Sémantický web – 10 let poté

Doc. Ing. Vilém Sklenák, CSc.
Vysoká škola ekonomická, fakulta informatiky a statistiky,
katedra informačního a znalostního inženýrství
sklenak@vse.cz

INFORUM 2011: 17. konference o profesionálních informačních zdrojích
Praha, 24.–26. 5. 2011

Abstrakt

V roce 2001 zveřejnil Tim Berners-Lee svou vizi sémantického webu. Co se podařilo za uplynulých 10 let? A co ještě ne. Produkty sémantického webu.

1 Úvod

Tento příspěvek volně navazuje na příspěvek [13], který zazněl na konferenci Inforum 2003. V tehdejším příspěvku byly shrnuty základní principy sémantického webu.

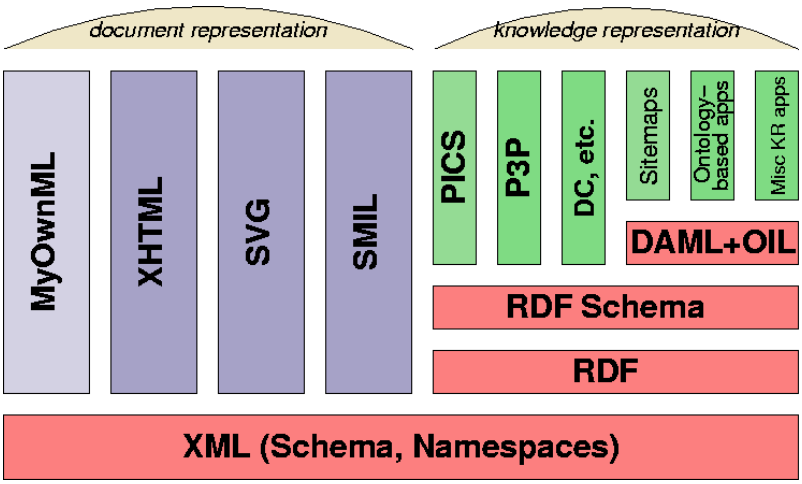
Vznik myšlenky a rozvoj základních principů sémantického webu není záležitostí posledních deseti let, jak by se mohlo na první pohled znát vzhledem k rostoucímu počtu publikací, konferencí, workshopů apod. Je však pravda, že k širší popularizaci sémantického webu došlo především zásluhou článku [3], který „otec“ webu T. Berners-Lee společně s dalšími spoluautory vydali v prestižním časopise Scientific American právě v květnu 2001. I proto je rok 2001 považován za symbolický počátek historie sémantického weu. V té době však práce na sémantickém webu trvaly již několik let. Od počátku T. Berners-Lee zdůrazňoval, že *„sémantický web je rozšířením současného webu, jež datům přiřazuje přesný význam, díky kterému bude možná kooperace jak lidí, tak softwaru“*.

2 Jak to začalo

Jak plyne z již citovaného článku [3] sémantický web není nějaký nový web, ale jde o rozšíření konceptu a doplnění dat toho stávajícího. Doplnění o metadata, která by měla popisovat sémantické informace webových zdrojů a která by měla být zápsána pomocí strojově srozumitelných jazyků. Součástí metadat by také byla použitá slovní zásoba a soubor vztahů mezi jednotlivými pojmy.

Na webu je však téměř nemožné prosadit jednotný jazyk a vymezit jakousi jednotnou slovní zásobu. Plyne to jednak z principu decentralizovanosti samotného webu, jednak z povahy zpřístupňovaných informací – jde vlastně o všechny oblasti znalostí.

O to se však sémantický web nesnaží. Jeho myšlenka spočívá především v nabídce takového flexibilního a otevřeného datového modelu a odpovídajících datových jazyků tak, aby vyhovoval nekonečně varietě webu.



Obrázek 1: Jazyky sémantického webu

3 Jak to pokračovalo

Rekonstrukce událostí předurčující další vývoj sémantického webu za uplynulých deset let by mohla vypadat takto:

1998
- první zmínka o *sémantickém webu* – Tim Berners-Lee na konferenci WWW v australském Brisbane

2000
- spuštěn web http://semanticweb.org

2001
- článek [3] v časopise *Scientific American*
- konsorcium W3C ustavilo pracovní skupinu *Web Ontology Working Group* s cílem vývoje jazyka OWL (Web Ontology Language)
- spuštěna webová stránka *W3C Semantic Web Activity* – http://www.w3.org/2001/sw/

2002
- první ročník celosvětové konference *International Semantic Web Conference*
- konsorcium W3C ustavilo pracovní skupiny *Web Services Description Working Group* a *Web Services Architecture Working Group* s cílem podpory vývoje webových služeb

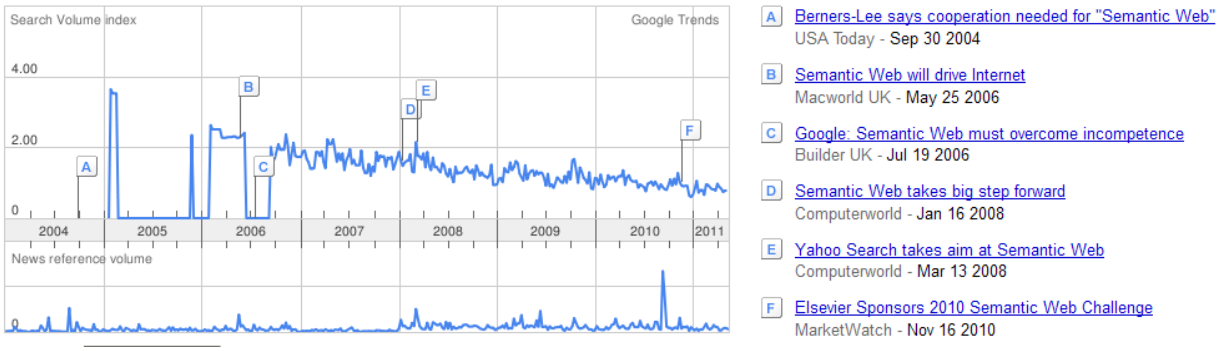
2004
- zveřejněny standardy OWL a RDF
- byl spuštěn *SWoogle* – vyhledávací stroj pro sémantický web
- zveřejněn standard *RDFS* jako jazyk pro reprezentaci RDF slovníků na webu

2007
- zveřejněn standard *GRDDL (Gleaning Resource Descriptions from Dialects of Languages)* jako technika pro získávání RDF dat z XML dokumentů
- zveřejněn standard *SAWSDL (Semantic Annotations for WSDL and XML Schema)* – definuje sadu rozšiřujících atributů pro jazyk WSDL

2008
- zveřejněn standard *SPARQL Query Language for RDF* pro dotazy nad RDF daty
- zveřejněn standard *RDFa* pro specifikaci atributů vyjařujících strukturu dat v XHTML

2009
- zveřejněn standard OWL 2
- zveřejněn standard *SKOS (Simple Knowledge Organization System* – datový model pro sdílení a propojování znalostí napříč webem
- zveřejněn standard *POWDER (Protocol for Web Description Resources* – mechanismus pro popis a hledání webových zdrojů

2010
- publikován standard *RIF (Rule Interchange Format* – umožňuje výměnu pravidel mezi systémy



Obrázek 2: Google – jak často se vyhledává „`semantic web`“

Zajímavým svědectvím doby je popularita dotazu „`semantic web`“ u služby Google. Jak ukazuje obr. 2, trend je lehce sestupný. Co však lze z toho odvodit? Že zájem o sémantický web klesá? Nebo že sémantický web nemá perspektivu? Určitě ne. Jde o to, že ve spojení se sémantickým webem se hledá řada jiných pojmů či frází, a ty do uvedeného grafu nejsou zahrnuty. Důvod je prostý – Google nehledá sémanticky.

Dalším ukazatelem vývoje v určité oblasti zkoumání jsou počty publikací. Ty se dají měřit různě, ale dobrou vypovídací hodnotu mají bezesporu počty z uznávaných databází Web Of Knowledge a Scopus, protože ty monitorují jen renomované časopisy, případně konference. Zde je u obou databází zřejmý rostoucí trend – viz obr. 3. I zde jsou však počty potenciálně zkresleny – jednak určitým zpožděním aktualizace databází oproti zveřejnění publikace jako takové, jednak také dotazem. Z komerčního hlediska jsou pak zajímavé počty titulů prodávaných prostřednictvím e-shopu Amazon.com – i zde je patrná vzestupná tendence.

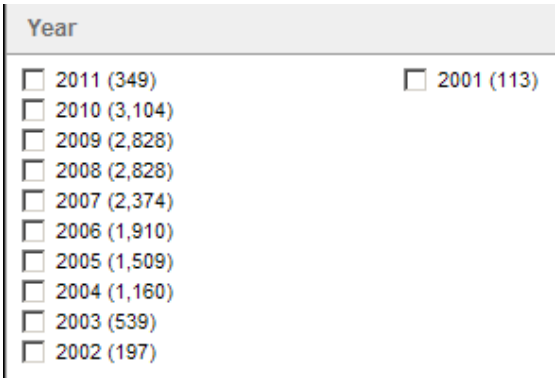
4 Jak to je nyní

Sémantickému webu a technologiím/principům s ním souvisejicím se věnuje řada konferencí. Prestižními konferencemi jsou zejména:

- *International Semantic Web Conference (ISWC)*, která se bude letos konat již podesáté: `http://iswc2011.semanticweb.org`

Field: Publication Year	Record Count	% of 10584	Bar Chart
1999	4	0.0378 %	
2000	27	0.2551 %	
2001	125	1.1810 %	
2002	297	2.8061 %	
2003	526	4.9698 %	
2004	842	7.9554 %	
2005	1150	10.8655 %	
2006	1326	12.5283 %	
2007	1548	14.6259 %	
2008	1805	17.0540 %	
2009	1813	17.1296 %	
2010	944	8.9191 %	
2011	177	1.6723 %	

(a) Web of Knowledge



(b) Scopus

2001	37	2006	227
2002	74	2007	244
2003	109	2008	284
2004	142	2009	256
2005	186	2010	409

(c) Amazon.com

Obrázek 3: Počty publikací k dotazu „`semantic web`“ podle let

- a *Extended Semantic Web Conference (ESWC)*, která se bude letos konat poosmé: http://www.eswc2011.org/.

Struktura konferencí nyní již pravidelně sleduje tři „proudy“:

1. výzkum (research papers),
2. použití (semantic web in use)
3. a využití v komerční sféře (industry track).

Již toto naznačuje postupné nejen pronikání sémantických technologií do komerčního využití, ale především zájem na spolupráci obou komunit – výzkumníků na straně jedné a firem na straně druhé. Pro zajímavost lze uvést témata, na která se soustředí konference ISWC 2011:

- Management of Semantic Web Data
  - Languages, tools, and methodologies for representing and managing Semantic Web data
  - Database, IR, and AI technologies for the Semantic Web
  - Search, query, integration, and analysis on the Semantic Web
  - Robust and scalable knowledge management and reasoning on the Web
  - Cleaning, assurance, and provenance of Semantic Web data, services, and processes
  - Principles and applications of very large Semantic Web data bases
  - Semantic wikis
  - Semantic Web Services
  - Evaluation of semantic web technology

- Natural Language Processing
  - Machine learning and information extraction for the Semantic Web
  - Semantic web population from the human web
  - Exploiting tags, categories, wikis for the semantic web
  - Application of semantic web to NLP
- Ontologies and Semantics
  - Specific ontologies and ontology patterns for the semantic web
  - Ontology methodology, evaluation, reuse, extraction, and evolution
  - Ontology modularity, mapping, merging, and alignment
  - Searching for and ranking ontologies
  - Reasoning over Semantic Web data
  - New formalisms for Semantic Web (such as probabilistic approaches)
  - Lightweight semantics (linked data, microformats, etc.)
- Semantic Web Engineering
  - Methods for Semantic Web application development
  - Tools for Semantic Web application development
  - Evaluation of Semantic Web technologies or data
  - Including legacy applications into the Semantic Web
  - Impact of specific application areas (e.g. e-science, e-gov, sensors) on semantic web design
- Social Semantic Web
  - Social networks and processes on the Semantic Web
  - Semantic Web technologies for collaboration and cooperation
  - Representing and reasoning about trust, privacy, and security
  - Modeling users and contexts in Semantic Web applications
- User Interfaces to the Semantic Web
  - Interacting with Semantic Web data
  - Semantic Web content creation and annotation
  - Mashing up Semantic Web data and processes
  - Novel interaction paradigms aimed at linked data
  - Semantic web applications to Web 2.0 sites
  - Natural language Semantic Web interfaces
  - Information visualization of Semantic Web data
  - Personalized access to Semantic Web data and applications

Co je však potěšující, že sémantický web a jeho technologie již zdaleka nejsou jen polem pro výzkum a různé experimenty, ale že začínají pronikat do softwarových produktů. Spektrum aplikací implementujících v menší či větší míře technologie sémantického webu je široké, lze na ně narazit jak ve webových službách pro běžné uživatele, tak se stávají součástí řešení pro firemní sféru.

Jeden z poměrně populárních technologických blogů *ReadWriteWeb*[1] již dva roky po sobě vytipoval nejlepší aplikace sémantického webu. Za rok 2010 [11] to byly:

1. *Freebase*[2] – tento produkt je dílem firmy Metaweb, jedné z předních firem v oblasti sémantickýh technologií. Firma Metaweb se stala během roku 2010 akvizicí firmy Google, která si tímto způsobem chce zajistit know-how pro „chytřejší“ vyhledávání [7].
2. *GetGlue*[3] – je to jedna ze služeb typu vytváření sociálních sítí v oblasti zábavy.
3. *FlipBoard*[4] – uvedení iPadu odstartovalo řadu nových firem (tzv. startupů), FlipBoard je „sociálně“ orientovaným časopisem, který má integrovány sémantické přístupy

[1] http://www.readwriteweb.com
[2] http://www.freebase.com/
[3] http://www.getglue.com/
[4] http://www.flipboard.com/

s cílem lepšího určování relevance informací.

4. *Hunch*[5] – služba Hunch, který byla dříve službou typu Q&A, se změnila v roce 2010 na personalizované doporučování v oblasti volného času (filmy, knihy, dovolená aj.) s využitím technik mapování a rozhodovacích stromů.

5. *Apture*[6] – jedná se o vyhledávací službu založenou na sémantice kontextu.

Z přehledu [11] rovněž stojí za pozornost upozornění na největší firmy/organizace, které implementují sémantické technologie – mj. FaceBook (díky protokolu *Open Graph*), Google (zásluhou služby *Google Squared*), *data.gov.uk* – jeden z největších počinů v oblasti *linked open data* (viz dále).

Lze však nalézt i další fakta svědčící o tom, že si sémantické technologie prorážejí cestu k stále širšímu uplatňování. Například firma Google využívá ontologii GoodRelations – jejím použitím (nemnoho řádků v RDFa) v rámci webové stránky lze výrazně zlepšit SEO stránky [9]. I další velcí „hráči“ jako jsou Oracle, IBM aj. [10] nechtějí zůstat pozpátku. Například firma Oracle nabízí sadu nástrojů pro správu RDf databází [12] jako podporu pro vývoj sémanticky orientovaných business aplikací.

5 Co dál

Jedním z nejaktuálnějších trendů či cest, jak směřovat k vytváření sémantického webu, jsou *linked data*. Sám Tim Berners-Lee v [2] říká:

> *„The value of your own information is very much a function of what it links to, as well as the inherent value of the information within the web page“*

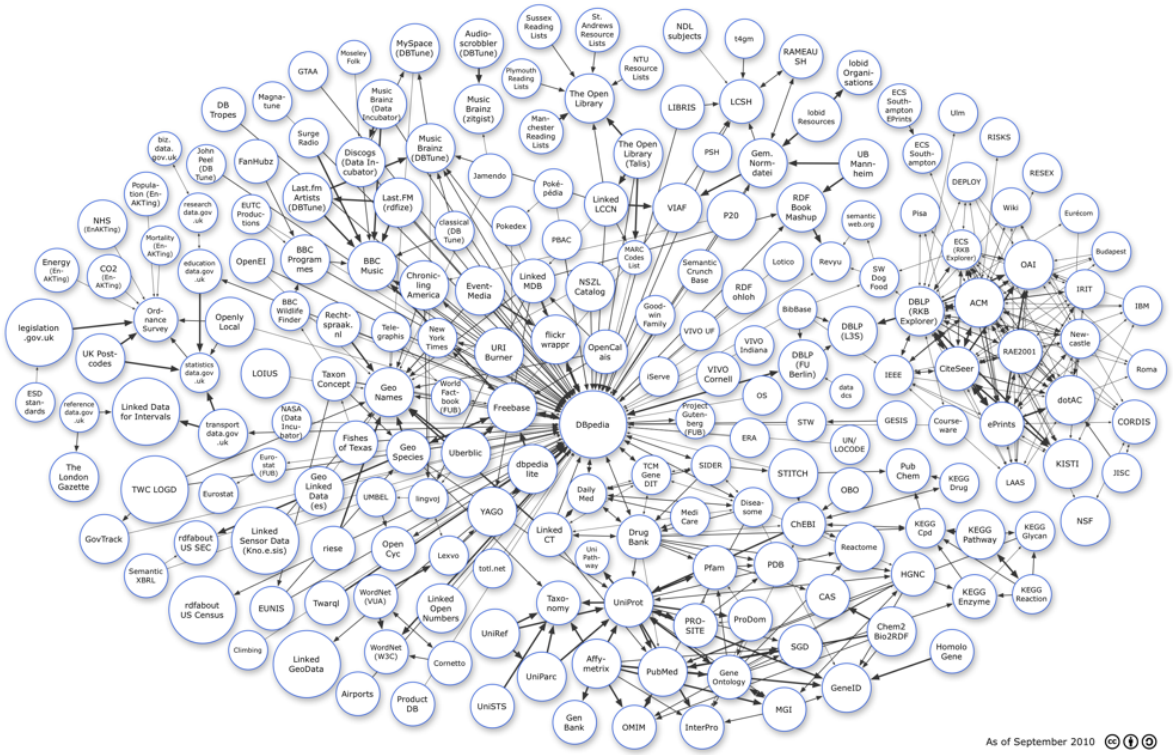
Iniciativa *linked data* se zaměřuje na propojování strukturovaných dat na webu pomocí odkazů. Termín linked data označuje styl publikování a propojování dat na webu, soubor doporučení, jak se tohoto stylu držet, a také data publikovaná podle tohoto modelu [5]. Hlavní ideou je vytvoření globálního datového prostoru, kde jsou propojeny a sdíleny nejen dokumenty, ale i data [4, 8]. Tento prostor bývá nazýván také *web dat* a představuje další vrstvu klasického webu dokumentů.

Iniciativa linked data vznikala v rámci aktivit výzkumné komunity sémantického webu a zvláště projektu konsorcia W3C *Linking Open Data Project (LOD)*[7], spuštěného roku 2007. Tento projekt si kladl za cíl vytipovat datové zdroje, publikované pod otevřenou licencí, a zpřístupnit je na webu za použití rámce RDF podle principů linked data. Projekt byl otevřeně přístupný všem zájemcům o publikování dat touto cestou. Nejspíš právě otevřenost tohoto projektu vedla k takovému rozmachu webu dat. Projektu LOD se zúčastnily nejprve menší výzkumné a univerzitní skupiny a malé společnosti, později se přidaly významné organizace, jako je BBC, Thomson Reuters nebo také Kongresová knihovna.

Všechny zdroje dat, které se projektu účastní a zároveň tvoří web dat, včetně jejich vzájemných propojení, znázorňuje pravidelně aktualizovaný *Linking Open Data Cloud* (viz

[5] http://www.hunch.com/
[6] http://www.apture.com/
[7] http://linkeddata.org

obr. 4). Tento diagram vzniká v rámci projektu LOD, existuje rovněž katalog *LOD Cloud Data Catalog*, který je vytvářen v rámci organizace *Comprehensive Knowledge Archive Network (CKAN)*[8], provozující registr zdrojů dat zveřejněných pod otevřenou licencí.



Obrázek 4: Linked Data – propojené zdroje[9]

Principy linked data stanovil Tim Berners-Lee ve čtyřech základních pravidlech [2]. V praxi představují návod, jak publikovat a propojovat data za využití stávající webové infrastruktury a při dodržení zavedených webových standardů.

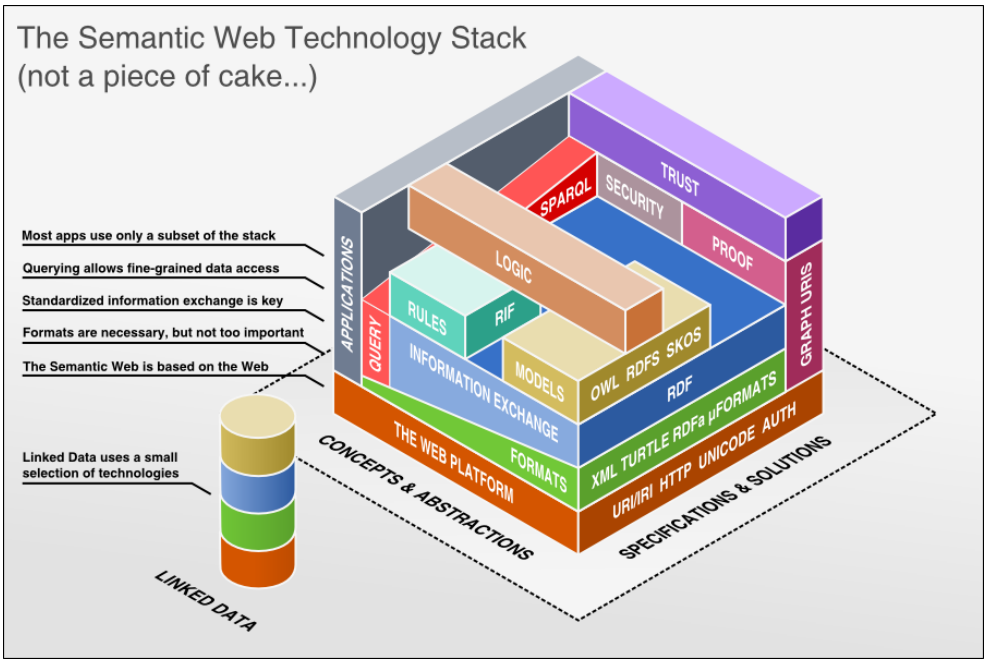
1. Pojmenovat věci URI identifikátory.
2. Používat HTTP URI, aby bylo možné je vyhledat.
3. Hledá-li někdo URI, poskytnout užitečné informace s využitím standardů (RDF, SPARQL).
4. Uvedením odkazů na jiná URI lze objevovat více.

Linked data i sémantický web se shodují v mnoha předpokladech: snaží se o rozšíření současného webu, propagují použití rámce RDF, identifikátorů URI a efektivnější zpracování dat webovými aplikacemi. Také propojování dat je společným bodem obou iniciativ.

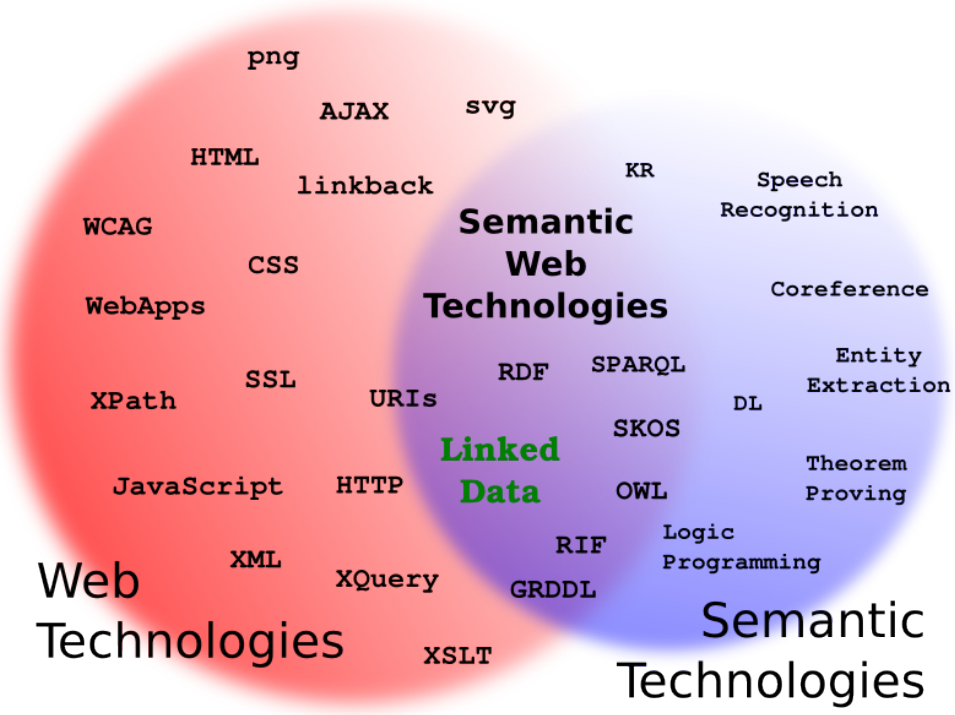
Iniciativa linked data původně vychází z konkrétního projektu, je tedy je zaměřena praktičtěji. Vztah mezi těmito termíny je vhodně vyjádřen v [5]: *„Vize sémantického webu je založena na budování globálního webu strojově zpracovatelných dat. Zatímco sémantický*

[8] http://www.ckan.net
[9] http://lod-cloud.net

web, nebo web dat je cílem či výsledkem, linked data představuje prostředek nebo způsob jeho dosažení.“



Obrázek 5: Linked Data jako podpora sémantického webu[10]

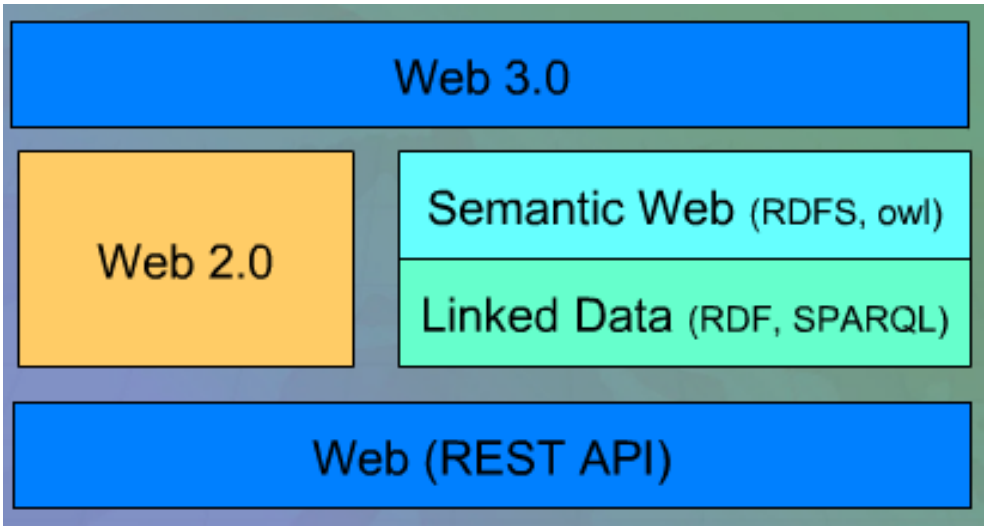


Obrázek 6: Linked Data jako podpora sémantického webu

[10] http://bnode.org/media/2009/07/08/semantic_web_technology_stack.png

6 Závěr

Závěrem lze konstatovat, že sémantický web a jeho technologie po deseti letech od památného článku začínají naplňovat očekávání. Někomu se to může zdát, že to trvá dlouho. Ostatně to potvrzují i některé průzkumy či ankety – viz třeba [1]. Na druhou stranu se ukázalo, že web založený na původních principech má svá omezení. Spekulace, že Web 2.0 je reklamní trik a že se jedná o další z řady „buzzwords“, snad ani není potřeba vyvracet. I proto, pokud se před pár lety zdálo, že Web 3.0 je cosi uměle vytvořeného, lze tvrdit, že vývoj k tomu směřuje. A nezastupitelnou roli v něm sehrávají principy Webu 2.0, linked data a sémantického webu (viz obr. 7].



Obrázek 7: Linked Data vs. Web 3.0 vs. sémantický web

Reference

[1] Anderson, Janna Quitney, Rainie, Lee. *The Fate of the Semantic Web*. Pew Research Center, 2010. URL: http://www.pewinternet.org/Reports/2010/Semantic-Web.aspx.

[2] Berners-Lee, Tim. *Linked Data*. Last change: 2009/06/18 URL: http://www.w3.org/DesignIssues/LinkedData.html.

[3] Berners-Lee, Tim, Hendler, James, Lassilla, Ora. The Semantic Web. *Scientific American*, 2001. vol. 284, no. May. str. 35–43. URL: http://www.sciam.com/2001/0501issue/0501berners-lee.html.

[4] Bizer, Chris, Cyganiac, Richard; Heath, Tom. *How to Publish Linked Data on the Web*. 2008. URL: http://www4.wiwiss.fu-berlin.de/bizer/pub/LinkedDataTutorial/.

[5] Bizer, Chris; Heath, Tom; Berners-Lee, Tim. 2009. Linked Data – The Story So Far. *International Journal on Semantic Web and Information Systems*. 2009, vol. 5, no.3, s. 1–22. URL: http://eprints.ecs.soton.ac.uk/21285/1/bizer-heath-berners-lee-ijswis-linked-data.pdf.

[6] Boutin, Greg. *Tying Web 3.0, the Semantic Web and Linked Data Together – Linked Data is a Medium*. 2009. URL: http://www.semanticsincorporated.com/2009/05/tying-web-30-the-semantic-web-and-linked-data-together-part-23-linked-data-is-a-medium.html.

[7] Corbin, Kenneth. *Google Snaps Up Metaweb in Semantic Web Play*. 2010. URL: http://www.internetnews.com/search/article.php/3893741/Google-Snaps-Up-Metaweb-in-Semantic-Web-Play.htm.

[8] Heath, Tom; Bizer, Chris. *Linked Data: Evolving the Web into a Global Data Space*. Morgan & Claypool, c2011. URL: http://linkeddatabook.com/book. ISBN: 9781608454310 (ebook).

[9] Hepp, Martin. *Semantic SEO for Google with GoodRelations and RDFa*. 2010. URL: http://www.heppresearch.com/gr4google.

[10] Lunn, Bernard. *Semantic Enterprise: What Are The Gorillas Doing? (Oracle, IBM, HP, Cisco, Microsoft and SAP)*. 2010. URL: http://semanticweb.com/semantic-enterprise-what-are-the-gorillas-doing-oracle-ibm-hp-cisco-microsoft-and-sap_b710.

[11] MacManus, Richard. *Top 10 Semantic Web Products of 2010*. 2010. URL: http://www.readwriteweb.com/archives/top_10_semantic_web_products_of_2010.php.

[12] Oracle. *Oracle Database Semantic Technologies*. c2010. URL: http://www.oracle.com/technetwork/database/options/semantic-tech/index.html.

[13] Sklenák, Vilém. Sémantický web. In *Inforum 2003*. Albertina icome Praha, 2003. URL: http://www.inforum.cz/inforum2003/prispevky/Sklenak_Vilem.pdf.