyć schematy?

- aby wyrazić strukturę informacji i współdzielić jej rozumienie pomiędzy ludźmi lub automatami (→ łatwe zbieranie danych, tworzenie podsumowań itp.),
- aby mieć możliwość wielokrotnego wykorzystania spójnych „paczek wiedzy” ,
- aby dokonać analizy wiedzy danej dziedziny w interesującym nas aspekcie.

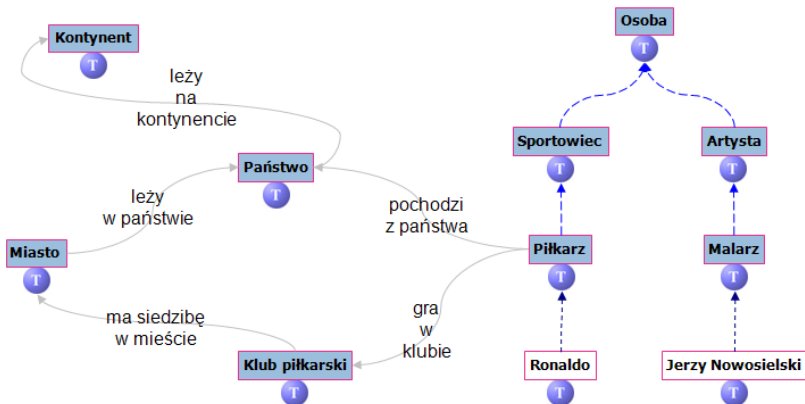
Uwaga:

- nie ma jedyne właściwego schematu dla danej dziedziny wiedzy!
- to sposób wykorzystania wiedzy wpływa na schemat i stopień jego szczegółowości.

Klasy możemy hierarchizować:



# Mapa wiedzy = schemat mapy wiedzy + instancje



## Użyteczność mapy wiedzy:

- wyszukiwanie:
  - konfrontacja zapytania z modelem wiedzy (Groclin – firma giełdowa i klub sportowy), możliwość uszczegóławiania zapytań na podstawie modelu wiedzy,
  - zawężanie zakresu poszukiwań na podstawie pojęć wybieranych z mapy,
- ułatwiona klasyfikacja: dołączanie dokumentów do mapy wiedzy na podstawie przekrojów mapy,
- unikalna nawigacja: dostęp do dokumentów poprzez sieć pojęć.

ISO 13250:2003 – standard reprezentacji i wymiany wiedzy.

Pomysł:

- utworzenie nad warstwą zasobów warstwy abstrakcyjnych **pojęć** (tematów, ang. *topics*) z możliwością tworzenia **powiązań** (ang. *associations*) między nimi,
- powiązanie obu warstw poprzez **wystąpienia** (ang. *occurrences*) pojęć w zasobach.

Najpopularniejsza notacja: XML Topic Maps (XTM) 2.0 z 2006 r.

- `<topicMap>` – korzeń dokumentu z definicją mapy pojęć,
- `<topic>` – nazwa i lista wystąpień pojęcia,
- `<instanceOf>` – informacja o powiązaniu pojęcia z klasą (pojęciem nadrzędnym); występuje w treści `<topic>`,
- `<topicRef>` – odwołanie do już zdefiniowanego pojęcia (np. w celu określenia klasy),
- `<occurrence>` – informacja o wystąpieniu pojęcia,
- `<resourceRef>` – odwołanie do zasobu (za pomocą URI),
- `<association>` – powiązanie między pojęciami,
- ...

<http://www.topicmaps.org/xtm/1.0/>

# XML Topic Maps – przykład

```
<topicMap>
  <topic id="kompozytor">
    <baseName><baseNameString>kompozytor</baseNameString></baseName>
  </topic>
  <topic id="chopin">
    <instanceOf><topicRef xlink:href="#kompozytor"/></instanceOf>
    <baseName><baseNameString>Fryderyk Chopin</baseNameString></baseName>
    <occurrence><resourceRef xlink:href="http://www.example.org/
      chopin.htm"/></occurrence>
  </topic>
  <topic id="polska">
    <instanceOf><topicRef xlink:href="#kraj"/></instanceOf>
    ...
  </topic>
  <association>
    <instanceOf><topicRef xlink:href="#urodzony-w"/></instanceOf>
    <member><roleSpec><topicRef xlink:href="#osoba"/></roleSpec>
      <topicRef xlink:href="chopin"/></member>
    <member><roleSpec><topicRef xlink:href="#kraj"/></roleSpec>
      <topicRef xlink:href="polska"/></member>
    </association>
</topicMap>
```

RDF – konkurencyjna (W3C, rekomendacja w 1999 r.) metoda definiowania wiedzy poprzez opis zasobów.

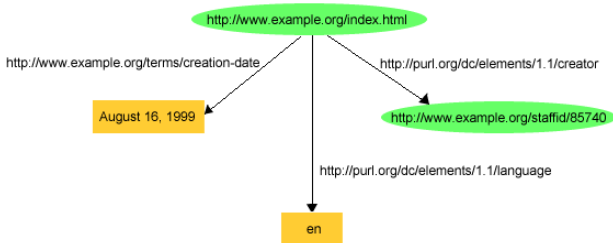
Reprezentacja wiedzy w RDF:

- zdania logiczne w postaci trójki (ang. *triple*) „podmiot-relacja-przedmiot” (np. „<Stanisław Lem> <jest-autorem> <Solaris>”),
- podmiot i przedmiot są zasobami,
- relacja (własność, ang. *property*) może być zasobem,
- skoro własność jest zasobem, można ją opisać inną własnością, czego wynikiem może być zaawansowany metagraf (węzły = zasoby, łuki = własności),
- rodzaje własności są nieograniczone.

Specyfikacja RDF definiuje sposób serializacji grafu do XML-a (RDF/XML). Zasoby identyfikowane są (oczywiście) URI.

<http://www.w3.org/RDF/>, <http://www.w3.org/TR/rdf-primer/>

# RDF – graf i jego serializacja

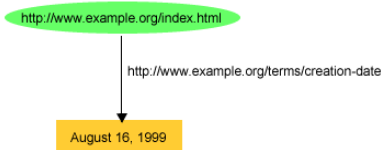


Notacja N3:

```
<http://www.example.org/index.html> <http://purl.org/dc/elements/1.1/
  creator> <http://www.example.org/staffid/85740> .
<http://www.example.org/index.html>
  <http://www.example.org/terms/creation-date> "August 16, 1999" .
<http://www.example.org/index.html>
  <http://purl.org/dc/elements/1.1/language> "en" .
```

Jeszcze prościej:

```
ex:index.html dc:creator exstaff:85740 .
ex:index.html exterms:creation-date "August 16, 1999" .
ex:index.html dc:language "en" .
```



Trójki:

`ex:index.html exterms:creation-date "August 16, 1999" .`

RDF/XML:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns"
  xmlns:exterms="http://www.example.org/terms/"
  <rdf:Description rdf:about="http://www.example.org/index.html">
    <exterms:creation-date>August 16, 1999</exterms:creation-date>
  </rdf:Description>
</rdf:RDF>
```

```
<!DOCTYPE rdf:RDF
  [<!ENTITY xsd
    "http://www.w3.org/2001/XMLSchema">]>
<rdf:RDF xmlns:rdf="http://www.w3.org/
  1999/02/22-rdf-syntax-ns"
  xmlns:prod="http://www.example.com/produkty/">
  <rdf:Description rdf:ID="item10245">
    <prod:model rdf:datatype="&xsd:string">Leader Price
      Magic Tent 2010</prod:model>
    <prod:osob rdf:datatype="&xsd:integer">2</prod:osob>
    <prod:waga rdf:datatype="&xsd:decimal">2,4</prod:waga>
    <prod:cena rdf:datatype="&xsd:decimal">9,99</prod:cena>
  </rdf:Description>
  ...
</rdf:RDF>
```

```
<rdf:RDF xmlns:rdf="http://www.w3.org/
          1999/02/22-rdf-syntax-ns"
          xmlns:prod="http://www.example.com/produkty/">
  <rdf:Description rdf:ID="item10245">
    <rdf:type rdf:resource="http://www.example.com
              /produkty/Namiot"/>
    <prod:model>Tesco Value Tent-0-Magic</prod:model>
    <prod:osob>2</prod:osob>
    <prod:waga>2,1</prod:waga>
    <prod:cena>19,99</prod:waga>
  </rdf:Description>
  ...
</rdf:RDF>
```

Albo w skrócie:

```
<rdf:RDF xmlns:rdf="http://www.w3.org/
          1999/02/22-rdf-syntax-ns"
```

RDF Schema to rekomendacja W3C z 2004 r. definiująca mechanizm opisu grup powiązanych zasobów oraz relacji między nimi za pomocą klas i własności:

- klasą jest zasób, dla którego zdefiniowano własność `rdf:type` o wartości `rdfs:Class`,
- własnością jest zasób, dla którego zdefiniowano własność `rdf:type` o wartości `rdfs:Property`,
- specjalizacja określana jest przechodnią relacją `subClassOf` (dla klas) i `subPropertyOf` (dla własności),
- wartości danej własności są instancjami określonej klasy, o ile zdefiniowano dla tej własności własność `rdfs:range` o wartości wskazującej tę klasę,
- dana własność może być przypisywana instancjom określonej klasy, o ile zdefiniowano dla tej własności własność `rdfs:domain` o wartości wskazującej tę klasę.

<http://www.w3.org/TR/rdf-schema/>

RDQL – język zapytań wzorowany na SQL.

Zapytanie:

```
SELECT ?x, ?fname
WHERE (?x, <http://www.w3.org/2001/vcard-rdf/3.0FN>,
      ?fname)
```

Wynik:

x		fname
=====		
<http://somewhere/JohnSmith/>		"John Smith"
<http://somewhere/RebeccaSmith/>		"Becky Smith"
<http://somewhere/SarahJones/>		"Sarah Jones"
<http://somewhere/MattJones/>		"Matt Jones"

Problem:

- w RDF można wyrazić dowolne własności,
- komunikacja przy pomocy RDF ma sens, jeśli partnerzy posługują się tym samym słownikiem.

RDF nie definiuje słownika, jedynie sposób zapisu metadanych!

Standardy oparte na RDF:

- Dublin Core,
- RSS (RDF Site Summary),
- OWL (Web Ontology Language).

Sformalizowany język do budowy ontologii; najnowsza wersja rekomendacji W3C z 27 października 2009 r.

## Podstawowe obiekty:

- Class,
- Property,
- Individual.

## Definiowanie własności:

- TransitiveProperty,
- SymmetricProperty,
- FunctionalProperty,
- inverseOf.

## Definiowanie klas:

- oneOf,
- intersectionOf,
- unionOf,
- własności instancji:
  - minCardinality,
  - maxCardinality.

Przykład ontologii: <http://www.w3.org/TR/2004/REC-owl-guide-20040210/wine.rdf>

Tim Berners-Lee, 2001:

*The Semantic Web will bring structure to the meaningful content of Web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users."*

Semantic Web to idea internetowej infrastruktury publikacji danych, neutralnej i umożliwiającej przetwarzanie informacji przez programy w celu automatyzacji, agregacji i wielokrotnego użycia.

To wciąż jedynie wizja:

- czy to w ogóle da się zrobić?
- kto opíše istniejące dane w lepszy, semantyczny sposób?
- czy to znaczy, że muszą istnieć dwie reprezentacje danych – dla ludzi i maszyn?
- a może zamiast dodatkowego opisu rozwinąć „rozumienie” istniejących opisów (np. HTML-a) przez maszyny?
- czy naprawdę chcemy, żeby automaty (= rządy, cenzura, ...) mogły zrobić wszystko to, co my?

Cory Doctorow, 2001: metadane nie pomogą w reprezentacji wiedzy, bo:

- 1 ludzie kłamią,
- 2 ludzie są leniwi,
- 3 ludzie są głupi,
- 4 poznać siebie: mission impossible,
- 5 schematy nie są neutralne,
- 6 metryka wpływa na wyniki,
- 7 zawsze jest więcej niż jeden sposób opisu.

<http://www.well.com/~doctorow/metacrap.htm>

Coś jednak się dzieje:

- [www.wolframalpha.com](http://www.wolframalpha.com),
- [www.trueknowledge.com](http://www.trueknowledge.com),
- [www.freebase.com](http://www.freebase.com),
- [www.dbpedia.org](http://www.dbpedia.org),
- <http://openmind.media.mit.edu/>,
- <http://www.mpi-inf.mpg.de/yago-naga/>,
- ...

A może Państwo zrobią to lepiej?

**There are three kinds of people:**

- ① those who make things happen,**
- ② those who watch things happen,**
- ③ those who wonder what happened.**

Mam nadzieję, że zawsze będę Państwo wśród pierwszych.