Metadata on the Web

Miroslav Milinovic
SRCE
Zagreb, Croatia
<miro@srce.hr>

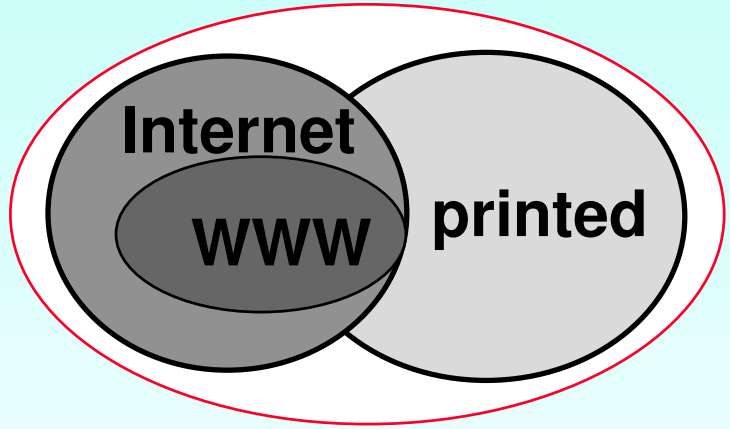
4th CARNet Users' Conference – CUC 2002, Zagreb, 25 – 27 September, 2002

Content

- Web information space
- Introduction to metadata
- Simple solutions (HTML META tag)
- Dublin Core
- RDF
- Other efforts
- Metadata deployment
- Metadata in Croatian Web space
- Summary

Internet information space

- is NOT unified
- many subjects
- different formats
- different resources (information services)
- various tools and techniques for searching and information retrieval
- some information is not (yet):
  - published electronically
  - available on the Net



Web information space

- growing rapidly since 1993
- e-commerce as a driving force
- publicly indexable Web (size estimates):
  - by Lawrence and Giles (1999)
    800 millions web pages; 15 TB (6 TB) data;
    83%-com; 6%-sci/edu; 1,5%-porn
  - by Inktomi and NEC Institute (2000)
    > 1000 million web pages;
    55% - .com, 8% - .net, 4% - .org, 1% - .gov
  - by mwp@SRCE.hr (2002)
    Croatian Web space (.hr TLD) ≈ 320 GB (≈ 6 millions resources)

Web information space

- 60% of the Web is indexed / catalogued by major search engines & catalogues
- 85% of users use search engines / catalogues to locate info

Steve Lawrence, Lee Giles (Nec Institute, February 1999)

- 30% of Web pages are copied or mirrored

Shivakumar and Garcia-Molina (1998)

- 40% of 800 million Web pages are duplicates

FAST (2000)

Web information space

- "deep" Web
  - public info 400 to 550 times bigger than "surface" Web
  - 7500 TB of data

    The Deep Web: Surfacing Hidden Value; White Paper; BrightPlanet.com, July 2000

- Users see Internet as key information source
  - 2/3 of users consider Internet as "important" or "extremely important" source of information
  - 53%(47%) ranked TV (radio) at the same level of importance

    Center for Communication Policy, UCLA, August 2000

Problems?

- high user expectations
- tools and mechanisms
  - still not good enough
  - under constant development
- information space is not (well) organised
- uncertain:
  - quality of info
  - integrity of info
  - trust in source of info

Web: an information service

- it is easy to publish
  - lower barriers / easier access
  - fast and efficient publishing (especially for dynamic sources)
- information is distributed
- information space manageability is reduced and complicated
- new relationships between authors, publishers, information intermediaries & distributors and users

Web ⇔ library

- search systems are motivated by advertising
- index/catalogue coverage is unpredictable and limited
- recall .vs. precision
- index spam (metadata spam)
- volatile resources (and their names - URLs)
- versions, editions, back issues, archiving, … ?
- authority and quality of service ?
- Intellectual Property Rights ?

Metadata: enabling higher quality

- structured data about data
  - helps to impose order on Web info space
  - enables automated discovery / manipulation
- many dimensions:
  - richness
  - functionality
  - standards and discipline in usage
  - language/culture

Metadata challenges

- metadata takes many forms:
  - resource discovery
  - content rating
  - rights management
  - security and authentication
  - …
- accommodate multiple varieties of metadata
- functionality ⇔ simplicity
- extensibility ⇔ interoperability
- human and machine creation and use
- meet the specific community needs

Modularity and interoperability

- modularity
  - enables distributed management
  - responsibility at the right place
- interoperability requires conventions about:
  - semantics (the meaning of the elements)
  - structure (human-readable, machine-parseable)
  - syntax (grammars to convey)

Metadata association models

- embedded (e.g. META tag)
- associated with (in HTTP header)
- trusted third party (explicit HTTP GET)

Syntax alternatives

- HTML
  - simple embedded metadata
- XML
  - rapidly developing, emerging tools
- RDF
  - dedicated for metadata

HTML META tag

- simple solution
- two main types of HTML META tag:
  - with **NAME** attribute
    - used to specify information about the resource (AUTHOR, KEYWORDS, DESCRIPTION, TITLE, ...)
  - with **HTTP-EQUIV** attribute
    - used as the equivalent of HTTP header

HTML META tag syntax

- with **NAME** attribute
  - used to specify information about the resource (AUTHOR, KEYWORDS, DESCRIPTION, TITLE, ...)
  - value of the **NAME** attribute is not standardised

```
<META NAME="value" CONTENT="value">
```

- with **HTTP-EQUIV** attribute
  - used as the equivalent of HTTP header

```
<META HTTP-EQUIV="value" CONTENT="value">
```

- attribute **CONTENT** defines actual metadata value

HTML META tag - examples

```
<HEAD>
<TITLE>title text</TITLE>
<META name="DESCRIPTION" content=”short text">
<META name="KEYWORDS" content=“keyword list">
</HEAD>
```

```
<HEAD>
<TITLE>title text</TITLE>
<META HTTP-EQUIV="Content-Type”
 CONTENT="text/html; charset=iso-8859-2”>
</HEAD>
```

Dublin Core (DC)

- proposed standard (revised version of RFC 2413)
- simple resource description semantics
- build on interdisciplinary consensus about a core element set for resource discovery
- core set has 15 elements
- DC elements are defined using a set of 10 attributes from the ISO11179 standard for the description of data elements
- specification (ver 1.1)
  - http://dublincore.org/documents/dces/

Dublin Core (DC)

- elements are optional and repeatable
- interdisciplinary (semantic interoperability)
- international aspect (DC in 29 languages)
- extensible:
  - starting place for richer description
  - improve description precision with substructure
- “lego bricks” concept - modular extensibility
  - adding elements to support specific requirements
  - complementary packages of metadata

DC elements

- Title
- Creator (author)
- Subject (keywords)
- Description
- Publisher
- Other Contributor
- Date
- Resource Type
- Format
- Resource Identifier
- Source
- Language
- Relation
- Coverage
- Rights Management

DC qualifiers

- element refinements
  - refine the semantics of an element
- encoding schemas
  - parsing rules or algorithms for interpreting a value
  - controlled vocabularies, classification schemas, enumerated lists
- specification:
  - http://dublincore.org/documents/dcmes-qualifiers/

DC in HTML

- META tag:
  - syntax:

<META name="PREFIX.ELEMENT_NAME" content="ELEMENT_VALUE">

  - example:

<META name="DC.Creator" content="December, John">

DC in HTML

- LINK tag (associate element name prefix with reference definition of element set)
  - syntax:

<LINK rel="schema.PREFIX" href="LOCATION_OF_DEFINITION">

  - example:

<LINK rel="schema.DC" ref="http://dublincore.org/documents/dces/">

DC in HTML

- repeating elements
  - example:

<META name="DC.Creator" content="Green, John">
<META name="DC.Creator" content="Brown, Fred">

DC in HTML

- Qualification:
  - syntax:

```
<META
    name="PREFIX.ELEMENT_NAME.SUBELEMENT_NAME"
    content="ELEMENT_VALUE"
    scheme="SCHEME"
    lang="LANGUAGE">
```

  - examples:

```
<META name="DC.Date.Created" content="2000-08-01" scheme="ISO8601">
<META name="DC.Relation.isPartOf" content="http://www.somewhere.net">
```

DC Metadata Initiative (DCMI)

- DCMI
- formal support for evolution of DC
- Working groups
- Domain-specific initiatives
  - DC-Education, DC-Libraries, DC-Government, …
- DC-Registry
  - http://www.schemas-forum.org/
- regular meetings/workshops
- DC home page: http://dublincore.org/

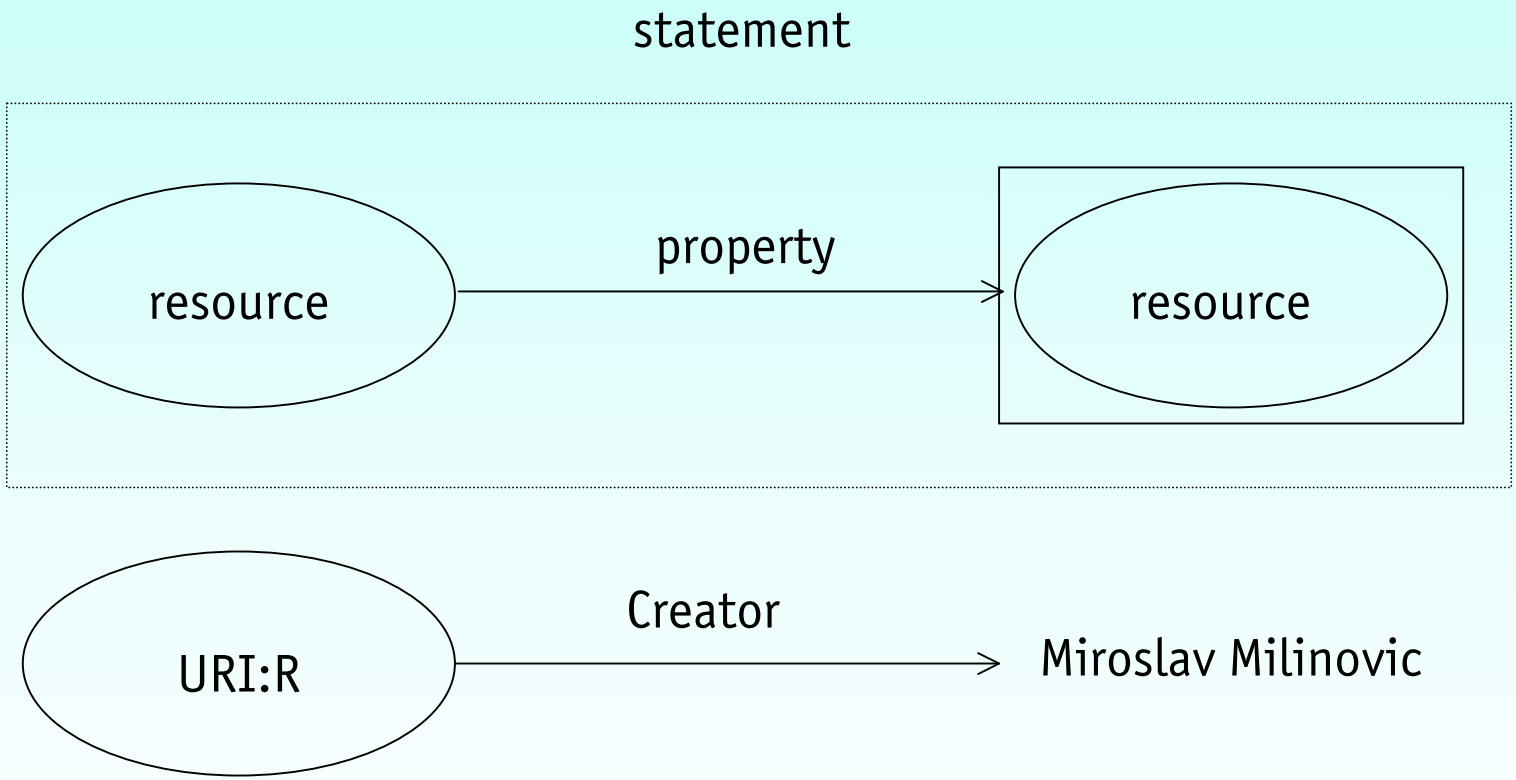
Resource Description Framework

- W3C RDF home page: http://www.w3.org/RDF/
- step ahead - generalised view on metadata
- improve on HTML, XML, …
- machine understandable metadata
- support structured values
- support metadata bureaus
- encourages authenticated metadata
- base for a variety of descriptions:
  - cataloguing, privacy, IPR, ...

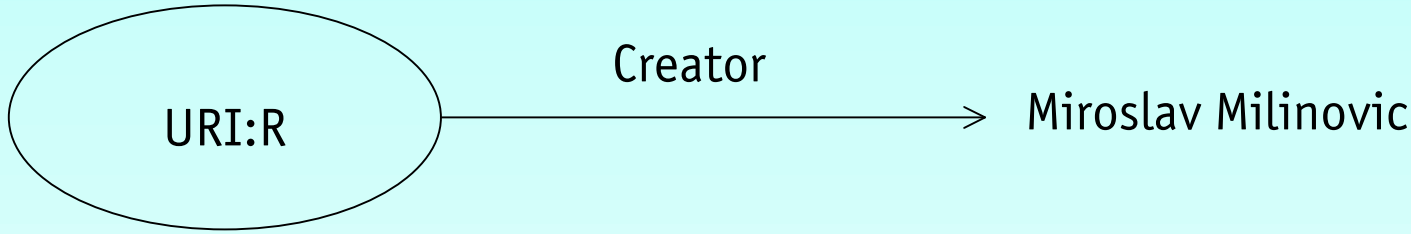
RDF concept

- formal data model
- schema model
- syntax (in XML)

RDF data model



RDF (XML) syntax




```
<RDF xmlns =“http://www.w3.org/TR/WD-rdf-syntax#”
   xmlns:dc=“http://dublincore.org/documents/dces/”>
   <Description about =“URI:R”>
      <dc:Creator>Miroslav Milinovic</dc:Creator>
   </Description>
</RDF>
```

RDF containers

- permit the aggregation of several values for a property
- 3 types of containers:
  - **Bag**
    - unordered grouping
  - **Sequence**
    - ordered grouping
  - **Alternatives**
    - alternative values
    - need to choose one
    - first value is default

RDF Bag

- unordered group
- example: Miro and Ivan are co-authors

```
<BIB:Author>
    <Bag>
            <li> Miro </li>
            <li> Ivan </li>
    </Bag>
</BIB:Author>
```

RDF Sequence

- ordered group
- example: Miro is first author, Ivan is second

```
<BIB:Author>
    <Seq>
            <li> Miro </li>
            <li> Ivan </li>
    </Seq>
</BIB:Author>
```

RDF Alternatives

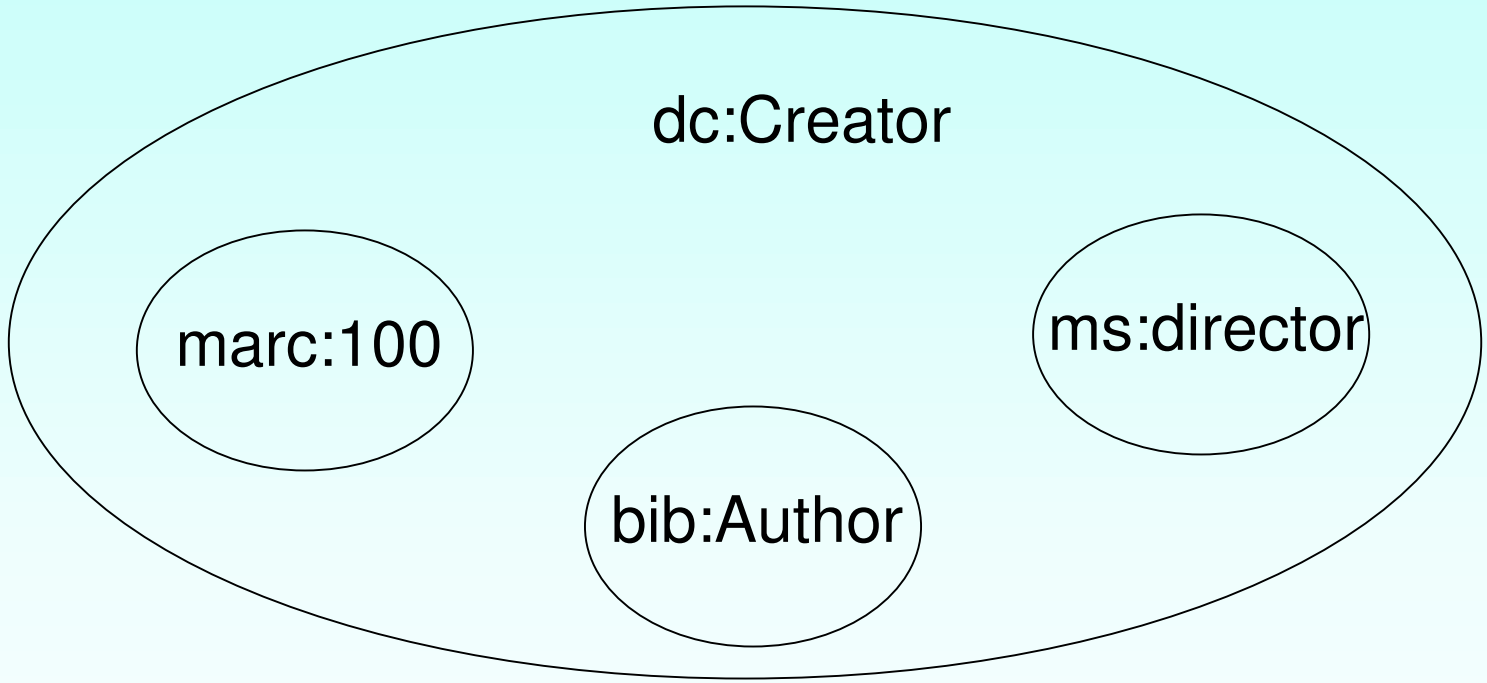
- client chooses one of several values (first is default)
- example: distance is 15 km or 9,3 miles

```
<DC:Coverage>
   <Alt>
         <li> 15KM </li>
         <li> 9.3M</li>
   </Alt>
</DC:Coverage>
```

RDF schemas

- declaration of vocabularies:
  - properties defined by a particular community
  - characteristics of properties and/or constrains on corresponding properties
- enable interoperability
  - communities can share machine readable tokens
- examples: dc, bib, admin, … marc, ms

RDF schemas



RDF in “real life”

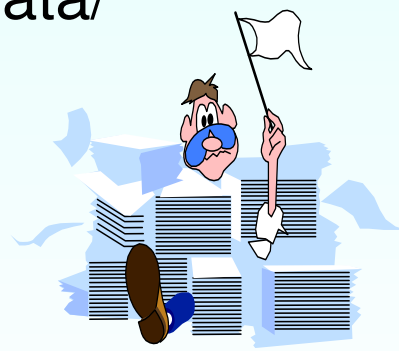
```
<?xml version="1.0"?>
<rdf:RDF xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#”
    xmlns:dc = "http://purl.org/dc/elements/1.1/">
  <rdf:Description rdf:about="http://dublincore.org/documents/dces/">
   <dc:title>Dublin Core Metadata Element Set</dc:title>
   <dc:description>This document is most recent version of the
    Dublin Core Metadata Element Set.</dc:description>
   <dc:publisher>Dublin Core Metadata Initiative</dc:publisher>
 </rdf:Description>
</rdf:RDF>
```

RDF - conclusion

- general purpose framework
- provides structured, machine-understandable metadata for the Web
- metadata vocabularies can be developed without central co-ordination
- RDF schemas describe the meaning of each property name
- signed RDF = basis for trust

Other metadata efforts

- P3P - Platform for Privacy Preferences Project
- PICS - Platform for Internet Content Selection
- DSig - Digital Signatures
- CC/PP - Composite Capabilities/Preference Profiles
- …
- further info at: http://www.w3.org/Metadata/



Metadata deployment

- standards are maturing
- several projects (DC-DOT, Reggie, CORC, MAENAD, Nordic metadata project, SAFARI, …)
- tools:
  - are here (even commercial) but are still under development
  - still lack integration with other SW applications
  - need to be more configurable (languages, schemas, formats, ...)
  - W3C RDF Validation Service: http://www.w3.org/RDF/Validator/
- prospect are improving ...

Actual situation

- no real standard, yet (?)
  - Dublin Core is a good candidate
- HTML has META tag
- use metadata / META tag (with care)!
- search engines:
  - make use of metadata (?)
  - have problems with META tag (?)

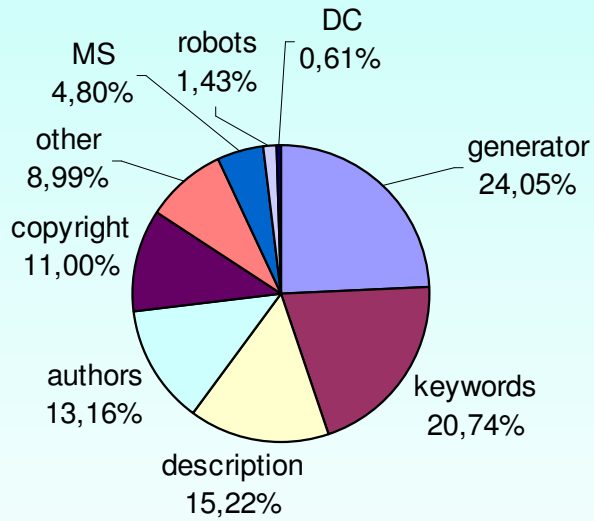
Actual situation

- about 800 million web pages
- 15 TB (6 TB) of data
- simple HTML META tag - on 34% of sites
- Dublin Core standard - only on 0,3 % of sites
- META tag diversity - 123 distinct tags noted

Steve Lawrence, Lee Giles (Nec Institute, February 1999)

Metadata in Croatian Web space

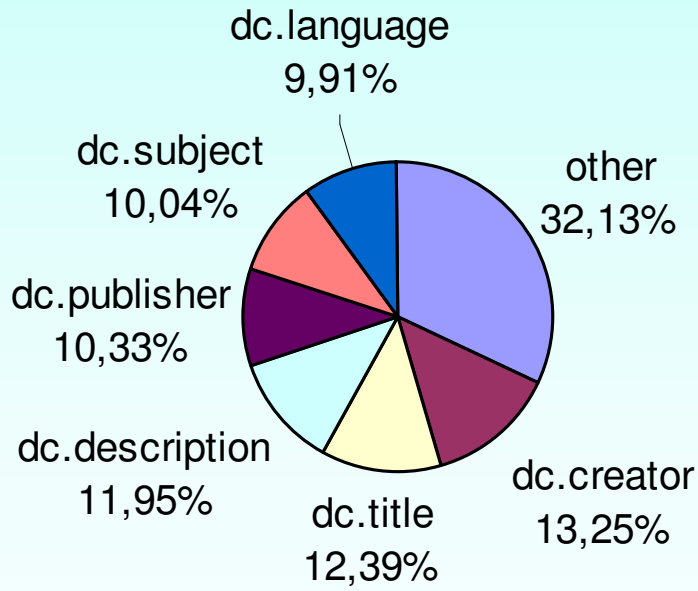
- 31% of HTML files have META tag
- 744 distinct values of NAME attribute in META tag
- Distribution of “standards”:
  - Dublin Core – 0,09%
  - HTML editors – 25%
  - Search engines – 19,7%
  - ROBOTS META tag – 1,35%



source: http://www.srce.hr/mwp/

Metadata in Croatian Web space

Frequency of various DC elements



source: http://www.srce.hr/mwp/

Summary

- Web information space
- Introduction to metadata
- Simple solutions (HTML META tag)
- Dublin Core
- RDF
- Other efforts
- Metadata deployment
- Metadata in Croatian Web space