Linked Data Driven Dynamic Web Services for Providing Multilingual Access to Diverse Japanese Humanities Databases

Biligsaikhan Batjargal
Ritsumeikan University, Japan
biligsaikhan@gmail.com

Takeo Kuyama
Ritsumeikan University, Japan
is017080@ed.ritsumei.ac.jp

Fuminori Kimura
Ritsumeikan University, Japan
fkimura@is.ritsumei.ac.jp

Akira Maeda
Ritsumeikan University, Japan
amaeda@is.ritsumei.ac.jp

Abstract

Several cultural domain resources in different languages have become available as Linked Open Data (LOD) in the last few years. However, there is little re-use of this data in multilingual information retrieval applications. The paper discusses Linked Data driven approaches in providing integrated multilingual access to diverse Japanese humanities databases by linking and re-using LOD resources dynamically. It proposes a method, which dynamically generates links across databases using Linked Data when users perform keyword searches. We built a prototype information retrieval system based on LOD resources, personal names authority data, subject headings, and links to other Linked Data resources. Furthermore, we demonstrate how this approach is integrated in real-life retrieval systems and how linking and accessing diverse databases can be enhanced to make use of the available LOD resources.

The proposed method also enables to access multiple databases in different languages by using the notations in various languages, which were obtained from the authority data resources. It allows to access to additional data not only in Japanese databases but also multilingual databases in other countries without depending on languages and formats of each database.

Keywords: Japanese humanities databases; Linked Data; authority data; information retrieval

1. Introduction

The paper discusses an approach in providing integrated multilingual access to diverse Japanese humanities databases, which combines the traditional information retrieval techniques and Linked Data driven techniques together, by linking LOD resources dynamically. Nowadays, many humanities and cultural organizations have provided their resources as LOD in the forms: 1) as interlinked RDFa (Resource Description Framework –in–attributes) pages or other RDF files with a predefined subset of subject/predicate/object triples, 2) as huge downloadable files in RDF/XML, 3) through publicly accessible SPARQL (a recursive acronym for SPARQL Protocol and RDF Query Language) query interfaces, i.e. endpoints.

Interlinked RDFa files are useful for digging of datasets further, but usually need some data conversion before integrating into applications, and such RDFa files cannot be queried at scale. The adoption and utilization of downloadable huge sized RDF/XML files in any application require a costly Linked Data/Semantic Web infrastructure e.g. Triplestores and SPARQL engines, which might be not easily integrated in the existing infrastructures. SPARQL endpoints are very effective in delivering the required pieces of data for a given inquiry. However, they require some knowledge of SPARQL and its syntax, and dataset-specific know-how is necessary.

In Linked Data, a triple may contain another URI, especially an external link, beyond the original dataset. This allows establishing easy linkage between datasets of totally different and distinct resources, forming a huge and virtually interlinked graph. However, one of the weaknesses of Linked Data is that the traditional metadata record, which contains all related information about a certain entity together, dissolves into a set of triples. A larger dataset may have billions of triples and these triples are merged with triples from other records, possibly different records that originated from different datasets. However, same as in any normalized relational database or traditional metadata based information retrieval system, meaningful information that can be delivered to users have to be retrieved from the triples in RDF datasets.

Therefore, to overcome these obstacles, in this paper we propose a method, which could help users to find additional relevant information for a given query input more efficiently using LOD.

2. Related Work

Several approaches have been proposed to deal with the diverse data of cultural institutions. In Linked Open Data for Academia (LODAC) Museum, which is a part of Lod.ac project, Kamura et al. (2011) aimed to integrate Japanese museums data and share them as Linked Data. Various Japanese museums data can be searched and browsed in Japanese as LOD through the LODAC Museum online system. In a similar research of the eCultura project, Cornejo et al. (2010) introduced a set of services and applications to access and integrate diverse web-based contents of the cultural domain. Hara (2011) aimed to develop information access methods for quantifying data and revealing new intellectual relations in area studies. In their resource sharing system, they have constructed 'shared meta-database' that can perform mapping of data items of each database into the semantically matching data items in 'shared metadata', which consists of common data items of various metadata. Meij et al. (2010) in their web-based repository service aimed to provide a uniform point of access (searching and browsing across vocabularies and collections or for collection indexing) to many collections at once using SKOS-based controlled vocabularies.

Beyond the cultural domain, in their Library Web Services for Economics, Neubert (2012) suggested autosuggest services by exploiting the SKOS data model; retrieval support that matches the query string against a full text in predefined local Lucene index; and mapping between datasets using SKOS vocabularies.

In general, all above approaches collect data or build indices in advance. However, we propose a different approach, which has the following three differentiations. First, instead of collecting data from each database in advance, we create links from the search results dynamically when a user performs a search. The links to other databases are generated by considering the initial query and returned data in the search results. Second, we access not only Japanese databases but also multilingual databases in other countries. Third, authority data is used for showing additional relevant information for a given query input.

3. The Proposed Method

We propose a Linked Data driven method as an extension to a Federated Search System for Ukiyo-e Prints (FeSSU) for realizing multilingual access by generating links dynamically across multiple databases as users perform keyword searches (Kuyama et al., 2012). Ukiyo-e, Japanese traditional woodblock printing, is known as one of the fine arts of the Edo period (1603–1868). The FeSSU is a federated search system for online databases in the Internet, which is capable of retrieving information from the Library of Congress, the National Diet Library (NDL) of Japan, the British Museum, the Boston Museum of Fine Arts, the Victoria and Albert Museum, the Ashmolean Museum, the Tokyo-Edo Museum, the Art Institute of Chicago, the New York Metropolitan Museum of Art, and the Ritsumeikan University Ukiyo-e database in parallel (Kimura et al., 2012; Batjargal et al., 2011).

3.1. Linked Data Utilization in the FeSSU

The proposed method utilizes existing LOD resources, personal names authority data, subject headings, and links to other Linked Data resources for generating links dynamically from the search results of the FeSSU. The name authorities and subject headings of NDL of Japan i.e. Web NDL Authorities and Virtual International Authority File (VIAF) are used as authority data. The

Web NDL Authorities provides centralized search on name authorities and subject headings in Japan. The VIAF contains the authority data of western national libraries that have been described in a variety of languages of western countries.

Using authority data has the advantages of 1) unifying different notation methods and synonyms, 2) distinguishing and searching individual names even if different names are referring to the same, and 3) containing additional information such as different notations and synonyms.

The reasons of using authority data in the proposed method are as follows. Firstly, simple string matching is not viable for appropriate matching of different representations of proper nouns. Secondly, databases in different languages could be accessed using the reading in authority data, which gives phonetic representation of a personal name written in Japanese kanji. A romanized representation (romaji) is commonly used in western databases when the Japanese proper nouns need to be translated into other languages, because kanji have several different possible readings. Lastly, some authority data could contain links to DBpedia as Linked Data.

The following Linked Data resources are utilized: 1) LODAC, 2) the thesaurus of Japanese art, 3) DBpedia, and 4) the Japanese DBpedia. The thesaurus of Japanese art contains information such as artwork, creator, work title, era, owner, schools, painting style, etc (Fukuda and Omuka, 2007). DBpedia exposes structured data extracted from Wikipedia freely to the public and Japanese contents are available separately in the Japanese DBpedia.

3.2. Generating Links Dynamically

Figure 1 illustrates the process of generating links. Databases inside the oval shapes are databases that already provide Linked Data. Solid lines denote the already available links between the databases. Dashed lines denote the dynamic links, which are generated automatically by the proposed method. SPARQL queries are created for searching artist names from the Web NDL Authorities and VIAF as well as generating links to other Linked Data resources.

The proposed method searches authority data of the proper nouns that have been found in the FeSSU search results. Any personal name written in Japanese will be searched from the Web NDL Authorities for retrieving its proper reading and then will be represented in romaji, which should be sent as a search keyword to the VIAF for obtaining personal name notations in various languages. By using the links and aliases in the obtained authority data, users will be allowed to access and view additional relevant data in other databases or to search DBpedia and LODAC.

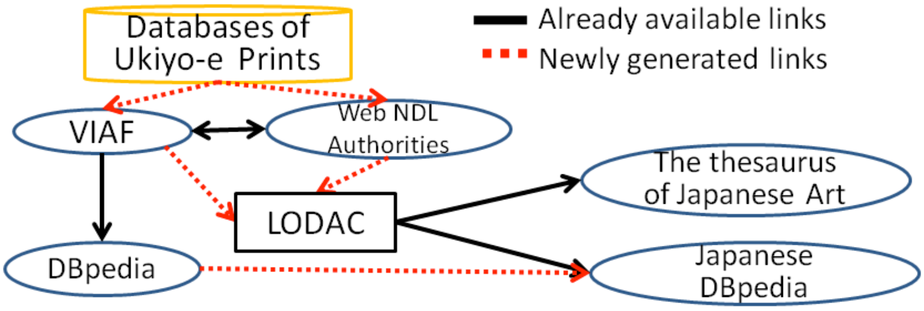

FIG. 1. The overview of dynamic links generation of the proposed method

In order to show additional information and links for other works of the same author in a given record, the author name will be sent to the VIAF as a keyword. If a link to DBpedia exists in the authority data obtained from VIAF then the additional information in that link will be shown. For instance, as shown in Figure 2, additional information in the link ‘URI of Hiroshige Utagawa’ of DBpedia will be shown to users using a link from ‘URI of 歌川広重’. Further search results for the aliases of ‘Hiroshige Utagawa’ e.g. ‘安藤広重(Hiroshige Ando)’ and ‘一立斎広重(Hiroshige

Ichiryusai)' will also be shown. The 'URI of dbpedia:Edo', and 'URI of dbpedia:Painting' in the example of Figure 2 will be sources of the further links. Using the 'URI of dbpedia:Painting', 'URI of dbpedia:Edo', and their values, additional information such as 'Hiroshige Utagawa is a painter, who was born in Edo' will be shown to users. Similarly, if a link to the Japanese DBpedia exists in the DBpedia, then additional information in that link will also be shown. Finally, search results from the LODAC will be shown by using the URI of Japanese DBpedia as a search key.

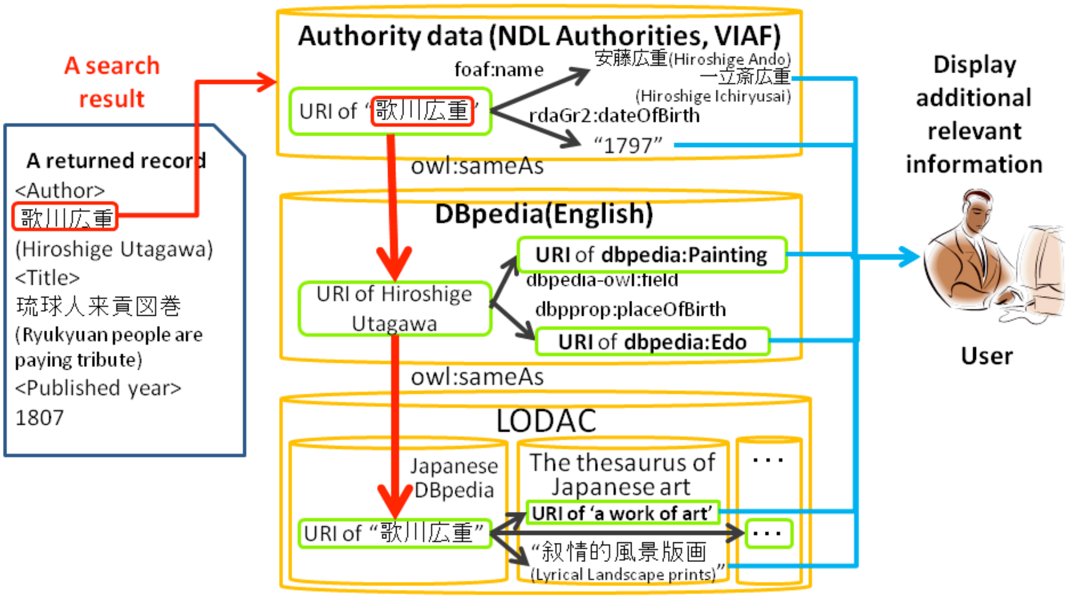

FIG. 2. An example of generating dynamic links in the proposed method

Furthermore, authority data in Web NDL Authorities has a link to Library of Congress Subject Heading (LCSH). Thus as shown in Figure 3, cross-language access between Japanese and English can also be implemented. Using a link in Web NDL Authorities, the proposed system can generate further links. In this way, for a given query input "浮世絵", users will be able to access to other records that contain "Hashirae", "Pillar Prints" and "Ukiyo-e" by using the links "Ukiyoe" to the LCSH. Even if the user has no knowledge about the term "Hashirae", he or she can access them using the proposed approach.

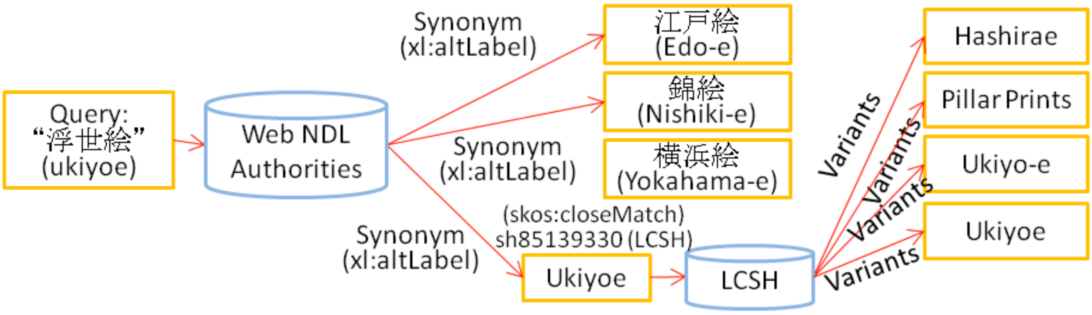

FIG. 3. An example of multilingual access using Subject Headings

4. Evaluation

We have conducted experiments to evaluate the reliability of the relevant information, which have found by the proposed method. Some irrelevant data to the original record might be found; even artist names in a record along with birth and death year are used as a search keyword for looking authority data in the proposed method. Since aliases in authority data are used, the proposed method could output irrelevant records if the artists have different names but the same alias. Therefore, we have calculated the precision of the proposed method by manually checking whether the data retrieved using the aliases are relevant to the artist in the original record or not.

Table 1 shows the precision of 4,008 newly retrieved records for several Ukiyo-e artists. The precision in this experiment expresses the degree of reliability of the newly returned search results. Aliases for randomly selected records from the FeSSU search results were generated using authority data resources VIAF and Web NDL. The proposed method achieved the average precision of 83.09%. The highest precision of 99.79% can be achieved for the aliases of Ukiyo-e artists with full names. The precision rate drops to 68.47% for the aliases that have a sole name such as surname, pen name or honored title of Ukiyo-e artists. The most of the cases of generating the links to irrelevant records were occurred when a sole name was used as an alias. For instance, the surname 英泉(Eisen) or 関月(Kangetsu) was used instead of the full name 池田英泉 (Ikeda Eisen) or 蔀 関月 (Shitomi Kangetsu). A low precision for the sole names might have occurred due to the high probability of matching irrelevant artists who have the same surname or given name due to their ambiguities in short strings.

TABLE 1: Average precision

Type of alias	Precision
Full name	99.79%
Sole name (surname, pen name or honored title)	68.47%
All the aliases	83.09%

5. Conclusion and Future Work

The FeSSU implementations are still under development. For each new database linked, we learned about new requirements, some changes or improvements. However, our general conclusions from the proposal of a Linked Data driven method and the development of the FeSSU are so far: 1) It is possible to provide multilingual integrated access to diverse databases by generating links between databases dynamically. 2) It is also possible to generate links to miscellaneous relevant data dynamically from the search results of multiple databases including those already provide Linked Data. However, some of data might be irrelevant because of the ambiguities in short strings. 3) Developing the FeSSU enables the offering of a single interface not only for diverse databases, but also allows the exploiting of additional knowledge, which could affect users' research efficiency.

Generalizing the proposed approach to other humanities databases and conducting a humanist-oriented evaluation on usefulness of the proposed method are the future tasks.

References

Batjargal, Biligsaikhan, Fuminori Kimura, and Akira Maeda (2011). Metadata-related Challenges for Realizing a Federated Searching System for Japanese Humanities Databases. Proceedings of the 11th International Conference on Dublin Core and Metadata Applications (DC2011). The Hague, Netherlands, 2011, pp. 80–85.

Cornejo, Carlos M, Ivan Ruiz-Rube, and Juan Manuel Dodero (2010). eCultura, a semantically-enriched web-based approach to manage cultural contents, Proceedings of the IEEE International Conference on Information Reuse and Integration (IRI 2010), Las Vegas, NV, pp. 126–131.

Fukuda, Hiroatsu, and Toshiharu Omuka (2007). Constructing Web-based Art Thesaurus, Focusing Around Data Conversion. The Bulletin of Japan Art Documentation Society, No.14, p.56-66.

Hara, Shoichiro (2011). Area informatics: concept and status. Proceedings of the Second International Conference on Culture and Computing, Kyoto, Japan, pp. 214–228.

Joachim Neubert (2012). Linked Data Based Library Web Services For Economics. Proceedings of the 12th International Conference on Dublin Core and Metadata Applications (DC2012), Kuching, Malaysia, pp. 12–22.

Kamura, Tetsuro, Fumihiro Kato, Ikki Ohmukai, Hideaki Takeda, Toru Takahashi, and Hiroshi Ueda (2011). Study support and integration of cultural information resources with Linked Data, Proceedings of the Second International Conference on Culture and Computing, Kyoto, Japan, pp. 177–178.

Kimura, Fuminori, Takushi Toba, Taro Tezuka, and Akira Maeda. (2009). Federated Searching System for Humanities Databases Using Automatic Metadata Mapping. Proceedings of the 9th International Conference on Dublin Core and Metadata Applications (DC2009), pp. 139–140.

Kuyama, Takeo, Biligsaikhan Batjargal, Fuminori Kimura, and Akira Maeda (2012). Integrated Multilingual Access to Diverse Japanese Humanities Digital Archives by Dynamically Linking Data. Conference Abstracts of Digital Humanities 2012 (DH2012), Hamburg, Germany, pp. 473–476.

Meij, Lourens, Antoine Isaac, and Claus Zinn. (2010). A Web-Based Repository Service for Vocabularies and Alignments in the Cultural Heritage Domain. Proceeding of the 7th international conference on the Semantic Web: Research and Applications, pp. 394-409.