Reprezentácia a Získavanie Znalostí

Sémantický Web - úvod

Dáta – informácie - znalosti

Dáta (údaje)

15.12.1912

Informácia - *The term 'information' is described as the **structured, organised** and processed **data**, presented **within context**, which makes it relevant and useful*

Titanic sa potopil 15.12.1912

Poznatok *- je poznatok spracovaná informácia, produkt poznávacej činnosti.*

Každá loď sa môže potopiť

Znalosť *– štruktúrovaný systém poznatkov a skúseností v určitej oblasti*

Difference between information and knowledge

1. **Information denotes the organised data about someone or something** obtained from various sources such as newspaper, internet, television, discussions, etc. **Knowledge refers to the awareness or understanding on the subject** acquired from education or experience of a person.

2.Information is **nothing but the refined form of data**, which is helpful to understand the meaning. On the other hand, **knowledge** is the relevant and objective information that **helps in drawing conclusions.**

3. Data compiled in the meaningful context provides information. Conversely, when information is combined with experience and intuition, it results in knowledge.

4. Processing improves the representation, thus ensures easy interpretation of the information. As against this, processing results in increased consciousness, thus enhances subject knowledge.

5. Information brings on comprehension of the facts and figures. Unlike, knowledge which leads to the understanding of the subject.

6. The transfer of information is easy through different means, i.e. verbal or non-verbal signals. Conversely, the transfer of knowledge is a bit difficult, because it requires learning on the part of the receiver.

7. Information can be reproduced in low cost. However, exactly similar reproduction of knowledge is not possible because it is based on experiential or individual values, perceptions, etc.

8. **Information alone is not sufficient to make generalisation or predictions** about someone or something. On the contrary, **knowledge has the ability to predict or make inferences**.

9. Every **information is not necessarily a knowledge, but all knowledge is an information.**

Teória poznania – *epistemológia*

Epistemology is the study of knowledge

- Zaoberá sa tým čo je poznanie/poznatky a ako sa získava
- Poznakty (aj informácie) majú formu tvrdení *(propositional knowledge)*
- **Tvrdenie** *(proposition)*
  - je niečo čo môže byť vyjadrené deklaratívnym spôsobom *(syntax)*

  je oznamovacia veta
  - môže byť pravdivé alebo nepravdivé *(sémantika)*

  má pravdivostnú hodnotu

Tradične sa poznanie chápe ako pravdivé zdôvodnené presvedčenie

Poznatky - tvrdenia, ktoré považujeme za pravdivé

Syntax vs. sémantika

Syntax	(from greek $\sigma\upsilon\nu\tau\alpha\xi\iota\varsigma$ composition, sentential structure) denotes the (normative) structure of data, i.e., it characterizes what makes data "well-formed"
Semantik	(greek $\sigma\eta\mu\alpha\upsilon\tau\iota\kappa o\varsigma$ belonging to the sign) denotes the meaning of data, i.e., it characterizes what conclusions can be drawn from it.

$4+) = ($	$3 + 4 = 12$	$3 + 4 = 7$
syntactically wrong	syntactically correct	syntactically correct
–	semantically wrong	semantically correct

Znalosti a Informačné systémy

IS dokážu formálne pracovať s dátami a uchovávať informácie

Adam má kamaráta Borisa.
Boris má kamaráta Cyrila.
Cyril má kamaráta Dušana.
Dušan má kamaráta Emila.

Uložené v RDB

ID	OSOBA	KAMARAT
1	Adam	Boris
2	Boris	Cyril
3	Cyril	Dusan
4	Dusan	Emil

Ale

Kto sú Borisovi kamaráti?

Implicitné a explicitné znalosti

Mnohé informácie nie sú vyjadrené explicitne ale vyplývajú nepriamo z daných faktov a vedomostí.

Využitie IT pre ich získanie si vyžaduje:

- Metódy formálnej logiky
- Automatizovanú dedukciu

Logické uvažovanie
Einsteinov hlavolam

Facts

There are 5 houses in five different colors.
In each house lives a person with a different nationality.
These five owners drink a certain type of beverage, smoke a certain cigar and keep a certain pet.
No owners have the same pet, smoke the same brand of cigar or drink the same beverage.
The Brit lives in the red house
The Swede keeps dogs as pets
The Dane drinks tea
The green house is on the left of the white house
The green house's owner drinks coffee
The person who smokes Pall Mall rears birds
The owner of the yellow house smokes Dunhill
The man living in the center house drinks milk
The Norwegian lives in the first house
The man who smokes blends lives next to the one who keeps cats
The man who keeps horses lives next to the man who smokes Dunhill
The owner who smokes BlueMaster drinks beer
The German smokes Prince
The Norwegian lives next to the blue house
The man who smokes blend has a neighbor who drinks water

The question is: *Who owns the fish?*

Integrácia dát a získavanie znalostí

Matrika

Jozef Mrkvička

narodený 11.12.2013

meno otca **Adam Mrkvička**

narodený 1.2.1983

meno matky **Eva Mrkvičková**

rodená Póriková

narodená 2.1.1984

Zdravotné strediská

Petržalka

Adam Mrkvička

krvná skupina A

Dlhe diely

Eva Mrkvičková

krvná skupina A

Jozef Mrkvička

krvná skupina B

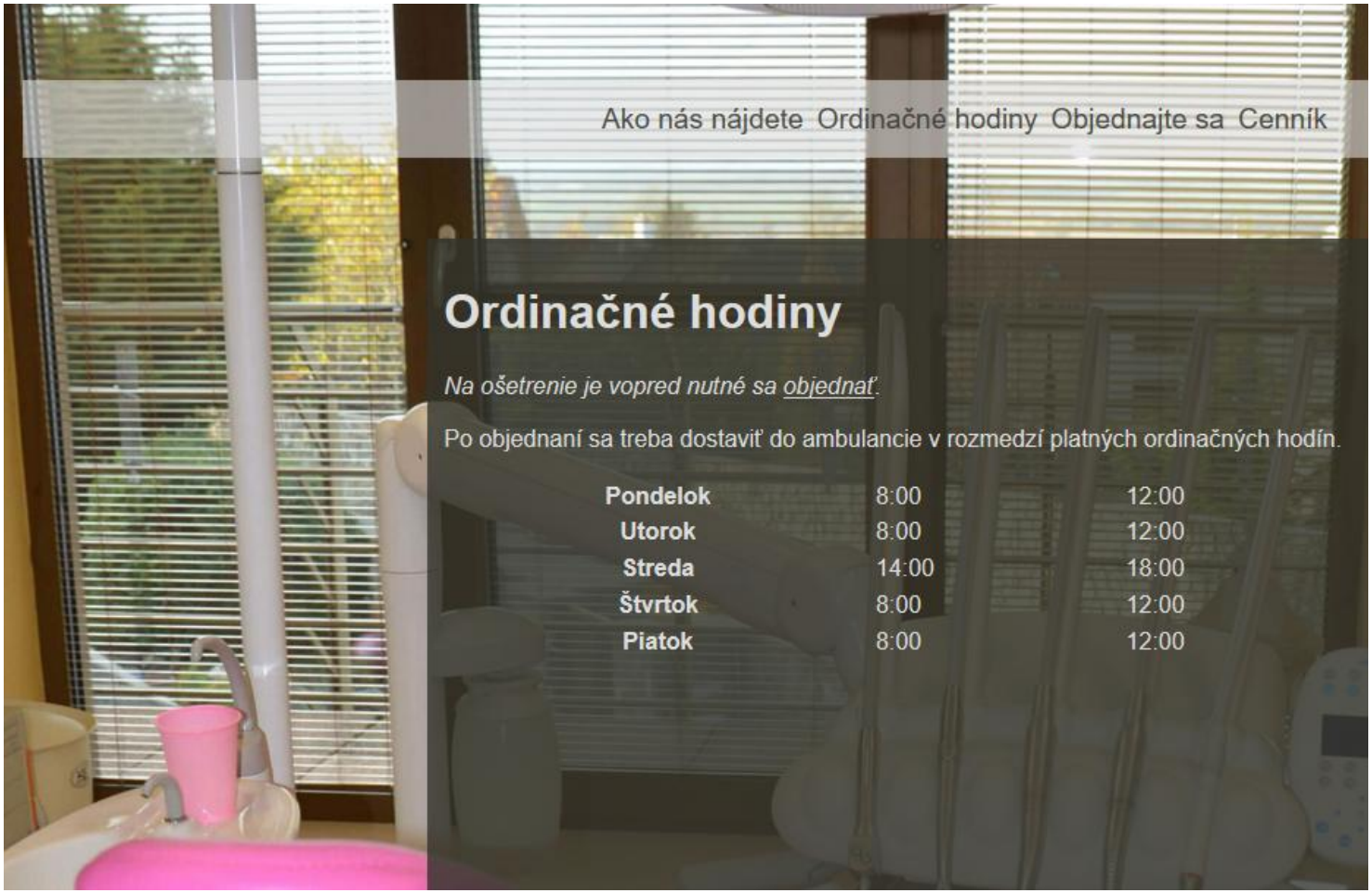
Medicínske znalosti

Krvnú skupinu dedí dieťa po jednom z rodičov.

Prečo Sémantický Web?

Na WEBe sa dnes nachádza obrovské množstvo informácií

- Štátne inštitúcie, administratíva... (eGoverment)
- Vzdelanie, školy, veda,... (eLearning, eEducation)
- Zdravotínctvo
- Sociálne siete



Technológie sémantického webu

Cieľ:

Sématický web by mal poskytovať metódy, algoritmy a formálne jazyky pre počítačovo spracovateľnu **reprezentaciu** informácii (údajov, znalostí), ktorá umožní:

- **integráciu informácií** v rámci celého internetu
- využitie **algoritmov pre** automatizované odhalovanie skrytých (implicitnych) informacii, z explicitných (faktov) - **získavanie znalostí**

Ako popisovať/modelovať informácie?

- **Relačné databázy**, ERM *(persistentné uchovávnie dát)*
- **Web**
  - **HTML** *(prezentácia dát)*
  - **XML, json** (content model) *(výmena dát)*
- OOM/OOD (UML)
- iné (rôznorodé dátové formáty)

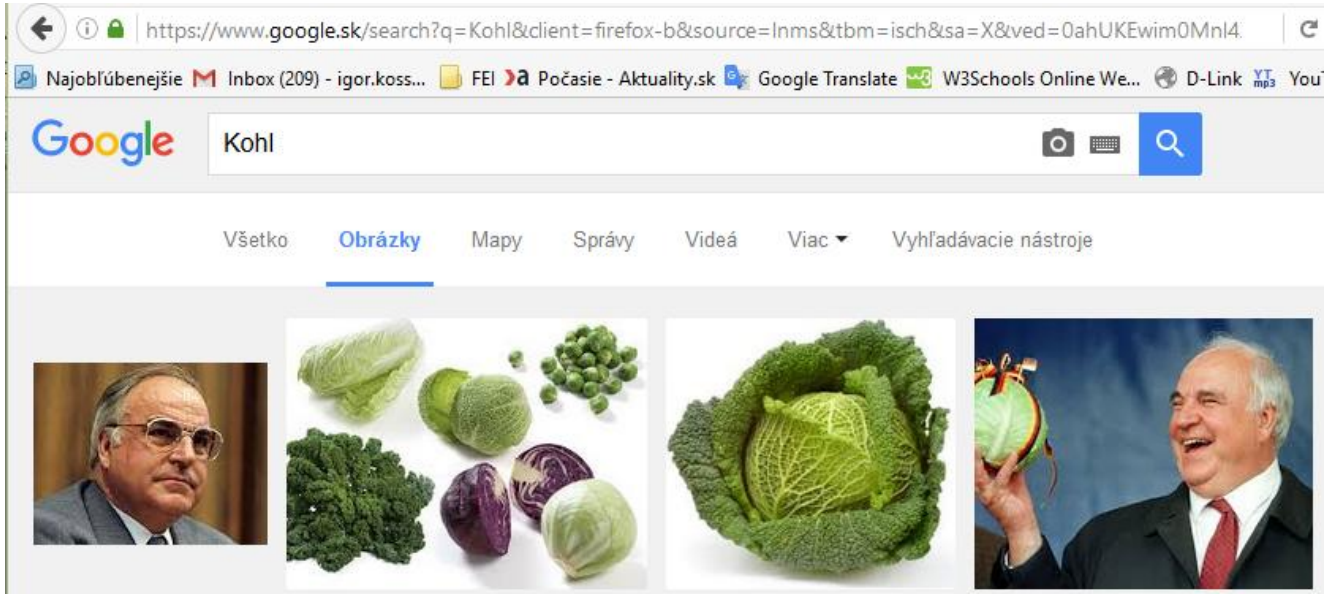
Problémy ER

- rozšíritelnosť
  - Pridávanie nových stĺpcov do DB
  - Modifikácie štruktúry
- integrácia
  - Dátový model je špecifický pre každú DB
    - Názvy tabuliek
    - Názvy stĺpcov
  - Klúče – jednoznačné identifiátory entít
    - Sú dané modelom, teda špecifické pre každú DB
    - Hodnoty väčšinou nemajú sémantiku a sú generované. T.j. nedajú sa použiť pre integráciu

Problémy webu

- Problém s vyhľadávaním - **nejednoznačnosť pojmov**

  Vyhľadávanie je založené zväčša na klúčových slovách.Jedno slovo však môže mať rôzne významy.



Cieľ → sémantické vyhľadávanie

- Heterogenita prezentácie informácií:

  Tá istá informácia môže byť vyjadrená
  - pomocou rôznych znakových sád
  - v rôznych prirodzených jazykoch
  - na rôznych miestach na stránke

  Cieľ → cross-web information integration

Problémy HTML

- Informácie na webe sú **primárne určené pre čítanie ľuďmi**
- nevhodné pre automatické spracovanie počítačom.

Počítače dokážu získať len informáciu o rozmiestnení.

```
<body>
    <div id="menu-container">
        <a href="/sk/ako-nas-najdete.html">Ako nás nájdete</a>
        <a href="/sk/ordinacne-hodiny.html">Ordinačné hodiny</a>
        <a href="/sk/objednajte-sa.html">Objednajte sa</a>
        <a href="/cennik.html">Cenník</a>
    </div>
    <div id="container">
        <h1>Ordinačné hodiny</h1>
        <p class="note">Na ošetrenie je vopred nutné sa <a href="objednajte-sa.html">objednať</a>.</p>
        <p>Po objednaní sa treba dostaviť do ambulancie v rozmedzí platných ordinačných hodín.</p>
        <table class='hours hour-range '>
            <!--tr><td></td><th>od</th><th>do</th></tr-->
            <tr>
                <th>Pondelok</th>
                <td>8:00</td>
                <td>12:00</td>
            </tr>
            <tr>
                <th>Utorok</th>
                <td>8:00</td>
                <td>12:00</td>
            </tr>
            <tr>
                <th>Streda</th>
                <td>14:00</td>
                <td>18:00</td>
```

Problémy XML a JSON

RZZ prednáša Kossaczký .

```
<Vyucujuci  Meno = “Kossaczky” >
       <Prednaska   Predmet=“RZZ” />
</Vyucujuci>
```

XML je navrhnutý primárne na modelovanie kompozicií, teda vzťahu celku a jeho častí.

Prednáška RZZ však nie je mojou súčasťou, ani mojím majetkom. (Predmet, môže v pohode prednášať aj niekto iný.)

Ale ani naopak

```
<Predmet   Meno=“RZZ” >
       <Prednaska  Vyucujuci =“Kossaczky” />
</Predmet>
```

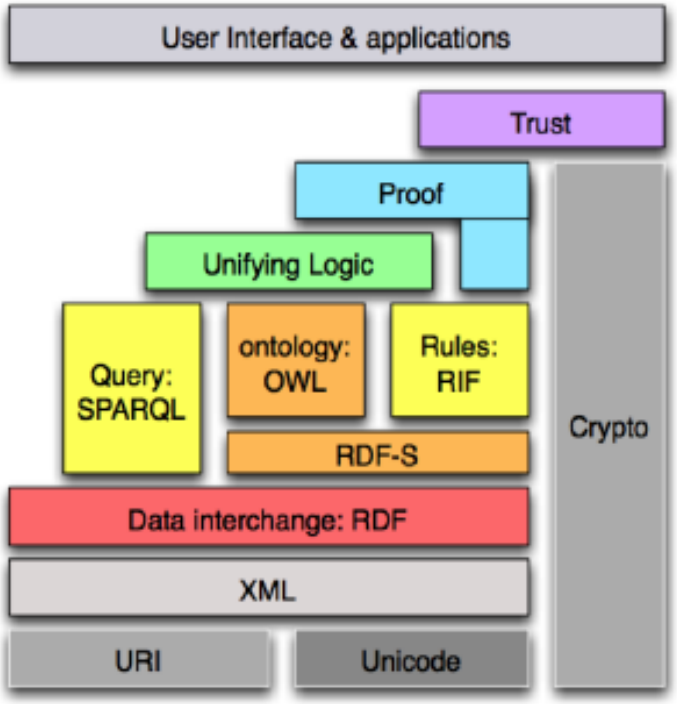
Ja tiež nie som súčasťou predmetu RZZ. (v pohode viem existovať aj bez neho)

Rovnako problematické je aj:

```
<Prednaska>
       <Vyucujuci> Kossaczky </Vyucujuci>
       <Predmet> RZZ </Predmet>
</Prednaska>
```

Riešenie

Semantic Web – Standards



1994	First public presentation of the Semantic Web idea
1998	Start of standardization of data model (RDF) and a first ontology languages (RDFS) at W3C
2000	Start of large research projects about ontologies in the US and Europe (DAML & Ontoknowledge)
2002	Start of standardization of a new ontology language (OWL) based on research results
2004	Finalization of the standard for data (RDF) and ontology (OWL)
2008	Standardization of a query language (SPARQL)
2009	Extension of OWL to OWL 2.0
2010	Standard Rule Interchange Format (RIF)