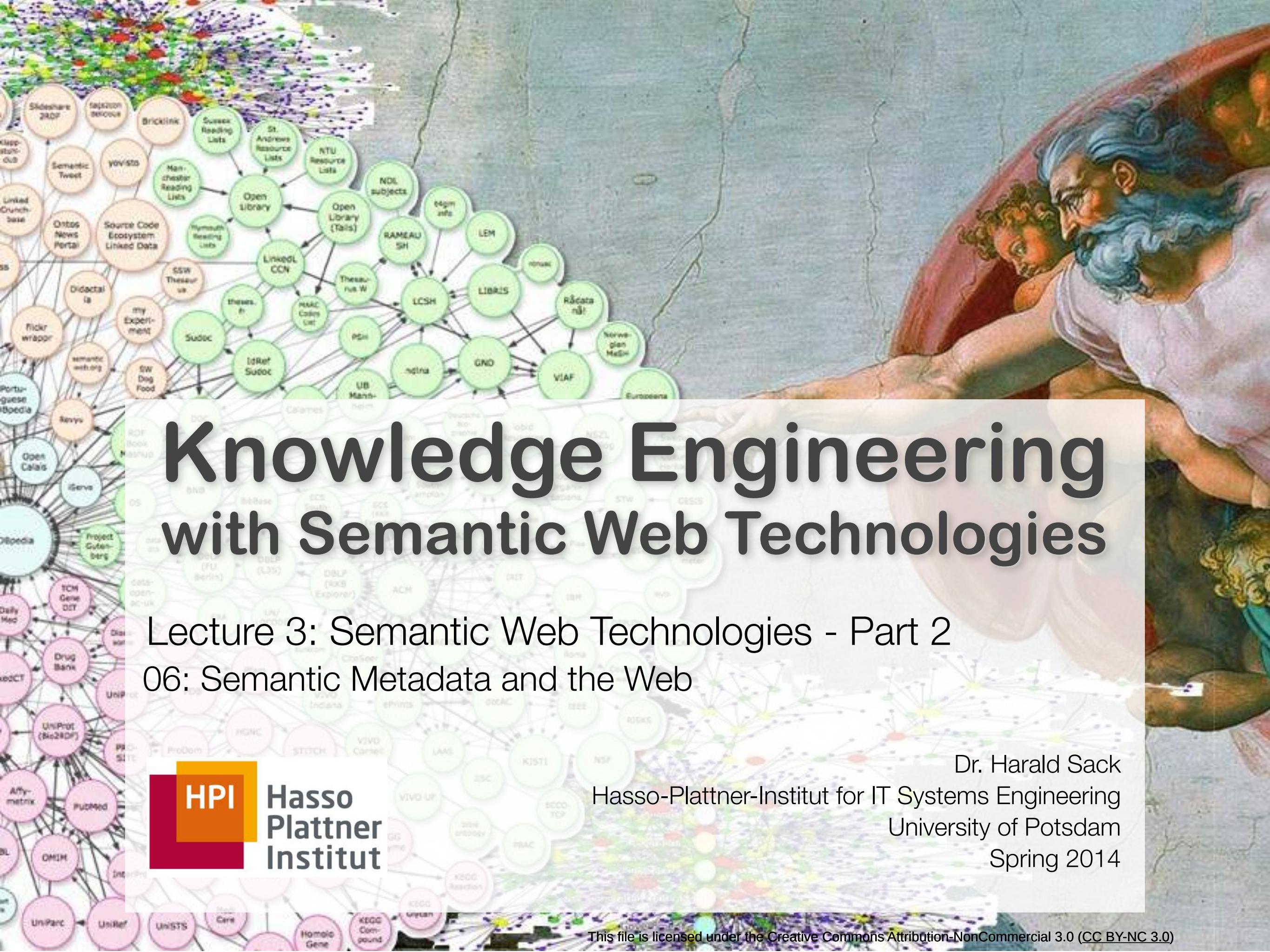


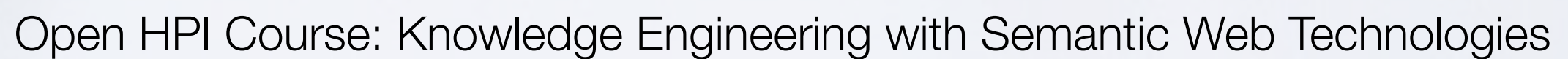


Knowledge Engineering with Semantic Web Technologies

Lecture 3: Semantic Web Technologies - Part 2
06: Semantic Metadata and the Web



Dr. Harald Sack
Hasso-Plattner-Institut for IT Systems Engineering
University of Potsdam
Spring 2014





06 Semantic Meta Data and the Web

Open HPI Course: Knowledge Engineering with Semantic Web Technologies
Lecture 03: Semantic Web Technologies Pt.2



3. Semantic Web Technologies Pt.2

Semantic Metadata and the Web

The Semantic Web Technology Stack (not a piece of cake...)

Most apps use only a subset of the stack

Querying allows fine-grained data access

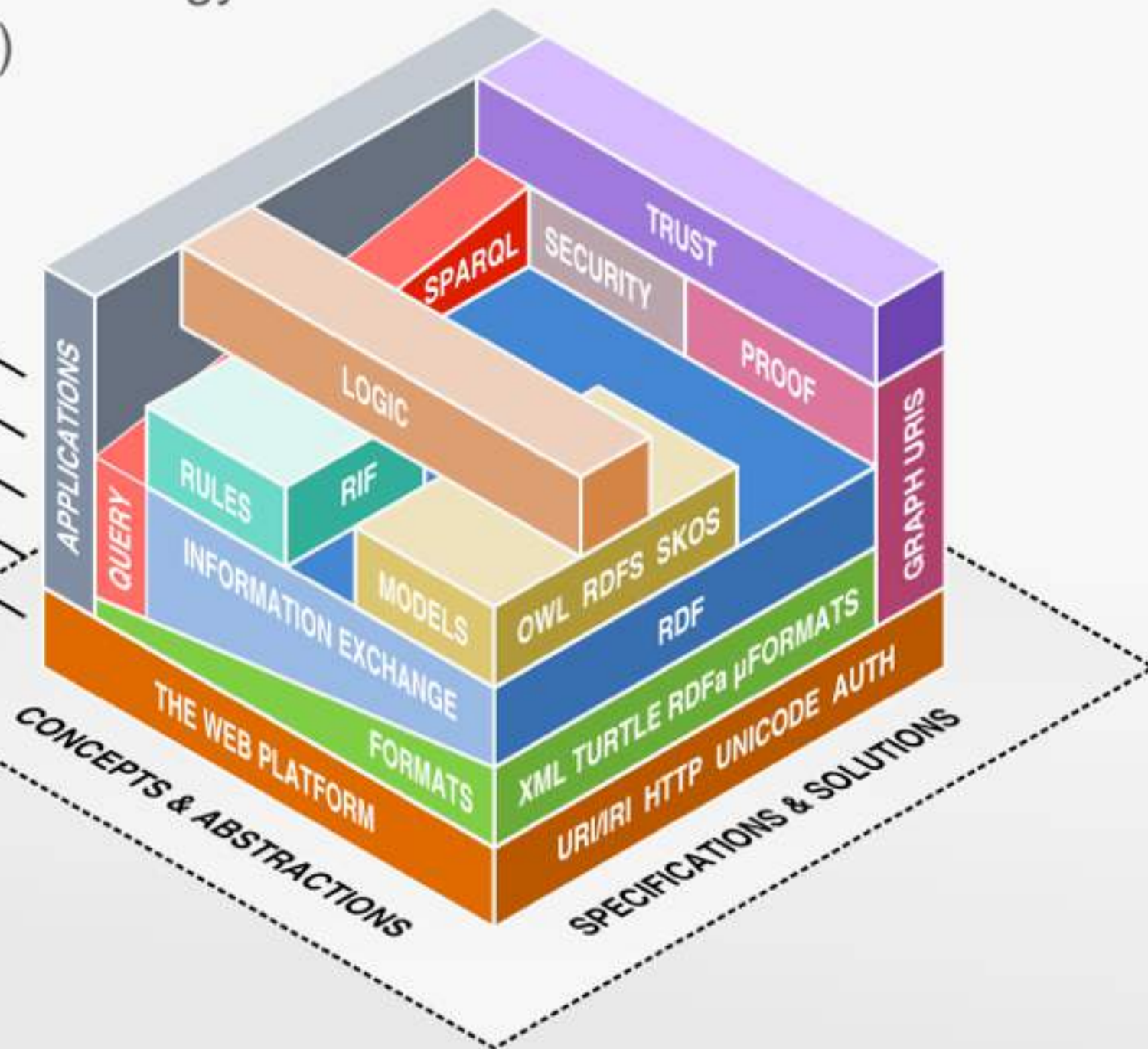
Standardized information exchange is key

Formats are necessary, but not too important

The Semantic Web is based on the Web

Linked Data uses a small
selection of technologies

LINKED DATA



- RDF/XML can directly be embedded in an HTML document via `<head>` and `<script>`
- RFC 3870 defines Mime-Type **application/rdf+xml**

```
<head>
  <title>My Document</title>
  <script type="application/rdf+xml">
    <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://www.w3.org/" dc:title="W3C Homepage"/>
    </rdf:RDF>
  </script>
</head>
```

- W3C recommended linking of external RDF documents via HTML `<link>` element in HTML `<head>`

```
<link rel="meta" type="application/rdf+xml"
      href="Meta-Data-for-Web-Page.rdf"/>
```

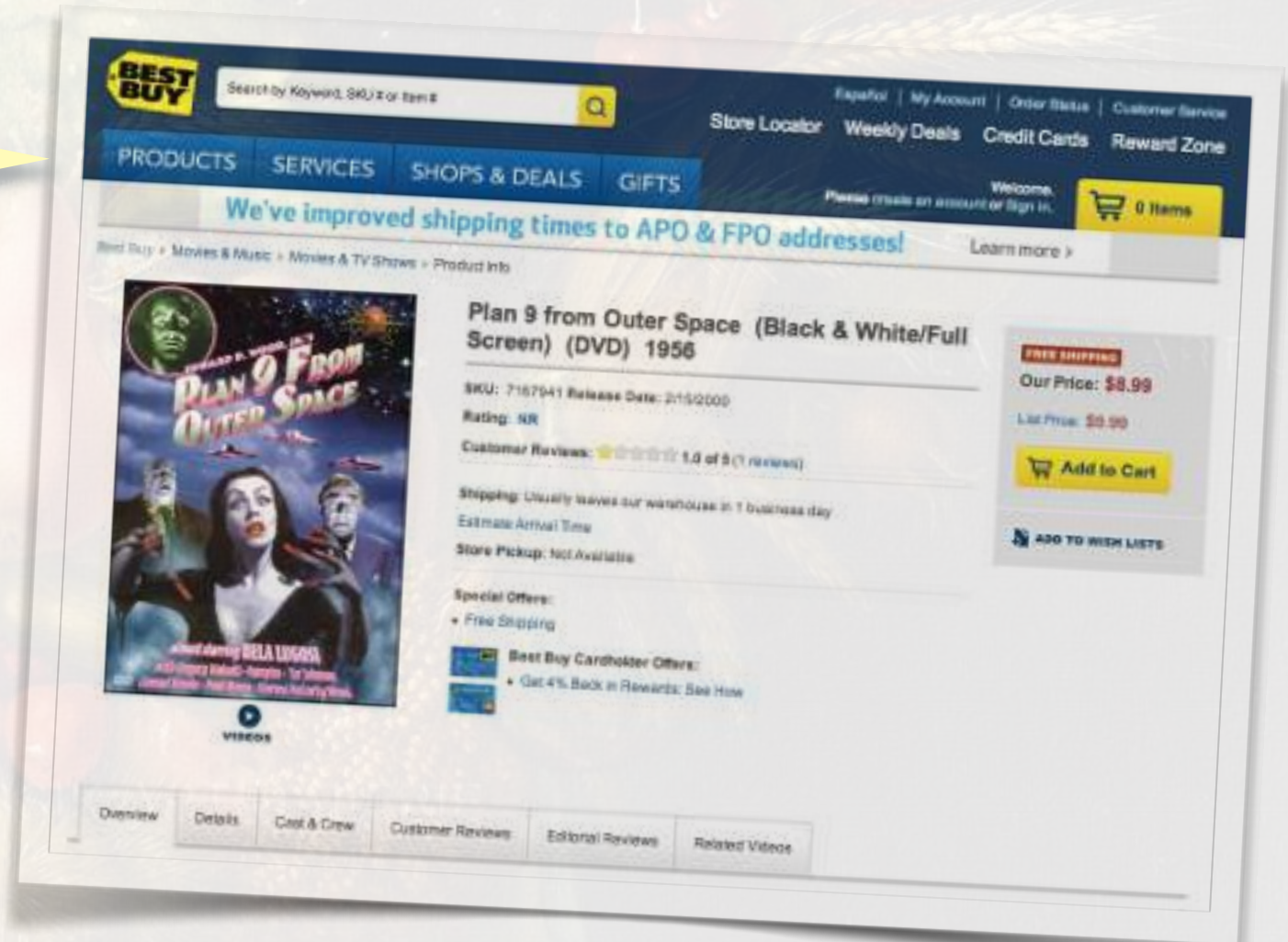
- RDF documents can also be embedded via `<a href=...>` and additional markup (e.g. RDF logo) within HTML `<body>`

- **Problem:** HTML and RDF data are separate

RDF data
Turtle

web page
HTML

```
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan  
%209%20from%20outer%20space&lp=6&cp=1> <http://  
ogp.me/ns#title> "Plan 9 from Outer Space - DVD" .  
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan  
%209%20from%20outer%20space&lp=6&cp=1> <http://  
ogp.me/ns#type> "product" .  
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan  
%209%20from%20outer%20space&lp=6&cp=1> <http://  
ogp.me/ns#url> "http://www.bestbuy.com:80/site/Plan-9-  
from-Outer-Space---DVD/6429224.p?  
id=23883&skuld=7167941" .  
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan  
%209%20from%20outer%20space&lp=6&cp=1> <http://  
ogp.me/ns#image> "http://images.bestbuy.com:80/  
BestBuy_US/images/products/6429/6429224s.jpg" .  
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan
```



- **Maintenance**
- **Verification**
- **Consistency**

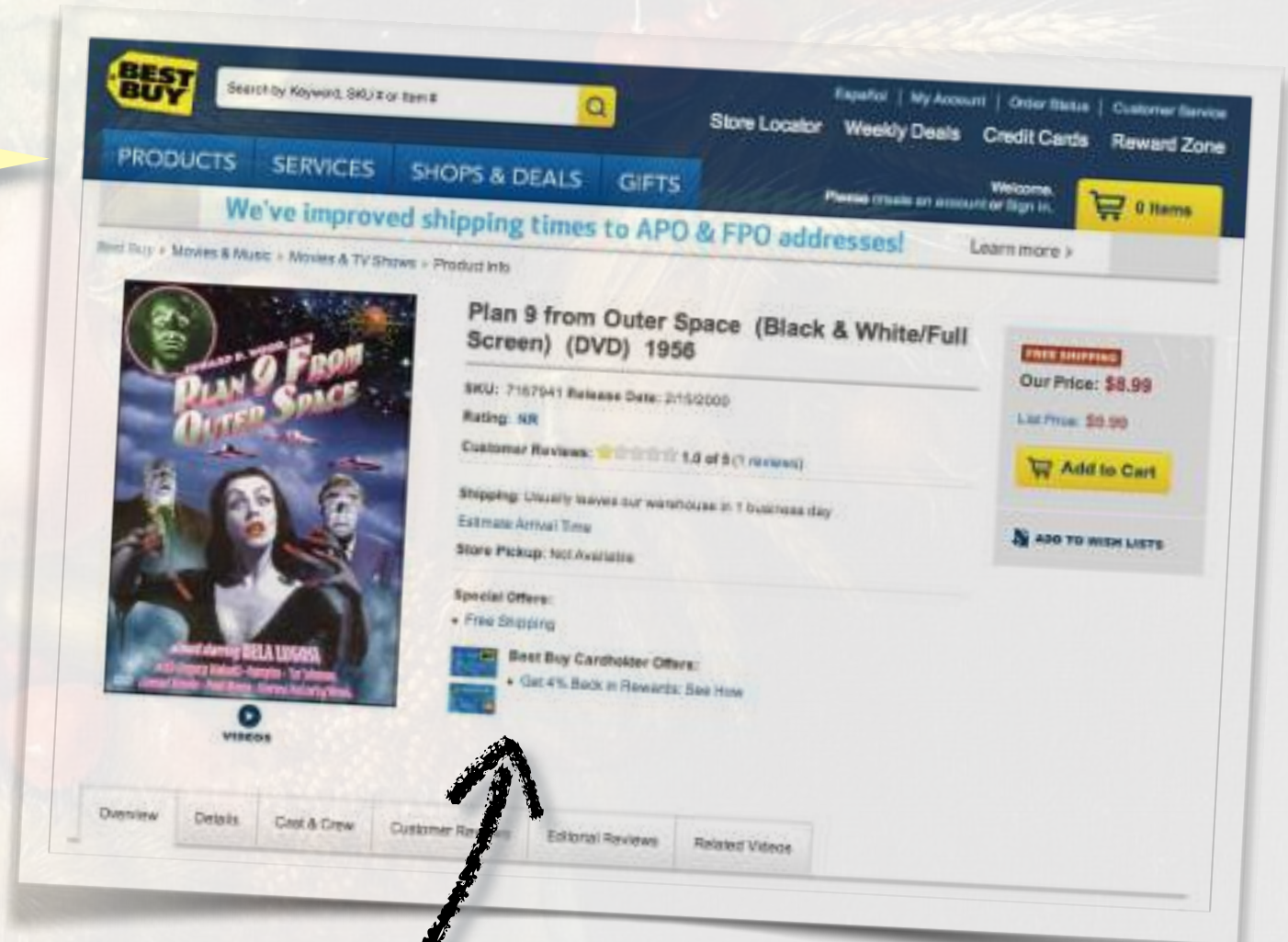
Semantic Annotation in the Web

→ ,Embed' RDF directly into HTML

RDF data
Turtle

web page
HTML

```
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan  
%209%20from%20outer%20space&lp=6&cp=1> <http://  
ogp.me/ns#title> "Plan 9 from Outer Space - DVD" .  
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan  
%209%20from%20outer%20space&lp=6&cp=1> <http://  
ogp.me/ns#type> "product" .  
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan  
%209%20from%20outer%20space&lp=6&cp=1> <http://  
ogp.me/ns#url> "http://www.bestbuy.com:80/site/Plan-9-  
from-Outer-Space---DVD/6429224.p?  
id=23883&skuld=7167941" .  
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan  
%209%20from%20outer%20space&lp=6&cp=1> <http://  
ogp.me/ns#image> "http://images.bestbuy.com:80/  
BestBuy_US/images/products/6429/6429224s.jpg" .  
<http://www.bestbuy.com/site/Plan+9+from+Outer+Space  
+-+DVD/7167941.p?id=23883&skuld=7167941&st=plan
```



- In principle there are three ways to embed structured data with explicit semantic annotations within HTML documents



Domain specific microformats



Generic RDFa and embedded RDF

- HTML5 Microdata including

schema.org



- Microformats (μformats) emerged about 2005
- (X)HTML Markup to express (limited) semantics in an HTML document
 - designed to solve simple, specific problems
 - designed for humans first, machines second
 - used in web pages to describe a specific type of information, as e.g. a person, an event, a product, a review, etc.
- Applications can easily extract data from HTML documents
- In general, Microformats use the **class** attribute in HTML tags (most times **** or **<div>** tags) and assign brief and descriptive names to entities and their properties



- Microformats reuse the following (X)HTML tag attributes:
 - **class**
 - **rel**
 - **rev**
- Predefined standard microformats:
 - **hCard** - personal data (vCard, RFC2426)
 - **hCalendar** – calendars and events
 - **rel-Tag** – tags, keywords, categories
 - **XFN** – XHTML friends network
 - **hReview, VoteLinks** -- opinions, ratings, and reviews
 - **XOXO** – lists and outlines
 - ...



- Simple Example: HTML marked up with **hCard** microformat

```
<div class="vcard">
  <span class="fn">Joe Blow</span>
  <span class="title">Senior Manager</span>
  <span class="org">The Example Company</span>
  <span class="adr">
    <span class="street-address">Hauptstr. 123, </span>
    <span class="postal-code">14482</span>
    <span class="locality">Potsdam</span>
  </span>
  Tel.<span class="tel">604-555-1234</span>
  
</div>
```



- Microformats can **easily be transcoded to RDF** via XSLT
- **New microformat** vocabularies **first must be consolidated** by the community, while always a new XSLT stylesheet has to be developed for extraction
- By using **more than one** microformat vocabulary in a single (X)HTML document the processing **complexity increases** rapidly
- **Conflicts** with used (X)HTML attributes might be possible



Embedding RDF in HTML Attributes



- RDFa = **RDF** in HTML **a**tttributes
- enables generic RDF annotation in (X)HTML documents by reusing existing (X)HTML attributes
- RDFa 1.0 based on XHTML (W3C Recommendation 2008)
- RDFa 1.1 based on HTML5 (W3C Recommendation June 2012)
 - **RDFa Lite 1.1**
 - RDFa 1.1



- RDFa reuses existing (X)HTML attributes (e.g. href, src) and introduces new HTML attributes
 - vocab,
 - typeof,
 - property,
 - resource,
 - prefix



- First we need a **vocabulary** to talk about things

```
<p vocab="http://xmlns.com/foaf/0.1/">  
  My name is Harald Sack and you can  
  give me a ring via 1-800-555-0527.  
</p>
```



- Then we have to define the **type of thing** we are talking about

```
<p vocab="http://xmlns.com/foaf/0.1/"  
  typeof="Person">  
  My name is Harald Sack and you can  
  give me a ring via 1-800-555-0527.  
</p>
```



- Now we can define all **properties** of the thing we are talking about

```
<p vocab="http://xmlns.com/foaf/0.1/" typeof="Person">  
  My name is  
  <span property="name">Harald Sack</span>  
  and you can give me a ring via  
  <span property="phone">1-800-555-0527</span>.  
    
</p>
```



- We can create **identifiers** for the things we are talking about

```
<p vocab="http://xmlns.com/foaf/0.1/"  
  resource="#harald" typeof="Person">  
  My name is  
  <span property="name">Harald Sack</span>  
  and you can give me a ring via  
  <span property="phone">1-800-555-0527</span>.  
    
</p>
```

- **resource** refers to the base URI of the webpage



- And if the vocabulary is not sufficient to describe all properties, we can use additional vocabularies by using **prefixes**

```
<p vocab="http://xmlns.com/foaf/0.1/"  
  prefix="ov: http://open.vocab.org/terms/"  
  resource="#harald" typeof="Person">  
  My name is  
  <span property="name">Harald Sack</span>  
  and you can give me a ring via  
  <span property="phone">1-800-555-0527</span>.  
  My favorite beverage is  
  <span property="ov:preferredBeverage">Green Tea  
  </span>.  
</p>
```



- Distinguish two different sorts of RDF triples
 - Triple with resource as object
 - Triple with literal as object

	Subject	Property	Object
Object is Literal	resource/about	property	content or #PCDATA
Object is Resource (URI)	resource/about	property/rel	href or resource

about / rel only for compatibility with RDFa 1.0



Differences with RDFa 1.0

<html>

...

<div xmlns:dc="http://purl.org/dc/elements/1.1/">

I'm currently reading

Programming Perl

by

Larry Wall

.

</div>

...

</html>

Namespace

Subject

Property

Object
(Literal)



Differences with RDFa 1.0

```
<?xml version="1.0" encoding="UTF-8"?>
<html xmlns="http://www.w3.org/1999/xhtml" version="XHTML+RDFa 1.0"
  xml:lang="en">
  <head>...</head>
  <body>...</body>
</html>
</xml>
```

XHTML

```
<!DOCTYPE html>
<html version="HTML+RDFa 1.1" lang="en"
  xmlns="http://www.w3.org/1999/xhtml"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#"
  xmlns:xsd="http://www.w3.org/2001/XMLSchema#"
  xmlns:cc="http://creativecommons.org/ns#"
  xmlns:dc="http://purl.org/dc/elements/1.1/"
  xmlns:foaf="http://xmlns.com/foaf/0.1/">
```

HTML5



RDFa



RDFa is already applied by

- Google, Yahoo
- Facebook, MySpace, LinkedIn
- Best Buy, Tesco, O'Reilly
- SlideShare, Digg
- WhiteHouse.gov, Library of Congress, UK government
- Newsweek, BBC
- ...



RDFa

...more RDFa Extractors

RDFa Play Documentation Tools Developers

Examples: Person Social Network Event Place Product SVG

```
xsd: http://www.w3.org/2001/XMLSchema#  
typeof="gr:Offering">  
<div>  
  <h1 property="gr:name">Canon Rebel T2i (EOS 550D)</h1>  
  <div rel="foaf:depiction">  
      
    The Canon Rebel T2i (EOS 550D) is Cannon's top-of-the-line consumer digita  
    It can shoot up to 18 megapixel resolution photos and features an ISO rang  
  </p>  
  <link rel="gr:hasBusinessFunction" href="http://purl.org/goodrelations/v1#Se  
  <meta property="gr:hasEAN_UCC-13" content="013803123784" />  
  Sale price:  
  <span property="gr:hasPriceSpecification" typeof="gr:UnitPriceSpecification"  
    <span property="gr:hasCurrency" content="USD">$</span>  
    <span property="gr:hasCurrencyValue" datatype="xsd:float">899</span>  
  </span>  
  <link rel="gr:acceptedPaymentMethods" href="http://purl.org/goodrelations/v1  
  <link rel="gr:acceptedPaymentMethods" href="http://purl.org/goodrelations/v1  
  [<a rel="foaf:page" href="http://shop.usa.canon.com/">more...</a>]  
</div>  
</div>
```

Canon Rebel T2i (EOS 550D)



The Canon Rebel T2i (EOS 550D) is Cannon's top-of-the-line consumer digital SLR camera. It can shoot up to 18 megapixel resolution photos and features an ISO range of 100-6400.

Sale price: \$ 899 [\[more...\]](#)

Visualization Raw Data

Web Page

- Item 1
 - type: Offering
 - name: Canon Rebel T2i (EOS 550D)
 - depiction: http://shop.usa.canon.com/wcsstore/eStore/images/t2ikit_1_1.jpg
 - description: The Canon Rebel T2i (EOS 550D) is Cannon's top-of-the-line consumer digital SLR camera. It can
 - hasBusinessFunction: http://purl.org/goodrelations/v1#Sell
 - hasEAN_UCC-13: 013803123784
 - hasPriceSpecification: Item 2
 - acceptedPaymentMethods: http://purl.org/goodrelations/v1#PayPal
 - acceptedPaymentMethods: http://purl.org/goodrelations/v1#MasterCard
 - page: http://shop.usa.canon.com/
- Item 2
 - type: UnitPriceSpecification
 - hasCurrency: USD
 - hasCurrencyValue: 899



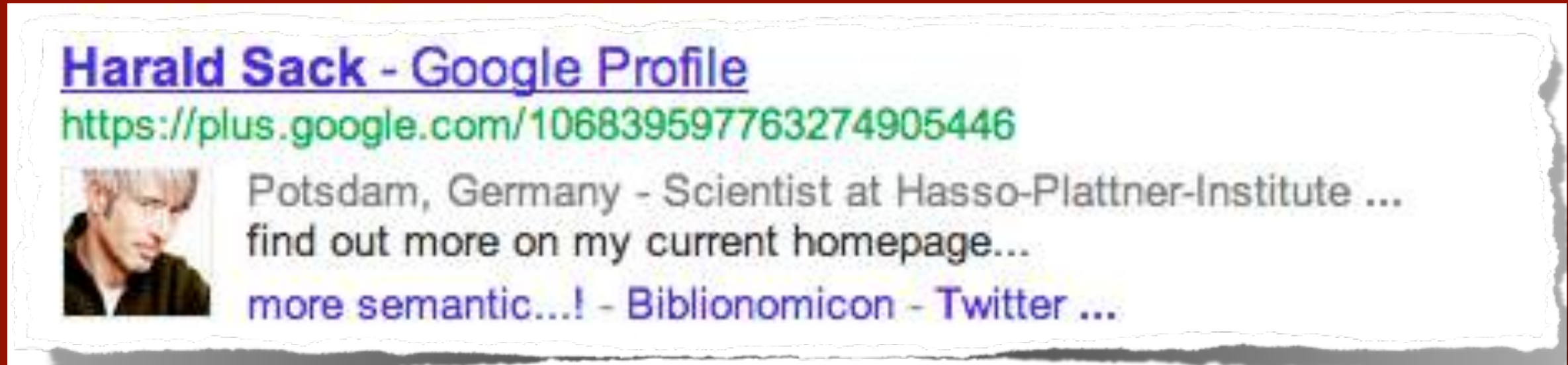
HTML5 microdata

- **HTML5 Microdata** extends microformats and addresses its shortcomings
- **schema.org** is the most prominent microdata vocabulary

schema.org

- Initiative started by Bing, Google, Yahoo! on June 2, 2011.
- **Goals:**
 - Semantic markup for web page content based on proprietary **HTML5 microdata schemata**
- **Search engines, web crawler, and Browser** recognize and understand schema.org markup
- small number of formats (extensible)
- RDF mapping available

schema.org and Rich Snippets



- **Snippets** = The few lines of text that appear under every search result
 - Designed to give users a sense for what's on the web page and why it's relevant to their query
- If Google understands the content on the web page, it can create **“Rich” Snippets**
 - = detailed information intended to help users with specific queries
 - You can help Google find this relevant information by adding additional microdata as HTML markup

HTML5 microdata

- **Microdata** extends microformats and addresses its shortcomings
- **Microdata** annotates the DOM with scoped name/value pairs from custom vocabularies
 - centered around **custom vocabularies**:
anyone can define a microdata vocabulary and start embedding custom properties in their own web pages
 - works with **name/value pairs**:
every microdata vocabulary defines a set of named properties
 - applies concept of “**scoping**”:
microdata re-use the hierarchical structure of the DOM itself to provide a way to say “all the properties within this element are taken from this vocabulary”

HTML5 microdata

- HTML5 microdata specification comprises
 - **microdata Vocabularies** (<http://data-vocabulary.org/>)
 - Persons, Events, Organisations, Products, Reviews, Breadcrumbs, Ads, aggregated Reviews/Ads
 - **microdata Global Attributes**
 - **itemscope** (creates resource)
 - **itemtype** (applied vocabulary)
 - **itemid** (identifies resource)
 - **itemprop** (property of a resource)
 - **itemref** (references on other resources)

HTML5 microdata

```
<section itemscope itemtype="http://data-vocabulary.org/Person">
  Hello, my name is
  <span itemprop="name">John Doe</span>,
  I am a
  <span itemprop="title">graduate research assistant</span>
  at the
  <span itemprop="affiliation">University of Dreams</span>.
  My friends call me
  <span itemprop="nickname">Johnny</span>.
  You can visit my homepage at
  <a href="http://www.JohnnyD.com" itemprop="url">www.JohnnyD.com</a>.
  <section itemprop="address" itemscope
    itemtype="http://data-vocabulary.org/Address">
    I live at
    <span itemprop="street-address">1234 Peach Drive</span>
    <span itemprop="locality">Warner Robins</span>
    ,
    <span itemprop="region">Georgia</span>.
  </section>
</section>
```

<http://yovisto.com/labs/swt1112/microdata01.html>

HTML5 microdata

Extracted rich snippet data from the page

Item

Type: `http://data-vocabulary.org/person`

`name` = John Doe

`title` = graduate research assistant

`affiliation` = University of Dreams

`nickname` = Johnny

`url`

`text` = `www.JohnnyD.com`

`href` = `http://www.johnnyd.com/`

`address` = *Item(1)*

Item 1

Type: `http://data-vocabulary.org/address`

`street-address` = 1234 Peach Drive

`locality` = Warner Robins

`region` = Georgia

<http://www.google.com/webmasters/tools/richsnippets>

schema.org

- extends microdata vocabulary with proprietary **category hierarchies**:
 - Creative works:
 - CreativeWork, Book, Movie, MusicRecording, Recipe, TVSeries ...
 - Embedded non-text objects:
 - AudioObject, ImageObject, VideoObject
 - Event
 - Organization
 - Person
 - Place, LocalBusiness, Restaurant ...
 - Product, Offer, AggregateOffer
 - Review, AggregateRating
- and according **properties** (relations) with different domain and range

schema.org



**13% of all HTML pages
contain structured data**

Source: WebDataCommons.org, February 2012

- 1.4 billion HTML pages parsed (Common Crawl corpus)
- 188 million pages contained Microformat, Microdata, RDFa data

**30% of all HTML pages
contain structured data**

Source: Yahoo! Research, January 2012

- 3.2 billion HTML pages parsed (Bing Crawl corpus)
- 973 million pages contained Microformat, Microdata, RDFa data

Structured Data in the Top k Alexa Websites

First x in AL	PLDs in CC		% containing structured data			
	#	% AL	overall	RDFa	Microdata	Microformats
100	99	99.00	74.75	34.34	55.56	68.69
1k	963	96.30	62.62	40.08	31.67	46.11
10k	9,294	92.94	47.34	30.47	15.55	29.75
100k	85,058	85.01	31.94	16.46	7.20	20.07
1m	734,882	73.49	20.56	7.55	3.04	14.18

Bizer et al: Deployment of RDFa, Microdata, and Microformats on the Web – A Quantitative Analysis, (ISWC 2013)

next Lecture:

Lecture 4: Knowledge Representations I

Open HPI Course: Knowledge Engineering with Semantic Web Technologies