Lined Open Data for Knowledgebase

Shoichiro HARA
Center for Integrated Studies, Kyoto University

Kyoto University, since its foundation in 1897, has collected, created and accumulated numerous and various materials, data and information as its academic resources, and it has developed databases for researchers to access these resources, i.e., KULINE (the university OPAC operated by the library), KURENAI (the university repository developed by the library), KURRA (the university research archives developed by the museum), University Open Course Ware (operated by the Academic Center for Computing and Media Studies), and various research databases (developed and maintained by research institutes and centers). However, as each database system is independent and heterogeneous, it is difficult to carry out even simple searches such as finding original experimental data related to the paper. Obviously, we cannot use such isolated databases for advanced research usages to discover hints and/or create new knowledge.

To overcome these drawbacks in databases, a new project has been launched to develop an innovative information platform to accumulate and link academic resources, and offered the platform as advanced research utilities. This platform will comprise three sublayers. The fist layer is "Open Data Layer" to accumulate and open heterogeneous data. This layer uses RDF (Resource Description Framework) that can describe data of different structures by uniform way. For example, this layer can accumulate thesauri (tree structure), bibliographic catalogues (table structure) and documents (XML) simultaneously. The second layer is "Data Link Layer." Academic data, especially humanities' data, are messy and ambiguous (i.e., a term "title" may be used as “book title” in a database but the same term may be used as “personal title” in another database, how to define “Tokyo Bay Area,” and how to calculate “ca. 1162”). This layer uses ontology techniques such as RDFS and OWL to link ambiguous notions and/or vocabularies and to create "Academic Big Data." Academic Big Data comprise small fragments of heterogeneous data, which will form complex structures. This is the different feature from ordinary big data comprising simple structure data from sensors and IoT devices. The third layer is "Application Layer." As Academic Big Data is too huge and complicate for researchers to retrieve, categorize and analyze by hands, and then applications to support these processes are necessary. The new project will develop some utilities, i.e., to estimate subjects of contents by natural language techniques, to categorize huge data sets by deep learning techniques, and to organize data according to topological spatiotemporal expressions. This platform will also provide APIs to create mashup applications easily. We expect the platform to reconstruct knowledgebase from databases, which is used to construct meaningful chunks from scattered data.

This presentation will describe overview and state of progress about the project to promote advanced usages of academic resources of Kyoto University as "linked open data" on Cloud environment.