Using the Heterogeneous Database and Linked Data Technologies with Case Study of Thai Local Government Planning Database

Lerluck Boonlamp

***Abstract*—These paper discuss about the comparison between two technologies for dealing with enormous quantities of data from relational database. The two technologies that were compared are heterogeneous database and linked data technologies by analyzing the government information. The experiments were done based on government project planning data. The results expressed that the advantages of linked data are to retrieve and relate data while consuming them. It is better to handling with text data than heterogeneous technology. However, visualizing and analyzing data heterogeneous database can do better than linked data where it is a challenge feature to be developed for linked data technology.**

***Index Terms*—Linked data, heterogeneous database, semantic search.**

I. Introduction

In this paper, we measure the ability in finding and comparing information over cross-database between two different technologies which are heterogeneous relational database and linked database. We use semantic search in the RDF model and keyword search in the relational model. These comparisons address the following questions: 1) what can semantic search achieve that keyword search cannot? 2) How is it difficult to find data needed with cross heterogeneous databases and cross linked database?

A critical requirement for the evolution of the current web of documents into the web of data is the vast quantities of data stored in relational databases (RDB)[1]. Local government of Thailand has many data stored in RDB. These data are not often to be used as important asset. While many countries have already realized the importance of linked data and publish their data into linked databases.

Linked data from a RDB usually has a simple browsing endpoint. When users click the URI of a resource, most browsers display all its relationships. Some popular resources may have hundreds of related resources; consequently users cannot easily find interesting resources if they are in a large number of relationships.

Data warehouses (DWs) are data repositories that integrate data from different sources, and keep historical data for analysis and decision support [2]. The usage of a data warehouse has evolved from reporting and decision support system to decision making operational systems [3]. Benefits of data warehouse such as time-saving for users, improved quantity and quality of information, informed decision-making, improvement of business processes and ultimately support for the accomplishment of strategic business objectives [4]. Data warehouse is increasingly used in health care to provide the tools for decision making and individualizing disease management [5].

In federated database systems, there are approaches and architectures which require no combination of multiple databases and system into one [6][1]. A federated database or virtual database, there is no actual data integration in the different databases as a result of data federation. Federated database systems can provide a user interface, enabling users to store and retrieve data in multiple databases with a single query even if components of databases are heterogeneous. In a federation of heterogeneous databases [7], there is the need for data sharing among the diverse databases, and for resource consolidation of all supporting software, hardware and personal, although each database has its own autonomy in terms of, for example, its integrity constraint, application specificity, and security requirements.

II. Searching Conceptual in the Linked Data and Relational Database

A. Linked Data Conceptual

The Semantic Web is not just putting data on the web. It is about making links, so that a person or machine can explore the web of data. With linked data, when you have some of it, you can find others, related data. The term "Linked Data" refers to a set of best practices for publishing and connecting structured data on the web [8]. Berners-Lee [9] outlined four basic rules for publishing Linked Open Data (LOD) on the web.

- Use URIs as names for things
- Use HTTP URIs so that people can look up those names.
- When someone looks up a URI, provide useful information, using the standards (RDF*, SPARQL)
- Include links to other URIs, so that they can discover more things.

Example as following four principles is presented in Fig. 1.

The four principles of the linked data at Fig. 1 have

Manuscript received August 25, 2014; revised October 15, 2014.

Lerluck Boonlamp is with Faculty of Technology and Environment, Prince of Songkla University, Phuket campus, Thailand (e-mail: lerluck.kuerklung@gmail.com)

[1]http://dig.csail.mit.edu/2009/AFOSR/papers/Federation%20Architecture.pdf

DOI: 10.7763/IJMO.2015.V5.428

presented the relationship between book and publisher domains. Book and publisher have *identifiers* as URI addresses and the *publishedby* property links presents the *relationship* between two entities. Following URIs, will see more entities and related. Vocabulary is used as property linkage between entities resources.

Fig. 1. Four principles of linked data represents by RDF graph.

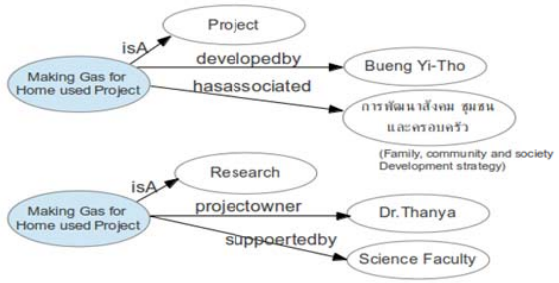


Fig. 2. Two different entities with the same name "produce gas for home use" reprinted in RDF graph. Each entity has a different set of properties and property values.

These four principles provide a framework for publishing data on the Web. The principles share a common data format based on URIs and RDF, as well as using SPARQL as a common language for data manipulation. In addition, the linked data [10] is required to identify an entity via a single HTTP schema based URI. The identified entity is represented by URI. According to the third principle, this data is represented using RDF and provides useful information when users access a URI. RDF is a generic, graph based data model that represents information based on triples. The *subject* and *predicate* in a statement must always be resources; the *object* can either be a resource or a literal. Querying RDF data uses SPARQL. SPARQL is the standard query language for accessing RDF data.

Furthermore, the forth linked data principle is to set RDF links into other data sources on the Web and enable applications to discover additional data sources. It is able to interlink between one data sources to others with *relationship* that points at related things in other data source. The *identify* point at URI aliases which enables the client to retrieve further description about an entity from other data sources.Moreover, *vocabulary* links point from data to the definition of vocabulary term that are used to represent the data and to the definition of related terms in other vocabularies[1].

Therefore, our domain will comply with the linked data principles. With the principles, it enables us to discover and retrieve more useful information of local government planning such as strategy, projects and organization. We can find more data which is related to our particular interests. These innovations for distributing, interlinking of data will encourage interoperability amongst organizations in local government. Collaboration plans amongst local governments are essential for sustainable development and economic growth in the regions, especially with a developing country like Thailand. It will require cooperation from several parties such as organizations, stakeholders and citizens. Therefore, evidence of information is extremely important as a data decision support system (DDSS). In addition, we can distribute our local government planning dataset across multiple data sources to the Linked Open Data (LOD) cloud.

B. Semantic Search in Linked Data

Consider if we want to know more about "การผลิกก๊าซหุงต้มในครัวเรือน" (Produce gas for home used). This entity of type *Project* in the Bueng-Yi-Tho*Organization,*which have many projectsin the local government dataset and belong to the Family, community and society development *Strategy*. Besides, it could be created by other person or organizations. Fig. 2, demonstrates the two different entities and the information related to "Produce gas for home used" entities, in linked data in form of the RDF data model.

With semantic search, in the RDF data model, user can refine their search, navigate through the initial results and filter out the results, which do not have the properties that they are looking for. In fact, the explicit representation of properties in RDF which does not exist in the relation model will facilitates this refinement of search results. Fig. 2, the user can search for "Produce gas for home used" and the instances, which match the search string will be shown to the user.

C. Keyword Search in Relational Database

Keyword search in relational databases [11] has been widely studied in recent years because it requires users neither to know a certain structured query language nor to know the database schema. Most of the existing keyword search methods assume that the database are static and focus on answering keyword queries. In reality, database is often updated frequently; new records are inserted into the tables. For example, we suppose to know producing gas for home use project by Dr. Thanya. For the keyword query can be "Produce gas by Dr. Thanya" consisting of three keywords "Produce gas" in project name and "Thanya" author. Therefore, finding results that are formed by th tuples containing the keywords is the keyword search in relational databases. Database matched keywords are underlined, and the tuple connections of keywords query matched a circle between tables (project table and owner table) and return query results that relevance with three keywords query. So, the return results can be list of projects that done by Dr. Thanya and the project name that included the word "Produce" and "Gas" in the project name. The results will rank by relevance of three keywords.

Summary, when users want to search data in linked data or in relational database. The users will specify their information need by a set of keywords. With the linked data searched results, users can retrieve list of relevance information of that keywords. On the other hand, the results

from relational database do not allow users to browse the other available properties and navigates through sets of entities as in semantic search where data is described in RDF model.

III. Finding Information Results Across Linked Data

Finding data across different linked datasets, we able to retrieve data from different organizations if their publish data with linked data technology and describe data in RDF data model as represented at Fig. 3. The entity search is possible when they use the same property name or standard vocabulary.

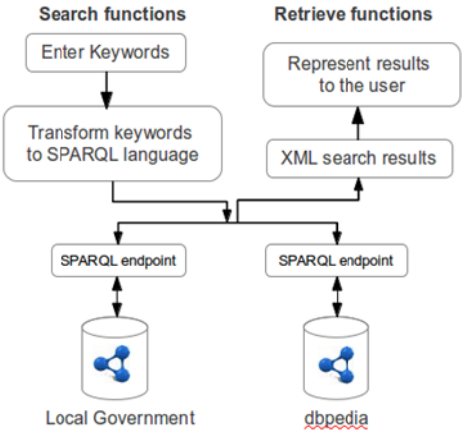


Fig. 3. Finding data different resources framework.

For example, if we would like to find information about "ภูเก็ต" (Phuket province) from existing local government dataset and DBpedia dataset. These two datasets stored in linked data format, therefore to query and retrieve data we need to send SPARQL query language to SPARQL endpoints and the system will return the results to the user. At Fig. 4, it represents entity ภูเก็ตsearch result from two SPARQL endpoints with a single query.

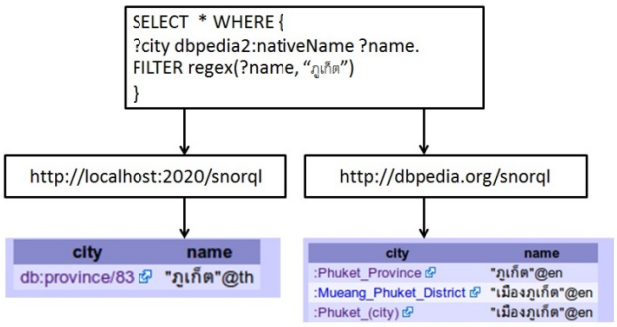


Fig. 4. Searching results from two dataset resources.

User cans browse other information that related with entity results by following URI links. At Fig. 4, user enters only the word "ภูเก็ต"(Phuket) and submits it to the system. The back end, it will transform the query into the sparql language as shown and send the query command to both local government and DBpediasparql endpoints. The system will return the entity that meet the user's entered.

IV. Experiment

In this paper, with the question how it is difficult to find data needs with heterogeneous databases and linked data. Our purpose is learning the linked data technology with existing local government database in planning that the linked data is a proper technology to use for finding and analyzing data or better use other applications. For experiment this, we use two applications name Tableau and D2R server to test our assumption.

A. Thai local Government Planning Database

The original data is store in MySQL database and we interested only in planning. The planning data consist of Thailand provinces, local government organizations, strategies, list of project plan and budgets. Data is updated once time a year for organization planning.

B. Heterogeneous Database Application

Tableau[2] is an application that can connect directly to databases, cubes, data warehouses, files and spreadsheet. It takes only a few clicks with no programming is required. In a minute, we are able to accessing data, visualizing results and business analytics and easily to combine data from different sources. The most important connecting between different data sources is an *attribute name*. The different data sources are connected if they have the same attribute name.

C. Linked Data Application

The D2R server [12], it is an application to publish relational database to RDF. Database schema is the most essential part which needs to be clarified before we start generating a mapping file. After that, ontology and data models are designed in order to state semantic vocabulary. The D2R server uses D2RQ mapping language to capture mapping between application specific databasesschemas. The D2RQ mapping specifies how resources are identified and how property values are generated from database content. In addition, the D2R Server enables HTML browsers to navigate the content and search data by using SPARQL endpoint and SPARQL language.

D. Experiment Results

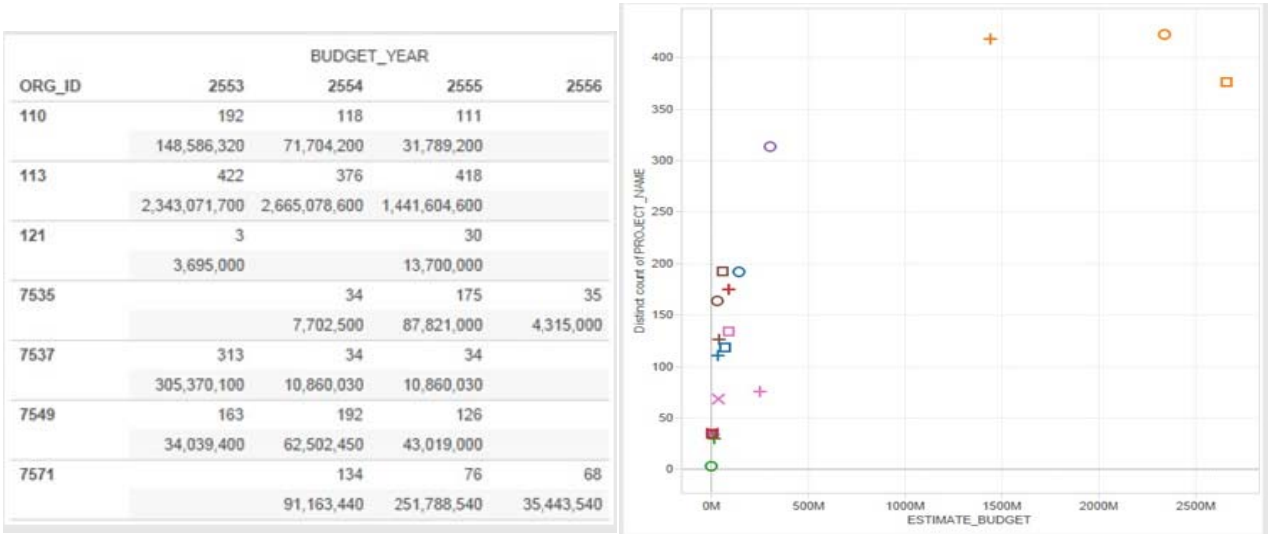
The experiment tests are representing the consuming data in the local database between RDB and RDF.

One: The provincial governor would like to know the total number of projects and budgets of each organization and able to see the details of projects.

This technology, the result of finding is represented at Fig. 5, it is able to answers the provincial governor with the information of total number of projects, total number of budgets but cannot see name of projects and cannot drill down to see project's details. But it is able to sees more details of organization with the organization id linkage. The reason why it represents only numbers not a text format because limited of multi-dimensions. In the other side, the technology support more usefulness feature of visualization data with varies of graph types and suggests the appropriated graph type with existing data. Therefore, easy for the provincial governor to distinguish and categorize

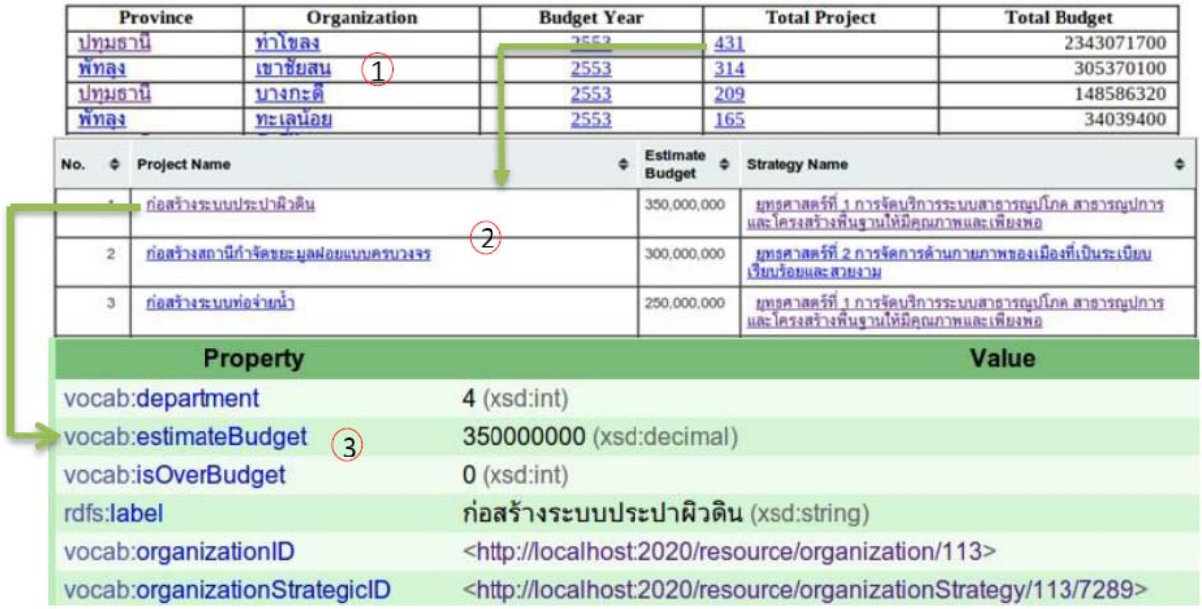
[2]http://www.tableausoftware.com/

group of data.



	BUDGET_YEAR			
ORG_ID	2553	2554	2555	2556
110	192	118	111	
	148,586,320	71,704,200	31,789,200	
113	422	376	418	
	2,343,071,700	2,665,078,600	1,441,604,600	
121	3		30	
	3,695,000		13,700,000	
7535		34	175	35
		7,702,500	87,821,000	4,315,000
7537	313	34	34	
	305,370,100	10,860,030	10,860,030	
7549	163	192	126	
	34,039,400	62,502,450	43,019,000	
7571		134	76	68
		91,163,440	251,788,540	35,443,540

Fig. 5. Case study 1 with tableau technology.



Province	Organization	Budget Year	Total Project	Total Budget
ปทุมธานี	ท่าโขลง	2553	431	2343071700
พัทลุง	เขาชัยสน	2553	314	305370100
ปทุมธานี	บางกะดี	2553	209	148586320
พัทลุง	ทะเลน้อย	2553	165	34039400

No.	Project Name	Estimate Budget	Strategy Name
1	ก่อสร้างระบบประปาผิวดิน	350,000,000	ยุทธศาสตร์ที่ 3 การจัดบริการระบบสาธารณูปโภค สาธารณูปการ และโครงสร้างพื้นฐานให้มีคุณภาพและเพียงพอ
2	ก่อสร้างสถานีกำจัดขยะมูลฝอยแบบครบวงจร	300,000,000	ยุทธศาสตร์ที่ 2 การจัดการด้านกายภาพของเมืองที่เป็นระเบียบเรียบร้อยและสวยงาม
3	ก่อสร้างระบบท่อจ่ายน้ำ	250,000,000	ยุทธศาสตร์ที่ 3 การจัดบริการระบบสาธารณูปโภค สาธารณูปการ และโครงสร้างพื้นฐานให้มีคุณภาพและเพียงพอ

Property	Value
vocab:department	4 (xsd:int)
vocab:estimateBudget	350000000 (xsd:decimal)
vocab:isOverBudget	0 (xsd:int)
rdfs:label	ก่อสร้างระบบประปาผิวดิน (xsd:string)
vocab:organizationID	<http://localhost:2020/resource/organization/113>
vocab:organizationStrategicID	<http://localhost:2020/resource/organizationStrategy/113/7289>

Fig. 6. Case study 1 with linked data.

At the Fig. 6, using linked data technology. It is able to discover all requirements for the provincial governor. The results are different between two technologies, the heterogeneous return the number and cannot query more dimensions. With the linked technology, it represents province name, organization name and can calculates the total number of project in different years. Every entity has URI address except the literal entity. For example if user follows province name link, it will opens the province page which shows all entities that contain in province table as shown in the number 3 and represent data in form of Subject (province name), Predicate and Object or Value (S, P, O). The user can drill up and drill down to find more data which related to his/her particular interests. The user is able to finds more data because data is stored in the RDF graph and used the linked data technology to represent the results.

Two: The provincial governor would like to find irregular budget with the similar project and details.

The RDB testing's results. Filtering data is not difficult to find the similar project with the project name but to linking between data is complicated and has a restriction. The linkage between data or find more data if that data is in the numerical data type or that data is the primary key. In the other hand, the linked data, to filtering and drill down to find more relevant data are easy.

TABLE I: COMPARING BETWEEN HETEROGENEOUS AND LINKED DATA TECHNOLOGIES

	Linked data technology	Heterogeneous technology
Data	• Data in different local database • Database can be in different data models and/or use different names • Data is store in RDF format.	• Data in different local database may be identified as logically representing. • Database can be in different data models and/or use different names • Data can be in different type of formats.
Query management	• Query management, sending single query to the individual database via the SPARQL protocol.	• Distributed query management provides the ability to combine data from different local database in a single retrieval operation.
Retrieval	• Return the results from different database in a single retrieval • Provide useful information, included other URIs links, so that user can discover more things by following it • Result of data is in form of XML, JSON that easy for developer to extend data in visualization graph	• Return the results from different database in a single retrieval • Limited to drill up and down the data, only with numeric type or data which represent as a primary key

E. Compare Technologies between Heterogeneous and Linked Data

Table I is represented that two technologies can access to different databases and different formats but specially with linked data. Linked data will know and see data inform of RDF format. In additional, both can send one query command through databases but the results are difference. Linked data returns data entity with URIs which able to discover more things when user following linkages. In the other hands, with heterogeneous, the user need more query to get a more information.

V. Conclusion

Thai local government database in planning domain, this data is almost in text format. Therefore, linked data is proper technology for text formatting and ease of consuming data. If following the link results in order to obtain further contextual information which increase the visibility of data. The experiment between heterogeneous and linked data can support all requirements but in a different results. The linked data, the user can get more usefulness of information with related data without more queries. Whereas, it is needs more queries with heterogeneous technology. Querying data cross database, we try to propose if different organizations working in the same domain, they should use same data model. Using same data model is supporting ease of consuming data from different datasets. However, linked data technology is not appropriate for monitoring or comparing text information, it is difficult to distinguish irregular data. Therefore, visualization data graph is needed.

With linked data technology, we can find more data which is related to our particular interests. This innovations for distributing, interesting of data will encourage interoperability among organizations in the local government.

References

[1] J. Zhang, C. Ma, C. Zhao, J. Zhang, L. Yi, and X. Mao, "A novel Ranking framework for linked data from relational databases," *Tsinghua Science and Technology*, vol. 15, no. 6, 2010.

[2] M. Caniupan, L. Bravo, and C. A. Hurtado, "Repairing inconsistent dimensions in data warehouses," *Data and Knowledge Engineering;* vol.79-80, pp. 17-39, 2012.

[3] R. Bhashyam, "Technology Challenges in a Data Warehouse," in *Proc. the 2004 VLDB Conference*, pp. 1225-1226, 2004.

[4] H. J. Watson, D. L. Goodhue, and B. H. Wixom, "The benefits of data warehousing: why some organizations realize exceptional payoffs," *Information and Management*, vol. 39, pp. 491–502, 2002.

[5] E. Roelofs, L. Persoon, S. Nijsten, W. Wiessaler, A. Deker, and P. Lambin, "Benefits of a clinical data warehouse with data mining tools to collect data for a radiotherapy trial," *Radiotherapy and Ontology*, vol. 108, pp.174-179, 2013.

[6] D. McLeod and D. Heimbigner, "A federated architecture for database systems," in *Proc. the national computer conference*, pp. 283-366, 2010.

[7] C. Bizer, T. Heath, and T. B. Lee, "Linked data – The story so far," *International Journal on Semantic Web and Information Systems,* vol. 5, no. 3, pp. 1-22, 2007.

[8] T. B. Lee. Linked Data Design Issues. (2006). [Online]. Available: http://www.w3.org/DesignIssues/LinkedData.html

[9] O. Hartig and A. Langeger, "A Database perspective on consuming linked data on the web," *Datenbankspek Trum*, vol. 10, no. 2. pp. 57-66.

[10] Y. Xu, J. Guan, and Y. Ishikawa, "Scalable continual top-k keyword search in relational databases," *Knowledge and Information Systems*, vol. 26, no. 2, 2011.

[11] C. Bizer and A. Seaborns, "D2RQ: treating non-RDF database as virtual RDF graphs," in *Proc. the 3rd International Semantic Web Conference (ISWC2004)*, 2004.

Lerluck Boonlamp is a Ph.D student in the field of information management, School of Engineering and Technology, Asian Institute of Technology, Thailand. She was born on 26 November 1972. She earned a bachelor degree in administration special in Business computer (in 1998) from Prince of Songkla University, Songkhla province, Thailand and she received her master degree in information management (in 2006) from the AsianInstitute of Technology, Pathumthani Province, Thailand.

She has currently worked as an instructor in Faculty of Technology and Environment, Prince of Songkla University, Phuket campus, Thailand since 2008. She was an instructor in Faculty of Hospitality and Tourism from 2006 to 2008. And she worked as a computer technician in Faculty of Hospitality and Tourism from 1998 to 2004. Her research interests are Web technologies, Semantic technologies, Linked Data, Social Web technologies and Data mining.