Towards Semantic Mashup Tools for Big Data Analysis

Azzakiy Hendrik, Amin Anjomshoaa, A. Min Tjoa

Towards Semantic Mashup Tools for Big Data Analysis

Hendrik[1], Amin Anjomshoaa[2], and A Min Tjoa[2]

[1] Department of Informatics, Faculty of Industrial Technology, Islamic University of Indonesia
hendrik@uii.ac.id
[2] Institute of Software Technology and Interactive Systems, Vienna University of Technology, Austria
{anjomshoaa,amin}@ifs.tuwien.ac.at

Abstract. Big Data is generally characterized by three V's: *volume*, *velocity*, and *variety*. For the Semantic Web community, the *variety* dimension could be the most appropriate and interesting aspect to contribute in. Since the real-world use of Big Data is for data analytics purposes of knowledge workers in different domains, we can consider mashup approach as an effective tool to create user-generated solution based on available private/public resources. This paper gives brief overview and comparison of some semantic mashup tools which can be employed to mash up various data sources in heterogenous data format.

Keywords: Mashup, Linked Data, Big Data, Semantic Web

1 Introduction

According to [16], Big Data is a common concept to define datasets whose size exceeds the processing capacity of traditional database systems. While this is not the commonly agreed definition, Big Data is generally characterized by three V's: *volume*, *velocity*, and *variety*[16,9,5,7]. ***Volume*** dimension relates to the size of data from one or more data resources in tera-, peta-, or exabytes. The ***velocity*** dimension focuses on the data streams and how to store near real-time data as well as handling the increasing rate of the data amount. The latter, ***variety*** dimension associates with the heterogeneity of data both at the schema-level and the instance-level.

For the Semantic Web community, the *variety* dimension could be the most appropriate and interesting aspect to contribute in. Here, Linked Data can be considered as an alternative solution for addressing the issues of *variety* dimension. Since its introduction by Tim Berners-Lee in 2006, Linked Data (LD) has emerged as a recent trend in the current era of the Web. Linked Data refers to an approach to publish and interlink structured data on the web using some principles forming a global database, called Web of Data [4]. The effort for adopting LD in real life was initiated by Linking Open Data Project[3]. Starting from 2007,

[3] http://www.w3.org/wiki/SweoIG/TaskForces/CommunityProjects/LinkingOpenData

a significant number of datasets are published and interlinked into Linked Open Data Cloud (LOD Cloud) by both individuals and private/public organisations. The LOD Cloud covers various domains such as government, life science, entertainment, etc. Moreover, today not less than 928 active datasets, comprise around 62 billion RDF triples are available as Linked Data[4].

The typical target group of Big Data solutions is knowledge workers in different domains who are not familiar with technical details of Big Data and data integration. As a result, there is a growing need to provide a solution with less learning curve for such users. We can consider mashup as an effective tool to support users in creating user-generated solutions based on available private/public resources and integrate several data sources with different formats easily. As a result both skilled programmers and non-skilled users are able to benefit from the large amount of data and solve their problems.

The aim of this paper is to compare the existing semantic mashup tools to help the readers in choosing the suitable tool to mash up data from various sources for data analytics purposes. Using such tools can help the end users to design a prototype or even visualize the data analysis results for their use cases.

The remainder of this paper is organized as follows. In section 2 the definitions and detailed description of semantic mashups will be introduced. Then, we give the overview of each existing semantic mashup tool in section 3. Finally, we conclude this study in section 4.

2 Traditional Mashups versus Semantic Mashups

Mashup approach allows users to build ad-hoc applications by combining several different data sources and services from across the web. The foundation of the approach consists of sharing, reusing, and combining applications, code, components and APIs which is not an innovation in computer science area. However, the approach is innovative by the fact that this approach is widely used to speed up the process of realizing creative ideas.

There are three approaches for development of mashup solutions [2]. First, **manual** approach, which requires programming or scripting skills of users to integrate the data sources, generate visualizations, and create new functionalities. Second, **semi-automatic**, which assists the users to build a mashup application using provided tools. Third, **automatic** approach which allows creation of mashups without user's involvement, as the resources (data, visualization, as well as functionality) are chosen and invoked automatically by the tool.

The semi-automatic approach is further categorized as follows:

1. spreadsheet-based tools, in which the users provide the data directly into a spreadsheet; The examples of this category are AMICO:CALC [5] and MashSheet[6].

[4] `http://stats.lod2.eu`
[5] `http://amico.sourceforge.net/amico-calc.html`

2. widget-oriented tools, allow users to create the mashup through a visual editor. Yahoo Pipes[6] and Intel Mash Maker[7] are examples of this category of mashups.
3. demonstration-based tools, allow users to mash up their data by providing examples and completing the data integration task via a visual step-by-step process. The instances of this category are Dapper[8] and Karma[18].

Semantic mashups can be seen as a complementary extension of the traditional mashup approaches. The Semantic Web and mashups can provide a solid basis for many interesting applications and boost each other. The Semantic Web and ontologies may facilitate the creation of mashup solutions for novice users. This application of Semantic Web has its roots in Semantic Web Service concept that is aiming to automate service discovery and composition without human intervention. The basic difference between Semantic Web Services and Semantic Mashups approaches is derived from different target users. The Semantic Web Services are mainly managed and used by IT experts who are aware of underlying data structures and corresponding services; however, the Semantic Mashups target group is novice users who need to combine the Mashup Widgets for their specific purposes [1,8]. Semantic web can be used to annotate combination of several APIs especially to automate the selection and composition of these APIs [14]. It also benefits the data integration process by adding semantic to the data using ontologies and semantic web languages (RDF), which enable the machine to automate data exchange. Combining mashups with Linked Data opens a lot of possibilities for data integration and more efficient use of distributed data.

3 The Existing Semantic Mashup Tools

As defined by [10], ideally there are three requirements that should be fulfilled by mashup tools: first it should be *generic* and able to address various application domains, secondly it should be *powerful* to manage complex logics of the problems, and thirdly it should be *simple* to be used by novice users. Based on these essentials, we classify the existing mashup tools into two different groups: data analytics tools and generic tools. The data analytics tools are targeting the analysis of large amount of specific data and derive required information. While the generic tools can be used for different kind of purposes. In the rest of this section, some examples of these two groups will be presented.

3.1 Data analytics tools

As mentioned before, this group of tools is targeting the analysis of large amount of data. The data is usually taken from a specific domain and using the mashup tools this data will be processed to derive the required results.

[6] `http://pipes.yahoo.com/pipes/`
[7] `http://software.intel.com/en-us/articles/intel-mash-maker-mashups-for-the-masses`
[8] `http://open.dapper.net`

3.1.1 Black Swan Events [9]

This mashup tool was inspired by the ***Black Swan Theory*** which represents a black swan as an unpredictable event that has massive effects [17]. The aim of this tool is to help domain experts to find important (*black swan*) events based on historical or statistical data [13]. The tool integrates statistical data and events data into a single repository, which come in various data formats including structured (e.g., CSV, RDF, or XML) and unstructured (e.g., plain text) data. Currently, it manages more than 400 statistical time series datasets which cover annual data of 200 countries for the past 200 years. The statistical data is gathered from international organisations such as World Bank or International Monetary Fund (IMF), as well as some projects such as Gapminder[10] and Correlates of War[11]. Furthermore the events data are collected from available sources such as DBpedia, Freebase, and BBC historical timeline.

The collected data is then analyzed using several methods such as regression techniques and rule mining to discover the interesting events and patterns in statistical data. In order to help users to investigate the correlation between an event and statistical data easily, the tool comes with interactive visualisation feature which depicts an 'annotated time line' using a graph chart. There are two methods to explore the data by using the visualization, i.e statistic-based method and rule-based method. The first method enables users to select any statistic indicators such as Economy, Health, Energy, etc., for a target country. This will help the users to find the events that match to country's statistical outliers and the rules that generate event-outlier pairs. For instance, Fig. 1 shows the Black Swan visualization for the effect of the German reunification and income growth in Germany. The second method aims to be used by the advanced users. It allows the users to select a rule and find the matching pairs of statistical outliers and historical events[12].

3.1.2 Super Stream Collider [13]

Super Stream Collider (SSC)[15] is a web-based mashup tool to aggregate live stream data (e.g., sensor stream data from Linked Sensor Middleware[14] for data streams such as weather, traffic, flight, etc., and social stream data such as twitter streams) and Linked Data resources such as DBpedia and Sindice. While the input data comes in various data formats, the output is only available as RDF.

This tool provides an easy to use interface for either novice users, who do not have any technical knowledge about Semantic Web, as well as advanced users, who have knowledge of Semantic Web standards and technologies. By using widget-based paradigm, the users can drag-n-drop any data source widget into

[9] `http://blackswanevents.org`
[10] `http://www.gapminder.org/`
[11] `http://www.correlatesofwar.org/`
[12] `http://blackswanevents.org/?page_id=179`
[13] `http://http://superstreamcollider.org/`
[14] `http://lsm.deri.ie`

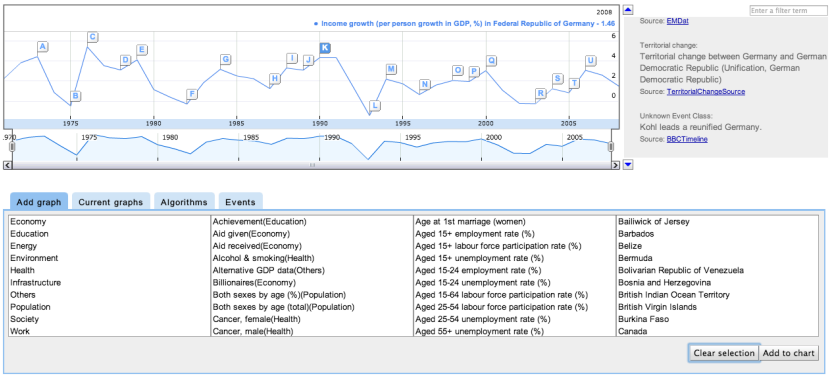

Fig. 1. The blackswan visualization to show the effect of the German reunification and income growth in Germany

the visual editor and connect them via several operator widgets such as merge, location and timer widgets. For the advanced users, SSC provides additional operators such as SPARQL/CQELS[15] editor. These can be used either by writing the query directly in the query text area or by using the visual editor which helps the users to learn writing SPARQL/CQELS interactively. SSC's user interface is depicted in Fig. 2.

The mashup process can be monitored by inspecting the flow of data from the sources to the final output via SSC's debugging component. Using this tool, the users may receive the result data as raw data, RDF data or even data visualisation in several types of charts. The final output, then can be queried, visualised, and published using supported stream protocols (i.e., PubSubHubbub, XMPP and WebSockets). For example, using WebSockets, the HTML and Android developers can embed the output widget in their application without extra efforts and knowledge about RDF or SPARQL query.

3.2 Generic tools

Unlike the data analytics tools which are focusing on specific domain and data processing pattern, the generic tools are equipped with generic functions which can be used for different kind of solutions.

3.2.1 DERI Pipes [16]

DERI Pipes which is also known as Semantic Web Pipes (SWP) is a Semantic Web tool which was inspired by Yahoo pipes. While the Yahoo pipes is mainly aimed to work with RSS feeds as data source, DERI pipes focuses on graph based data model [12], i.e, RDF data.

[15] `https://code.google.com/p/cqels/`
[16] `http://pipes.deri.org/`

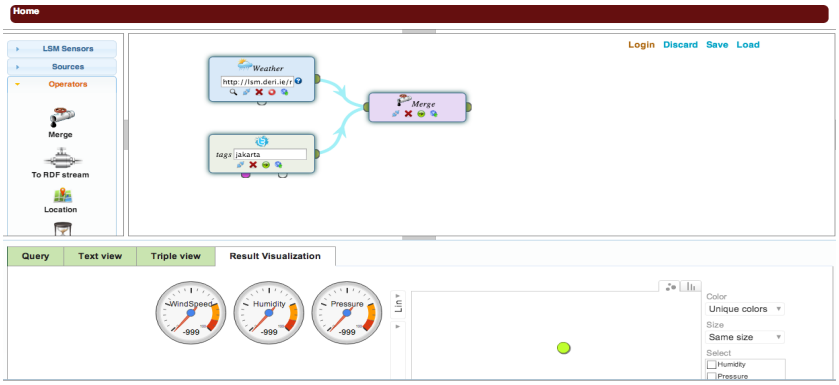

Fig. 2. The Super Stream Collider User Interface

The tool has an easy to use web based GUI called pipe editor to compose workflow of connected operators which form the data pipes. These operators are visualised as widgets and have input and output ports (Fig. 3). There are two kinds of supported operators, namely general operators and base operators. The general base operators only support merge and split operations. The latter operators comprise of getRDF and getXML operator for fetching data from web URL and converting it to RDF or XML formats, XSLT operator to execute XSL transformation of XML input, RDFS and OWL operator to materialize RDFS or OWL inference rule for a specific input, and SPARQL and CONSTRUCT operator to query and align RDF data.

By adopting the concept of UNIX pipeline, the output of the pipeline workflow can be fed directly as an input for other pipeline workflows. The pipeline output is in RDF or JSON format and for visualizing the results, SWP uses SIMILE exhibit[17]. It also provides RSS feeds output thus can be used as an input for other mashup tools such as Yahoo Pipes which only accept RSS as input. Finally, the users can store and publish their pipes to be reused by the other users.

3.2.2 MashQL [18]

This tool enables users to exploit the benefits of Web of Data without prior knowledge of semantic web technologies such as RDF and SPARQL. By using the query-by-diagram paradigm, it allows users to query and mash up a massive amount of structured data on the web intuitively [11].

The core element of MashQL system is a visual editor that processes the input data and generate the required output. Here, the users merely choose the attributes of input concepts that should appear in the widgets output. It also enables the users to filter data with some arithmetic and relational operators for string and numeric attributes. The widget output can be then piped as a

[17] `http://www.simile-widgets.org/exhibit/`
[18] `http://sina.birzeit.edu/mashql/`

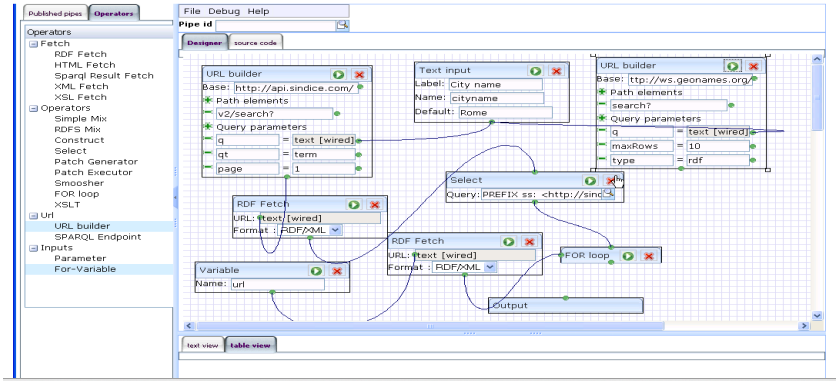

Fig. 3. DERI Pipes User Interface

new input for other MashQL widgets and mashed up with other data inputs as shown in Fig. 4. The system then translates this process into a SPARQL query which is transparent to the users.

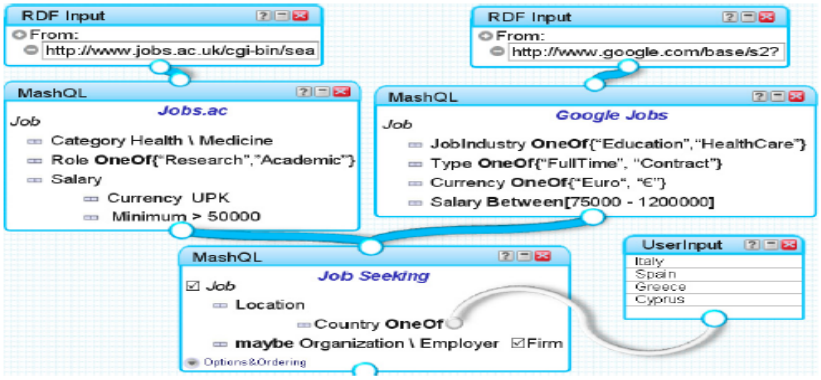

Fig. 4. MashQL User Interface

3.2.3 Information Workbench [19]

Among the other semantic web tools explained before, Information Workbench (IWB) has more comprehensive features as a mashup tool. It supports collaborative knowledge management among the end users and integrates both structured and unstructured data coming from internal or external resources. The IWB provides a framework to develop, maintain and deploy applications and supports Big Data analysis scenarios by providing a comprehensive SDK

[19] `http://www.fluidops.com/information-workbench/`

(Solution Development Kit). Furthermore, the IWB also provides solutions for business intelligence and data analytics in an integrated environment.

Its *data provider* component collects, integrates, and maintains data from several data resources into a central triple data repository based on datawarehousing techniques. Alternatively, it provides a federation layer called FedX, to virtually integrates local and public Linked Data sources. The benefit of the latter is to provide capability of on-demand access to up-to-date data [3].

The users can develop an application by composing available common purpose widgets such as visualisation and exploration widgets, social media widgets, authoring and content creation widgets, and analytics and reporting widgets. It is also possible for the advanced users to create customized widgets for their special purposes. The user interface of Information Workbench can be seen in Fig. 5.

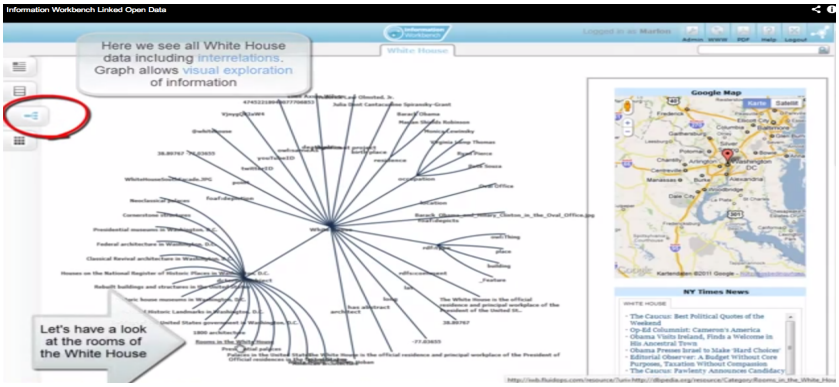

Fig. 5. Information Workbench User Interface

4 Conclusion

This paper provides a survey of existing semantic mashup tools. Using such tools the knowledge workers will be able to create ad-hoc data integration solutions based on the available structured and unstructured data resources such as Linked Open Data (LOD), Open Government Data (OGD), and private datasets. The results of this survey is provided in Table 1 which demonstrate the features of these Semantic Mashup tools including supported input data formats, data source registration, mashup approach, and visualization form.

The data gathering, processing, and integration tasks in Big Data domain are the main challenging issues for the knowledge workers and solution providers. Lowering such entrance barriers is, therefore, essential for the evolution and development of Big Data solutions. In our belief the mashup solutions have the potential to address these requirements and to empower the solution providers

Table 1: Comparison of Existing Semantic Mashup Tools

	Input Data Format	Data Source Registration	Mashup Approach	Visualization Form
BlackSwan Events	CSV, RDF, XML, Linked Data, Plain Text	manual	manual	line chart
Super Stream Collider	Stream Data,Linked Data, Social Stream Data, RDF	input widgets	semi automatic (Widget-based)	various chart
Semantic Web Pipe	RDF, XML, HTML, Linked Data	input widgets	semi automatic (Widget-based)	faceted browser
MashQL	RDF, XML, HTML	input widgets	semi automatic (Widget-based)	table data
Information Workbench	Linked Data, RDF, HTML, CSV, XML, Relational	data provider, federation layer	semi automatic (Widget-based)	various chart

and novice users to create and adapt individual Big Data applications based on elaborated and domain-specific widgets in a user-friendly environment without worrying about technical challenges of data integration.

Acknowledgments. This research has received support from ***Ernst-Mach-Stipendien granted by the OeAD*** - Austrian Agency for International Cooperation in Education & Research, financed by BMWF.

References

1. Anjomshoaa, A., Tjoa, A.M., Hubmer, A.: Combining and integrating advanced it-concepts with semantic web technology mashups architecture case study. In: Intelligent Information and Database Systems, pp. 13–22. Springer (2010)
2. Fischer, T., Bakalov, F., Nauerz, A.: An overview of current approaches to mashup generation. In: Proceedings of the International Workshop on Knowledge Services and Mashups (2009)
3. Haase, P., Schmidt, M., Schwarte, A.: The information workbench as a self-service platform for linked data applications. In: COLD (2011)
4. Heath, T., Bizer, C.: Linked Data: Evolving the Web into a Global Data Space. Morgan & Claypool Publishers (2011)

5. Hendler, J.: Broad Data: Exploring the Emerging Web of Data. Big Data 1(1), 18–20 (Mar 2013), http://online.liebertpub.com/doi/abs/10.1089/big.2013.1506
6. Hoang, D., Paik, H.Y., Ngu, A.: Spreadsheet as a generic purpose mashup development environment. In: Service-Oriented Computing, Lecture Notes in Computer Science, vol. 6470, pp. 273–287. Springer Berlin Heidelberg (2010), http://dx.doi.org/10.1007/978-3-642-17358-5_19
7. Hopkins, B., Evelson, B., Hopkins, B., Evelson, B., Leaver, S., Moore, C., Cullen, A., Gilpin, M., Cahill, M.: Expand Your Digital Horizon With Big Data. Tech. rep. (2011)
8. Hoyer, V., Stanoevska-Slabeva, K.: The changing role of it departments in enterprise mashup environments. In: Service-Oriented Computing–ICSOC 2008 Workshops. pp. 148–154. Springer (2009)
9. IBM: Analytics : The real-world use of big data. Tech. rep. (2012)
10. Imran, M., Kling, F., Soi, S., Daniel, F., Casati, F., Marchese, M.: ResEval Mash : A Mashup Tool for Advanced Research Evaluation. In: World Wide Web Conference. pp. 361–364 (2012)
11. Jarrar, M., Dikaiakos, M.D.: Mashql: A query-by-diagram topping sparql. In: Proceedings of the 2Nd International Workshop on Ontologies and Information Systems for the Semantic Web. pp. 89–96. ONISW '08, ACM, New York, NY, USA (2008), http://doi.acm.org/10.1145/1458484.1458499
12. Le-Phuoc, D., Polleres, A., Hauswirth, M., Tummarello, G., Morbidoni, C.: Rapid prototyping of semantic mash-ups through semantic web pipes. In: the 18th international conference on World wide web - WWW '09. p. 581. ACM Press, New York, New York, USA (2009), http://portal.acm.org/citation.cfm?doid=1526709.1526788
13. Lorey, J., Mascher, A., Naumann, F., Retzlaff, P., Forchhammer, B., Zamanifarahani, A.: Black Swan : Augmenting Statistics with Event Data. In: 20th ACM Conference on Information and Knowledge Management (2011)
14. Malki, A., Benslimane, S.M.: Building semantic mashup. In: ICWIT. pp. 40–49 (2012)
15. Nguyen, H., Quoc, M., Serrano, M., Le-phuoc, D., Hauswirth, M.: Super Stream Collider Linked Stream Mashups for Everyone. In: Proceedings of the Semantic Web Challenge co-located with ISWC2012. vol. 1380 (2012)
16. Oracle: Information Management and Big Data A Reference Architecture. Tech. Rep. February (2013), http://www.oracle.com/technetwork/topics/entarch/articles/info-mgmt-big-data-ref-arch-1902853.pdf
17. Taleb, N.N.: The Black Swan:: The Impact of the Highly Improbable Fragility. Random House LLC (2010)
18. Tuchinda, R., Szekely, P., Knoblock, C.A.: Building mashups by example. In: Proceedings of the 13th international conference on Intelligent user interfaces. pp. 139–148. ACM (2008)