ISSN (Print) : 0974-6846
Indian Journal of Science and Technology, *Vol 8(23), DOI: 10.17485/ijst/2015/v8i23/79342 September 2015* ISSN (Online) : 0974-5645

Ontology-based Integration and Refinement of Evaluation-Committee Data from Heterogeneous Data Sources

Heeseok Jeong and Hanjo Jeong*

NTIS Center, KISTI, Korea; hsjeong@kisti.re.kr, hanjo.jeong@kisti.re.kr

Abstract

Korean National Science and Technology Information Service (NTIS) provide a service of searching national R&D projects and their participating researcher information. It also provides a service of recommending and selecting evaluation committees for the R&D projects. Such R&D data and information are collected from 17 Korean government ministries and agencies and integrated into NTIS. Therefore, the duplicates of the R&D data and researcher information can be inserted because the titles of a researcher's R&D accomplishment data can be differently inserted from the different organizations. Furthermore, the names of researchers and other related objects such as organizations and journals can also be inserted vairously as the names have various aliases in general. In this research, we present an ontology-based data integration and refinement system for integrating such researcher information and their R&D accomplishment data, which would be useful for the recommendation and selection services. Also, we also used Jaro-Winkler distance algorithm to find and eliminate the duplicated accomplishment data. Furthermore, incorrectly entered data are also corrected from the duplicate elimination process with the information obtained from some authoritative science libraries.

Keywords: Data Integration, Data Refinement, Jaro-Winkler Distance, National R&D Data, Ontology

1. Introduction

The most of data integration systems has a purpose of aggregating and analyzing the data as in the business intelligence area[1–4]. However, such approaches using a star schema integrate data by simply aggregating data. On the appearance of Semantic Web technologies such as RDF[5], RDFS[6] and Web Ontology Language (OWL)[7], data can be integrated in more semantic ways. The Semantic Web technologies represent data as knowledge with a machine-readable format and a standardized representation framework. Thus, the data integration and management would be performed in a more automatic way with the semantics of data[8–13].

In this research, we dealt with problems of the entity identification and the data refinement through the ontology-based data integration and an automatic text comparison algorithm. The National Science and Technology Information Service (NTIS)[14] has goals to build a more complete, robust and trustworthy information database for researchers. That is because many of government ministries and agencies and industrial organizations use NTIS information for selecting evaluation committees and verifying participating researchers for their R&D projects.

In this paper, we present an overall architecture of the ontology-guided data integration and refinement system along with a method of integrating, refining and verifying the researcher information using ontology and authoritative science libraries. For a pre-refinement, we used Jaro-Winkler distance algorithm[15–18] to find and eliminate the duplicates of the researchers'

* *Author for correspondence*

accomplishment data. With the information obtained from the authoritative science libraries, we also correct the wrong data.

2. Data Integration and Pre-Refinement

To provide a committee selection service for the evaluation of the national R&D projects, the demographic information of the evaluation committee candidates such as their affiliations and their R&D accomplishment data should be managed and maintained as up to date. Figure 1 represents the overall system architecture of the ontology-based data integration and refinement system. At first, the researchers' demographic data and their R&D project data along with their accomplishment data are collected from 17 Korean government ministries and agencies. The collected data are integrated and refined via using a duplicate-finding algorithm using Jaro-Winkler distance algorithm. The bibliographies provided external authoritative science libraries are also used to standardize and correct the data. Lastly, the integrated and refined R&D data for each researcher is stored and maintained by the Researcher Information DB.

2.1 Data Collection and Integration

The most of Korean government ministries and agencies have their own R&D project management systems, and the researchers are required to submit their R&D accomplishment to apply for performing and participating in an R&D project. Furthermore, the participating researchers are required to submit the results of the R&D projects after they completed the project works. The researchers' R&D accomplishment data and the outputs of the current R&D project data are collected to the Researcher Information DB and R&D Project DB respectively via the NTIS information link framework. The information link fraemwrok provides APIs for extracting, transforming and loading the researchers' information and R&D data.

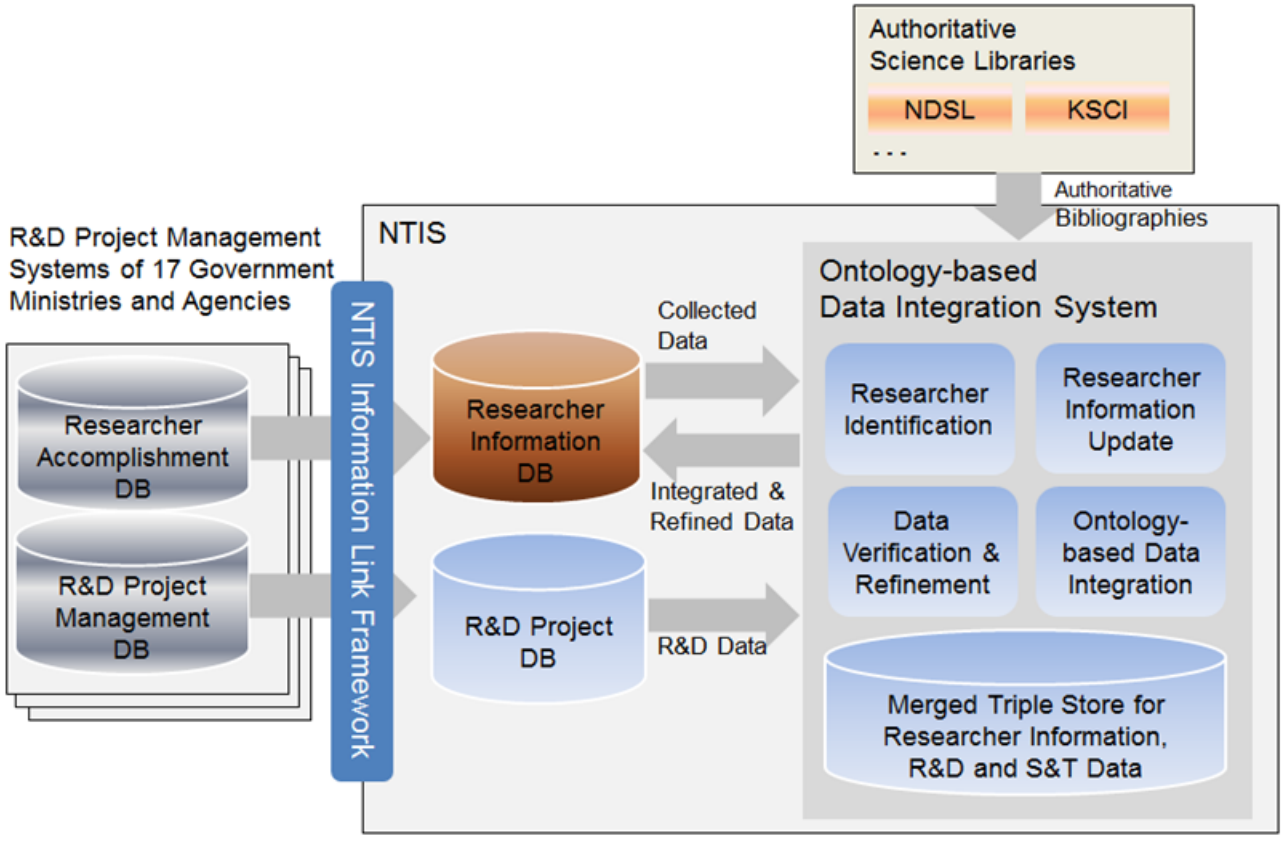

Figure 1. Overall system architecture.

The collected researcher information is stored in the Researcher Information DB and the information is integrated and refined with a process consisting of three statuses: 1. A status of register that indicates the data is collected and registered in the Researcher Information DB. 2. A status of verified that indicates the data is verified with the authoritative science libraries such as NDSL (National Digital Science Library)[19] and KSCI (Korea Science Citation Index)[20]. 3. A status of integrated that the data is integrated and merged with the R&D project data. 4. A status of refined that the data is pre-refined using Jaro-Winkler distance algorithm described in the next section. Figure 2 shows an example of table schema for managing the integration and refinement process for research papers. THR_PAPER Table contains the integrated and merged data along with its related researcher demographic information in THR_BASE table and its registration information in IM_REG_NO. Remaining CL_ Tables are interim tables to refine the integrated paper data using Jaro-Winkler distance algorithm. Such table schema is similarly created for each R&D data such as research projects, patents, and reports.

2.2 Pre-Refinement using Jaro-Winkler Distance

The researcher accomplishment data such as their published research papers and patents is often not correct and duplicated since a research paper or a patent is inserted as different researchers as they co-authored it, and the data is collected from 17 government ministries and agencies.

Figure 3 shows a duplicated accomplishment-data detection algorithm using Jaro-Winkler distance. It basically checks the duplicates on a researcher because a researcher's accomplishments data is collected various sources, and the accomplishment data is not exactly identical for the entire attributes of the data. Such data

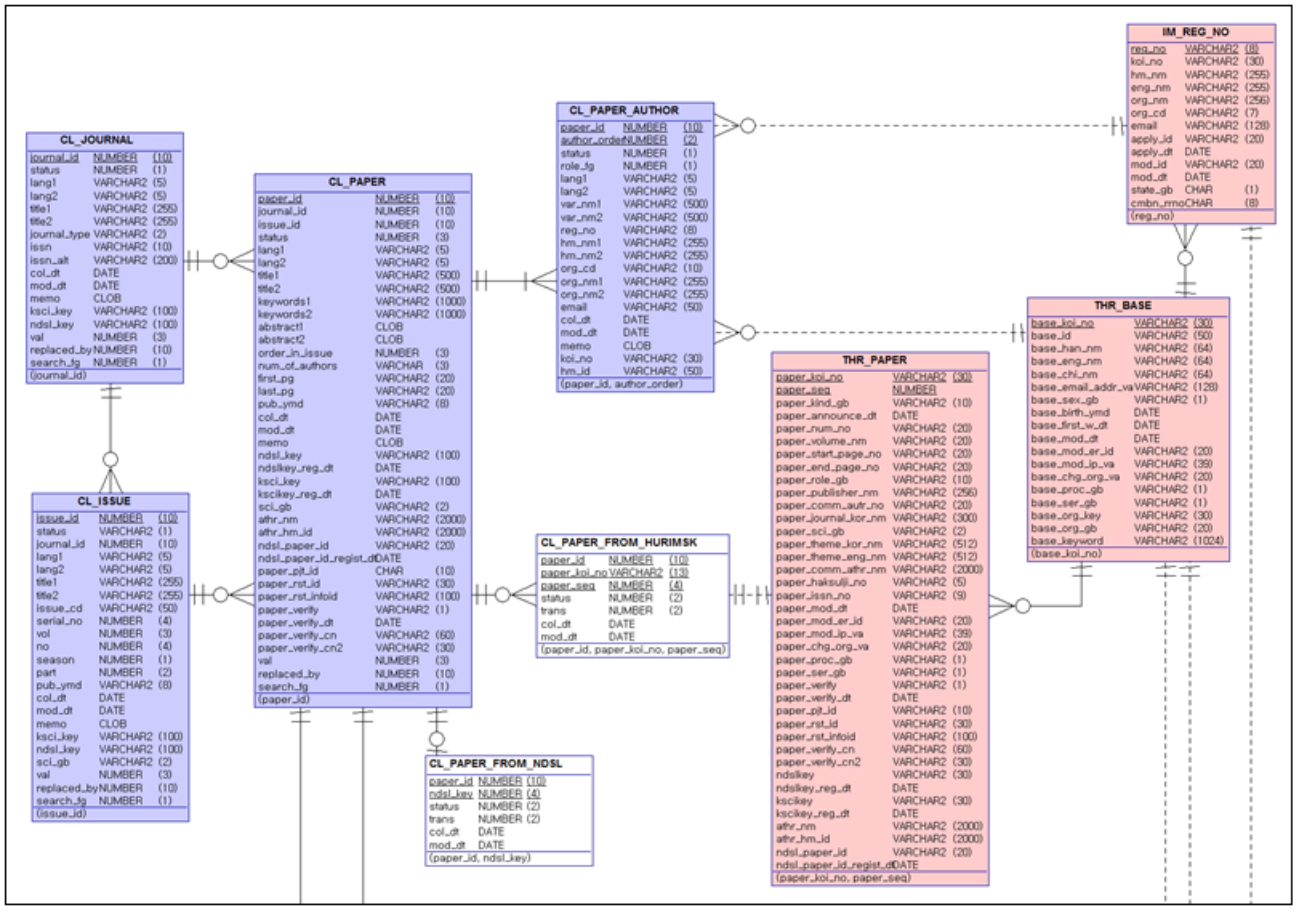

Figure 2. An example of table schema for research papers.

is treated as a different data at the initial collection process since the data is regarded as identical if they have exact values on the entire attributes such as title, journal, year, etc. In addition, the titles are sometimes a bit different because of some mistakes and errors on the data insertion. Therefore, we examine every possible pair of the accomplishment data inserted for a researcher, and extends the creation of the possible pair with the researcher's colleagues who wrote the accomplishment data together. The possible pair can be found by checking the co-author set. In the algorithm, the visited researchers are removed from the Un-Scanned Researcher Set thereby preventing re-scanning of the accomplishment data of the previously visited researchers.

This scanning algorithm prevents us to scan every pair of the accomplishment data by using a naive exhaustive comparison, which compare every accomplishment data with every other accomplishment data and it would require processing $O(n^2)$ in the number of the accomplishment data that is more than 3 million. Such scanning policy is applied because of the assumption that the duplicates can occur only in the cases: 1. A researcher's accomplishment data can be differently inserted from the various sources. 2. An accomplishment data can be differently inserted by its co-authors. The Jaro-Winkler distance measure is calculated for the every possible pair of the accomplishment as in Equation 1, and the pairs having the distance value higher than a threshold is added to a candidate duplicates set. The candidate duplicates set is manually checked and replaced with an identical data obtained from the authoritative sources such as NDSL and KSCI.

```
Set of Duplicated Data Set D ← an empty set
Set of Un-scanned Researchers U ← Entire Set of Researchers
while U is not empty
  for each researcher r in U
    create a co-author set of Researchers C ← an empty set
    add r to C, and remove r from U
    add co-authors in U to C
    create r's accomplishment data set A_r
    for each c in C
      create c's accomplishment data set A_c
      for each a_r in A_r
        for each a_c in A_c
          compute Jaro-Winkler Distance (a, a')
          if the distance is bigger than threshold,
            add the pair of a, a' to D
        end for
      end for
    end for
  end for
end while
```

Figure 3. Duplicated accomplishment-data detection algorithm using Jaro-Winkler distance.

$$F_j(s_1,s_2) = \frac{1}{3}\left(\frac{N_c}{|S_1|} + \frac{N_c}{|S_2|} + 0.5\frac{N_t}{N_c}\right) \quad (1)$$

Where s_1 and s_2 are the comparing strings with lengths $|s_1|$ and $|s_2|$, respectively. N_c is the number of matching common characters between the two strings and N_t is the number of transpositions, which is the number of transpositions of the common characters to make the order of the common strings identical for the two strings[15,16].

Figure 4 shows a partial result of duplicated paper detection process performed using Jaro-Winkler distance. The *distance* column represents the Jaro-Winkler distance measure, and an accomplishment data is identified with its author as a *researcher_id* and the researcher's accomplishment sequence number as *seq*. The table shows the values of the distance of the research paper pairs identified by researcherID_seq# and researcherID_seq# for the titles strings. The titles are pre-processed before the distance measure is calculated as follows: 1. Capitalization, 2. Space Normalization, and 3. Symbol Elimination. The value 1 means the titles are exactly identical, and we found that the value higher than 0.95 is mostly identical. We used 0.90 as a threshold because it is worth to look up the pairs for the duplicates check if the distance measure is higher than 0.90. The table also shows how the scanning are performed among the researchers and their colleagues.

Distance	Researcher_ID1	seq1	Researcher_ID2	seq2	Title 1	Title 2
1	K430802100006	280	K430802100006	385	Terfenadine-β-Cyclodextrin Inclusion Complex v	Terfenadine-β-Cyclodextrin Inclusion Complex
0.98	K430802100006	61	K430802100006	685	The Enhancement of Liver Targetability of [3H]-	The enhancement of liver targetability of[3H]-
1	K430802100006	521	K430802100006	540	Trials of in situ-gelling and mucoadhesive aceta	Trials of in situ-gelling and mucoadhesive ace
0.99	K430802100006	744	K590210100007	666	Preparation and Characterization of Cytarabine-	Preparation and Characterization of Cytarabin
0.99	K430802100006	552	K590210100007	567	Preparation and Characterization of Poly(γ-Ben	Preparation and Characterization of Poly(γ-be
0.98	K430802100006	702	K590210100007	492	Preparation of and Drug Release From S/O/W E	Preparation of and Drug Release from S/O/W
0.99	K430802100006	564	K590210100007	567	Preperation and characterization of poly(γ-benz	Preparation and Characterization of Poly(γ-be
0.98	K430802100006	541	K590210100007	3	Prolonged Release of Tegafur from S/O/W Muti	Prolonged release of tegafur from S/O/W Mu
0.96	K430802100006	736	K590210100007	491	Release of Tegafur From Microsphere-in Oil in \	Release of Tegafur from Microsphere-in-Oil-in
1	K590210100007	646	K590210100007	850	Biodistribution and Genotoxicity of Transferrin-(	Biodistribution and Genotoxicity of Transferrin
0.96	K590210100007	308	K590210100007	598	Bioequivalence evaluation of risperidone 2 mg t	Bioequivalence evaluation of finasteride 5 mg
1	K590210100007	613	K590210100007	746	Bioequivalence of boryung torsemide tablet to t	Bioequivalence of boryung torsemide tablet t
1	K590210100007	245	K590210100007	641	Determination of tiropramide in human plasma	Determination of tiropramide in human plasm
1	K590210100007	245	K590210100007	858	Determination of tiropramide in human plasma	Determination of tiropramide in human plasm
1	K590615100010	21	K590615100010	519	Right hemisphere abnormalities in major depre	Right hemisphere abnormalities in major depr
1	K590615100010	457	K590615100010	764	The correlational study between event-related p	The Correlational Study between Event-Relate
0.97	K590615100010	40	K590615100010	425	The effect of immediate and delayed word repe	The effect of immediate and delayed work rep
1	K590615100010	476	K590615100010	824	The event-related potential (ERP) study of recog	The event-related potential (ERP) study of rec
0.98	K590615100010	469	K590615100010	768	Time-frequency analysis of 40Hz in oddball par	Time-frequency analysis of 40-Hz in oddball p
0.96	K590615100010	471	K590615100010	710	Volumetric analysis of frontal substructures in	Volumetric analysis of frontal substructures ir
1	K590615100010	639	K710806100002	910	An MEG study of alpha modulation in patients	An MEG study of alpha modulation in patient
1	K590615100010	639	K710806100002	1133	An MEG study of alpha modulation in patients	An MEG Study of Alpha Modulation in Patien

Figure 4. Examples of duplicated-paper detections using Jaro-Winkler distance.

3. Ontology-based Data Integration

In this research, we used ontology to integrate the researchers' R&D data collected from the number of heterogeneous sources. After we pre-processed the information by eliminating the duplicated data and substituting the data with the authoritative data, we identify and create the objects such as researcher objects, researcher's organization objects, researcher's accomplishment objects, journal objects, etc. An ontology model is created with the identified objects and their relationships as described in the following subsections.

3.1 Ontology Model for Researcher's R&D Accomplishment

Figure 5 shows the ontology model for representing the researcher information and the researcher's R&D accomplishments such as national R&D projects, research

papers, patents, and research reports. The model represents researchers' human & resource information by specifying researchers' demographic information and organization information that the researchers are currently affiliated. The R&D accomplishment-data objects excepting the R&D-project objects are connected as outputs of the R&D-project objects since the researchers insert their produced R&D data as their accomplishment on the national R&D projects.

This ontology model guides the data integration, and the data can be refined with such ontological knowledge if a collected data has wrong and un-matching information.

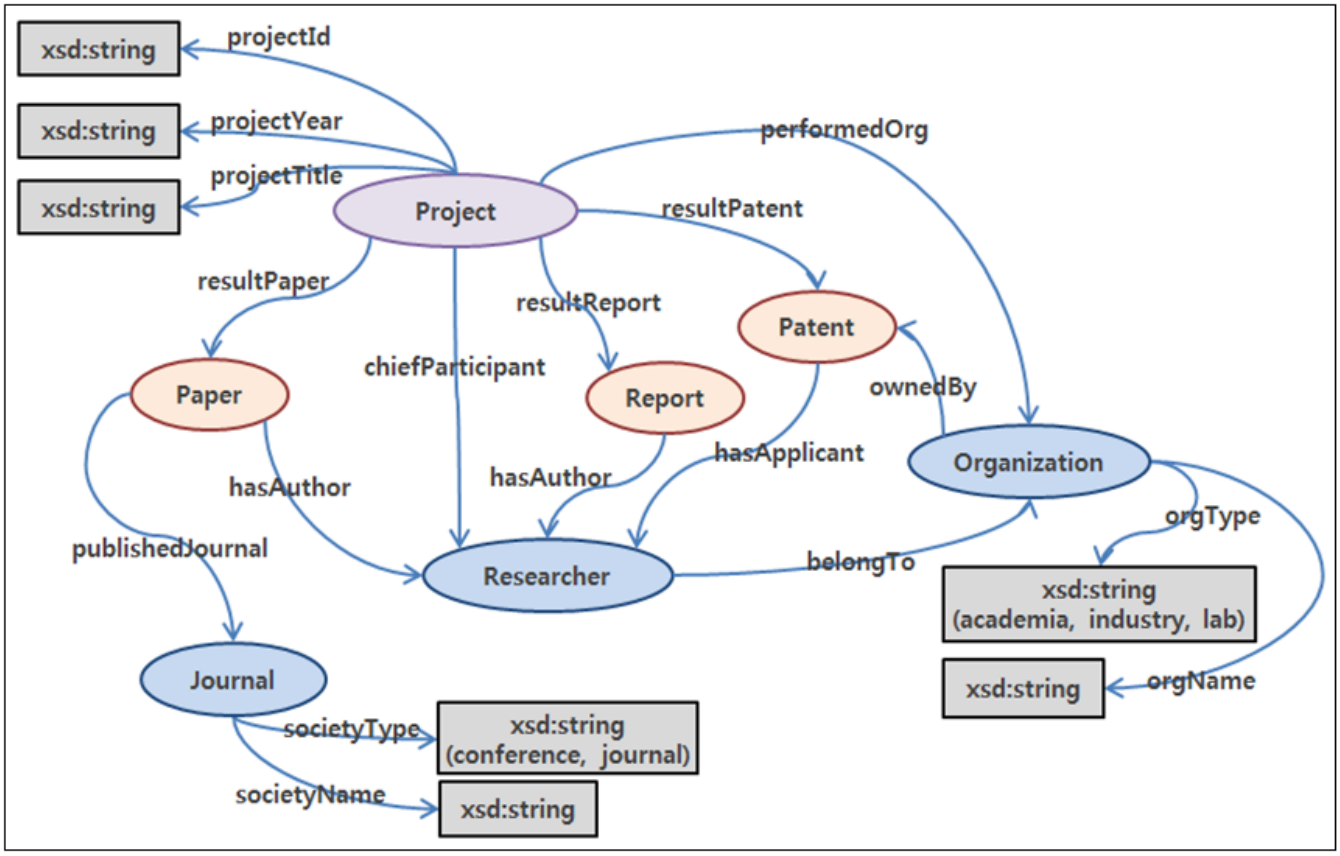

Figure 5. Ontology Model for Researcher's R&D Accomplishment.

To perform such refinement, we used the RDF triple format and SPARQL (SPARQL Query Language for RDF)[21]. The un-matching data can be found through some SPARQL query patterns and eventually refined by admin for managing the researcher information as describe in the next section.

We also used the bibliographies obtained from the authoritative science libraries such as NDSL and KSCI to verify and refine the researcher information and the R&D data. Using such authoritative information enables us to automatically refine such error-prone data entered by researchers themselves and produced by the insufficient automatic author and organization identification. The automatic author and organization identification cannot be perfect since there are many researchers having same names and the names of humans and organizations have many aliases. Also, such information is recorded as in multiple languages. The object identification problems can be alleviated using such authoritative data sources.

3.2 Object Comparison and Merge

The pre-refinement of the accomplishment data with the Jaro-Winkler distance algorithm using the title strings described in Section 2.2 eliminates the duplicates in a researcher's accomplishment and substitutes the titles with the authoritative data. However, other data such as researchers, organizations and journals is still ambiguous and duplicated as the names of such data is too vague in

general. As we created the data as the objects along with the ontology schema, the ontological object structure and relationships can be useful for identifying the objects thereby being able to eliminate the duplicate objects.

At first, a SPARQL query is posed to find the related objects such as researchers, organizations and journals with the identical accomplishment data. If a different set of objects having the identical accomplishment data is found, we perform a merge process for the objects and their relationships by using the authoritative data. The merge process will remove the duplicate objects having alias names by changing the names using the authoritative data.

4. Conclusion

In this research, we presented an ontology-based information integration and refinement system for researcher information and their R&D project data along with their accomplishment data. The collected and refined information would be useful as the basis information for recommending and selecting the evaluation committees for evaluating the national R&D projects. We also provided a duplicate data detection algorithm using Jaro-Winkler distance for processing the free text data.

For the future research, more sophisticated and detailed methods for comparing ontological objects will be elaborated with a combination of a free-text processing approach and an ontology reasoning and inference approach. Such approach would enable us to automatically identify objects during the object comparison process.

5. Acknowledgment

This research was supported by Maximize the Value of National Science and Technology by Strengthen Sharing/Collaboration of National R&D Information funded by the Korea Institute of Science and Technology Information (KISTI).

6. References

1. Dayal U, et al. Data integration flows for business intelligence. ACM Proceedings of the 12th International Conference on Extending Database Technology; 2009. p. 1-11.
2. Chen H, Chiang RH, Storey VC. Business intelligence and analytics: From big data to big impact. MIS Quarterly. 2012; 36(4):1165-88.
3. Tank DM. Reducing ETL load times by a new data integration approach for real-time business intelligence. IJEIR. 2012; 1(2):1-5.
4. Petermann A, Junghanns M, Muller R, Rahm E. Graph-based data integration and business intelligence with BIIIG. Proceedings of the VLDB Endowment. 2014; 7(13):1577-80.
5. Resource Description Framework (RDF). Available from: http://www.w3.org/RDF/
6. RDF (Resource Description Framework) Schema. Available from: http://www.w3.org/TR/rdf-schema/
7. OWL2 Web Ontology Language. Available from: http://www.w3.org/TR/2012/REC-owl2-overview-20121211/
8. Noy NF. Semantic integration: A survey of ontology-based approaches. ACM Sigmod Record. 2004; 33(4): 65-70.
9. Kondylakis H, Plexousakis D. Exelixis: Evolving ontology-based data integration system. Proceedings of the 2011 ACM SIGMOD International Conference on Management of Data; 2011. p. 1283-6.
10. Lenzerini M. Ontology-based data management. Proceedings of the 20th ACM International Conference on Information and Knowledge Management; 2011. p. 5-6.
11. Shan G, Rui G, Wenjun W, Xiankun Z. Research on Ontology-based Emergency Situation Data Integration and Sharing. Journal of Convergence Information Technology. 2012; 7(9).
12. Calvanese D, Giese M, Hovland D, Rezk M. Ontology-based integration of cross-linked datasets. Proceedings of the 14th International Semantic Web Conference (ISWC); 2015.
13. Livingston KM, Bada M, Baumgartner WA, Hunter LE. KaBOB: Ontology-based semantic integration of biomedical databases. BMC Bioinformatics. 2015; 16(1):126.
14. National Science and Technology Information Service (NTIS). Available from: http://www.ntis.go.kr/
15. Jaro MA. Probabilistic linkage of large public health data files. Statistics in Medicine. 1995; 14(5-7): 491-8.
16. Porter EH, Winkler WE. Approximate string comparison and its effect on an advanced record linkage system. US Bureau of the Census. Research Report; 1997.
17. William C, Ravikumar P, Fienberg S. A comparison of string metrics for matching names and records. KDD Workshop on Data Cleaning and Object Consolidation. 2003; 3:73-8.
18. Winkler WE. Overview of record linkage and current research directions. Bureau of the Census; 2006.
19. National Digital Science Library (NDSL). Available from: http://www.ndsl.kr/
20. Korea Science Citation Index (KSCI). Available from: http://ksci.kisti.re.kr/main/main.ksci
21. SPARQL Query Language for RDF. Available from: http://www.w3.org/TR/rdf-sparql-query/