

# XML i nowoczesne metody zarządzania treścią

Wykład 15: Od treści do wiedzy

Maciej Ogrodniczuk

MIMUW, 21 stycznia 2009

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 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)?

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

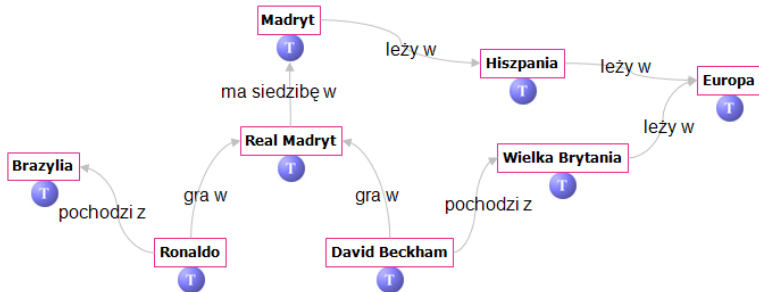
- DITA?
- moduł synonimów w wyszukiwarce (szukam „błękitnej bluzeczki”, znajduje się „niebieski sweterek”)?
- model podobieństw (szukam Toyoty Yaris, znajduje się Opel Corsa)?

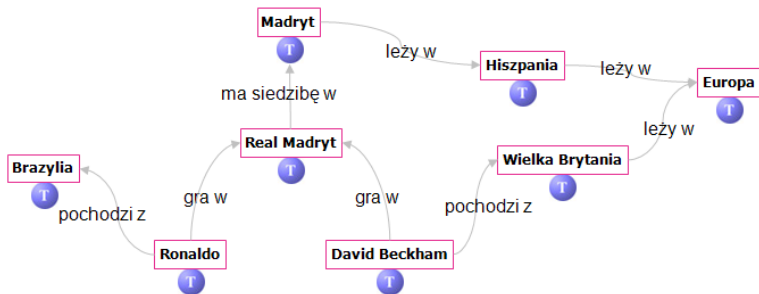
Co można rozumieć jako zarządzanie wiedzą?

- 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

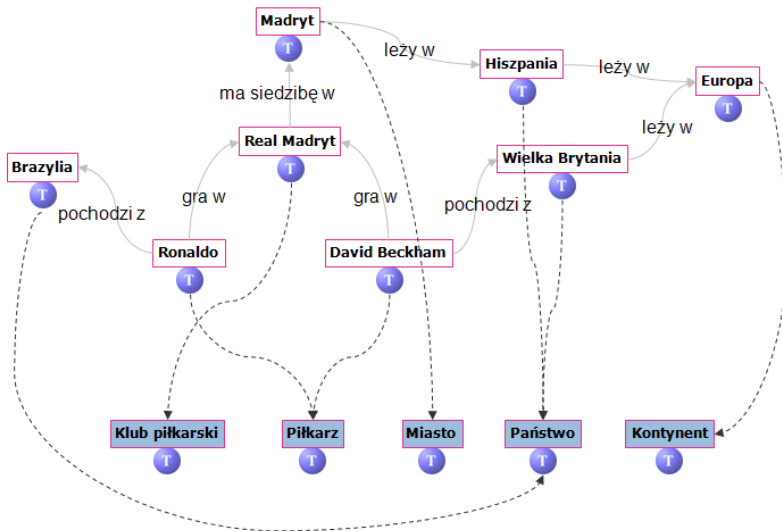




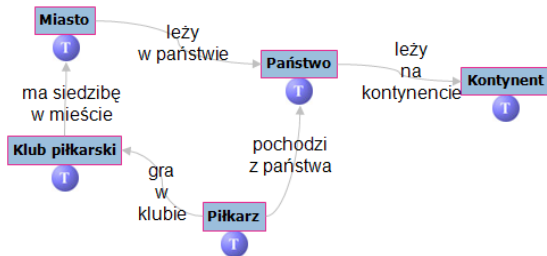
Niektóre problemy:

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

# Klasy pojęć



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.

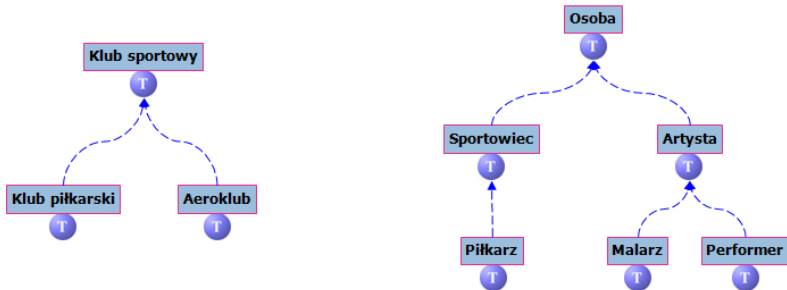
Po co tworzyć schematy?

- 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.

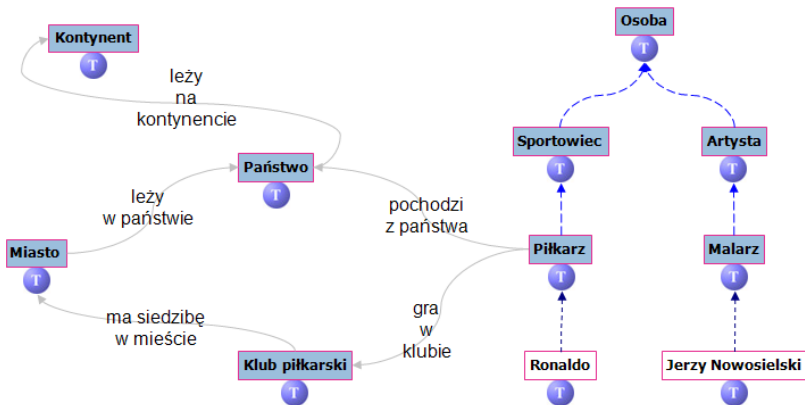
Uwaga:

- 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/>

```
<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.

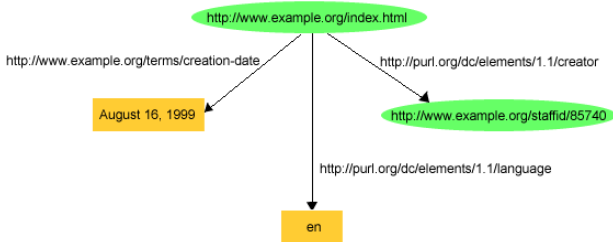
Reprezentacja wiedzy w RDF:

- 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.

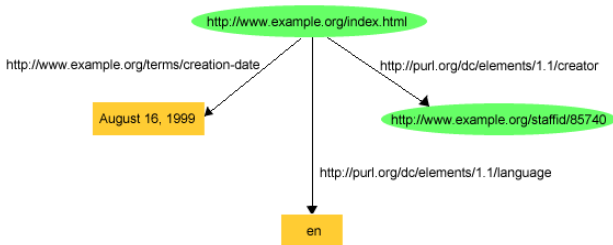
Specyfikacja RDF definiuje sposób serializacji grafu do XML-a (RDF/XML). Zasoby identyfikowane są (oczywiście) URI.

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

# RDF – graf i jego serializacja



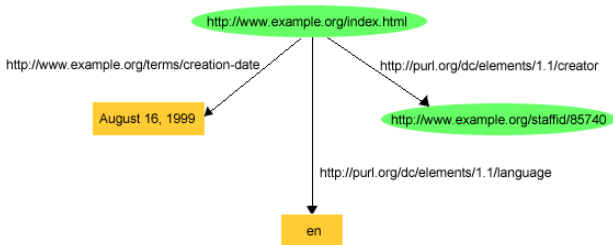
# RDF – graf i jego serializacja



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" .
```

# RDF – graf i jego serializacja

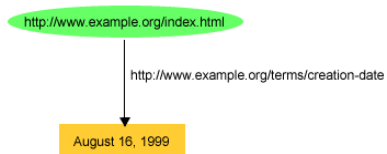


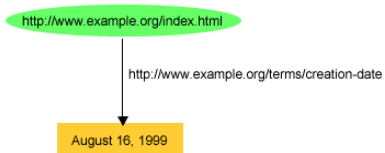
Notacja N3:

```
<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:

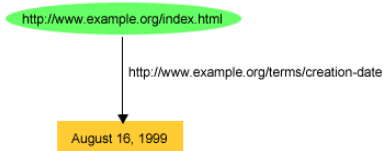
```
ex:index.html dc:creator exstaff:85740 .
ex:index.html exterms:creation-date "August 16, 1999" .
ex:index.html dc:language "en" .
```





Trójki:

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



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"
          xmlns:prod="http://www.example.com/produkty/">
  <prod:Namiot rdf:ID="item10245">
    <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>
  </prod:Namiot>
  ...
</rdf:RDF>
```

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.

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

Coś jednak się dzieje:

- [www.cuil.com](http://www.cuil.com),
- [www.wolframalpha.com](http://www.wolframalpha.com),
- [www.hakia.com](http://www.hakia.com),
- [www.trueknowledge.com](http://www.trueknowledge.com),
- [www.freebase.com](http://www.freebase.com),
- inne warte obejrzenia:
  - [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.**

Egzamin odbędzie się **2 lutego** br. (wtorek) w godz. 14-16 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,
- na ocenę z przedmiotu przekłada się suma punktów z egzaminu i pracowni.

O czym jeszcze nie mówiłem (ale jest jak dawniej):

- egzamin będzie się składać z 16 pytań testowych wielokrotnego wyboru (prawdziwa co najmniej jedna odpowiedź z czterech) lub opisowych,
- 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.

## 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).

# Czego się spodziewać (na przykładzie edycji 2008/09)?

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).

# Czego się spodziewać (na przykładzie edycji 2008/09)?

## 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*).

# Czego się spodziewać (na przykładzie edycji 2008/09)?

## 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*).

Z listu prof. Marciszewskiego:

*W tym semestrze ankiety oceniające zajęcia ponownie będą przeprowadzone w formie elektronicznej — nie będzie ankiet w wersji papierowej.*

*Ankiety będą dostępne dla studentów w USOS-ie do 24 stycznia.*

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