XML, Semantic Web and Content Analytics

XML Prague Pre-conference 2014

Felix Sasaki

DFKI / W3C Fellow

What do you need to follow this session?

- Ideal: a computer with internet access, to be able to provide feedback
- At least: curiosity

What is LIDER?

http://www.lider-project.eu/

- Project funding this session
- Aim: bring people like you together with
  - Linked data experts
  - Language technology people

What is Content Analytics?

- A set of technologies to make sense of data
- Quite broad – like usage scenarios
  - Sentiment analysis
  - Business intelligence
  - “Intelligent” web search
  - …

What is Content Analytics?

- Some basic technologies
  - Named entity recognition
    - "Welcome to Prague!" (=city)
  - Relation extraction
    - "Prague is the capital of Czech Republic"

Demo

- Dbpedia spotlight
- Basic named entity recognition
- Data extracted from wikipedia
- Available as a RESTful service:

http://tinyurl.com/dbspotlight-demo

- Deployed in oXygen automatic annotation implementation; see oXygen users session and

http://www.youtube.com/watch?v=F6zIW6blF5k

Issues with content analytics

- Tooling is language specific
  - Most support of course for English
  - Task: “Get content analytics for your language!”
- Content = not only text!
  - E.g. more and more multimedia content
  - Signal analysis has its limits

    “Find my all movies with a kiss scene at the end!”: Won’t work ☹
  - Again textual metadata (e.g. subtitles, closed captions) may help

WHY AM I TELLING YOU ABOUT CONTENT ANALYTICS?

Vision: better content analytics via **structured data!**

- More and more structured data on the Web
  - Textual data with additional information (markup, metadata)
- Huge knowledge bases
  - Wikipedia / DBpedia is just an example
  - Also: Wikidata, Freebase, BabelNet, ...
- Not necessarily created with content analytics in mind – you just need to bring things together

What is XML?

- Details can be skipped here ...
- From the point of view of content analytics:
  - A way to store (semi) structured data
  - A way to store annotations of content that may link to structured data

```
<span ...
its-ta-class-ref="http://nerd.eurecom.fr/ontology#Location"
its-ta-ident-ref="http://dbpedia.org/resource/Prague">
Prague</span>
```

XML aware content analytics

- Content analytics algorithms process pure text
- Knowledge about XML structures improves content analytics quality

```
<para>Mr. Obama<footnote><para>Barack Obama is ...</para></footnote> came to Prague yesterday.</para>
```

Without the knowledge a content analytics tool “sees”:

“Mr. ObamaBarack Obama is ... came to Prague yesterday.”

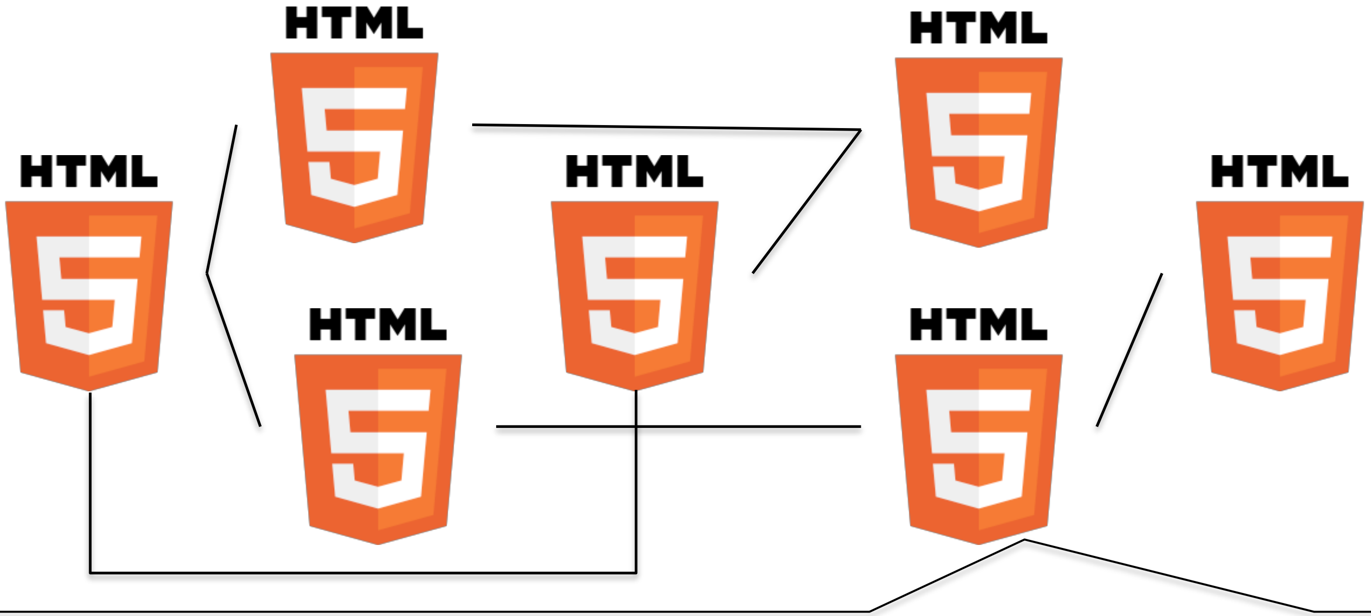
XML aware content analytics

- Need: process different parts of XML content differently
  - Headings: Often not sentences in the linguistic sense
  - Footnotes: embedded in text (see last slide)
  - Non textual content: XML data base structures constitute a mix of structured (non)textual + semi structured data

XML knowledge & tooling could lead to new content analytics approaches

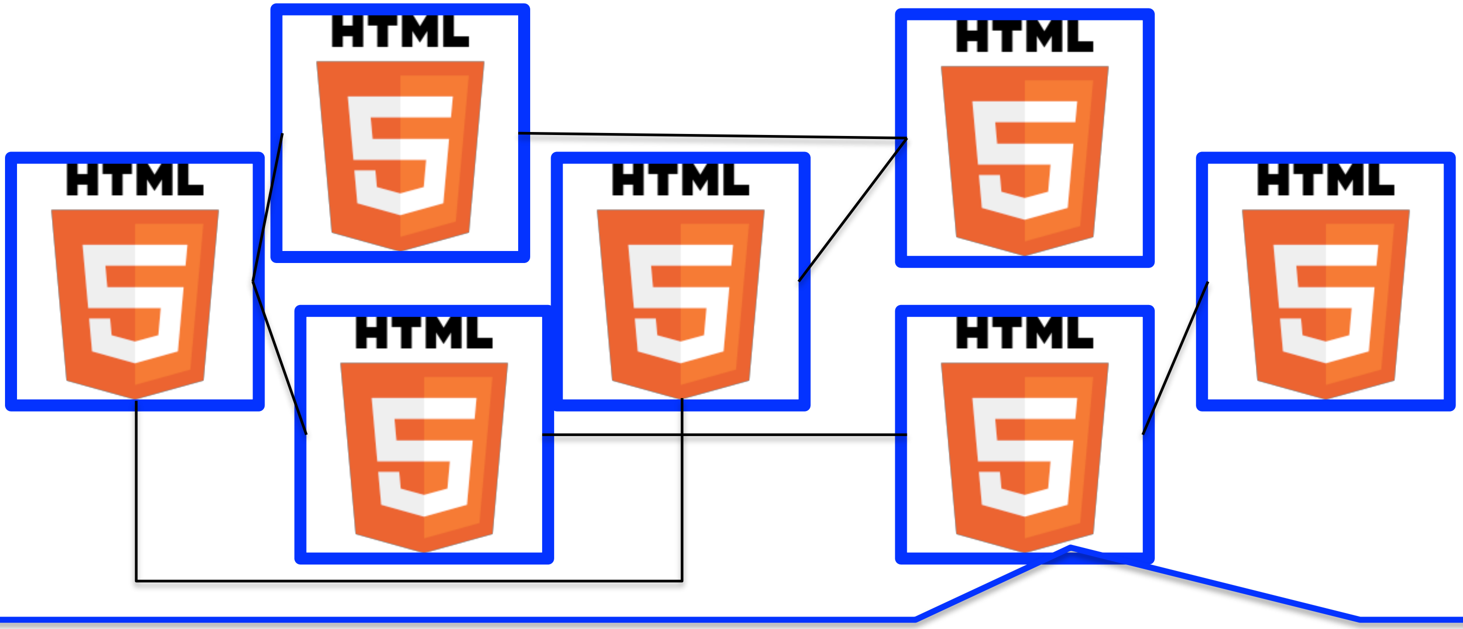
WHAT IS SEMANTIC WEB? SHORT INTRODUCTION

Building blocks of the Web



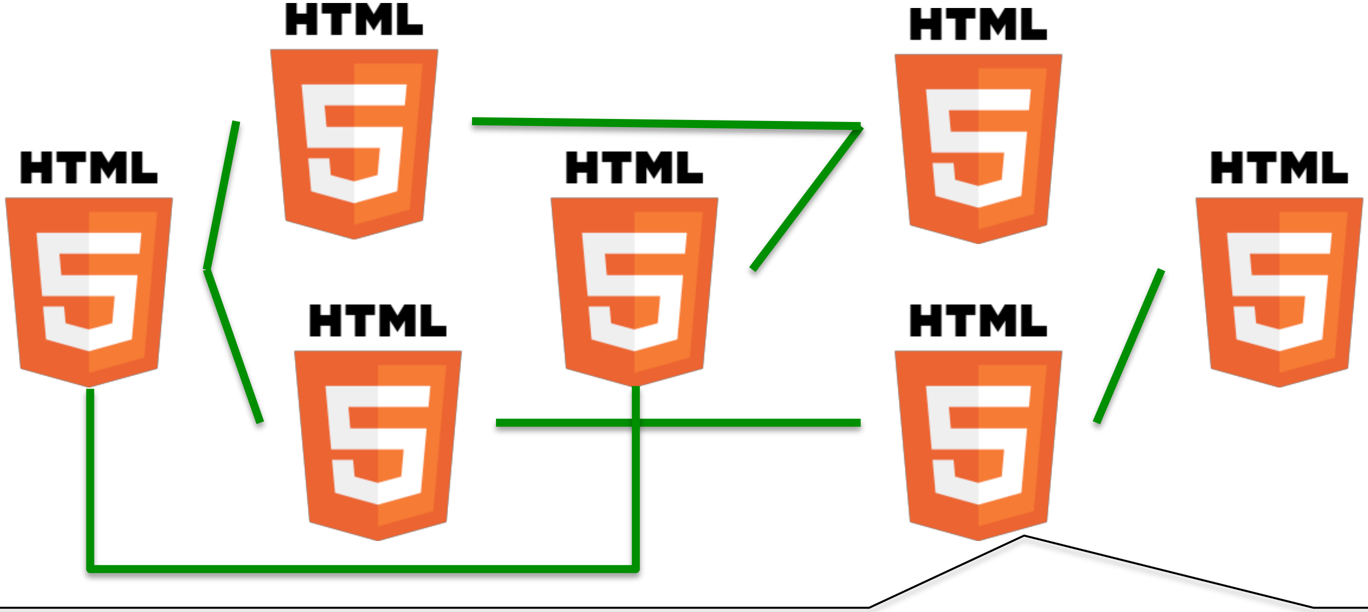
```
<p>All content on this site is licensed under
<a
 href="http://creativecommons.org/licenses/by/3.0/">
   a Creative Commons License</a>. </p>
```

Content



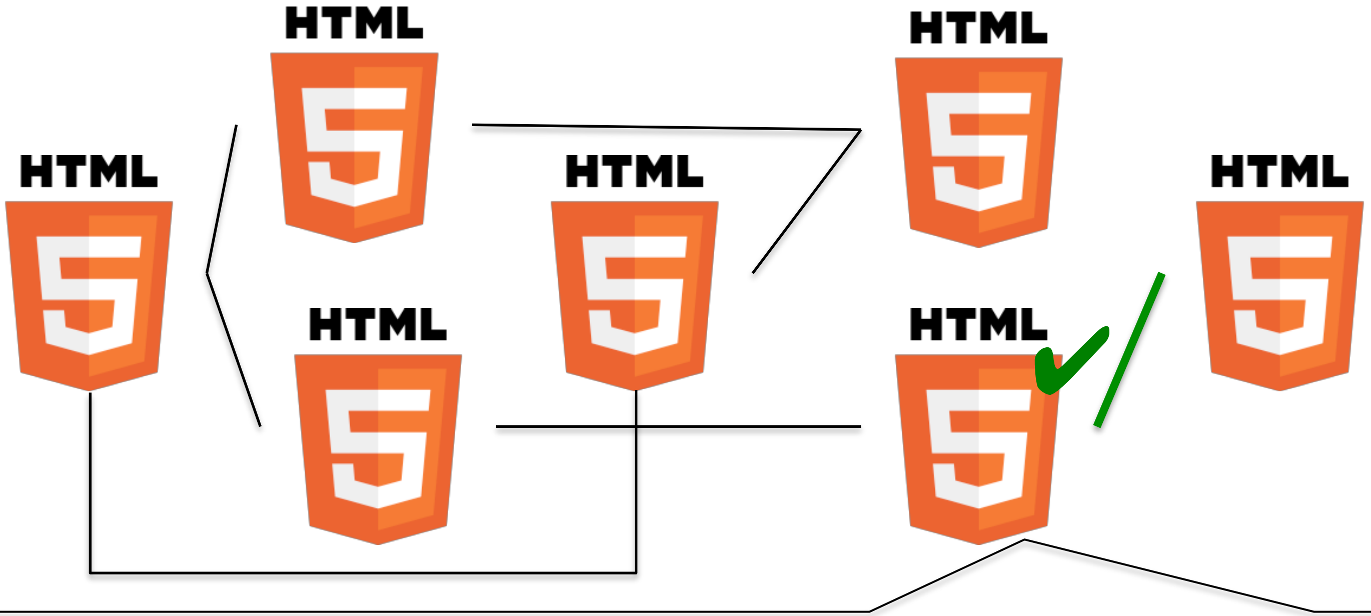
```
<p>All content on this site is licensed under
<a
 href="http://creativecommons.org/licenses/by/3.0/">
   a Creative Commons License</a>. </p>
```

Links (or “identifiers”)



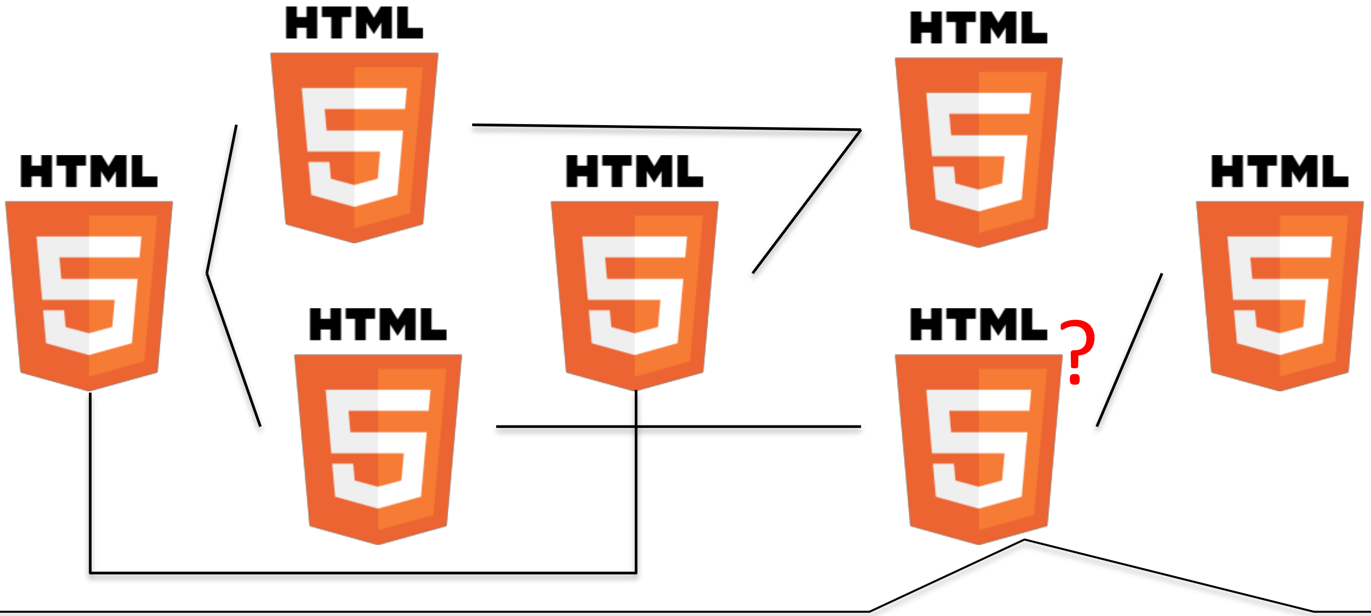
```
<p>All content on this site is licensed under
<a
 href="http://creativecommons.org/licenses/by/3.0/">
   a Creative Commons License</a>. </p>
```

Easy: “Find all content that links to http://creativecommons.org/licenses/by/3.0/“



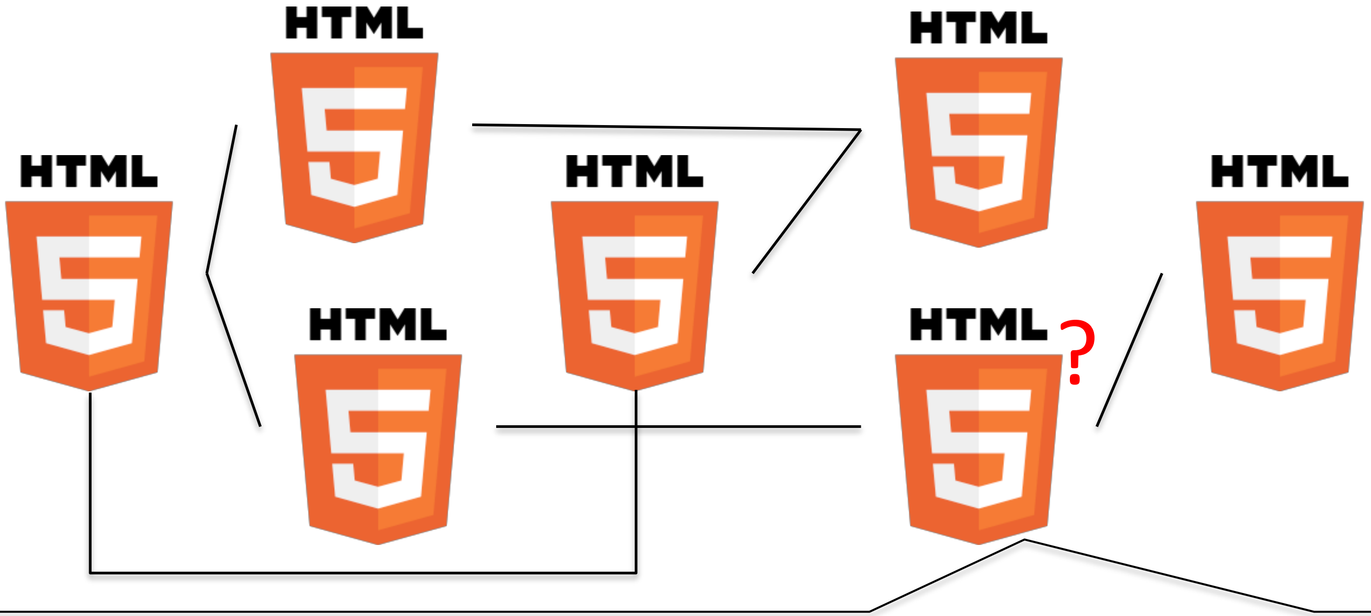
```
<p>All content on this site is licensed under
<a
  href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License</a>. </p>
```

Still difficult: “Find all content that links to a creative commons license“



```
<p>All content on this site is licensed under
<a
 href="http://creativecommons.org/licenses/by/3.0/">
   a Creative Commons License</a>. </p>
```

Semantic Web to the rescue = Providing machine readable information on the Web

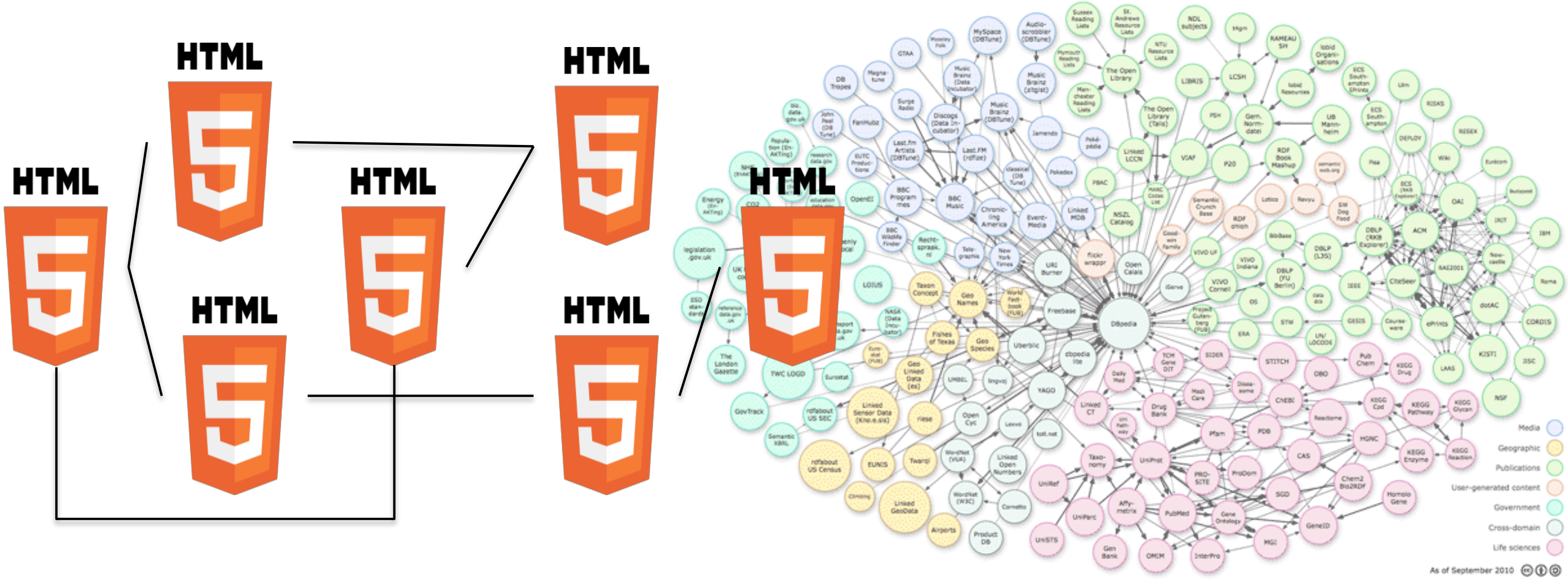


```
<p>All content on this site is licensed under
<a property="http://creativecommons.org/ns#license"
  href="http://creativecommons.org/licenses/by/3.0/">
    a Creative Commons License</a>. </p>
```

Semantic Web = Providing machine readable information on the Web

Web of documents

Web of data



What is Semantic Web - summary

- A set of technologies to work with interlinked data:
  - 1) represent; 2) model; 3) store; 4) process.
  - 1) RDF; 2) RDF Schema, OWL; 3) Turtle / RDFa / ...; 4) SPARQL.
- Also (but not critical here): a means to infer new information from data
  - "X is father of Y" -> "Y is child of X"

INTERACTIVE PART

Aim of this session

- Gather your thoughts on the relation between XML, Semantic Web and content analytics
- Together: go through a LIDER survey & find out:

  “What are

  your use cases for content analytics?”

  http://tinyurl.com/co-an-survey
- That may be related to XML – or not
- Also: gather feedback in free text form: please join me at http://tinyurl.com/ca-gdocs