LINKING WEB DATA

JELENA JOVANOVIC

EMAIL: JELJOV@GMAIL.COM
WEB: HTTP://JELENALJOVANOVIC.NET
Quick reminder:
Gigantic Global Graph &
Web of (linked) data
Gigantic Global Graph

International Information Infrastructure (III)
- network/graph of computers known as Internet or Net
- “It isn't the cables, it is the computers which are interesting”

World Wide Web (WWW)
- network/graph of documents known as Web
- “It isn't the computers, but the documents which are interesting”

Gigantic Global Graph (GGG)
- network/graph of entities (resources) and data that describe the entities
- “It's not the documents, it is the things they are about which are important”

TBL’s blog post on GGG: http://dig.csail.mit.edu/breadcrumbs/node/215
Web of (Linked) Data: Main Features

Web of data is based on the assumption that data have:

- well-defined structure (= structured data),
- explicitly defined meaning (= associated with machine intelligible semantics)
Web of (Linked) Data: Main Features

Two major data sources on the Web of Data:

- data embedded in Web pages
- data accessible through various kinds of program interfaces, including:
  - RESTful APIs (e.g. Knowledge Graph API)
  - SPARQL query endpoints (will be covered next week)
DATA EMBEDDED IN WEB PAGES
Main idea:

Make the content of the Web ‘legible’ to computers, by presenting it in the language they ‘understand’: using structured data with explicitly defined semantics.
Example:

Traditional Web page

Jane Doe
Professor
20341 Whitworth Institute
...
Graduated from <a href="http://www.umbc.edu/">UMBC</a>
...
Research associates:
<a href="http://www.xyz.edu/students/alicejones.html">Alice Jones</a>
...

Web page with embedded structured data

```html
<div vocab="http://schema.org/" typeof="Person">
  <span property="name">Jane Doe</span>
  <span property="jobTitle">Professor</span>
...
Graduated from <a href="http://www.umbc.edu/" property="alumniOf">UMBC</a>
...
Research associates:
<a href="http://www.xyz.edu/students/alicejones.html" property="colleague">Alice Jones</a>
...
</div>
```

Links are presented in the same way regardless of the meaning of the established connection.

Links have associated meaning:

- [https://schema.org/jobTitle](https://schema.org/jobTitle)
- [https://schema.org/colleague](https://schema.org/colleague)

Defining the type of an entity and its attributes:

- `vocab="http://schema.org/"
- `typeof="Person"`
DATA EMBEDDED IN WEB PAGES

Numerous Web pages already contain structured data with explicitly defined semantics

A few examples:

- movies at RottenTomatoes.com
- events at Ticketmaster.com
- products at BestBuy.com
- recipes at AllRecipes.com
Extracting data from Web pages

Structured data testing tool

- [https://search.google.com/structured-data/testing-tool](https://search.google.com/structured-data/testing-tool)
- Provides Web administrators with an insight into the data intelligible to programs that access the given Web page
- However, it does not allow for direct, program-based access to the embedded data
  - In other words, it cannot be called from a program to extract data from a Web page
extracting data from web pages

Microdata Distiller

- [http://www.w3.org/2012/pyMicrodata/](http://www.w3.org/2012/pyMicrodata/)
- provides program-based access to the data embedded in Web pages
- main advantage: it allows you to easily pull data from Web pages – without page scraping or any other similar efforts – and use the extracted data in your program
- it can be called as a RESTfull service or installed and run locally
ADDING DATA TO WEB PAGES

To embed structured data in Web pages, we need:

- **vocabularies** for describing the content of the page in a machine-intelligible format
- a way to *extend HTML* to make those machine-intelligible descriptions an integral part of the Web page

To address the 1\textsuperscript{st} requirement, we can use Schema.org or some other RDFS vocabulary

To address the 2\textsuperscript{nd} requirement, we can use RDFa, Microdata, or JSON-LD – W3C recommendations for extending HTML with machine processable descriptions
**SCHEMA.ORG**

Vocabulary for describing Web data in machine intelligible form; currently, the most widely used vocabulary of this kind

The initiative for its development came from the major Web companies: Google, Yahoo, Microsoft (Bing), Yandex

It is now developing within the W3C as a community effort: [https://www.w3.org/community/schemaorg/](https://www.w3.org/community/schemaorg/)

Initially it defined only a handful of types, but has significantly evolved over time

- list of all the types currently supported by Schema.org: [http://schema.org/docs/full.html](http://schema.org/docs/full.html)
Recommendation:

• Video of the keynote talk by Google’s R. Guha – leader of the W3C WebSchemas group – on the topic of structured data, Schema.org, and associated open technologies: http://videolectures.net/iswc2013_guha_tunnel/

• alternatively, or in addition, check this article: “Schema.org: Evolution of Structured Data on the Web” http://queue.acm.org/detail.cfm?id=2857276
EXTENDING HTML WITH MACHINE INTELLIGIBLE DESCRIPTIONS

W3C recommendations (de-facto standards) for embedding structured data in HTML pages:

- RDFa (RDF in attributes)
- Microdata
- JSON-LD
Jane Doe Professor  
20341 Whitworth Institute 405 N. Whitworth Seattle, WA 98052 
(425) 123-4567  
jane-doe@xyz.edu  
Research associates: Alice Jones Bob Smith

This code sample is created using the [http://rdfa.info/play/](http://rdfa.info/play/) editor
GRAPH OF DATA EXTRACTED FROM THE WEB PAGE SHOWN IN THE PREVIOUS EXAMPLE

Jane Doe Professor
20341 Whitworth Institute 405 N. Whitworth Seattle, WA 98052
(425) 123-4567
jane-doe@xyz.edu
Research associates: Alice Jones Bob Smith
The same example in Microdata notation; In fact, everything is (more or less) the same, the only difference is in the names of the used HTML attributes
Unlike RDFa and Microdata, JSON-LD data are not embedded in the body of an HTML page (the part that is rendered in the browser), but in the head of the page (the part for programs access only).
RDFa, Microdata, JSON-LD

RDFa:
- Relevant info, code, materials, etc. about RDFa: http://rdfa.info/
- Specification: http://www.w3.org/TR/xhtml-rdfa-primer/

Microdata:
- Specification: http://dev.w3.org/html5/md/

JSON-LD:
- Relevant info, code, materials, etc. about JSON-LD: http://json-ld.org/
- Specification: http://www.w3.org/TR/json-ld/

Good source of examples is Schema.org site where for each class, there is at least one example code segment given in each of the 3 standards
MORE ABOUT VOCABULARIES

Schema Actions

- Schema.org classes and attributes aimed at
  - describing (in machine intelligible way) different kinds of actions that a Web site offers to its visitors, and how those actions can be invoked by a (third party) program
  - integrating data about users’ actions from different Web sites
SCHEMA ACTIONS IN EMAIL MARKUP

RSVP action

Example code for embedding Schema Action markup in email (Gmail) messages: https://developers.google.com/gmail/markup/reference/
GOODRELATIONS

GoodRelations

- [http://www.heppnetz.de/projects/goodrelations/](http://www.heppnetz.de/projects/goodrelations/)
- vocabulary for describing products, offers, shops, and the like
- already in wide use in the e-commerce domain
  - E.g. Kmart.com, Sears.com, BestBuy.com
- a number of tools have been developed to facilitate its use
- it has been integrated into Schema.org
  - [http://schema.org/Product](http://schema.org/Product) ; [http://schema.org/Offer](http://schema.org/Offer) ...
MORE ABOUT VOCABULARIES

Open Graph Protocol (OGP)

- [http://ogp.me/](http://ogp.me/)
- introduced by Facebook to obtain more information about the things people ‘Like’ outside the Facebook’s domain
  - RDFa + OGP data embedded in the page provide a formal description of the “liked” item
  - thus obtained information is used for further extending Facebook’s Entity Graph
- OGP supports the description of several popular domains: music, video, articles, books, websites and user profiles
TOOLS FOR WORKING WITH EMBEDDED STRUCTURED DATA

Google has developed a set of supporting tools for
  • embedding structured data in Web pages
  • tracking and management of Web pages with embedded data,
  • detecting errors in how the data is embedded in a page

Some of these tools:
  • Structured Data Dashboard (https://goo.gl/V8NZ8L)
  • Data Highlighter (https://goo.gl/P5SZOc)
  • Structured Data Markup Helper (https://goo.gl/1Ywtfg)

Video from Google IO 2013 conference introduces and describes these tools: https://developers.google.com/events/io/sessions/351340935
Linking data on the Web
5 Star (Linked) Open Data

Linked Open Data star scheme by example: [http://5stardata.info/](http://5stardata.info/)
LINKED DATA PRINCIPLES

1) Use URIs as names for things
   ISBN: 9781775411840

2) Use HTTP URIs so that people can look up those names
   <http://www.worldcat.org/title/north-and-south/oclc/606818482>

3) Provide useful information for each URI, using the Linked Data standards (RDF(S), SPARQL)
   <http://www.worldcat.org/title/north-and-south/oclc/606818482>
   rdf:type schema:Book

4) Include links to other URIs so that more things could be discovered.
   <http://www.worldcat.org/title/north-and-south/oclc/606818482>
   schema:author <http://viaf.org/viaf/39377536/>

This URI uniquely identifies the authoress Elizabeth Gaskell
DATA PUBLISHED ON THE WEB FOLLOWING THE LINKED DATA PRINCIPLES

Source: http://lod-cloud.net/
APPLICATION EXAMPLES
Rich Snippets

- Richer rendering of Google search results for those pages that embed structured and semantically described data
- Try it yourself by searching Google for any movie, concert, recipe, mobile phone app, Sourceforge project, ...
- Detailed insight into Rich Snippets is available at: https://goo.gl/6JBY9k

Boyhood (2014) - IMDb
www.imdb.com/title/tt1065073/
★★★★★ Rating: 8/10 - 217,851 votes
Boyhood -- Clip: Talk To Me Boyhood -- Featurette: Behind the Scenes ..... The actress that portrays Samantha is the movie's director's daughter, Lorelei Linklater ...
Full Cast & Crew - Ellar Coltrane - Lorelei Linklater - Patricia Arquette
**INTERACTIVE SNIPPETS (YANDEX ISLANDS)**

- Feature of the Yandex search engine, similar to Rich Snippets, but more interactive
  - E.g., allows one to check-in for a flight
- The following article provides more information: [http://goo.gl/uxjb6g](http://goo.gl/uxjb6g)
Rich Pins are pins with additional, advanced features

- E.g., for products, they provide information about the current price, availability, current discounts, special deals

- An overview of different types of Rich Pins is available at: https://business.pinterest.com/en/rich-pins

- Documentation for developers gives detailed information on the use of structured data for the creation of Rich Pins: https://developers.pinterest.com/docs/rich-pins/
PERSONAL DIGITAL ASSISTENTS

Some well known examples:

- Google Now ([https://www.google.com/search/about/](https://www.google.com/search/about/))
  - bought by Google; ceased to exist as a standalone app in Sept 2015; the technology behind it is integrated into Google Now
“Thanks to Google Now, as I stroll around San Francisco, live bus times are offered to me whenever I pull my phone from my pocket at a bus stop. And when I get up in the morning, Google Now presents a panel summarizing my optimum transit journey to work along with specific buses and an estimate of the time the trip will take…”

Source: Google’s Answer to Siri Thinks Ahead, MIT Technology Review
Google Now

“Google Now provides updates to restaurant and hotel reservations or flight information received in Gmail. By marking up email notifications to your users, you can use Google Now to bring them similar updates about your services and products”
“Airline flight providers can add schema.org markup to their outgoing emails to enable flight tracking for customers that use Microsoft Cortana”

WEB (RESTful) APIS FOR DATA ACCESS

Domain models and program logic

User interface (dialog)

Data source: Web APIs

Web (RESTful) APIs for Data Access

Sole reliance on Web APIs for data collection has drawbacks:

- To collect data from different sources, one needs to
  - get acquainted with the specificities of different APIs that provide access to the required data
  - resolve the heterogeneity in the data collected from different sources

- APIs tend to change, causing the need to update the code that relies on them for data access

- Change in the conditions for accessing the data
  - changes in the quantity and/or kinds of the data that can be pulled through the API
NEW (ADDITIONAL) DATA SOURCE: LINKED (OPEN) DATA ON THE WEB

Data sources:
Web APIs + Linked Data

Domain models
(RDFS vocabularies)
and program logic

User interface (dialog)