Browsing Semantic Data in Slovakia

Ján Mojžiš
Institute of Informatics, SAS, Bratislava, Slovakia
upsyjamo@savba.sk

Michal Laclavík
Magnetic Media Online, New York, USA
laclavik@magnetic.com

Abstract

Semantic data browsing is important task for open and governmental data in behalf of public control. There are many projects and solutions regarding semantic data browsing and navigation, but despite the fact, in Slovakia, the availability of such data is poor. It is a shame, because projects like National Action Plan of Open Government and the site data.gov.sk are already operating for several years. In this work we would like to point out key aspects of semantic data and detail the Slovak market of semantic data. We design and propose our solution of semantic data browsing, evaluate the implementation in our AGECRT NET tool.

Keywords: semantic data, open data, navigation, visualization, RDF

1. Introduction

Data on the Web are published in various formats. Either for common users, who tends to use popular web browsers in order to browse online data. In such case, browsers use HTTP protocol to browse HTML syntax formatted data. Other kind of data is semantic data. Here we can include semantic web formats like RDF. Actually, it is only a standard and the concrete implementation is then RDF/XML format. As for HTML data, also for RDF data, there are browsers. They are different browsers, but in general, serve the same goal of browsing.

Nowadays, there is an initiative which press on publishers (the ones who put the data online) and encourage them to format their data in one of RDF formats. The benefits are apparent; machine readable format, clear syntax, suitable for batch and bulk machine processing. Here, the data is separated from the content. For comparison, the data in HTML format have the data mixed with the content. RDF format contains no content, but the data alone. We have found that most simple RDF format is RDF/N-Triple.

The USA governmental project data.gov hosts datasets of different departments and states of the USA. However, while the USA governmental project Data.gov hosts a decent volume of 189,998 datasets, its Slovak counterpart data.gov.sk appears just as a poor cousin, offering "only" 624 datasets. Beside of that, the potential for Slovakia is promising, yet it seems to have completely missed the chance. Like U.S. governmental data.gov website service, which was launched in late May 2009, Slovak counterpart, data.gov.sk, is controlled by the Open Government Partnership Action Plan of the Slovak Republic (OGSK)[1]. The intention is common for both projects, to publish governmental open data. Yet data.gov.sk is missing several vital governmental datasets. In comparison to data.gov.sk, which grew from 47 datasets[2] to more than 188,000 in just 6 years[3], its Slovak counterpart, data.gov.sk, available from 2012, contains only 624 datasets. Because data.gov.sk was proposed as part of OGSK[1], all governmental institutions (including cities) should participate with publishing. To

[1] Open Government Initiative Action Plan of the Slovak Republic (2012). Retrieved from http://www.otvorenavlada.gov.sk/data/files/1853_ogp-action-plan-slovakia.pdf

[2] Data.gov Tunns Six! (2015). Retrieved from https://gsablogs.gsa.gov/gsablog/2015/05/26/data-gov-turns-six/

[3] Data.gov datasets. (2015). Retrieved from http://catalog.data.gov/dataset

illustrate the clumsiness of official institutions with data.gov.sk, there is a city of Prešov, a bright example of initiative for open data, has its datasets published with data.gov.sk. Here we can find street names dataset, school-like institutions or statistics about citizen counts. That would be great, but Prešov is the only city, actually publishing with data.gov.sk, to the day of writing this work. In Slovakia, there are 8 countries, each with 1 regional capital (including a city of Prešov), so 7 regional capital cities are still absent in data.gov.sk.

Regarding OGSK, the measure of its fulfillment can be verified using either report by independent research institution (Kurian, 2013) or that by the government itself[4], although reports seems a bit outdated (last reports are dated to 2013, based on year period, 2012/2013, completely missing 2013/2014 and recent). But based on (Kurian, 2013), prior to the publication, only 8 out o 22 commitments was completed. To compare with Czech Republic, a close neighbor to Slovakia, Czech government supports SPARQL endpoints for structured data querying through their governmental portal portal.gov.cz[5] as well as they support bulk download.

Slovak Insurance Agency (SIA) is publishing its debtors lists composed of natural persons in CSV format[6], which is a kind of structured, machine-readable format. Although SIA is governmental, its dataset is not included in data.gov.sk. Interestingly, another governmental institute, the Všeobecna Zdravotná Posťovňa (VsZP), governmental health institution, also publishes its debtors online, but in comparison to SIA, this dataset, actually, is listed under data.gov.sk[7]. What is important to say, that there are no legal obstacles in laws in Slovakia, the ones, which would prevent institutions from publishing with data.gov.sk. Processing of personal data in Slovakia is regulated by an Act No. 122/2013 Coll. on Protection of Personal Data and on Changing and Amending of other acts, resulting from amendments and additions executed by the Act. No. 84/2014 Coll. Here, Section 10, Paragraph 3, Point e clearly states, that (personal) information made public by controller may be processed legally and that information is, for example, also SIA's debtors list. Therefore, it should have been included in data.gov.sk.

Problems and delay, regarding the process of data publication, is often caused by departments alone. From OGSK we can see that there is an action plan of data publishing, which involves various departments and kinds of data[8]. A worth of notice, however, is that there are no debtors datasets listed neither for VsZP nor SIA. But from this report, it is clear, that the Slovak Business Register (SBR), which lists all legal and natural persons, who actually do (legal) business in Slovakia, its dataset publication is only partial and, it is still under the subject of negotiation. Moreover, the only format that user is able to get out of SBR is HTML, which complicates the steps of further machine processing, for instance, person titles are cast into the name field together with surname and the address can be noted in many different ways (with postal no., without it, city first, or street first, street numbering variability, abbreviations and etc.).

Even that the data gets published and bulk download is present, the data itself is in inappropriate format. For example Financial department of Slovakia also publish debtors

[4] Report on Action Plan fulfillment for years 2012 and 2013. (2013). Retrieved from http://www.otvorenavlada.gov.sk/hodnotenie-plnenia-uloh-z-akcneho-planu-ogp-pre-roky-2012-a-2013/

[5] Dataset of Czech Administration of Social Insurance. (2014). Retrieved from https://portal.gov.cz/app/RejData/rec.jsp?id=1646070&id_rej=97898&y=2015&m=11&doctype=idx

[6] Debtors list of Social Insurance Agency in Slovakia, in CSV format. (2015). Retrieved from http://www.socpoist.sk/index/open_file.php?file=dlznici/2015-11-06_SP_dlznici_CSV.zip

[7] Health Insurance Agency Všeobecká zdravotná poisťovňa, listed under data.gov.sk. (2015). Retrieved from http://data.gov.sk/dataset?_organization_limit=0&organization=vseobecna-zdravotna-poistovna

[8] Dataset listing of Slovak government. (2012). Retrieved from www.otvorenavlada.gov.sk/data/files/2651_statna-sprava-datasety.xls

lists[9], but offer only PDF formatted files on its website page. So do commercial health insurance agency UNION[10], which also has its debtors, but again, only PDF is available. We wonder such wastefulness, because an effort was clearly present; export to PDF. But how much more work it would take to provide CSV for instance?

For the field of statistics information publishing, there is Statistical Office of the Slovak Republic (SLOVSTAT)[11]. It collects (local) macro-economical, social and other statistics, very similar to Eurostat[12]. But in comparison to Eurostat, it does not offer the bulk downloading function, so whole datasets can only be browsed manually via their on-line web interface. Eurostat datasets, however, could easily be downloaded and then analyzed and processed.

The commercial, non-profit or government-independent organizations or projects in Slovakia, are partially serve as complements to, somewhat paralyzed, governmental institutions, which rather do their bureaucracy, strictly following various acts and are backside facing towards citizens. We can name foaf.sk, a project intended to online SBR data browsing and the collection of personal and business data out of public available datasets (like SIA and VsZP). Other projects include vorsr.sk, Fair-Play Alliance[13], Transparency International Slovakia[14] and, perhaps, Slovak Open Data Initiative[15]. We shall discuss them further.

In this paper we would like to point out the situation with semantic data availability in Slovakia and suggest improvements. As it seems, that official governmental institutions tends rather to not publish their data openly, user needs to navigate through inappropriate HTML formatted pages and other projects emerges lame, we propose our solution to semantic data browsing. But the lack of vital SPARQL endpoints and bulk downloads forcing us into, normally avoidable, workarounds. Here inaccuracy may occur during information extraction, while parsing HTML via scripts or regular expressions. In any case, we are able to browse the data of SBR, refine original structure of data as much as possible and create graphs. We are able to provide graph visualizations in advance. SBR dataset is the first we have picked, because the lacking of sufficient solutions is noticeable. As we have already stated, SBR contain all legal and natural persons, who do business in Slovakia and in comparison to the Trade Register of the Slovak Republic, it contains connections essential for social network creation.

2. Related Work

In this section we discuss projects, solutions and services which aim to semantic data browsing improvement, either supplying relationship visualization, data structure refining or information extraction, kind of added value. Thus, we exclude data.gov.sk, SBR, VsZP, SIA and other governmental institutions.

Dokulil and Katreniaková suggest visualize for navigation in RDF data based on user's mental map. They are able to reduce graph by removing nodes or a whole subtree. Animation is available during restructuring operation (Dokulil & Katreniaková, 2008).

[9] Debtors list of Finance Department of Slovak Republic in PDF format. (2015). Retrieved from https://www.financnasprava.sk/sk/elektronicke-sluzby/verejne-sluzby/zoznamy/zoznam-danovych-dlznikov

[10] Debtors list of Health insurance agency UNION in PDF format. (2015). Retrieved from http://www.union.sk/documents/51150/Zoznam_dlznikov_FO

[11] http://slovak.statistics.sk/

[12] Bulk download of data published by Eurostat. (2015). Retrieved from http://ec.europa.eu/eurostat/data/bulkdownload

[13] Fair-Play Alliance Slovakia. (2015). Retrieved from http://www.fair-play.sk/

[14] Transparency International Slovakia. (2015). Retrieved from http://www.transparency.sk/en

[15] Open Data Initiative. (2015). Retrieved from http://opendata.sk/liferay/o-nas

Slovak project foaf.sk is intended to SBR data browsing primary. With the help of regular expressions and with the public-available data of SBR, authors extracted information and created graph. As we have mentored above, SBR does not publish its data in any structured format, nor does it offer bulk download and the data is missing on data.gov.sk also. Authors thus need to parse the HTML, filter-out the content and refine the data on their own. They have handled the duplicate entries and missing person's unique identifiers like personal number (regarding personal privacy see aforementioned 122/2013), so they had used as much as name and address. In their task of relationship discovery, they use spread activation and offer full-text search with Sphinx tool (Suchal & Vojtek, 2009). Graph visualization is powered by Adobe Flash.

Foaf.sk, originally founded as a non-profit initiative of individuals, proud of graph visualization, later sold-off, we have found, that now (Nov/5/2015), it is missing its original visualization, stagnating and, perhaps, slowly started to decline. Only a brief report about reconstruction and plans for visualization is provided, but 5 months already, there is a no change.

M. Laclavík proposes a set of tools for information extraction (Laclavík et al., 2012), annotation and graph creation (Laclavík et al., 2010).. Ontea is a text annotation tool with the use of ontology and regular expressions. It can work in a cluster with MapReduce, extracts information from text, creating objects and, finally, semantic trees. The result is a graph, which can be searched with their next tool gSemSearch (former Email Social Network Search Prototype). To perform a relationship discovery task, gSemSearch is using spread activation (similar to foaf.sk). Spread activation is computed upon their database store SGDB, where they also perform graph traversal. In his former works (Laclavík at el., 2014), (Laclavík et al., 2011), (Laclavík et al., 2012), (Laclavík et al., 2011), he further use this concept and he performs distributed parallel computing with MapReduce only in initial stage, where he extracts and annotate entities needed for graph creation. This project, however, seems dead, as the graph is not visualized and to this day (Nov/5/2015) the official gSemSearch tool website is unavailable[16].

Another SBR browsing project and service is vorsr.sk, a virtualization of SBR with a great potential. Now, only a shell, an encapsulation or SBR, there is only Google search box and just a few direct links to SBR official website. In Czech Republic, for comparison, there is one visualization website, to serve as a kind of visual SBR browser, obchodni-rejstrik.podnikani.cz, with which, users can, still, see visualizations and find is appealing and functioning.

Fair-Play Alliance (FPA) and Transparency International Slovakia (TIS) are non-governmental organizations, aimed mainly to the fight against corruption. They publish various reports and offer data browsing. Although FPA provides data browsing through their Datanest service[17], search is supported and the menu is rich, again, no bulk download is possible and so neither is querying with SPARQL. The results of TIS are important and their purpose is reasonable. Manly they publish particular reports about corruption, statistics and audits. They serve as public control, but they do not focus on structured data availability, like SPARQL or structured data bulk download.

An initiative for Open Data is a project, "a group of people, which wants to carry through a plan for open and modern open government." On their site[18], however, it is

[16] GSemSearch tool, search online for relations in Enron Email corpus. (2013). Retrieved from http://try.ui.sav.sk:7070/enron/gSemSearch.html. Address now unavailable (2015)

[17] Lawyer employee register by Fair-Play Initiative. (2015). Retrieved from http://datanest.fair-play.sk/datasets/56#/data

[18] Open Data Initiative. (2015). Retrieved from http://opendata.sk/liferay/o-nas

markedly outdated with last news dated from 2012. Again, in comparison to opendata.cz in Czech, Slovak counterpart is rather dull with several links even dead. And this initiative shows a bit of fragmentation, being distributed, beside its official website opendata.sk, across group Facebook[19] and another website Utópia[20], also. On one of their branching sites, Utópia, we find more recent updates, but the effect of the initiative is rather unfilled. Also, the capital city of Bratislava supports various projects among which, a project about dataset publishing in standardized structured format was pending. But, unfortunately, this project was not supported[21]. Despite that, the datasets are now available[22], but again, neither structured data nor bulk downloading is provided.

Similar to data.gov.sk is Czech counterpart opendata.cz, which is but an independent initiative for open data publishing, with but a few datasets.

Abroad, we can find RelFinder. It is, perhaps, one of the most promising and known relationship discovery tools. The main difference from foaf.sk is its support of virtually any SPARQL endpoint (Linked Data publisher). RelFinder is capable of visualization as well. It is designed and created by P. Heim as a search and graph visualization tool. The solution is implemented as a web service and powered by Adobe Flex. The service allows user to specify SPARQL endpoint. RelFinder allows user to search for 2 vertexes for which the relationships are about to be discovered. To simplify search process, an auto-complete functionality is available, providing user with the list of matched entries. On the official website of the RelFinder[23], there is an implementation and user can test the service and search for relationships. Among the visualization advantage, the data is fetched and presented online, no data is stored therefore the data is never outdated. Also the ability to support virtually any SPARQL endpoint makes RelFinder a universal tool. For instance, when searched for ”Bill Gates” and ”Virginia”, the main relationship is returned as ”United States”. RelFinder offers filtering as well (for example based on connectivity). A filtering is available; it hides vertexes which are simply too far above threshold. On the graph visualization, edges and vertexes are colored (based on filters selected in tabs). Layout is maintained with force-directed algorithm. Its code is published under GNU and hosted by Google Code[24].

Similarly to the web browsers, Semantic Web browsers (intended to browse Linked Data), browse the semantic data online. Semantic Web browser can connect various endpoints, does support SPARQL query language and combines various visualization techniques (facets, graph visualizations). Among many browsers, those well-known are DBpedia Mobile (Becker & Bizer, 2008), Exhibit (Huynh, Karger, & Miller, 2007), MSpace (Smith et al., 2005) or Tabulator (Berners-Lee et al., 2006) and RelFinder. They help with displaying often offering various visualization techniques like lenses (Furnas, 1986). But because in Slovakia, the situation with semantic data is poor, the use of linked data browsers in Slovak environment is discouraging.

So, to conclude this section, many projects are launched and were executed already, but results are very poor, unification, which should involve data.gov.sk, is nowhere, rather fragmentation and distribution of the data is present, some websites are offline, others were never launched. Often a project, which was successfully launched actually, is just a one-shot

[19] Facebook group of Open Data Initiative Slovakia. (2015). Retrieved from http://datanest.fair-play.sk/datasets/56#/data

[20] Utópia, a branch of Open Data Initiative Slovakia. (2015). Retrieved from https://utopia.sk/liferay/home

[21] City of Bratislava: Project about selected datasets publication; unsupported. (2014). Retrieved form http://pr.banm.sk/liferay/datasetybanm

[22] Datasets of city of Bratislava; only in PDF format. (2015). Retrieved from http://zverejnovanie.banm.sk/

[23] RelFinder demo application. (2015). Retrieved from http://www.visualdataweb.org/relfinder/relfinder.php

[24] RelFinder source code hosted on Google Code. (2015). Retrieved from http://code.google.com/p/relfinder/

item, dissolving and eventually found dead. Together with a fact, that many governmental institutions still provide non-machine readable, structured data, the report on OGSK is a disappointment and Semantic Web standards like Notation 3 [25] or RDF/XML Syntax Specification[26] were published at least 10 years ago, we see the current status of semantic data publishing as a shame. And the potential there is, we are in the European Union, ready for Eurofunds collecting, we have action plans (OGSK), commissions and initiatives (Open Government Partnership OGP[1]), there are academic and research institutions like Slovak Academy of Sciences[27] and Slovak University of Technology[28] and more, yet results are just partial and effects are surely delayed. In order to evolve, we should always compare to such successful projects like data.gov, commercial-like obchodni-rejstrik.podnikani.cz or governmental portal of Czech Republic portal.gov.cz.

3. Our solution

With the aim primary on unstructured information extraction and refining, relationship discovery and visualization, we propose our solution for SBR in the first place. The reason for this is, primary, that HTML formatted results of SBR are very jerky and uncertainty regarding the structure of information is very high. Readers can also be pointed by J. Suchal and P. Vojtek (2009), that care should be taken towards type errors. We discuss that later.

In this work, we try to fill-up the gap of visualization and, somehow limited data access offered by SBR, adapting to the problems disclaimed above. We suggest a new client-side paradigm, which does not depend on a particular website like foaf.sk.

Figure 1 describes the schema briefly and the key elements are parsers with other tools on the top and structured formats, for datastore, on the bottom.

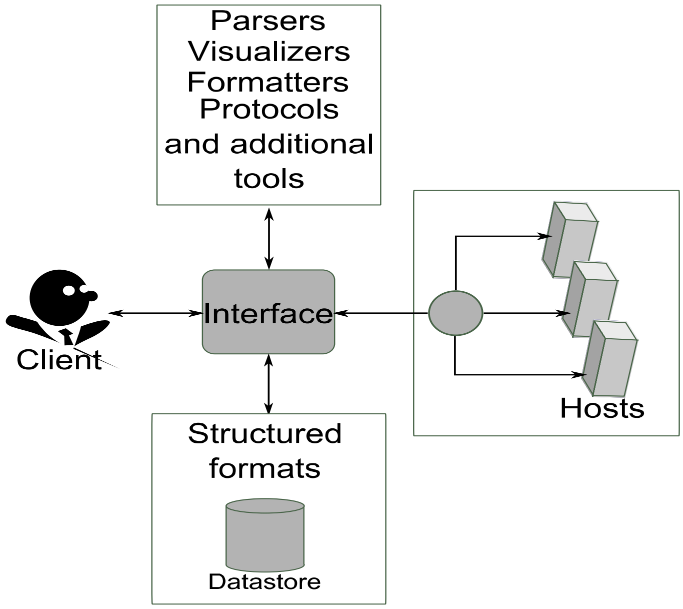


Figure 1. Schema of client-side application for semantic browsing.

[25] Notation 3, W3C. (2011). Retrieved from http://www.w3.org/TeamSubmission/n3/

[26] RDF/XML Syntax specification. (2004). Retrieved from http://www.w3.org/TR/REC-rdf-syntax/

[27] Slovak Academy of Sciences. (2015). Retrieved from http://www.sav.sk/?lang=en

[28] Slovak University of Technology in Bratislava. (2015). Retrieved from http://www.stuba.sk/english.html?page_id=132

We do not consider incremental machine querying as the best choice, because it both burdens the server for the duration of the process and it creates copies of records, which should then be updated periodically and it also induce possible errors. However, as Suchal & Vojtek (2009), we see no alternation to regular expression usage, because HTML syntax is unavoidable part of the solution, when refining data of SBR.

As our solution calculates with modules (wrapped as tools in Figure 1), we design solution modularly with plug-in possibility of other modules. Other modules may process different data sources from different hosts, but they should return results in common format. Each module should implement its own protocol, because of variability and particular purpose to serve. Each module should function as thread-safe implementation, using main thread idle time in order to keep user interaction seamless.

Structured format may be any standard structured format like RDF/XML or N-Triple, or particular modules can use their own format, in which case a description of such format must have been given clearly.

Client interacts with the application through interface, using functions of modules, loading or saving the data and performing tasks like visualization.

A choice to store the data was made to ease of later usage and enable data sharing. Because the data may still be unavailable in machine-readable form (like SBR) and there was an effort to gather it (parsing, refining), we find this option useful. Although we are aware that the data may became outdated after a while.

4. Implementation

In order to implement our solution we choose Java programming language as platform independent developing backend. The language and especially developing environment (Eclipse, Netbeans) are also platform independent and at minimum, Linux is supported, where .Net is only partially available. We design our GUI with the Swing library and for visualization, we use Jung graph library[29], which is very flexible and offers rich customizability. Regular expressions are already built inside Java, so the only problem is actual writing of expressions. For the purpose of quick testing, we have developed a testing webpage, where we can evaluate regular expressions prior their implementation in the module. Because we focus on SBR and HTML output, we have found, that best suitable regular expression can be derived from "[^<>]". Hence the negation tells the engine, that neither < nor > could be encountered during string evaluation. In SBR, there are common symbols like dash, parenthesis, numbers as well as diacritical marks like carons or acutes.

5. Data

To evaluate our solution, implemented in AGECRT NET tool, we used SBR browsing module of AGECRT NET and have connected to SBR dataset. Although we do not have direct numbers, SBR contains around 86.500 firm records. SBR website does not offer statistics regarding their records count, but on its main page[30], only update timestamps are displayed. Indirectly, however, it is possible to create an imagination of the size based on firm record identifiers. As J. Suchal states (Suchal, & Vojtek, 2009), they have harvested SBR dataset using continuous incrementing of (aforementioned) identifiers. We do not know if Suchal was aware of that, but identifiers (and thus firm records) are distributed across, at least, 8 different datasets. Parameters of GET URI include SID, which varies from 1 up to 7, so each firm ID must be queried with each different SID in order to ensure, that all firm records are searched.

[29] Java Universal Network/Graph Framework, JUNG. (2010). Retrieved from http://jung.sourceforge.net/

[30] Slovak Business Register. (2015). Retrieved from http://www.orsr.sk/Default.asp?lan=en

SBR contains firm records which are composed of firm detailing items. Those items include Partners, Management body, Stockholders or Supervisory board. Additionally, there are items on Branch of the enterprise or Restructuring trustees. Items can be either for natural persons or legal persons and other firms. With the use of such items from firm details, a social network (of firms and persons) may be created. However, unlike Facebook, there are no direct person-to-person edges, rather person-firm and firm-firm type edges.

There are 2 possible query types, we use, in general. Search for a firm. Here a firm name and address may be typed or its identification number supplied. Or search for a person, only name and surname are allowed. No address information is supported.

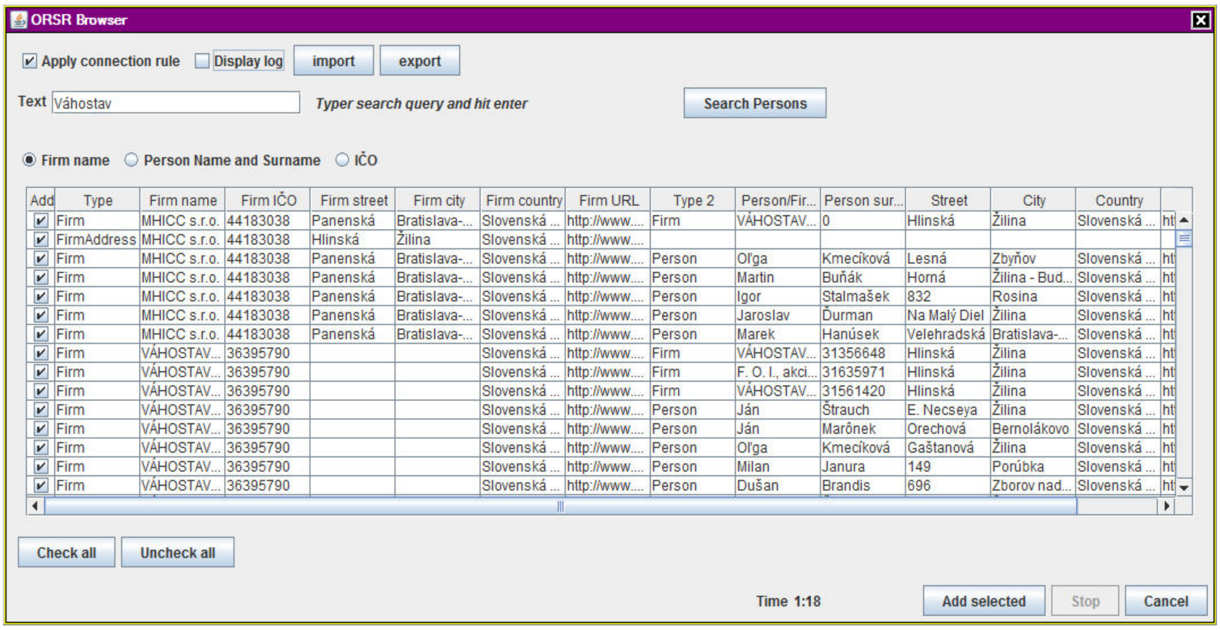
Firm records are divided into 2 sets. One is actual, containing information actual to a generated timestamp. Another is full, marking historical information and changes also.

6. Results

We have searched for firm "Váhostav", which is a rather big firm in Slovakia, with many press articles published about[31].

On Figure 2 there is a browsing window, for SBR data results, displaying tabular structure, which was refined from SBR dataset by continuous querying.

Resulting graph is rather complex. There are 175 vertices and 201 edges, which were created directly, containing 158 persons and 17 companies. As we see on Figure 3, some filtering methods are required in order to create suitable overview of relations. Figure 4 thus shows the visualization of the same graph, but with tens of vertices merged. Now it contains 22 persons and 17 companies. A merging was performed for clarification and is built inside the visualization module. A simple condition says that a merging is performed if persons are unique, thus if a person is connected only to 1 firm. More formally, person vertices are merged, if their vertex degree equals 1 (each).



☑ Apply connection rule ☐ Display log

Text Váhostav *Typer search query and hit enter*

● Firm name ○ Person Name and Surname ○ IČO

Add	Type	Firm name	Firm IČO	Firm street	Firm city	Firm country	Firm URL	Type 2	Person/Fir...	Person sur...	Street	City	Country	
☑	Firm	MHICC s.r.o.	44183038	Panenská	Bratislava-...	Slovenská ...	http://www....	Firm	VÁHOSTAV...	0	Hlinská	Žilina	Slovenská ...	ht
☑	FirmAddress	MHICC s.r.o.	44183038	Hlinská	Žilina	Slovenská ...	http://www....							
☑	Firm	MHICC s.r.o.	44183038	Panenská	Bratislava-...	Slovenská ...	http://www....	Person	Oľga	Kmecíková	Lesná	Zbyňov	Slovenská ...	ht
☑	Firm	MHICC s.r.o.	44183038	Panenská	Bratislava-...	Slovenská ...	http://www....	Person	Martin	Buňák	Horná	Žilina - Bud...	Slovenská ...	ht
☑	Firm	MHICC s.r.o.	44183038	Panenská	Bratislava-...	Slovenská ...	http://www....	Person	Igor	Stalmašek	832	Rosina	Slovenská ...	ht
☑	Firm	MHICC s.r.o.	44183038	Panenská	Bratislava-...	Slovenská ...	http://www....	Person	Jaroslav	Ďurman	Na Malý Diel	Žilina	Slovenská ...	ht
☑	Firm	MHICC s.r.o.	44183038	Panenská	Bratislava-...	Slovenská ...	http://www....	Person	Marek	Hanúsek	Velehradská	Bratislava-...	Slovenská ...	ht
☑	Firm	VÁHOSTAV...	36395790			Slovenská ...	http://www....	Firm	VÁHOSTAV...	31356648	Hlinská	Žilina	Slovenská ...	ht
☑	Firm	VÁHOSTAV...	36395790			Slovenská ...	http://www....	Firm	F. O. I., akci...	31635971	Hlinská	Žilina	Slovenská ...	ht
☑	Firm	VÁHOSTAV...	36395790			Slovenská ...	http://www....	Firm	VÁHOSTAV...	31561420	Hlinská	Žilina	Slovenská ...	ht
☑	Firm	VÁHOSTAV...	36395790			Slovenská ...	http://www....	Person	Ján	Štrauch	E. Necseya	Žilina	Slovenská ...	ht
☑	Firm	VÁHOSTAV...	36395790			Slovenská ...	http://www....	Person	Ján	Marônek	Orechová	Bernolákovo	Slovenská ...	ht
☑	Firm	VÁHOSTAV...	36395790			Slovenská ...	http://www....	Person	Oľga	Kmecíková	Gaštanová	Žilina	Slovenská ...	ht
☑	Firm	VÁHOSTAV...	36395790			Slovenská ...	http://www....	Person	Milan	Janura	149	Porúbka	Slovenská ...	ht
☑	Firm	VÁHOSTAV...	36395790			Slovenská ...	http://www....	Person	Dušan	Brandis	696	Zborov nad...	Slovenská ...	ht

Time 1:18

Figure 2. Results for firm name "Váhostav" are in table. Each row defines a firm with its name, identification number and address. Then a connection is specified (whether it be a person or another firm).

[31] Press publications about firm Váhostav (in Slovak). (2015). Retrieved from https://www.google.sk/search?q=v%C3%A1hostav&oq=v%C3%A1hostav#q=site:sme.sk+v%C3%A1hostav

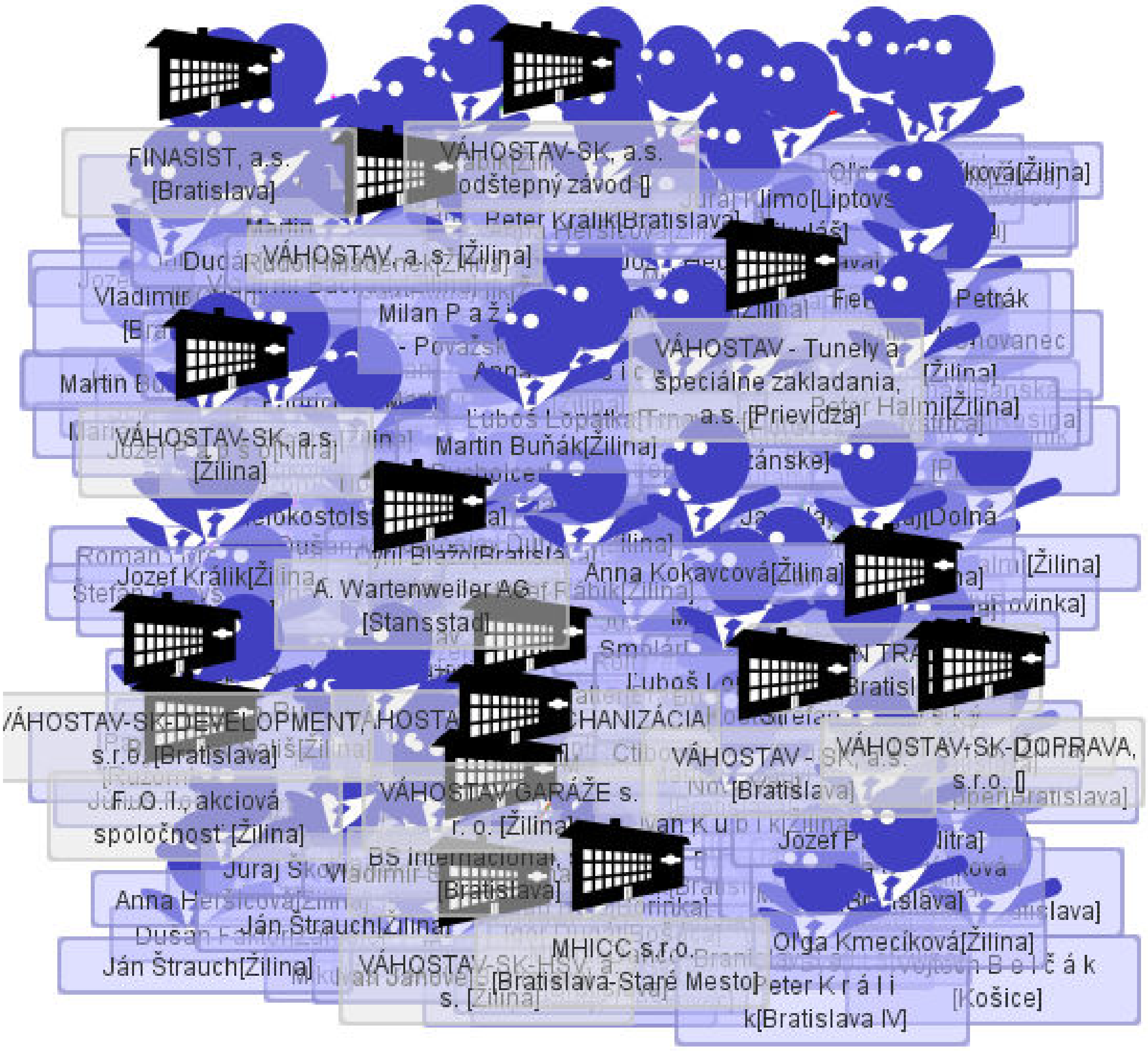


Figure 3. Graph created directly from the results.

The usefulness of such visualization has its key points regarding connections. Thanks to SBR browsing module, we were able to get 22 firm records for “Váhostav” query. Between any 2 companies, connections may be (and often are) not bidirectional, so, in order to navigate through connections, we have refined all 22 records. Although, even being filtered, graph is still complex.

And it is possible to further navigate and search for outgoing connections, for example firm “MERLIN TRADE, a.s.” on Fig.4 contains item on “Ján Kato”, which is already included in our graph and connected to “VÁHOSTAV-SK-DEVELOPEMENT” on bottom left side and “VÁHOSTAV-SK, a.s.” in the center.

Edge coloring and drawing is helpful with overlapped edges. For methods of visualization, including coloring, we refer to studies of H. Omote and K. Sugiyama (2006), and I. Herman, G. Melanon, and M. S. Marshall (2000) or our study on graph clutter filtering and connectivity distance (Mojzis & Laclavik, 2014).

We also use edge coloring and drawing methods to ease edge following and to help with overlapping. Person “Ján Kato” is connected with 2 different colors. It is thus possible to follow edge from its start to end. Or “Oľga Kmecíková”, who is connected with 4 edges, each one colored with different color and pattern. Edge coloring and patterning is helpful as we have already proposed in our previous work (Mojzis & Laclavik, 2014).

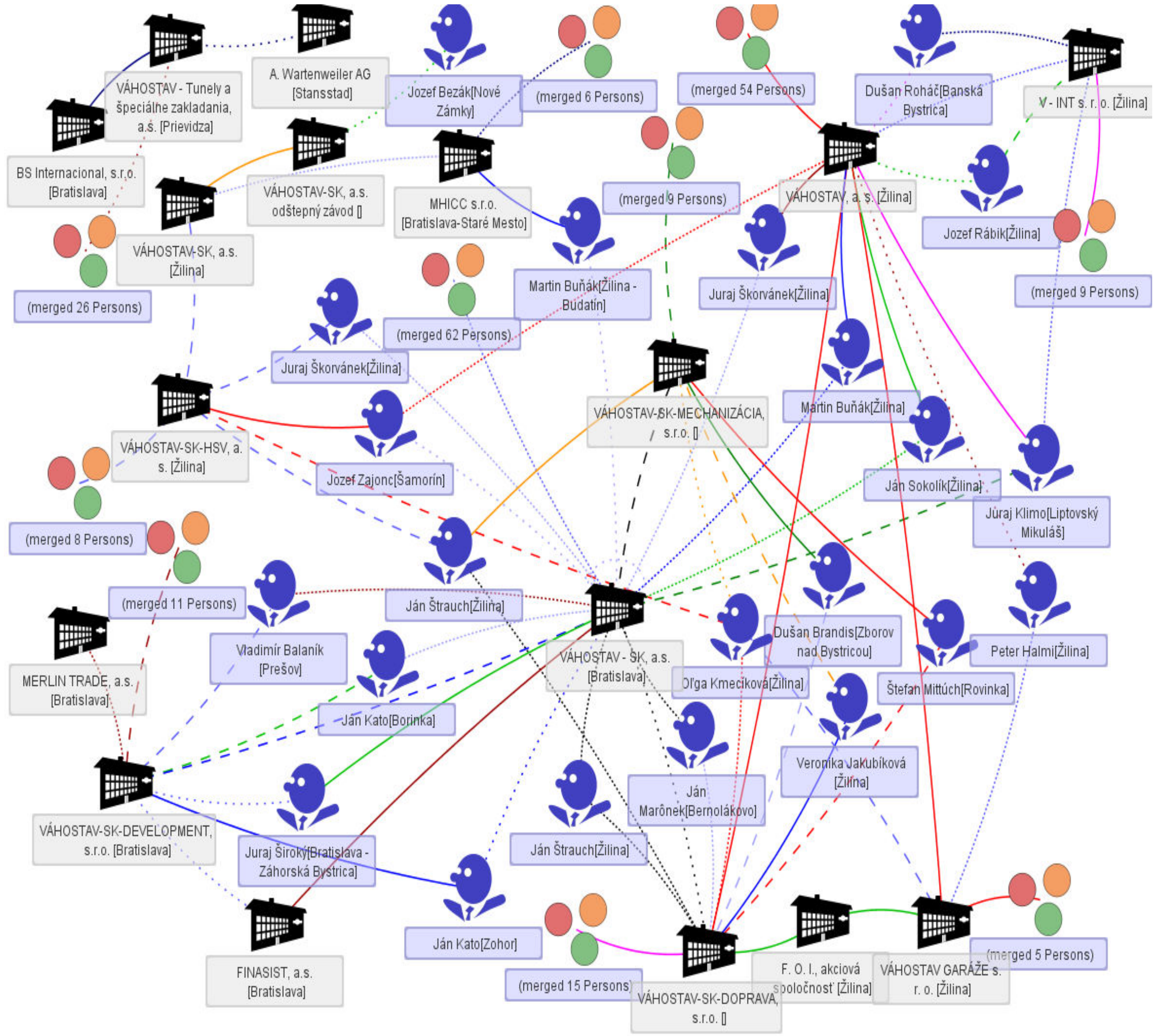

VÁHOSTAV - Tunely a špeciálne zakladania, a.s. [Prievidza]
A. Wartenweiler AG [Stansstad]
Jozef Bezák[Nové Zámky]
(merged 6 Persons)
(merged 54 Persons)
Dušan Roháč[Banská Bystrica]
V - INT s. r. o. [Žilina]
BS Internacional, s.r.o. [Bratislava]
VÁHOSTAV-SK, a.s. odštepný závod []
MHICC s.r.o. [Bratislava-Staré Mesto]
(merged 9 Persons)
VÁHOSTAV, a. s. [Žilina]
Jozef Rábik[Žilina]
VÁHOSTAV-SK, a.s. [Žilina]
(merged 26 Persons)
Martin Buňák[Žilina - Budatín]
Juraj Škovránek[Žilina]
(merged 9 Persons)
(merged 62 Persons)
Juraj Škovránek[Žilina]
VÁHOSTAV-SK-MECHANIZÁCIA, s.r.o. []
Martin Buňák[Žilina]
VÁHOSTAV-SK-HSV, a. s. [Žilina]
Jozef Zajonc[Šamorín]
Ján Sokolík[Žilina]
Juraj Klimo[Liptovský Mikuláš]
(merged 8 Persons)
(merged 11 Persons)
Ján Štrauch[Žilina]
Dušan Brandis[Zborov nad Bystricou]
MERLIN TRADE, a.s. [Bratislava]
Vladimír Balaník [Prešov]
VÁHOSTAV - SK, a.s. [Bratislava]
Oľga Kmecíková[Žilina]
Štefan Mittúch[Rovinka]
Peter Halmi[Žilina]
Ján Kato[Borinka]
Veronika Jakubíková [Žilina]
Ján Marônek[Bernolákovo]
VÁHOSTAV-SK-DEVELOPMENT, s.r.o. [Bratislava]
Juraj Široký[Bratislava - Záhorská Bystrica]
Ján Štrauch[Žilina]
FINASIST, a.s. [Bratislava]
Ján Kato[Zohor]
(merged 15 Persons)
VÁHOSTAV-SK-DOPRAVA, s.r.o. []
F. O. I., akciová spoločnosť [Žilina]
VÁHOSTAV GARÁŽE s. r. o. [Žilina]
(merged 5 Persons)

Figure 4. After merging, overview is more transparent. Tens of persons were merged together into clusters in order to clarify the visualization. Firms and persons are recognized based on their icons.

Comparing to Suchal (Suchal & Vojtek, 2009), we use city name as additional identifier even during visualization. We do not share his suggestion “If two persons with the same name, but different address occur inside one firm, consider that as one person”. There is a record of “Váhostav-SK, a.s.” with 2 persons, namely both with the same name “Juraj Široký” but with different addresses. Although the street is the same, numbering is different (8137/137 vs. 137). This is rather a complex problem and, as it, is very suitable for the field of machine-learning. Despite Suchal, we actually recognize persons based on their addresses and also (academic) titles. It is uncommon, that 2 different persons, named equally, live in the same street and city, if it is father and son. Yet they may have 2 different titles. Regarding different addresses, we encourage to use firm’s actual record as a complement, in order to repair typed errors and address changes, because changes and error fixes are performed by the competent staff of SBR. We do not dare to mark 2 person items as one and the same person. It is up to SBR, that they keep their person items updated every time a person gets re-housed. They however, use personal numbers and ID card numbers and other data, not available to public, which they may rely on. Because that, we have to tolerate that inaccuracies in personal items are unavoidable. However, we prefer rather to not recognize a relationship, as if the relation should be falsely marked, due to 2 to 1 person conversion.

Although we use clustering with person merging, this operation is completely reversible and does not affect connections at all. Personal information is keep intact.

7. Conclusion

An offer of semantic data market is poor and non-governmental or commerce sector does not improve it either. It is an unfortunate situation, because the potential is big and the hole is noticeable. Whether we look in Czech Republic, we find governmental SPARQL endpoints and datasets[32] or non-governmental project for business register browsing with obchodni-rejstrik.podnikani.cz. And availability of open data is vital on behalf of public control.

Data.gov is only partially implemented and, as the report says, only 8 out of 22 submitted tasks were finished. Government-independent initiatives and projects are, also, present in Slovakia, but their effect is marginal and they are often one-time events. Other projects we find slowly declining (including foaf.sk) and unavailability of the data is still very intensive. It was a great step forward to publish debtors, thanks to SIA and VsZP. Even in the case the data is provided online, like Financial department of Slovakia or commercial health insurance agency UNION do, only PDF formatted file with tables is available.

Basically, we divide open data publishing in Slovakia into 2 categories. First is category of clear and open data, machine readable. This category is very important and the data highly usable. But it is the smallest out of 2 groups. Second, although being published, the data itself is in inappropriate format (like PDF). Yet so small market of open data in Slovakia is divided and fragmented once more. As if it was not enough already, that data.gov.sk does not cover all of the few available machine readable sources.

By proposal of this paper, we try to fill up the gap of semantic tools availability, currently present in Slovakia. Because data sources are not always machine readable, we advance further the concept similar to Suchal and Vojtek (2009) and use regular expressions. But instead of online based service, we propose client based solution. The advantages of availability are preserved intact, but no additional costs, with funding online service, are present. Additionally, when application is about to be updated, it is up on a user, whether he wishes to do so.

The modular concept is counting with the possibility of new modules creation and inclusion. Proved already functioning and useable, we would like to recommend its usage.

Currently, there are negotiations in process, regarding offline dataset of SBR. We consult with support member of SBR and SOAP[33] query is under construction. Should we have succeeded, we publish and detail about how to obtain offline dataset of SBR. Unfortunately, prior to publishing of this work, we do not have any more information. We are almost sure, that it is possible to successfully connect to their SOAP server in order to get dataset.

Acknowledgement

This work is partially supported by following grants: CLAN (APVV-0809-11) and VEGA (2/0154/16).

[32] Dataset of Czech Administration of Social Insurance. (2014). Retrieved from https://portal.gov.cz/app/RejData/rec.jsp?id=1646070&id_rej=97898&y=2015&m=11&doctype=idx

[33] XML Soap. (2015). Retrieved from http://www.w3schools.com/xml/xml_soap.asp

References

Becker, Ch. & Bizer, Ch. (2008). DBpedia Mobile: A Location-Enabled Linked Data Browser. LDOW, 369, 2008

Berners-Lee, T., Chen, Y., Chilton, L., Connolly, D., Dhanaraj, R., Hollenbach et al. (2006). Tabulator: Exploring and analyzing linked data on the semantic web. In Proceedings of the 3rd International Semantic Web User Interaction Workshop, volume 2006. Athens, Georgia, 2006.

Dokulil, J. & Katreniaková, J. (2008). Navigation in RDF data. In Information Visualization, 2008. IV'08. 12th International Conference (pp. 26-31). Institute of Electrical and Electronics Engineers.

Furnas, G. W. (1986). Generalized fisheye views (Vol. 17, No. 4, pp. 16-23). ACM.

Herman, I., Melanon, G., & Marshall, M. S. (2000). Graph visualization and navigation in information visualization: A survey. Visualization and Computer Graphics, Institute of Electrical and Electronics Engineers Transactions on, 6(1), 24-43.

Huynh, D., Karger, D., & Miller, R. (2007). Exhibit: lightweight structured data publishing. In Proceedings of the 16th international conference on World Wide Web, pages 737–746. ACM, 2007.

Kurian M. Independent Reporting Mechanism SLOVAKIA: Progress Report 2012-13. (2013). Retrieved from http://www.opengovpartnership.org/files/slovakia-ogp-irm-public-comment-engpdf-0/download

Laclavík, M., Dlugolinský, Š., & Ciglan, M. (2014). Discovering relations by entity search in lightweight semantic text graphs. Computing and Informatics, 33:877–906, 2014.

Laclavík, M., Dlugolinský, Š., Kvassay, M., & Hluchý, (2011). L. Email social network extraction and search. In Proceedings of the 2011 IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology-Volume 03, pages 373–376. Institute of Electrical and Electronics Engineers Computer Society, 2011.

Laclavík, M., Dlugolinský, Š., Kvassay, M., & Hluchý, L. (2010). Use of E-mail Social Networks for Enterprise Benefit. In Web Intelligence/IAT Workshops, pages 67–70. Citeseer, 2010.

Laclavík, M., Dlugolinský, Š., Šeleng, M., Ciglan, M., & Hluchý, L. (2012). Emails as graph: relation discovery in email archive. In Proceedings of the 21st international conference companion on World Wide Web, pages 841–846. ACM, 2012.

Laclavík, M., Šeleng, M., Ciglan, M., & Hluchý, L. (2012). Ontea: Platform for pattern based automated semantic annotation. Computing and Informatics, 28(4):555–579, 2012.

Laclavík, M., Šeleng, M., Ciglan, M., Dlugolinský, Š., & Hluchý, L. (2011). gSemSearch: Objavovanie relácií v kolekciách textových a grafových dát. In 6th Workshop on Intelligent and Knowledge Oriented Technologies: WIKT, pages 1–5, 2011.

Mojzis, J. & Laclavik, M. (2014). Graph clutter filtering based on connectivity distance and visibility. In Science and Information Conference (SAI), 2014 (pp. 153-158). Institute of Electrical and Electronics Engineers.

Omote, H. & Sugiyama, K. (2006). Method for drawing intersecting clustered graphs and its application to web ontology language. In Proceedings of the 2006 Asia-Pacific Symposium on Information Visualisation-Volume 60 (pp. 89-92). Australian Computer Society, Inc.

Smith, D., Owens, A., Russell, A., Harris, C., Wilson, M. et al. (2005). The evolving mSpace platform: leveraging the Semantic Web on the Trail of the Memex. In Proceedings of the sixteenth ACM conference on Hypertext and hypermedia, pages 174–183. ACM, 2005.

Suchal, J. & Vojtek, P. (2009). Navigácia v sociálnej sieti obchodného registra SR. DATAKON, Srní, Czech Republic.