How Google is using Linked Data Today and Vision For Tomorrow

Thomas Steiner*[1], Raphaël Troncy[2] and Michael Hausenblas[3]

[1] Google Germany GmbH, ABC-Str. 19, 20354 Hamburg, Germany,
<tomac@google.com>
[2] EURECOM, Sophia Antipolis, France, <raphael.troncy@eurecom.fr>
[3] DERI, NUI Galway IDA Business Park, Lower Dangan Galway, Ireland,
<michael.hausenblas@deri.org>

Abstract. In this position paper, we first discuss how modern search engines, such as Google, make use of Linked Data spread in Web pages for displaying Rich Snippets. We present an example of the technology and we analyze its current uptake. We then sketch some ideas on how Rich Snippets could be extended in the future, in particular for multimedia documents. We outline bottlenecks in the current Internet architecture that require fixing in order to enable our vision to work at Web scale.

1 Introduction

The Web is the seminal part of the *Application Layer* of network architectures. Two major trends are currently opening huge perspectives and challenges on the Web in terms of research: the Web of Data (also called Semantic Web) and the Social Web (also called Web 2.0). The Web of Data is the idea of using Semantic Web standards to provide a common framework that allows data to be shared and reused across application, enterprise, and community boundaries. It is a collaborative effort led by the W3C with participation from a large number of researchers and industrial partners. Linked Data is about exposing, sharing, and connecting data on the Web, allowing application and data integration at scales never seen before. In parallel, Web 2.0 applications are a new trend in Web development and design that facilitates communication, secures information sharing, interoperability, and collaboration.

The fundamental shift triggered by these trends is that, while previously the Internet has been concerned about sending bits from one host of the network to another, new applications now require to make sense out of those bits. In other words, the Internet architecture needs a new layer, that takes care of data interoperability for interconnecting pieces of machine-processable data. Some have proposed to add a new data layer to the Systems Interconnection (OSI) stack, a so called "Linked Data Layer" located between the application layer and

* While the author is affiliated with Google, this paper is in no way to be seen as an official product strategy, nor as a future roadmap. Opinions are solely the respective author's.

the presentation layer and that aims to make sense of the data in such a way that it establishes interoperability between different applications [2].

In this paper, we first describe the Google Rich Snippet technology as a way of adding more semantics to Web applications, the variety of formats supported, and some rough analysis of its usage (section 2). We then present some ideas of future extension of this technology, in particular for multimedia content (section 3). Finally, we take the initial steps towards a complete overhaul of parts of the current Internet architecture by introducing the thought-experiment of Triple-centric Networking, inspired by Jacobson's Content-Centric Networking (section 4).

2 Google Rich Snippet Formats and Support

What are the costs and benefits of publishing data and semantic markup on the Web[4]? One needs to carefully examine these issues from different perspectives: consumer and publisher. We are starting to see a number of domains, such as the Public Sector Information/eGov area, Life Sciences, eCommerce, and others to not only be aware of the benefits of publishing Linked Data, but start to exploit it on a large scale. In this section, we present XHTML code samples with embedded RDFa mark-up and we show how Google uses this information in its Rich Snippets technology.

2.1 Which Formats are Supported in Rich Snippets?

Google introduced Rich Snippets on May 12, 2009 as a means of displaying structured data in search result pages with the objective of highlighting the searched-for properties to the user in a visually outstanding way [4]. The Rich Snippet feature was built from the beginning on open standards or community-agreed-on approaches such as RDFa [1], Microformats[5], and more recently Microdata[6]. While there is no guarantee that Rich Snippets will be displayed by Google if a Web page contains semantic mark-up, there is an expression-of-interest form available[6] where webmasters can indicate their consent and interest for Rich Snippets to be shown for their pages.

Multiple semantic mark-up formats are supported by Google (RDFa, Microformats, Microdata), the search company taking a very practicable approach: *A lot of previous work on structured data has focused on debates around encoding. Even within Google, we have advocates for microformat encoding, advocates for various RDF encodings, and advocates for our own encodings. But after working on this Rich Snippets project for a while, we realized that structured data on the web can and should accommodate multiple encodings: we hope to emphasize this by accepting both microformat encoding and RDFa encoding. Each encoding*

[4] http://lab.linkeddata.deri.ie/2010/star-scheme-by-example/

[5] http://microformats.org/wiki/Main_Page#Specifications

[6] http://www.google.com/support/webmasters/bin/request.py?contact_type=rich_snippets_-feedback

has its pluses and minuses, and the debate is a fine intellectual exercise, but it detracts from the real issues [4]. Together with the May 12 announcement on public Rich Snippets, an additional feature was launched: Rich Snippets in Custom Search[7] which allows for creating even richer snippets based on custom mark-up and snippet creation rules to Custom Search users, somewhat similar to Yahoo!'s BOSS[8] (Build your Own Search Service) initiative.

2.2 Rich Snippets by Example

The Rich Snippets technology support various vocabularies. For example, details about an offering such as the ticket booking of an upcoming event can be marked up in the body of a Web page in order to help understanding the location, schedule, price or reviews of the event. The code example in Figure 1 illustrates the mark-up for an event at a certain business location.

The example begins with a namespace declaration using `xmlns`. In the first line, `typeof="v:Event"` indicates that the marked-up content describes an Event. The dimensions that composed the event (description, type, starting time) are described with properties. The property name is prefixed with `v:` (`<span property="v:description">`). Google does not display information that isn't visible to the user, with a few exceptions. Geo information (latitude and longitude of the location) can be included in the HTML markup. Typically, this information is not visible on a Web page about an event, but providing it can help ensure that the location is accurately mapped.

Figure 2 shows how a search result using this semantic markup is then displayed on Google search result page.

2.3 How Much Semantic Markup is out There?

Goel et al. have compiled some statistics with regards to semantic mark-up on the Web in June 2010 [3]. A random sample of one million Web pages have been harvested in order to compare the use of Microformats and RDFa markup. Then, they examined how much of this mark-up data was actually used for Rich Snippets. It is remarkable and surprising how few semantic mark-up was live on the Web overall, and even more, that only a tiny fraction of all this semantic mark-up was then used for Rich Snippets at the time of this experiment (Table 1).

Further analysis of the dataset crawled has shown a number of pitfalls: incorrect labeling (e.g. marking up the date of an event as part of the event description), or incorrect inclusion of unrelated words in the structured mark-up (e.g. marking up "written by John Doe" rather than just "John Doe" as value of the property `v:reviewer`). Furthermore, they observe a general confusion with what parts of a document should be marked up at all. Although some web pages include RDFa event markup, none of them are used by the Rich Snippet technology as of today.

[7] `http://googlecustomsearch.blogspot.com/2009/05/enabling-rich-snippets-in-custom-search.html`

[8] `http://developer.yahoo.com/search/boss/boss_guide/`

```
<div xmlns:v="http://rdf.data-vocabulary.org/#" typeof="v:Event">
  <a href="http://www.example.com/events/poisel_offenback.hmtl"
     rel="v:url"
     property="v:summary">Philipp Poisel in Offenbach</a>
  <span property="v:description">See Philipp Poisel in Offenbach</span>
  When:
  <span property="v:startDate" content="2011-01-16T19:00-01:00">
    Jan 16, 7:00PM</span>
  <span property="v:endDate" content="2011-01-16T21:00-01:00">
    9:00PM</span>
  Where:
  <span rel="v:location">
    <span typeof="v:Organization">
      <span property="v:name">Capitol</span>,
      <span rel="v:address">
        <span typeof="v:Address">
          <span property="v:street-address">Kaiserstraße 106</span>,
          <span property="v:locality">Offenbach am Main</span>,
        </span>
      </span>
      <span rel="v:geo">
        <span typeof="v:Geo">
          <span property="v:latitude" content="50.10945"></span>
          <span property="v:longitude" content="8.76579" ></span>
        </span>
      </span>
    </span>
  </span>
  Category: <span property="v:eventType">Concert</span>
</div>
```

Fig. 1. RDFa markup for the upcoming Philipp Poisel Concert to be held in Offenbach on January 16th, 2011

Fig. 2. Rich Snippet preview for the Philipp Poisel Concert held in Berlin on September 6th, 2010

	Microformats	RDFa
Total pages	40,091	2,514
hCard / People	33,675 (13%)	1,160
Reviews	1,950 (88%)	872 (66%)
Recipe	152 (53%)	–
hCalendar / Event	126 (41%)	–
Products	519	77

Table 1. One million Web pages sampled from the Internet in June 2010. Percentages in parenthesis: actually used for generating Rich Snippets. Source: [3]

2.4 What is the Business Impact Of Rich Snippets?

Improving how search results are displayed in snippets preview involve a huge market. Tickets aggregator Web sites such as Giga-Music.de place so-called affiliate links to the final ticket vendors. These sites, very often, make their living from accumulating useful metadata around a certain event, and then, provide only links to vendors in order to finally get paid, e.g. on a pay-per-click model. In [3], Goel and Gupta gave some insights into the click and impression behavior for Rich Snippets. The overall tendency being an increasing number of impressions for Rich Snippets-enabled pages, and a higher click-through rate for pages with Rich Snippets. Taking into account the prior remark, it is thus clear that adding or removing Rich Snippets as a whole, or Rich Snippets features especially, has a huge impact on the labile search ecosystem. Web sites can suffer significant sales collapses by going down a position in their natural search ranking.

3 A Vision for Tomorrow's Rich Snippets in Search Engines

Search engines serve mainly as entry points to the Web. Let us imagine a user wants to see a concert of the German singer and songwriter Philipp Poisel. A straightforward query would be "philipp poisel konzerte[9]". At the time of writing, this search results in: the concerts section of the artist's official Website as the first result[10], a couple of fan club sites[11], some concert review sites[12], and some ticket aggregator sites[13].

In the following, we purely focus on how Linked Data could enrich the search experience, for example, for a Web search for Philipp Poisel concerts. We anticipate a huge uptake of semantic mark-up by Web site operators over the coming

[9] "konzerte" is the German word for "concerts".

[10] http://www.philipp-poisel.de/termine/

[11] E.g. http://www.philipppoiselfanclub.de/?tag=konzert

[12] E.g. http://www.ciao.de/Philipp_Poisel_Konzert_Tickets__8025011

[13] E.g. http://www.giga-music.de/philipp-poisel-konzerte-live-tour/

months. Especially, new business-related vocabularies such as the Tickets Ontology[5], are expected to see broader and broader usage and implementation. In the following, we assume that ticket vendors had implemented the Tickets Ontology on their Web pages. Therefore, we consider a couple of pages with the following triples (for the sake of clarity we omit the necessary prefixes and simplify the `xsd:dateTime` format):

```
foo:ticket a tio:TicketPlaceholder ;
  rdfs:label "Tickets for Philipp Poisel"@en ;
  tio:accessTo <http://data.events.example.org/123> .

foo:ExampleTicketVendor gr:offers foo:offer .

foo:offer a gr:Offering ;
  gr:name "Tickets for Philipp Poisel in Bremen"@en ;
  gr:description "Philipp Poisel in Bremen"@en ;
  gr:includes foo:ticket ;
  gr:hasBusinessFunction gr:Sell ;
  gr:hasPriceSpecification
    [a gr:UnitPriceSpecification ;
     gr:hasCurrency "EUR"@en ;
     gr:hasCurrencyValue "24.10"^^xsd:float ;
     gr:validThrough "2010-11-11T23:59"^^xsd:dateTime].
```

Each ticket instance has access to a `tio:Event`, which subclasses a `lode:Event`[8], an `event:Event`, and a `dul:Event`, and where the particular concert event dates are described. We assume that the Web site also implements `v:Event`. This would allow for comparative Rich Snippets. The obvious next step would be to bring the social experience into play, granted the user has given access to her social graph . This would mean to carry part of the "Facebook experience" right into the search experience. Prior research by Troncy et al. has shown that users generally start an event search on a general search engine, and that the decision criteria whether or not to attend an event is often dependent on whom of the user's friends plan to attend [10, 9].

An entirely different class of extended Rich Snippets could be based on multimedia semantics in order to provide richer video search results. We believe that there is high potential for semantically annotated multimedia content to improve content search. In Figure 4, we show a mock-up of a person highlighted, which could be based on media fragment URIs. Such media fragment URI could look like:

`http://example.org/video.webm?t=428,434#xywh=150,60,50,70&xywh=240,50,50,70`

The components of this URI are first a temporal dimension (t=428,434), which selects seconds 428 to 434 of the whole video, and then a spatial dimension (xywh=150,60,50,70 and xywh=240,50,50,70), which creates two bounding boxes at the x, y parameters with a width w and a height h.

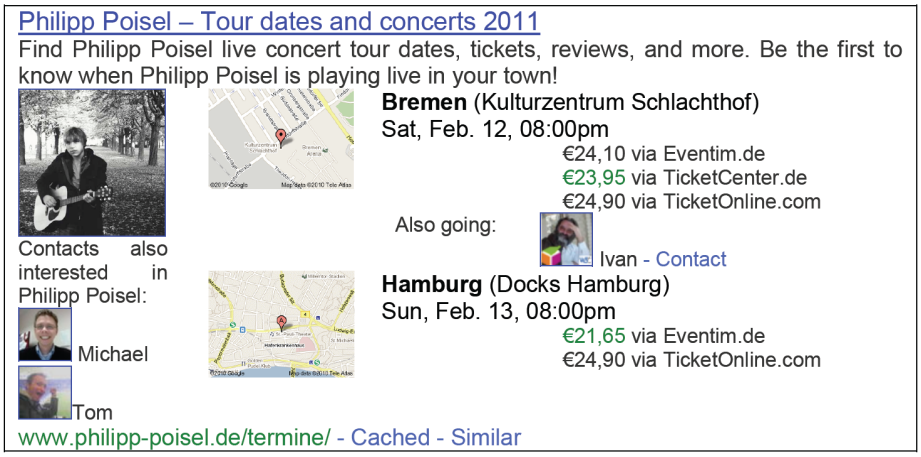

Fig. 3. Sketch of an extended Rich Snippet featuring maps previews, image preview, event and price information, including cheapest offer highlighted. In addition to that, the user's social graph is processed in order to find people interested in the same artist and to display who else plans to attend a particular event.

Fig. 4. Sketch of an extended Rich Snippet featuring semantically highlighted video preview (still frame or moving images), actor information, and in-video links.

4 Triple-centric Networking

The vision of extended Rich Snippet outlined above features information from more than just one data source which is different from today's Rich Snippets where the content is exclusively determined by the information in one particular Web page. It is obvious that in order to combine information coming from various data sources, an information sharing mechanism must be established. In the following, we sketch a thought-experiment derived from Content-centric Networking introduced by Jacobson [7]. In Content-centric Networking, there are two notions of packages involved: `Interest` and `Data` packages. Interests get broadcast by consumers, and as soon as a node can satisfy an interest, it responds with the data. Otherwise, it rebroadcasts the interest. The main advantage over common host-based networking is that data packages are not only exclusively thought for the initially interested node, but can be shared between nodes with common interests. A certain piece of information satisfies an interest if the content name in the interest package is a prefix of the content name in the data packet. Applied to RDF triples, this could mean that the content name would correspond to the subject. Figure 5 is adapted from Figure 2 in [7] and illustrates how the interest and data packages could look like if we applied the principle of Content-centric Networking to Triple-centric Networking.

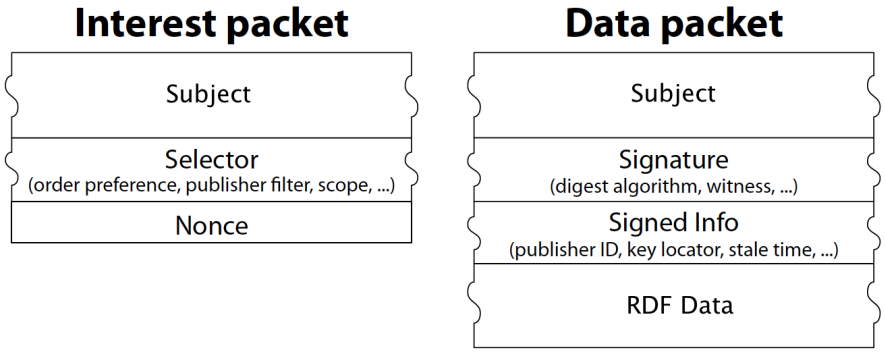

Fig. 5. The principles of Content-centric Networking applied to Triple-centric Networking. The figure shows the two possible packet types.

Experiments by Jacobson et al. have shown that Content-centric Networking is useful when many parties are interested in the same content. This is definitely the case if we think Web scale triple propagation for popular Web pages. One of the potential bottlenecks of rolling out multi data sources Rich Snippets could thus be avoided by moving triple propagation to this new Triple-centric Networking layer. We are at the early stages of this thought-experiment and have not carried out any experimentation to justify our assumption. However,

we believe that there is a potential improvement to the host-to-host networking infrastructure that dominates today's Internet traffic.

5 Conclusion

We have shown why and how Rich Snippets formats are currently supported by Google and we have provided code samples in RDFa. We have then briefly discussed the potential business impacts of Rich Snippets to Web site operators. It has become visible that Rich Snippets are a very sensible element in the Linked Data value chain due to the high visibility and the confirmed change of user click-through behavior. We have outlined potential extensions to Rich Snippets, driven by a concrete use case of an event-based Web search. We have interlinked the social graph of a user with common event-related data in the Linked Data cloud. In our mock-up, we neglected the business impact of Rich Snippets entirely. However, it is obvious that for the online ticket search example, the decision what ticket vendor to include, and what vendor to exclude from the vendors shown in the Rich Snippets is a crucial one.

The suggested addition of a Linked Data layer [2] as layer 7a between the current application and presentation layer to the ISO/OSI 7-layer architecture could help establish the links between the data providers and facilitate the work of, e.g., search engines, to make sense of these data and present them in an efficient way. In addition to that, we have also introduced the thought-experiment of Triple-centric Networking inspired by Jacobson's Content-Centric Networking.

Acknowledgments

The research leading to this paper was partially supported by the project AAL-2009-2-049 "Adaptable Ambient Living Assistant" (ALIAS) co-funded by the European Commission and the French Research Agency (ANR) in the Ambient Assisted Living (AAL) programme, and by the projects FP7-216444 "Peer-to-peer Tagged Media" (Petamedia), FP7-248296 "I-SEARCH" and FP7-256975 "LOD Around-The-Clock" (LATC) Support Action.

References

1. B. Adida, M. Birbeck, S. McCarron, and S. Pemberton. RDFa in XHTML: Syntax and Processing. W3C Recommendation, October 14, 2008. http://www.w3.org/TR/rdfa-syntax/.
2. S. Decker, M. Hauswirth, and S. Auer. The Future Internet Assembly. Linked Data in the Future Internet, December, 2010. http://www.future-internet.eu/home/future-internet-assembly/ghent-dec-2010/session-i-linked-open-data-i.html.
3. K. Goel and P. Gupta. Google Rich Snippets. Semantic Technology Conference, 2010. http://semtech2010.semanticuniverse.com/sessionPop.cfm?confid=42&proposalid=2745.
4. K. Goel, G. Ramanathan, and H. Othar. Introducing Rich Snippets. Google Webmaster Central Blog, May 12, 2009. http://googlewebmastercentral.blogspot.com/2009/05/introducing-rich-snippets.html.

5. M. Hepp. Tickets Ontology v1.0, November 17, 2010. http://purl.org/tio/ns.
6. I. Hickson. HTML Microdata. W3C Working Draft, October 19, 2010. http://www.w3.org/TR/microdata/.
7. Jacobson, D. Smetters, J. Thornton, M. Plass, N. Briggs, and R. Braynard. Networking Named Content. In *5th ACM International Conference on Emerging Networking Experiments and Technologies (CoNEXT'09)*, Rome, Italy, 2009.
8. R. Shaw, R. Troncy, and L. Hardman. LODE: Linking Open Descriptions Of Events. In *4th Asian Semantic Web Conference (ASWC'09)*, pages 153–167, Shanghai, China, 2009.
9. R. Troncy, A. Fialho, L. Hardman, and C. Saathoff. Experiencing Events through User-Generated Media. In *1st International Workshop on Consuming Linked Data (COLD'10)*, Shanghai, China, 2010.
10. R. Troncy, B. Malocha, and A. Fialho. Linking Events with Media. In *6th International Conference on Semantic Systems (I-SEMANTICS'10)*, Graz, Austria, 2010.