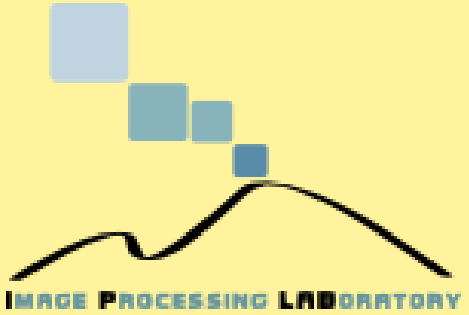
Estrazione ontologie da sentenze di Mafia

LORENZO DI SILVESTRO

www.dmi.unict.it/~disilvestro
disilvestro@dmi.unict.it

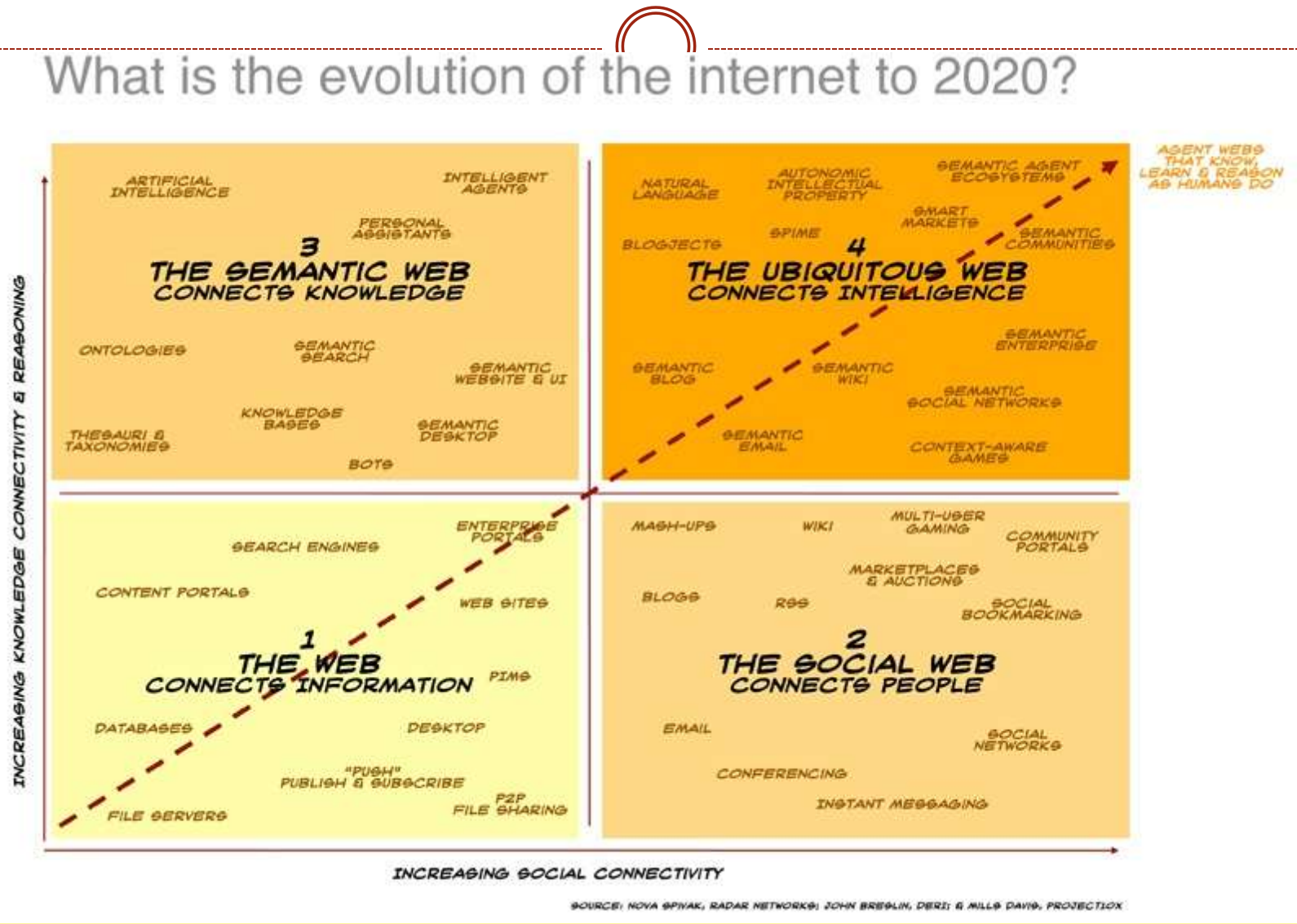


I dati

- Circa 900 sentenze penali
- Diventate irrevocabili in ogni ordine e grado per almeno un imputato
- Dal 2000 al 2006
- Nei 4 Distretti giudiziari siciliani, per i delitti di competenza delle Direzioni Distrettuali Antimafia

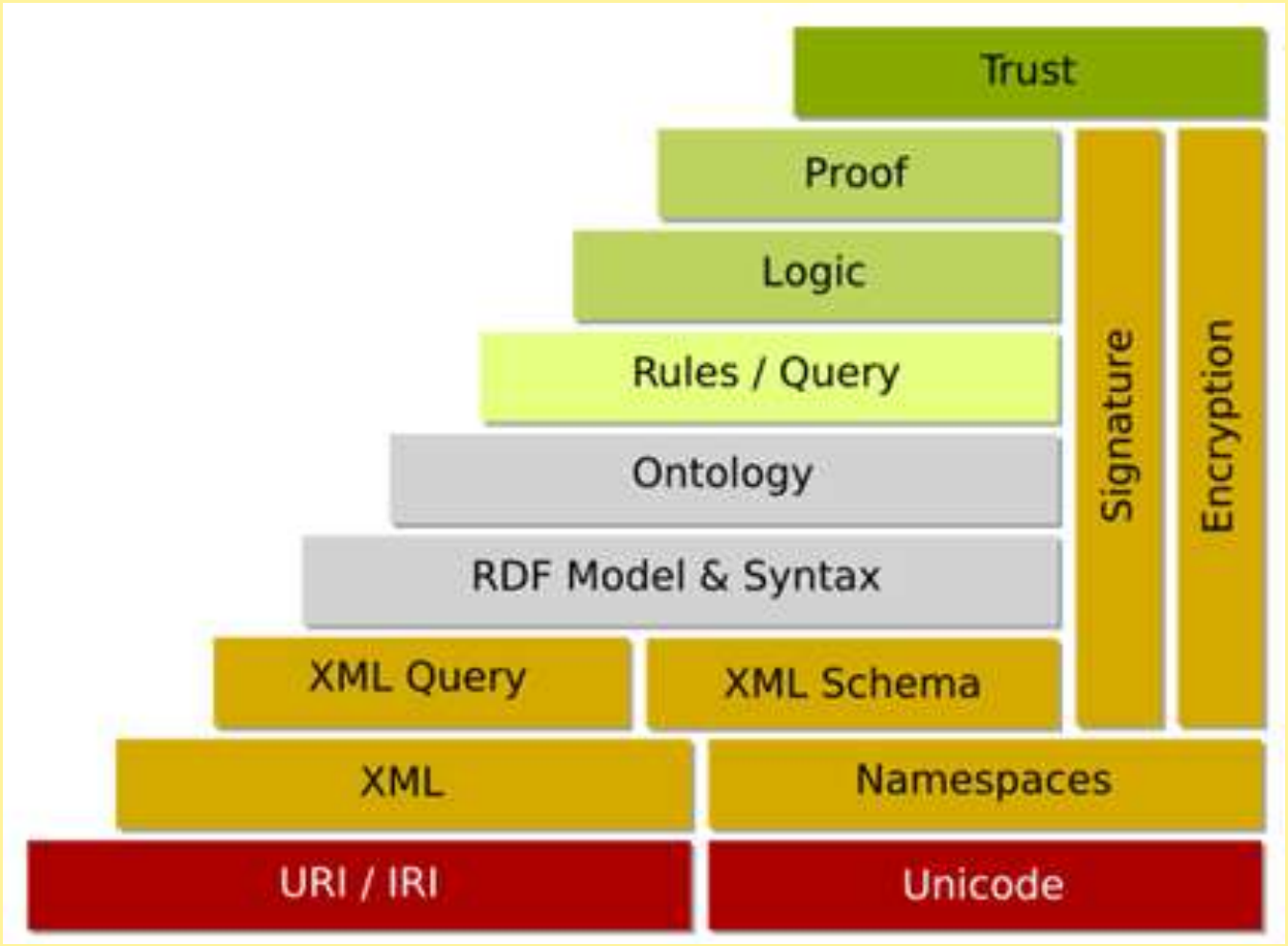
disilvestro@dmi.unict.it

Semantic Web



disilvestro@dmi.unict.it

Semantic Web Cake



disilvestro@dmi.unict.it

Unicode, URI e XML

- Unicode: standard internazionale per la definizione di una rappresentazione univoca dei caratteri
- URI: stringa che identifica univocamente una risorsa
- XML: standard *de facto* per la condivisione di documenti sul web

XML

- Meno oneroso il lavoro di un parser per l'individuazione di relazioni tra i soggetti della sentenza
- Facilita la pubblicazione sul web e l'esportazione in diversi formati
- Interoperabilità con altre banche dati a tema giuridico

disilvestro@dmi.unict.it

XML giuridico

- Progetto NormeInRete, finanziato dall'AIPA (Autorità per l'Informatica nella Pubblica Amministrazione)
- Standard XML per documenti giuridici
  - Leggi
  - Decreti legislativi
  - Decreti ministeriali
- Definiti da opportuni DTD
- I dati sono disponibili sul portale www.normattiva.it

disilvestro@dmi.unict.it

XML per sentenze penali

- Non esiste alcuno standard

Ma …

- www.processotelematico.giustizia.it

Però …

- Modello sperimentale
- Per la redazione di vari documenti prodotti durante il processo

Quindi …

- Schema *ad hoc*

disilvestro@dmi.unict.it

Conversione XML

Tecnologie:

- Jlex (generatore di analizzatori lessicali)
- CUP (generatore di parser LALR)

Fasi:

- Individuazione di una semantica formale (?)
- Definizione di uno XML Schema

disilvestro@dmi.unict.it

Conversione XML

```
n.03051/2008 Reg.SEN.
n.01940/2007 REG.ric.
n.01979/2007 REG.RIC.
n.02004/2007 REG.RIC.

REPUBBLICA ITALIANA IN NOME DEL POPOLO ITALIANO

Il Tribunale Amministrativo Regionale per la Toscana

(Sezione Seconda)

ha pronunciato la presente

SENTENZA

Sul ricorso numero di registro generale 1940 del 2007, proposto da:
Cammelli Andrea, rappresentato e difeso dall'avv. Mauro Montini, con domicilio
eletto presso il suo studio in Firenze, via dei Rondinelli, 2;

contro
```


disilvestro@dmi.unict.it

Conversione XML

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<?xml-stylesheet type="text/xsl" href="Sentenze.xsl"?>
<GA xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:h="http://www.w3.org/HTML/1998/html4">
<Provvedimento>
    <meta id="20070194020081125173628813" descrizione="" gruppo="20070194020081125173628813" >
    <descrittori>
        <registro anno="2007" n="01940"/>
        <fascicolo anno="2008" n="03051"/>
        <urn>urn:nir:tar.toscana;sezione.2:sentenza:00000-0000</urn>
        <registro n="01979" anno="2007"/>
        <registro n="02004" anno="2007"/>
    </descrittori>
    <file>20070194020081125173628813.xml</file>
    <pdffile>20070194020081125173628813.pdf</pdffile>
    <tipologia>Sentenza</tipologia>
    <h:div>Il Tribunale Amministrativo Regionale per la  Toscana</h:div>
    <h:div>(Sezione Seconda)</h:div>
    <h:div>ha pronunciato la presente</h:div>
    <h:div>SENTENZA</h:div>
    <ricorrenti>
        <h:div>Sul ricorso numero di registro generale 1940 del 2007, proposto da: </h:div>
        <h:div>Cammelli Andrea, rappresentato e difeso dall'avv. Mauro Montini,
        con domicilio eletto presso il suo studio in Firenze, via dei Rondinelli,</h:div>
    </ricorrenti>
    <resistenti>
        <h:div> ... </h:div>
    </resistenti>
```


disilvestro@dmi.unict.it

Conversione XML

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<?xml-stylesheet type="text/xsl" href="Sentenze.xsl"?>
<GA xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:h="http://www.w3.org/HTML/1998/html4">
<Provvedimento>
    <meta id="20070194020081125173628813" descrizione="" gruppo="20070194020081125173628813" >
    <descrittori>
      <urn>urn:nir:tar.toscana;sezione.2:sentenza:00000-0000</urn>
      <registro n="01979" anno="2007"/>
      <registro n="02004" anno="2007"/>
    <file>20070194020081125173628813.xml</file>
    <pdffile>20070194020081125173628813.pdf</pdffile>
    <tipologia>Sentenza</tipologia>
    <h:div>Il Tribunale Amministrativo Regionale per la  Toscana</h:div>
    <h:div>(Sezione Seconda)</h:div>
    <h:div>ha pronunciato la presente</h:div>
    <h:div>SENTENZA</h:div>
    <ricorrenti>
        <h:div>Sul ricorso numero di registro generale 1940 del 2007, proposto da: </h:div>
        <h:div>Cammelli Andrea, rappresentato e difeso dall'avv. Mauro Montini,
        con domicilio eletto presso il suo studio in Firenze, via dei Rondinelli,</h:div>
    </ricorrenti>
    <resistenti>
        <h:div> ... </h:div>
    </resistenti>
```


disilvestro@dmi.unict.it

Conversione XML

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<?xml-stylesheet type="text/xsl" href="Sentenze.xsl"?>
<GA xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:h="http://www.w3.org/HTML/1998/html4">
<Provvedimento>
    <meta id="20070194020081125173628813" descrizione="" gruppo="20070194020081125173628813" >
    <descrittori>
        <registro anno="2007" n="01940"/>
        <fascicolo anno="2008" n="03051"/>
        <urn>urn:nir:tar.toscana;sezione.2:sentenza:00000-0000</urn>
        <registro n="01979" anno="2007"/>
        <registro n="02004" anno="2007"/>
        <file>20070194020081125173628813.xml</file>
        <pdffile>20070194020081125173628813.pdf</pdffile>
    <h:div>Il Tribunale Amministrativo Regionale per la Toscana</h:div>
    <h:div>(Sezione Seconda)</h:div>
    <h:div>ha pronunciato la presente</h:div>
    <h:div>SENTENZA</h:div>
    <ricorrenti>
        <h:div>Sul ricorso numero di registro generale 1940 del 2007, proposto da: </h:div>
        <h:div>Cammelli Andrea, rappresentato e difeso dall'avv. Mauro Montini,
        con domicilio eletto presso il suo studio in Firenze, via dei Rondinelli,</h:div>
    </ricorrenti>
    <resistenti>
        <h:div> ... </h:div>
    </resistenti>
```


disilvestro@dmi.unict.it

Conversione XML

```
<?xml version = '1.0' encoding = 'UTF-8'?>
<?xml-stylesheet type="text/xsl" href="Sentenze.xsl"?>
<GA xmlns:xlink="http://www.w3.org/1999/xlink" xmlns:h="http://www.w3.org/HTML/1998/html4">
<Provvedimento>
    <meta id="20070194020081125173628813" descrizione="" gruppo="20070194020081125173628813" >
    <descrittori>
        <registro anno="2007" n="01940"/>
        <fascicolo anno="2008" n="03051"/>
        <urn>urn:nir:tar.toscana;sezione.2:sentenza:00000-0000</urn>
        <registro n="01979" anno="2007"/>
        <registro n="02004" anno="2007"/>
    </descrittori>
    <file>20070194020081125173628813.xml</file>
    <pdffile>20070194020081125173628813.pdf</pdffile>
    <tipologia>Sentenza</tipologia>
    <h:div>Il Tribunale Amministrativo Regionale per la  Toscana</h:div>
    <h:div>(Sezione Seconda)</h:div>
<ricorrenti>
    <h:div>Sul ricorso numero di registro generale 1940 del 2007, proposto da: </h:div>
    <h:div>Cammelli Andrea, rappresentato e difeso dall'avv. Mauro Montini,
    con domicilio eletto presso il suo studio in Firenze, via dei Rondinelli,</h:div>
</ricorrenti>
<resistenti>
    <h:div> ... </h:div>
```


disilvestro@dmi.unict.it

XSL

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0"xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
 <xsl:output method="html"/>
 <xsl:template match="//GA/Provvedimento">
  <html>
   <head><link rel="stylesheet" type="text/css" href="Provvedimento.css" /></head>
   <body class="corpo">
    <p class="registri">N.
      <xsl:value-of select="meta/descrittori/fascicolo/@n"/>/
      <xsl:value-of select="meta/descrittori/fascicolo/@anno"/>
      <xsl:if test="epigrafe/adunanza/h:div[4]='SENTENZA'"> REG.SEN.</xsl:if>
      <xsl:if test="epigrafe/adunanza/h:div[4]='DISPOSITIVO DI SENTENZA'"> REG.DISP.</xsl:if>
    </p>
      <xsl:for-each select="meta/descrittori/registro">
            <p class="registri">N. <xsl:value-of select="@n"/>/
                        <xsl:value-of select="@anno"/> REG.RIC.
            </p>
      </xsl:for-each>
    <p class="repubblica"><img border="0" src="stemma.jpg"/></p>
    <p class="repubblica">REPUBBLICA ITALIANA</p>
    <p class="innome">IN NOME DEL POPOLO ITALIANO</p>
     <p class="sezione"><xsl:value-of select="epigrafe/adunanza/h:div[1]"/></p>
     <p class="popolo">
     <xsl:for-each select="epigrafe/ricorrenti/h:div">
      <xsl:value-of select="."/><br/>
     </xsl:for-each>
     </p>
```


disilvestro@dmi.unict.it

XSL

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0"xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
 <xsl:output method="html"/>
 <xsl:template match="//GA/Provvedimento">
  <html>
   <head><link rel="stylesheet" type="text/css" href="Provvedimento.css" /></head>
   <body class="corpo">
    <p class="registri">N.
     <xsl:value-of select="meta/descrittori/fascicolo/@n"/>/
     <xsl:value-of select="meta/descrittori/fascicolo/@anno"/>
     <xsl:if test="epigrafe/adunanza/h:div[4]='SENTENZA'"> REG.SEN.</xsl:if>
     <xsl:if test="epigrafe/adunanza/h:div[4]='DISPOSITIVO DI SENTENZA'"> REG.DISP.</xsl:if>
<xsl:for-each select="meta/descrittori/registro">
      <p class="registri">N. <xsl:value-of select="@n"/>/
                          <xsl:value-of select="@anno"/> REG.RIC.
      </p>
    <p class="repubblica"><img border="0" src="stemma.jpg" /></p>
    <p class="repubblica">REPUBBLICA ITALIANA</p>
    <p class="innome">IN NOME DEL POPOLO ITALIANO</p>
    <p class="sezione"><xsl:value-of select="epigrafe/adunanza/h:div[1]"/></p>
    <p class="popolo">
    <xsl:for-each select="epigrafe/ricorrenti/h:div">
     <xsl:value-of select="."/><br/>
    </xsl:for-each>
    </p>
```


disilvestro@dmi.unict.it

XSL

```
<?xml version="1.0"?>
<xsl:stylesheet version="1.0"xmlns:xsl="http://www.w3.org/1999/XSL/Transform">
 <xsl:output method="html"/>
 <xsl:template match="//GA/Provvedimento">
  <html>
   <head><link rel="stylesheet" type="text/css" href="Provvedimento.css" /></head>
   <body class="corpo">
    <p class="registri">N.
      <xsl:value-of select="meta/descrittori/fascicolo/@n"/>/
      <xsl:value-of select="meta/descrittori/fascicolo/@anno"/>
      <xsl:if test="epigrafe/adunanza/h:div[4]='SENTENZA'"> REG.SEN.</xsl:if>
      <xsl:if test="epigrafe/adunanza/h:div[4]='DISPOSITIVO DI SENTENZA'"> REG.DISP.</xsl:if>
    </p>
      <xsl:for-each select="meta/descrittori/registro">
            <p class="registri">N. <xsl:value-of select="@n"/>/
                        <xsl:value-of select="@anno"/> REG.RIC.
            </p>
      </xsl:for-each>
<p class="repubblica"><img border="0" src="stemma.jpg"/></p>
<p class="repubblica">REPUBBLICA ITALIANA</p>
<p class="innome">IN NOME DEL POPOLO ITALIANO</p>
    <xsl:for-each select="epigrafe/ricorrenti/h:div">
     <xsl:value-of select="."/><br/>
    </xsl:for-each>
    </p>
```


disilvestro@dmi.unict.it

Output

N. 03051/2008 REG.SEN.
N. 01940/2007 REG.RIC.
N. 01979/2007 REG.RIC.
N. 02004/2007 REG.RIC.

R E P U B B L I C A I T A L I A N A

IN NOME DEL POPOLO ITALIANO

Il Tribunale Amministrativo Regionale per la Toscana

(Sezione Seconda)

ha pronunciato la presente

SENTENZA

Sul ricorso numero di registro generale 1940 del 2007, proposto da:
Cammelli Andrea, rappresentato e difeso dall'avv. Mauro Montini, con domicilio eletto presso il suo studio in Firenze, via dei Rondinelli, 2;

contro

disilvestro@dmi.unict.it

Namespace, XML Schema, XML Query

- **Namespaces:** serve a qualificare e disambiguare tag e attributi in un documento XML mediante degli URI, rendendoli quindi univoci sul web.
- **XML Schema:** serve a definire la struttura di un documento XML. Oggi il W3C consiglia di adottarlo al posto della DTD stessa, essendo una tecnica più recente ed avanzata.
- **XML Query:** è un linguaggio di query concepito per essere applicabile a qualsiasi sorta di documento XML e si basa sull'utilizzo di XPath per la specificazione di percorsi all'interno di documenti. XQuery ha funzionalità che consentono di poter attingere da fonti di dati multiple per la ricerca, per filtrare i documenti o riunire i contenuti di interesse.

eXist

- Sviluppato da *Wolfgang Meier* nel 2000
- Open source
- NXD (Native XML Database)
- XQuery 1.0/XPath 2.0/XSLT 1.0 e XSLT 2.0
- Apache Lucene
- Portabile (Java + web based)
- Modulare
- Ben documentato
- Aggiornato

disilvestro@dmi.unict.it

RDF

- Il **Resource Description Framework** (**RDF**) è lo strumento base proposto da W3C per la codifica, lo scambio e il riutilizzo di metadati strutturati e consente l'interoperabilità tra applicazioni che si scambiano informazioni sul Web.

- È costituito da due componenti:
  - **RDF Model and Syntax**: espone la struttura del modello RDF, e descrive una possibile sintassi.
  - **RDF Schema**: espone la sintassi per definire schemi e vocabolari per i metadati.

RDF (2)

- L'unità base per rappresentare un'informazione in RDF è lo statement.

Uno statement è una tripla del tipo

Soggetto – Predicato – Oggetto

- il soggetto è una risorsa
- il predicato è una proprietà
- l'oggetto è un valore (e quindi anche un URI che punta ad un'altra risorsa)

disilvestro@dmi.unict.it

Esempio RDF

Esempio RDF


Sentenza
N.0351/2006
Giudice
DSLLNZ28F12C594J
Nome
email
Nato a
Lorenzo Di Silvestro
disilvestro@dmi.unict.it
Catania

OWL

- Ontology Web Language (OWL) è un linguaggio di markup per rappresentare esplicitamente significato e semantica di termini con vocabolari e relazioni tra gli stessi.
- Lo scopo di OWL è descrivere delle basi di conoscenza, effettuare delle deduzioni su di esse e integrarle con i contenuti delle pagine web.
- La rappresentazione dei termini e delle relative relazioni è chiamata ontologia.

Ontologia

- Rappresentazione formale, condivisa ed esplicita di una concettualizzazione di un dominio di interesse.
- Teoria assiomatica del primo ordine esprimibile in una logica descrittiva.
- Applicate nel campo dell'intelligenza artificiale, nella rappresentazione e condivisione della conoscenza.
- Ragionamento deduttivo, classificazione, tecniche di problem solving

Tassonomie

- L'induzione di tassonomie da testi è oggetto di notevole interesse, tale processo prevede lo sviluppo di sistemi finalizzati all'inferenza di regole e assiomi da testo.

- L'apprendimento automatico procede attraverso l'utilizzo di metodologie per l'acquisizione di conoscenza alle quali corrispondono algoritmi di complessità crescente, il cui stato dell'arte è lontano dall'avere trovato una soluzione definitiva.

disilvestro@dmi.unict.it

Fasi dell'induzione automatica

- Estrazione automatica di terminologia
- Riconoscimento dei sinonimi, in genere effettuato utilizzando la misura della similitudine distribuzionale (Dagan, 2000)
- Popolazione della base di conoscenza per cui è possibile utilizzare tecnologie basate su bootstrap da esempi (Fleischman e Hovy, 2002) o metodolgie basate sul riconoscimento della similitudine distribuzionale (Cimiano e Völker, 2005)
- Sviluppo di sistemi finalizzati all'inferenza di regole e assiomi da testo

Problematiche

- Questo campo è stato ancora scarsamente esplorato, ma sta divenendo oggetto di crescente interesse in letteratura, in quanto l'acquisizione di regole di inferenza è certamente un aspetto fondamentale legato al riconoscimento dell'implicazione nel testo (Dagan et al., 2005)

- L'identificazione di tali proprietà è di importanza cruciale per attivare processi di ragionamento automatico

disilvestro@dmi.unict.it

Fasi

- Traduzione secondo standard XML
- Estrazione automatica delle informazioni
- Definizione di conoscenza
  - RDF
  - OWL
  - Reasoning

disilvestro@dmi.unict.it

Stage e Tesi

- Conversione dati
- Classificazione documenti
- Interfaccia web d'interrogazione
- D2R
  - a tool for publishing relational databases on the Semantic Web
- Algoritmi sui grafi
- Reasoning automatico

disilvestro@dmi.unict.it

Riferimenti

- **Berners-Lee, Hendler , Lassila.** 2001. The Semantic Web: a new form of Web content that is meaningful to computers will unleash a revolution of new possibilities (Scientific American)
- **Dagan**. 2000. Contextual word similarity. In Handbook of Natural Language Processing, chapter 19, pages 459–476 (Marcel Dekker Inc.)
- **Deerwester, Dumais, Furnas, Landauer.** 1990. Indexing by latent semantic analysis. Journal of the American Society of Information Science
- **Gliozzo, Strapparava, Dagan.** 2005. Investigating Unsupervised Learning for Text Categorization Bootstrapping. In Proceedings of EMNLP-2005
- **Staab, Studer**. 2004. Handbook on Ontologies (Springer)
- **Signore.** 2002. RDF per la rappresentazione della conoscenza. KM2002, Conference Proceedings.

disilvestro@dmi.unict.it