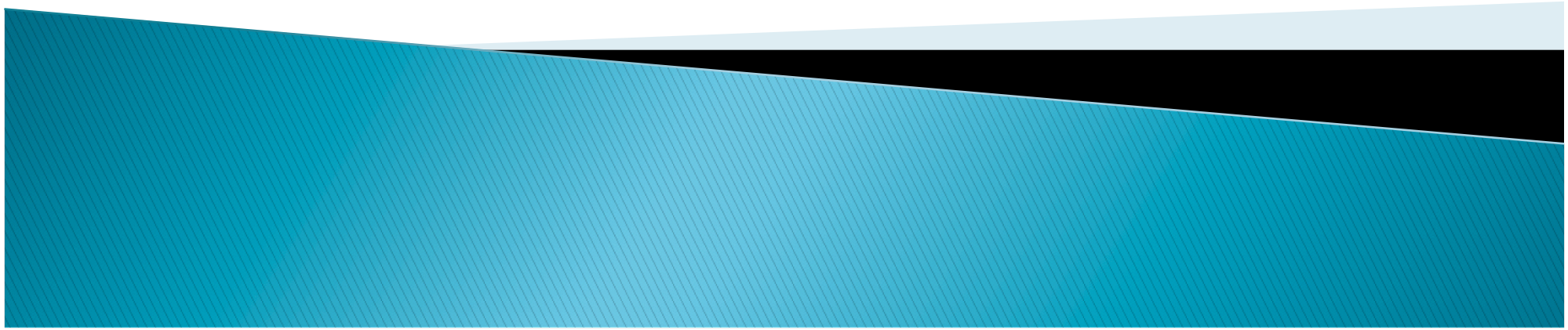
Combine the Web of Data and the Web of Documents Part 1: RDFa

ISWC 2010, Shanghai, 8th November, 2010

Ivan Herman (郝易文), W3C



What is RDFa?

- For RDF people, it *sounds* very simple:
  - RDFa is a serialization of RDF embedded in XHTML, HTML, or XML in general

So why bother? Why should we care?
Why is that of any importance?

RDFa may become the single biggest source of RDF triples on the Web after direct database access!

Data for a Web of Data

- Apart from relational databases, most of the data on the Web are in… (X)HTML content
- New content is generated every day
- How would one get structured data from that information?

Authors of the "traditional Web"…

- Do *not* generate RDF/XML files separately
  - RDF/XML is complex
  - it requires a separate storage, generation, etc mechanism
    - that is also valid for, e.g., Turtle
    - even when authoring with, say, Emacs, creating an extra file is a load

Solution

- Add extra structured content to the (X)HTML pages
- Let processors extract those and turn into RDF

Existing approaches

- Microformats
  - reuses HTML attributes like @class, @title
  - separate vocabularies (address, CV, …)
  - difficult to mix microformats (no concept of namespaces)
  - possible to transform via, e.g., XSLT + GRDDL, but all transformations are vocabulary dependent

Existing approaches

- Microdata
  - adds new attributes to HTML5 to express metadata
  - can use URI-s, it also fixes some vocabulary mappings (e.g., to Dublin Core elements)
  - has no notion of datatypes, namespaces
  - generic processing becomes possible to generate RDF

Existing approaches

- RDFa
  - adds new (X)HTML/XML attributes
  - has namespaces and URIs at its core; i.e., mixing vocabulary is just as easy as in RDF
  - complete flexibility for using Literals or URI Resources
  - *is a complete serialization of RDF*
  - generic processing becomes possible to generate RDF

RDFa is a complete bridge between the Web of Documents and the Web of Data

Therefore…

- It is very important for RDF experts to
  - know RDFa
  - parse it alongside Turtle, RDF/XML or other
  - when appropriate, generate RDFa pages

What does RDFa look like ?

Main principles of RDFa

- RDFa means "*RDF in attributes*". Ie:
  - all RDF contents are defined through XML attributes (no elements)
  - the XML/HTML *tree structure* is used
  - many of the attributes are defined by RDFa
    - some attributes (@href, @rel) are also reused
  - if possible, the text content is also reused (for literals) as well as @href values

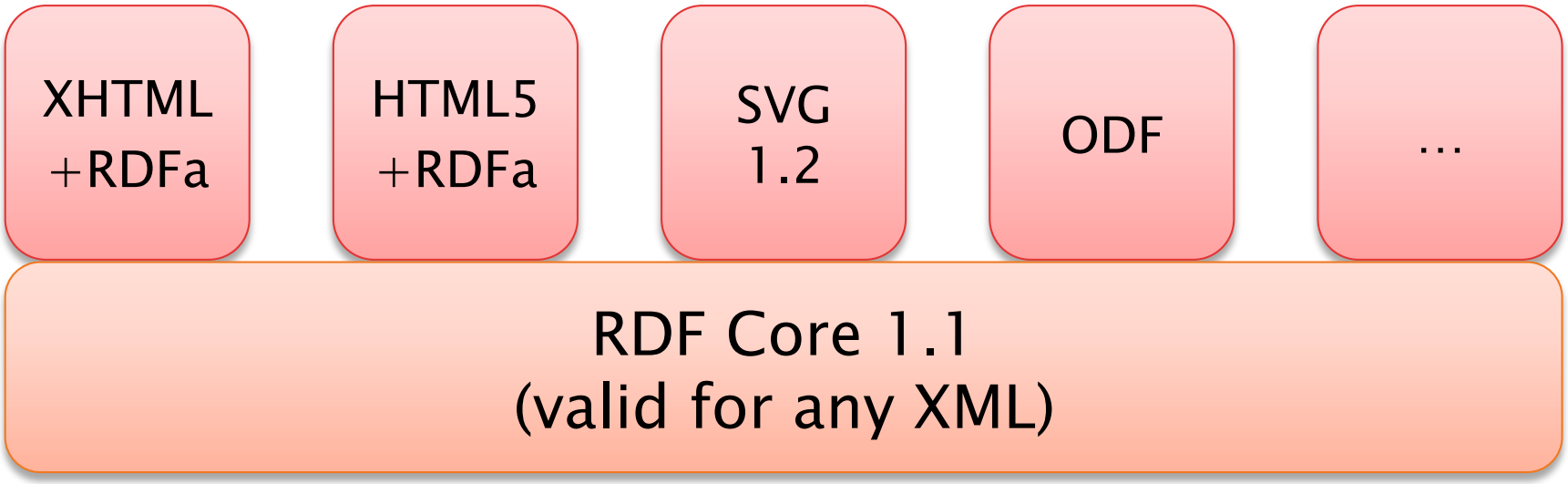
What does this mean in practice?

- The same (X)HTML file:
  - is used, unchanged, by browsers
    - they ignore attributes they do not know
  - can be used by specialized processors (or APIs) to extract RDF triples

Before getting into details…

- The current Recommendation is RDFa 1.0
- There is an RDFa1.1 in the making, almost ready
- *I will talk about RDFa1.1* and warn when the feature is not available in RDFa1.0

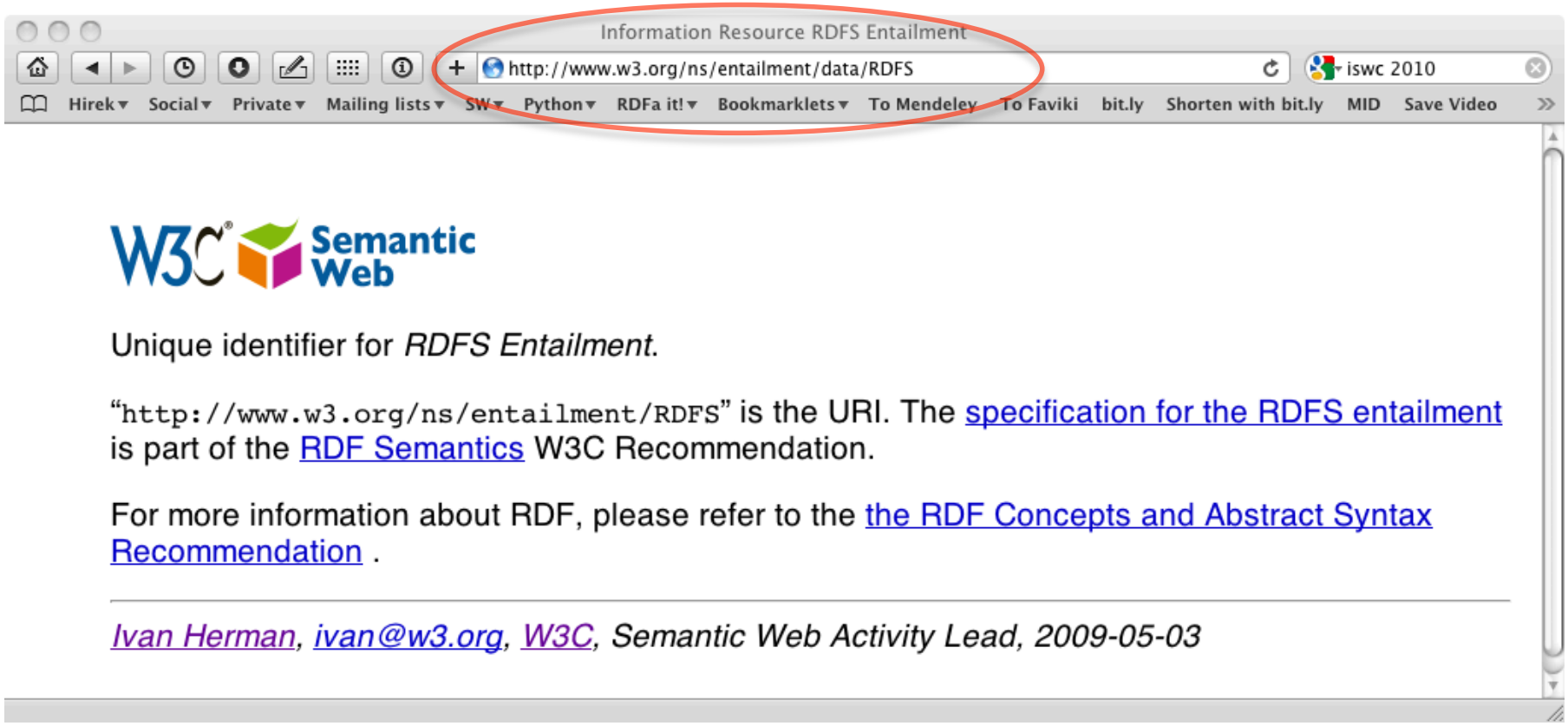
XML or (X)HTML?



- Formally:
  - RDFa WG defines Core and XHTML
  - HTML WG defines HTML5
- this tutorial uses XHTML examples

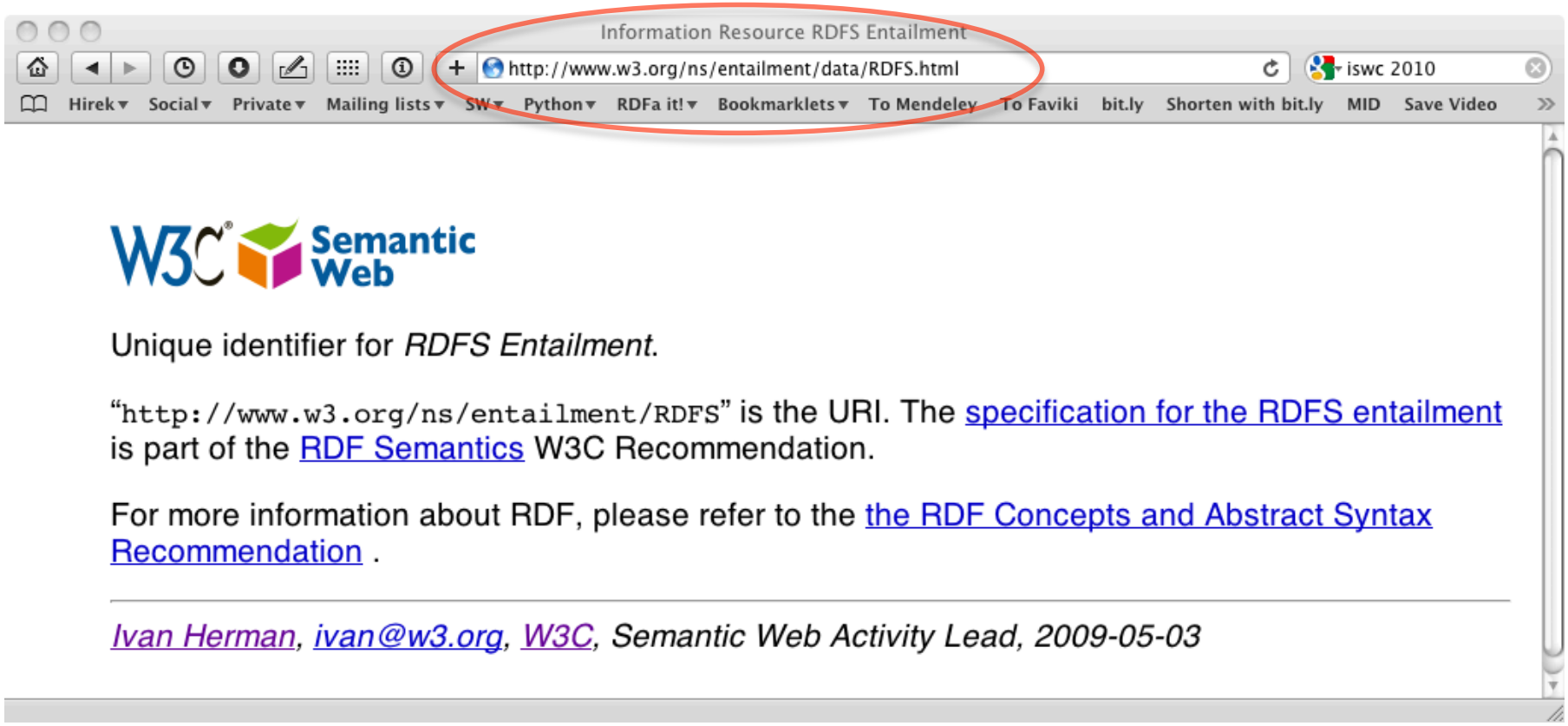
A typical usage pattern

- A browser usually asks for an HTML content:



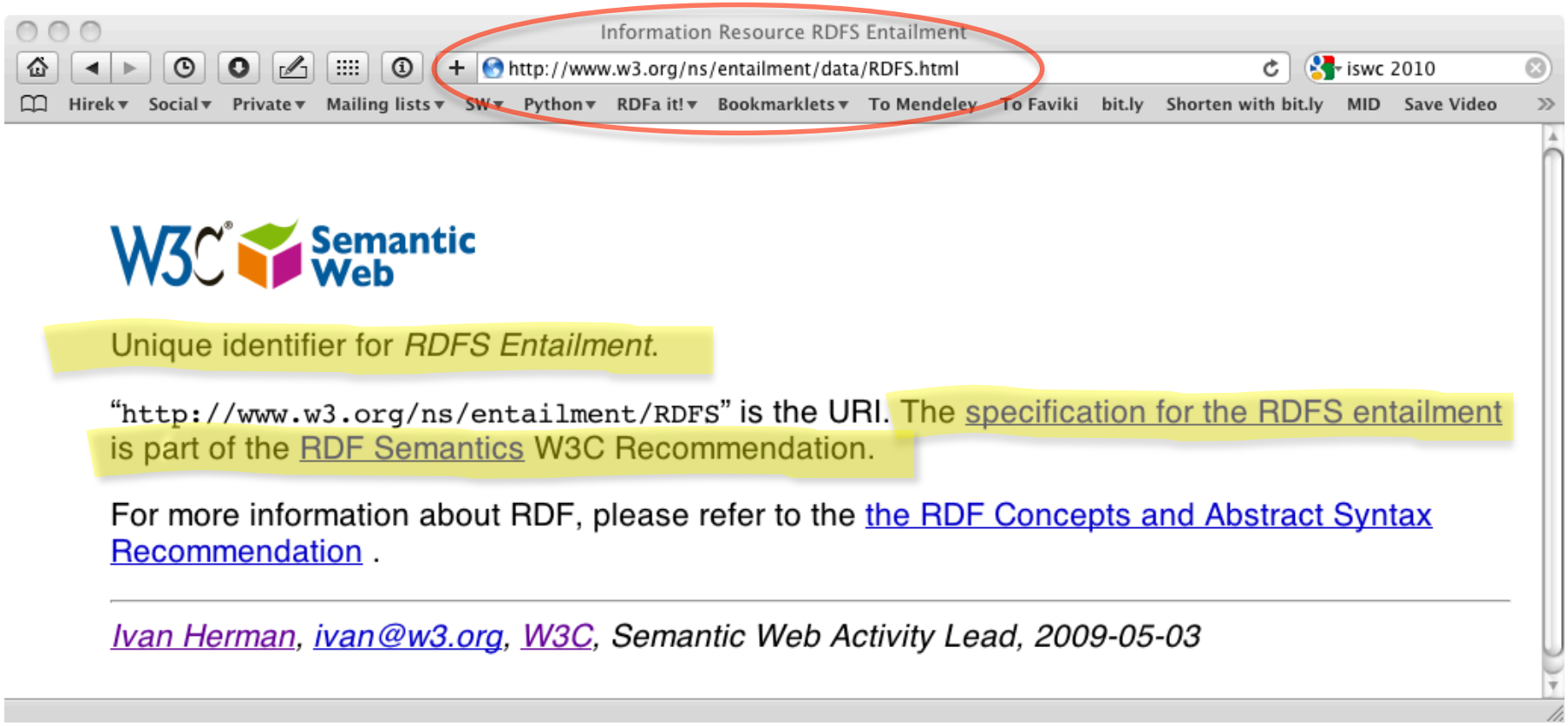
A typical usage pattern

- Via content negotiations this goes to:



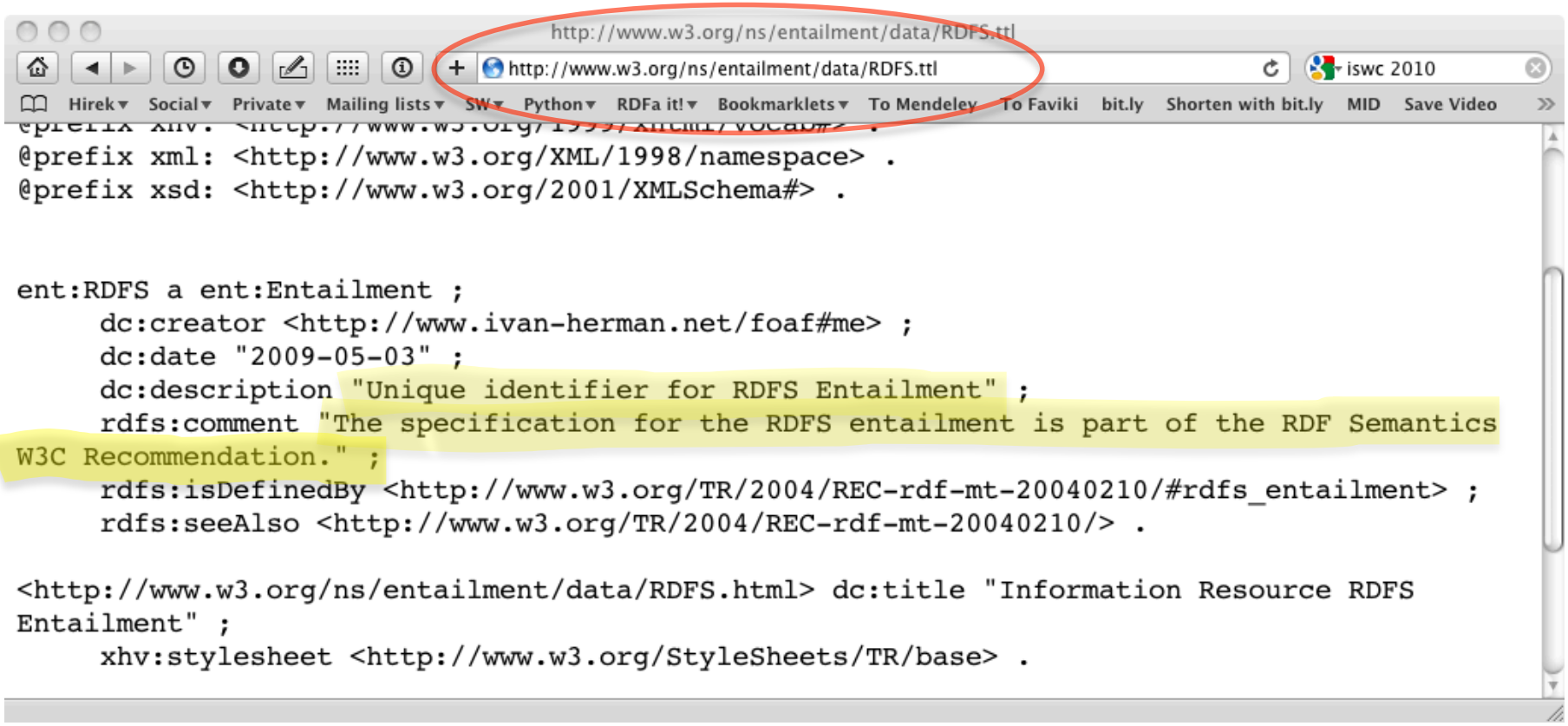
A typical usage pattern

- Via content negotiations this goes to:



A typical usage pattern

- But a client could ask for, say, Turtle:



Where does the Turtle content come from?

- The triples are embedded in the HTML file
  - a client may know how to extract RDF triples directly from that file; or
  - an online “distiller” service is used; or
  - the server is set up to generate the Turtle file automatically

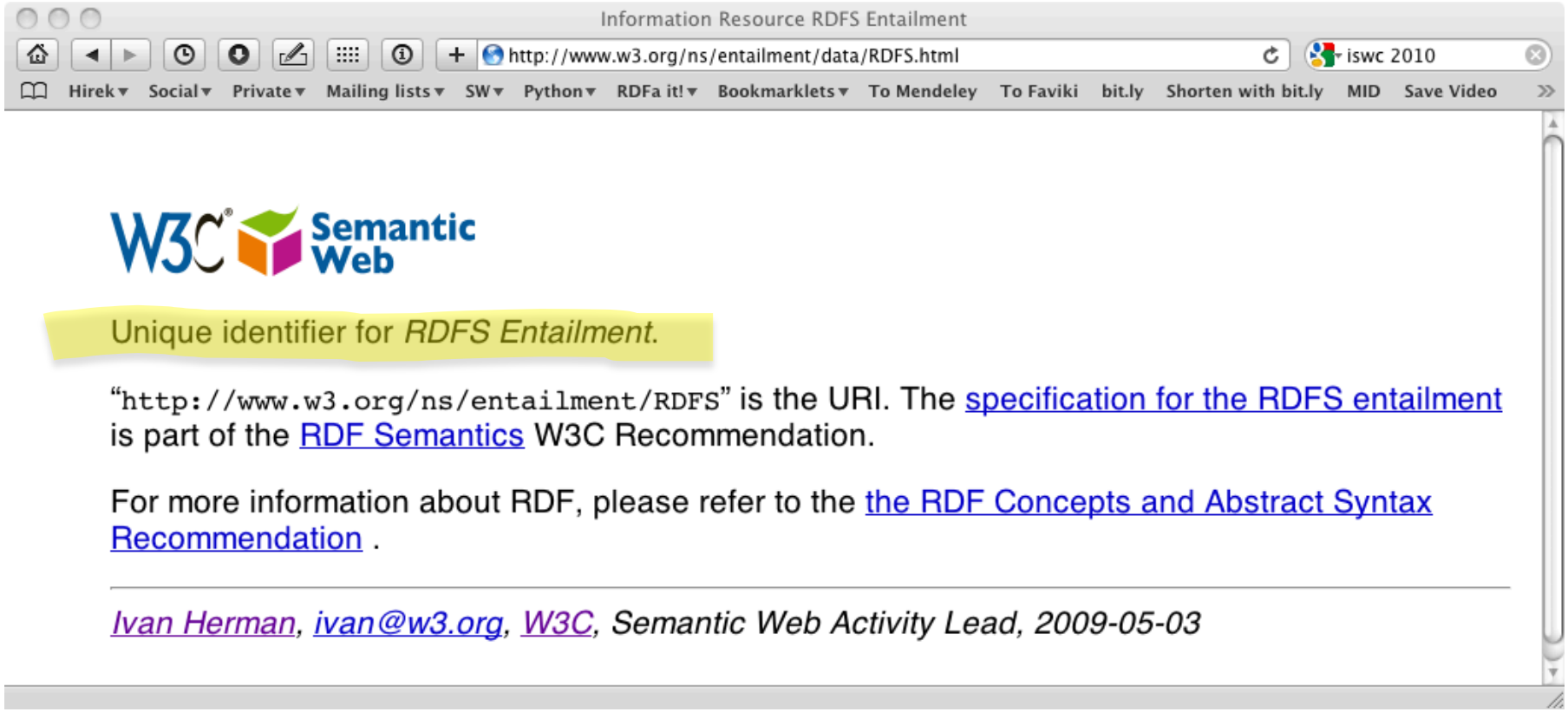
For example, use an online service…

… or set up the server…

```
RewriteEngine On
RewriteBase /ns/entailment/data/

RewriteRule RDFS.ttl
   /2007/08/pyRdfa/extract?format=turtle&amp;
       uri=http://www.w3.org/ns/entailment/data/RDFS.html [L]
```

The important point: the *content* is created only once

Enough talk; how does it work?


Information Resource RDFS Entailment
http://www.w3.org/ns/entailment/data/RDFS.html
iswc 2010
Hirek Social Private Mailing lists SW Python RDFa it! Bookmarklets To Mendeley To Faviki bit.ly Shorten with bit.ly MID Save Video
W3C Semantic Web
Unique identifier for RDFS Entailment.
"http://www.w3.org/ns/entailment/RDFS" is the URI. The specification for the RDFS entailment is part of the RDF Semantics W3C Recommendation.
For more information about RDF, please refer to the the RDF Concepts and Abstract Syntax Recommendation .
Ivan Herman, ivan@w3.org, W3C, Semantic Web Activity Lead, 2009-05-03

The source and generated RDF...

```
<p about="http://www.w3.org/ns/entailment/RDFS"
   property="http://purl.org/dc/terms/description">
     Unique identifier for <em>RDFS Entailment</em>.</p>
```

The source and generated RDF…

```
<p about="http://www.w3.org/ns/entailment/RDFS"
   property="http://purl.org/dc/terms/description">
     Unique identifier for <em>RDFS Entailment</em>.</p>
```

```
<http://www.w3.org/ns/entailment/RDFS>
   … .
```

The source and generated RDF…

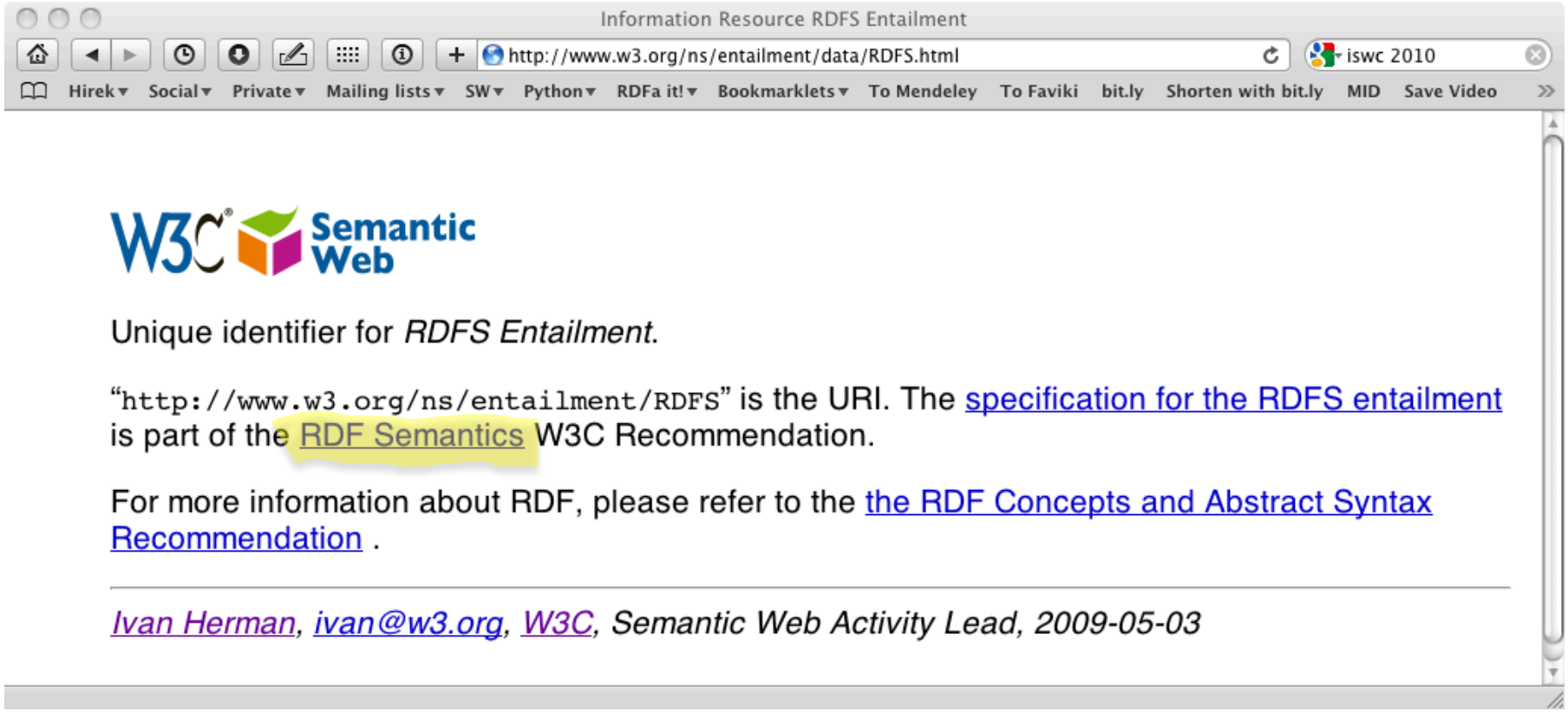
```
<p about="http://www.w3.org/ns/entailment/RDFS"
   property="http://purl.org/dc/terms/description">
      Unique identifier for <em>RDFS Entailment</em>.</p>
```

```
<http://www.w3.org/ns/entailment/RDFS>
   <http://purl.org/dc/terms/description>
      … .
```

The source and generated RDF...

```
<p about="http://www.w3.org/ns/entailment/RDFS"
   property="http://purl.org/dc/terms/description">
     Unique identifier for <em>RDFS Entailment</em>.</p>
```

```
<http://www.w3.org/ns/entailment/RDFS>
   <http://purl.org/dc/terms/description>
      "Unique identifier for RDFS Entailment." .
```


Information Resource RDFS Entailment
http://www.w3.org/ns/entailment/data/RDFS.html
iswc 2010
Hirek Social Private Mailing lists SW Python RDFa it! Bookmarklets To Mendeley To Faviki bit.ly Shorten with bit.ly MID Save Video
W3C Semantic Web
Unique identifier for *RDFS Entailment*.
"http://www.w3.org/ns/entailment/RDFS" is the URI. The specification for the RDFS entailment is part of the RDF Semantics W3C Recommendation.
For more information about RDF, please refer to the the RDF Concepts and Abstract Syntax Recommendation .
Ivan Herman, ivan@w3.org, W3C, Semantic Web Activity Lead, 2009-05-03

The source and generated RDF...

```
<a about="http://www.w3.org/ns/entailment/RDFS"
   rel="http://www.w3.org/2000/01/rdf-schema#seeAlso"
   href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
     RDF Semantics.
</a>
```

The source and generated RDF…

```
<a about="http://www.w3.org/ns/entailment/RDFS"
   rel="http://www.w3.org/2000/01/rdf-schema#seeAlso"
   href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
     RDF Semantics.
</a>
```

```
<http://www.w3.org/ns/entailment/RDFS>
  ….
```

The source and generated RDF…

```
<a about="http://www.w3.org/ns/entailment/RDFS"
   rel="http://www.w3.org/2000/01/rdf-schema#seeAlso"
   href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
     RDF Semantics.
</a>
```

```
<http://www.w3.org/ns/entailment/RDFS>
   <http://www.w3.org/2000/01/rdf-schema#seeAlso>
      … .
```

The source and generated RDF…

```
<a about="http://www.w3.org/ns/entailment/RDFS"
   rel="http://www.w3.org/2000/01/rdf-schema#seeAlso"
   href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210/">
     RDF Semantics.
</a>
```

```
<http://www.w3.org/ns/entailment/RDFS>
   <http://www.w3.org/2000/01/rdf-schema#seeAlso>
      <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> .
```

Is that it?

What we have is… Ntriples in HTML

- The combination of @about with @rel/ @property and possibly @href covers most of we need…
- But this is too complex for authors

Just compare

```
<http://www.w3.org/ns/entailment/RDFS>
   <http://purl.org/dc/terms/description>
      "Unique identifier for RDFS Entailment." .
<http://www.w3.org/ns/entailment/RDFS>
   <http://www.w3.org/2000/01/rdf-schema#seeAlso>
      <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> .
```

- with

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc: <http://purl.org/dc/terms/> .

<http://www.w3.org/ns/entailment/RDFS>
    rdfs:seeAlso
         <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> ;
    dc:description "Unique identifier for RDFS Entailment." .
```

The “Turtle” aspects of RDFa

- Use compact URI-s when possible
- Make use of XML structure for
  - shared subjects
  - shared predicates
  - create blank nodes
  - …

Compact URIs ("CURIE"s)

- Just like in Turtle:
  - define a prefix via @prefix
  - use prefix:reference to abbreviate a URI

CURIE definition and usage

```
<html>
  …
  <p about="http://www.w3.org/ns/entailment/RDFS"
     property="http://purl.org/dc/terms/description">
       Unique identifier for <em>RDFS Entailment</em>.</p>
  …
</html>
```

- can be replaced by:

```
<html prefix="dc: http://purl.org/dc/terms/">
  …
  <p about="http://www.w3.org/ns/entailment/RDFS"
     property="dc:description">
       Unique identifier for <em>RDFS Entailment</em>.</p>
  …
</html>
```

Some details on @prefix

- Can be anywhere in the XML tree and is valid for the whole sub-tree
  - i.e., the html element is not the only place to have it
- The same @prefix attribute can hold several definitions:
  - prefix="dc: http://purl.org… rdfs: http://…"

Some details on @prefix

- An alternative (deprecated) syntax is
  - xmlns:dc="http://purl.org/dc/terms/"
- CURIEs and “real” URIs can be mixed
  - if an attribute value can be interpreted as a CURIE, fine
  - alternatively, it is considered as a URI
- CURIEs can be used on RDFa attributes only!
  - e.g., not for @href

RDFa 1.0 Warnings on CURIEs

- In RDFa 1.0
  - only the xslt:XXX syntax is usable
  - CURIEs on @about can only be used with the syntax: about="[pref:ref]"
  - Only CURIEs can be used on, e.g., @property or @rel (no fallback on URIs)

Sharing subjects

- The basic principle: @about is inherited by children nodes
  - i.e., no reason to repeat it

Shared subject example

```
<html prefix="dc: http://purl.org/dc/terms/
              rdfs: http://www.w3.org/2000/01/rdf-schema#">
  …
  <body about="http://www.w3.org/ns/entailment/RDFS">
    …
    <p property="dc:description">
     Unique identifier for <em>RDFS Entailment</em>.</p>
    <p>…<a rel="rdfs:seeAlso"
     href="http://www.w3.org/TR/2004/REC-rdf-mt-20040210">
     RDFS Semantics</a>…</p>
```

… yielding

```
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix dc: <http://purl.org/dc/terms/> .

<http://www.w3.org/ns/entailment/RDFS>
    rdfs:seeAlso
       <http://www.w3.org/TR/2004/REC-rdf-mt-20040210/> ;
    dc:description "Unique identifier for RDFS Entailment." .
```

Intricacies of literals

Date examples

```
<body about=".." prefix="dc: http://…"
  <address>
    <p property="dc:date">2010-07-05</p>
  </address>
</body>
```

- This leads to:

```
@prefix dc: <http://…>
<..> dc:date "2010-07-05" .
```

Datatypes

```
<body about=".." prefix="dc: http://… xsd: http://…"
  <address>
    <p property="dc:date" datatype="xsd:date">2010-07-05</p>
  </address>
</body>
```

- This leads to:

```
@prefix dc: <http://…>
@prefix xsd: <http://…>
<..> dc:date "2010-07-05"^^xsd:date .
```

Controlling the literal

- The basic rule says: the (RDF) Literal is the enclosed text from the HTML content
- This is fine in 80% of the cases, but…
- It may not be natural in all cases! E.g.,
  - 2010-07-05 is the "official" ISO format (for xsd:date)
  - but "July 5, 2010" looks much more natural for a human…

Usage of @content

```
<body about=".." prefix="dc: http://… xsd: http://…"
  <address>
    <p property="dc:date" datatype="xsd:date"
       content="2010-07-05">July 5, 2010</p>
  </address>
</body>
```

- Also leads to:

```
@prefix dc: <http://…>
@prefix xsd: <http://…>
<..> dc:date "2010-07-05"^^xsd:date .
```

Subjects, and objects, and subjects again…

The rules until now

- What we said is:
  - @about sets the subject
  - @href sets the object
- But that is not always good enough
  - we do not always want active links (i.e., the "a" element)
  - what about other links in HTML?

@src also sets the object

```
<body about="…">
   …<img rel="foaf:depiction"
         src="http://www.ex.org/img.png"/>…
</body>
```

- yields:

```
<…> foaf:depiction <http://www.ex.org/img.png> .
```

We may not always want links…

- The RDFa @resource attribute is equivalent to @href
  - it sets the object, just like @href
  - but it is ignored by a browser, i.e., not a link!
  - e.g.,:

```
<span about="http://www.ivan-herman.net/foaf#me">
  <span rel="rdfs:seeAlso"
    resource="http://www.ivan-herman.net/foaf">
       Activity Lead</span>
</span>
```

“Chaining”

- Here is what we would like to have in RDFa

```
<http://www.w3.org/ns/entailment/RDFS>
    dc:creator <http://www.ivan-herman.net/foaf#me> .

<http://www.ivan-herman.net/foaf#me>
    foaf:mailbox <mailto:ivan@w3.org> ;
    foaf:workplaceHomepage <http://www.w3.org> .
```

"Chaining"

- A straightforward way:

```
<body about="http://www.w3.org/ns/entailment/RDFS">
  …
  <address>
    <span rel="dc:creator"
       resource="http://www.ivan-herman.net/foaf#me"/>
    <span about="http://www.ivan-herman.net/foaf#me">
      <a rel="foaf:mailbox"
          href="mailto:ivan@w3.org">ivan@w3.org</a>,
      <a rel="foaf:workplaceHomepage"
          href="http://www.w3.org">W3C</a>
    </span>
  </address>
```

"Chaining"

- A straightforward way:

```
<body about="http://www.w3.org/ns/entailment/RDFS">
  …
  <address>
    <span rel="dc:creator"
       resource="http://www.ivan-herman.net/foaf#me"/>
    <span about="http://www.ivan-herman.net/foaf#me">
      <a rel="foaf:mailbox"
          href="mailto:ivan@w3.org">ivan@w3.org</a>,
      <a rel="foaf:workplaceHomepage"
          href="http://www.w3.org">W3C</a>
    </span>
  </address>
```

“Chaining”: when objects become subjects…

- An alternative:

```
<body about="http://www.w3.org/ns/entailment/RDFS">
  …
  <address>
    <span rel="dc:creator"
       resource="http://www.ivan-herman.net/foaf#me">
         <a rel="foaf:mailbox"
             href="mailto:ivan@w3.org">ivan@w3.org</a>,
         <a rel="foaf:workplaceHomepage"
             href="http://www.w3.org">W3C</a>
    </span>
  </address>
```

Chaining means

- @resource (or @href) becomes a subject *for the sub-tree*
- This feature is a bit like in RDF/XML

Some extra features

Some extra features we do not have time for…

- Blank nodes can be created using "_:XX"
- Shorthand for rdf types
- An API has been defined for Web Applications

Some extra features we do not have time for…

- Prefix declarations can be collected in a separate file and referred to via a @profile attribute
  - the “profile file”
  - RDFa1.0 warning: this is an RDFa1.1 feature!

“Term” declarations

- A profile file can also define a *term*:
  - assign a URI to a simple string
- The term can be used directly by authors, without prefixes

Term Example

- Say, file “http://ex.org/prof” defines a mapping:
  - "desc" → "http://purl.org/dc/terms/description"

Term Example

```
<html prefix="dc: http://purl.org/dc/terms/
              rdfs: http://www.w3.org/2000/01/rdf-schema#">
  …
  <body profile="http://ex.org/prof.html" about="…">
    …
    <p property="desc">
      Unique identifier for <em>RDFS Entailment</em>.</p>
```

- yields

```
@prefix dc: <http://purl.org/dc/terms/> .

<…> <http://purl.org/dc/terms/description>
      "Unique identifier for RDFS Entailment."
```

Terms are important…

- Usage of CURIEs and URIs is intuitive for RDF people…
- It is *not* for average HTML authors!
- Profile files can be published by major publishers:
  - Dublin Core, FOAF, …
  - FaceBook, Google, …
- … and users can simply refer to the profiles

Typing

- Typing can of course be done using @rel="rdf:type"
- But that is a widely used combination, so there is a separate @typeof attribute for that

Typing example

```
<span about="http://www.ivan-herman.net/foaf#me"
      typeof="foaf:Person">
    <span property="foaf:name">Ivan Herman</span>
</span>,
```

- yields

```
<http://www.ivan-herman.net/foaf#me> a foaf:Person ;
   foaf:name "Ivan Herman" .
```

Publishing RDFa

- RDFa gives an easy way of publishing RDF data on the Web
- Often, the same RDF data is available in different formats, including RDFa
  - it is up to the client to choose which one to use
  - *Web Applications* would rely on RDFa, though…

LOC example

Semantic Web

From Library of Congress Subject Headings

Details | Visualization | Suggest Terminology

Semantic Web

URI
<http://id.loc.gov/authorities/sh2002000569#concept>

Type
Topical Term

Broader Terms
- Semantic integration (Computer systems)
- Semantic networks (Information theory)
- World Wide Web

Related Terms
- Microformats

Sources
- Engr. index online, May 6, 2002 (identifier: Semantic Web)
- ASTI on FirstSearch, May 6, 2002: in titles (semantic Web)
- Work cat.: 2002070545: The Semantic Web--ISWC 20002, 2002.

LC Classification
TK5105.88815

LOC example

Source of http://id.loc.gov/authorities/sh2002000569

```
                    <p rel="skos:inScheme" resource="http://id.loc.gov/authorities#topicalTerms">Topical
Term</p>
                </li>


                <li>
                    <h3>Broader Terms</h3>
                    <ul class="std">

                        <li>
                            <a href="http://id.loc.gov/authorities/sh2004000479#concept" rel="skos:broader">
                                <span property="skos:prefLabel" xml:lang="en">Semantic integration (Computer
systems)</span>
                            </a>
                        </li>

                        <li>
                            <a href="http://id.loc.gov/authorities/sh92004914#concept" rel="skos:broader">
                                <span property="skos:prefLabel" xml:lang="en">Semantic networks (Information
theory)</span>
                            </a>
                        </li>

                        <li>
                            <a href="http://id.loc.gov/authorities/sh95000541#concept" rel="skos:broader">
                                <span property="skos:prefLabel" xml:lang="en">World Wide Web</span>
                            </a>
                        </li>

                    </ul>
                </li>
```

Consuming RDFa

- Various search engines begin to consume RDFa
  - Google, Yahoo, …
    - they may specify which vocabularies they "understand"
    - this is still an evolving area
- Facebook's "social graph" is based on RDFa

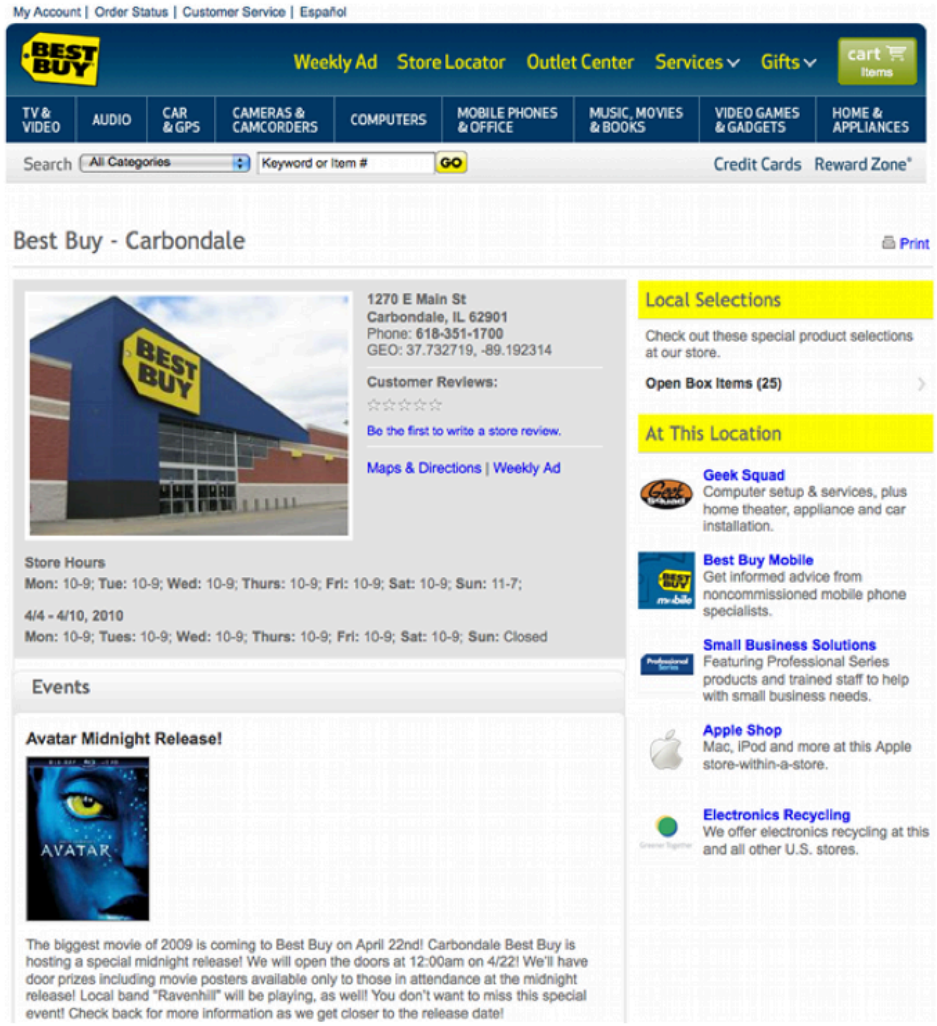
Google's rich sniplet

- Embedded metadata (microformat or RDFa) is used to improve search result page
  - at the moment only a few vocabularies are recognized, but that will evolve over the years

Effects of, e.g., Google

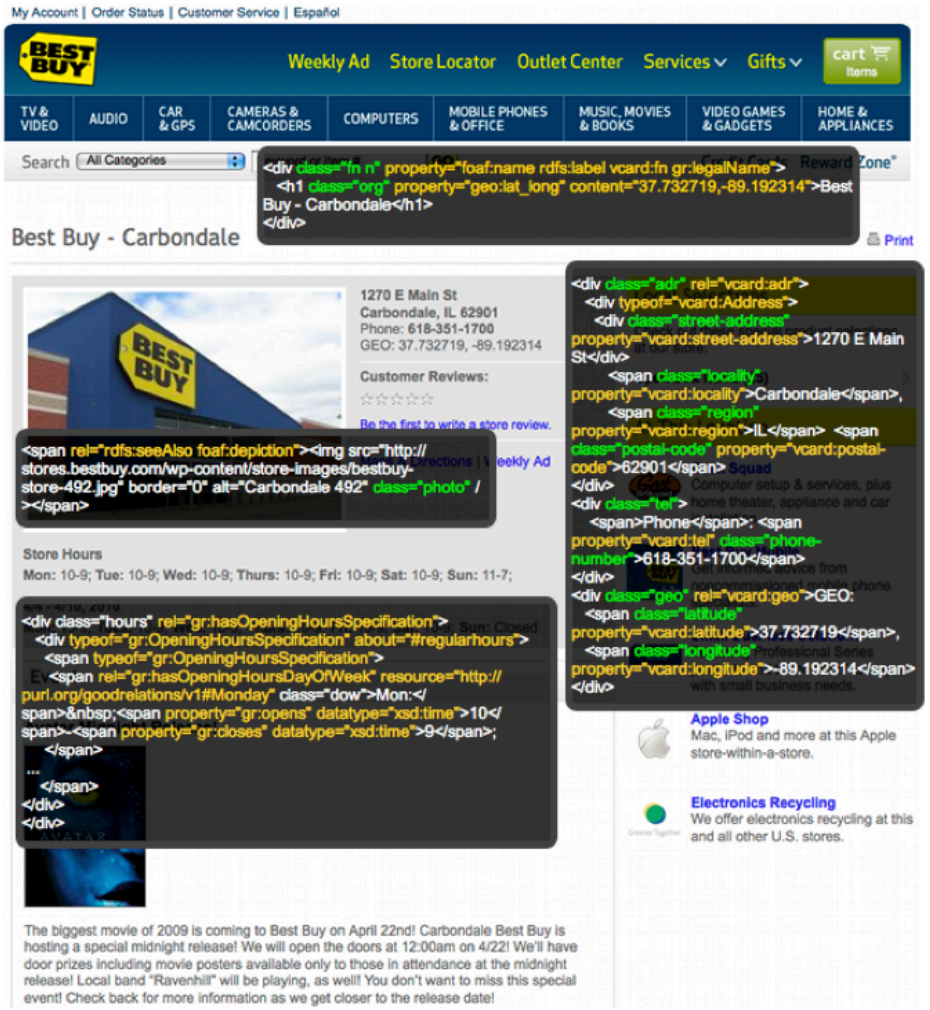
- A number of popular sites publish RDFa as part of their normal pages:
  - Tesco, BestBuy, Slideshare, The London Gazette, Newsweek, MSNBC, O'Reilly Catalog, the White House…
  - Creative Commons snippets are in RDFa

BestBuy Example for RDFa Usage



(78) *Courtesy of Jay Myers, BestBuy, SemTech2010 Presentation*

BestBuy Example for RDFa Usage

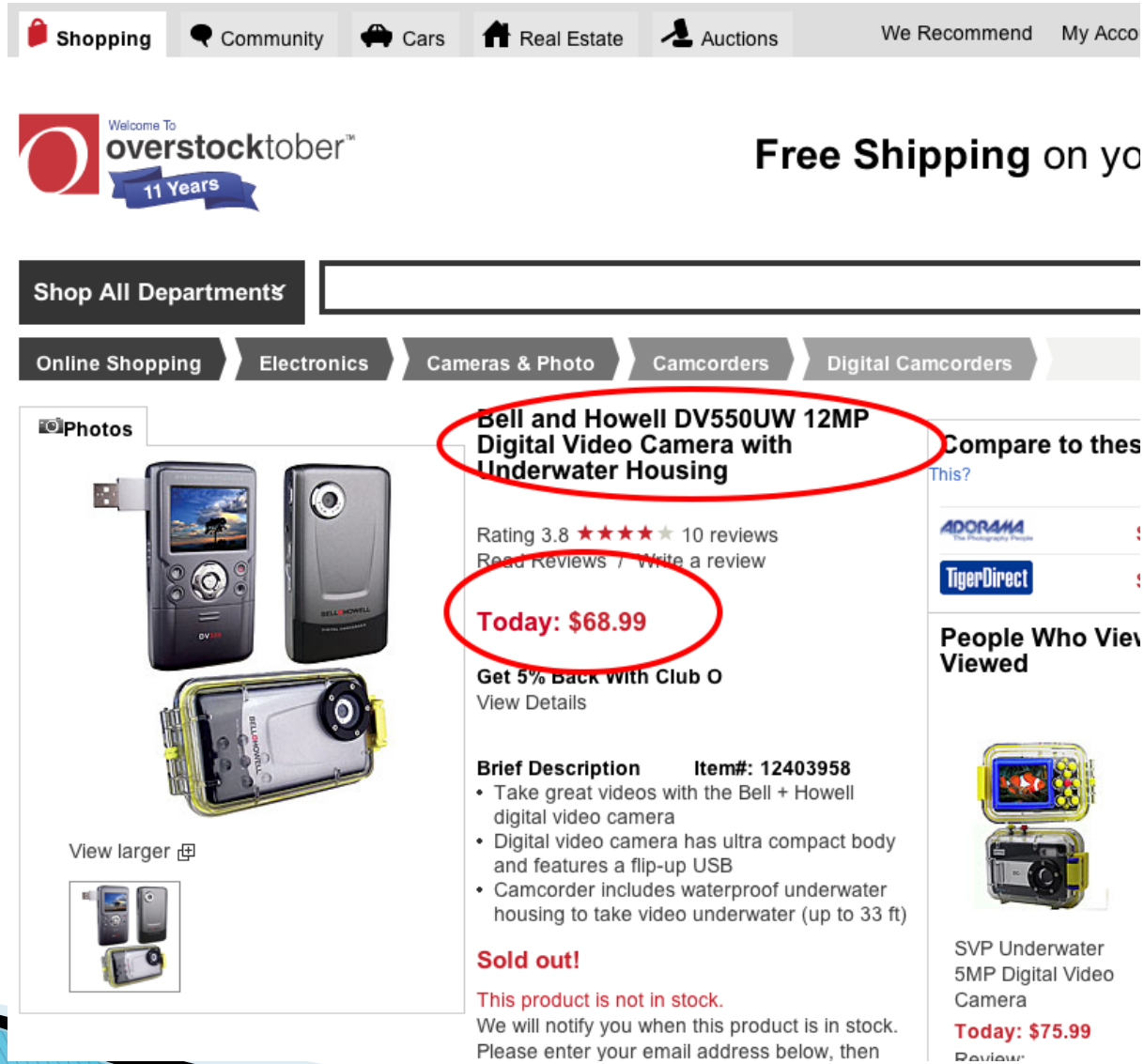


Courtesy of Jay Myers, BestBuy, SemTech2010 Presentation

Effects on BestBuy

- Reported in a BestBuy blog:
  - GoodRelations+RDFa improved Google rank tremendously
  - 30% increase in traffic on BestBuy store pages
  - Yahoo observers a 15% increase in click-through rate
- Not bad…

Overstock.com example



Overstock.com example

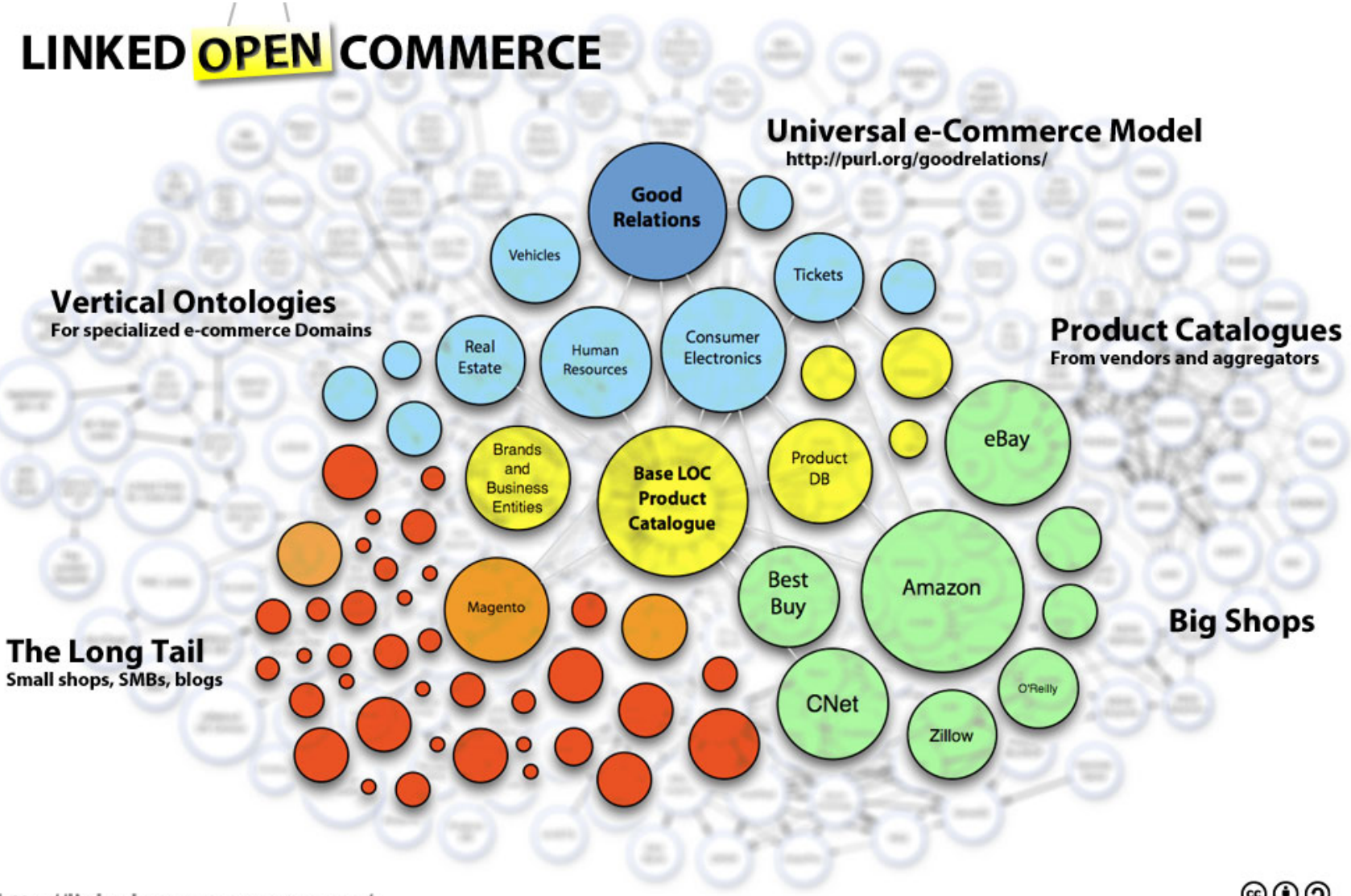
http://www.w3.org/2007/08/pyRdfa/extract?format=turtle&uri=http%3A//www.overstock.co...d-Howell-DV550UW-12MP-Digital-Video-Camera-with-

http://www.w3.org/2007/08/pyRdfa/extract?format=turtle&uri=http%3A//www.overstock.com/Electronic

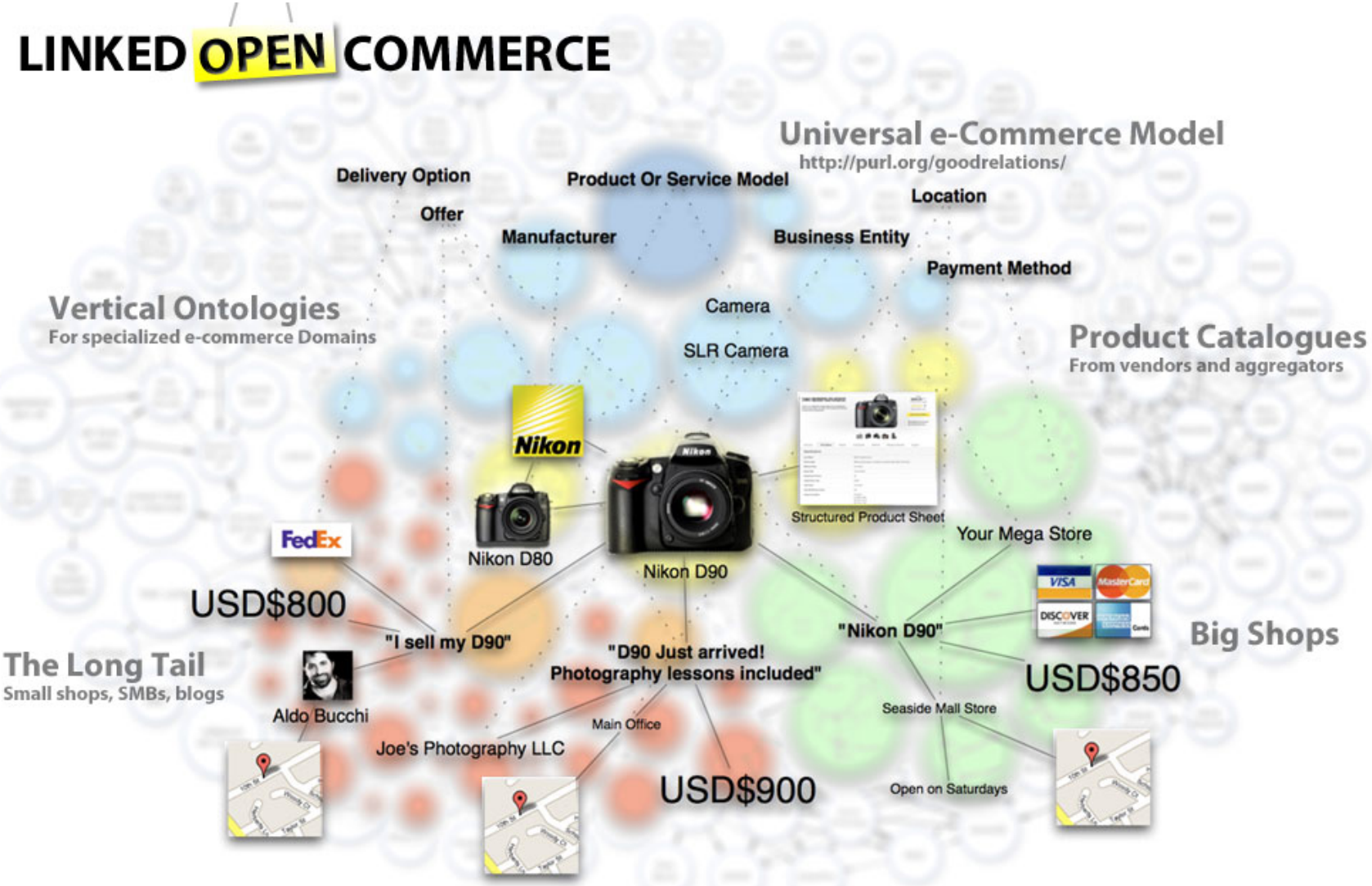
Social ▾ Private ▾ Mailing lists ▾ SW ▾ Python ▾ RDFa it! ▾ Bookmarklets ▾ To Mendeley To Faviki bit.ly MID Add Zemanta SOGo

```
with-Underwater-Housing/4450313/product.html#myoffer> .

<http://www.overstock.com/Electronics/Bell-and-Howell-DV550UW-12MP-Digital-Video-Camera-with-Underwater-Hou
reviews.html> v:count "10" .

<http://www.overstock.com/Electronics/Bell-and-Howell-DV550UW-12MP-Digital-Video-Camera-with-Underwater-Hou
xhv:icon <http://ak1.ostkcdn.com/favicon.ico> ;
     xhv:stylesheet <http://ak1.ostkcdn.com/css/ProductPageC.css>, <http://ak1.ostkcdn.com/css/os_master.le
<http://ak1.ostkcdn.com/css/productPageImgGallery.css> .

<http://www.overstock.com/Electronics/Bell-and-Howell-DV550UW-12MP-Digital-Video-Camera-with-Underwater-
Housing/4450313/product.html#TypeAndQuantityNode> a gr:TypeAndQuantityNode ;
     gr:hasUnitOfMeasurement "C62"^^xsd:string ;
     gr:typeOfGood <http://www.overstock.com/Electronics/Bell-and-Howell-DV550UW-12MP-Digital-Video-Camera-
Housing/4450313/product.html#product> .

<http://www.overstock.com/Electronics/Bell-and-Howell-DV550UW-12MP-Digital-Video-Camera-with-Underwater-
Housing/4450313/product.html#UnitPriceSpecification> a gr:UnitPriceSpecification .

<http://www.overstock.com/Electronics/Bell-and-Howell-DV550UW-12MP-Digital-Video-Camera-with-Underwater-
Housing/4450313/product.html#myoffer> a gr:Offering ;
     gr:hasBusinessFunction gr:Sell ;
     gr:hasPriceSpecification
         [ gr:hasCurrency "USD"^^xsd:string ;
             gr:hasCurrencyValue "68.99"^^xsd:float
         ] <http://www.overstock.com/Electronics/Bell-and-Howell-DV550UW-12MP-Digital-Video-Camera-with-Ur
Housing/4450313/product.html#UnitPriceSpecification> ;
```

More general: “linked open commerce”



More general: “linked open commerce”



- Publishing RDFa is an important step in combining the Semantic Web and the "traditional" Web
- But publishing is not always straightforward for a lambda Web designer and user
- This is where the role of Drupal 7 is huge!
  - make the publication of data in RDFa part of the normal CMS operation

So let us see what Drupal 7 can offer!

Thank you for your attention!

These slides are also available on the Web:

http://www.w3.org/2010/Talks/RDFa-Drupal-Tutorial/RDFa.pdf