Contribution ID: **29** Type: **not specified**

Platform for Humanities Open Data

Wednesday, 8 March 2017 14:00 (30 minutes)

Construction of humanities databases have difficulties due to some reasons. First, construction process requires expert knowledge and techniques of database systems, which impedes database construction by humanities researchers. Second reason is diversity of resource media, which enriches the humanities researches but is an obstacle to metadata standardization, and brings about heterogeneous databases. Third reason is this heterogeneity of metadata, which makes sharing data difficult.

Center for Integrated Area Studies, Kyoto University (CIAS) has developed two information tools, named MyDatabase (MyDB) and Resource Sharing System (RSS), to solve these difficulties. The main component of MyDB is a database builder, allowing humanities researchers to construct and revise databases without expert knowledge. MyDB stores metadata and accepts any vocabulary of metadata, including nonstandard one. This enables humanities researchers to use their own metadata vocabulary according to their own purpose. On the other hand, those metadata varieties make the integration processes difficult. RSS is developed to integrate heterogeneous databases on the Internet and to provide users with a uniform interface to retrieve databases seamlessly in one operation. Thus, MyDB and RSS have contributed to accelerate humanities open data, but there are still two problems to solve, especially of RSS: small coverage of databases and initial costs of integration. First, for example, Kyoto University releases KULINE (OPAC), KURENAI (repository), KURRA (archive), Open Course Ware and various databases developed by each research institute in the university, but RSS does not integrate these databases. Second, it is time consuming to integrate new databases into RSS and impossible to trace links automatically, that is, for now, RSS is not the appropriate tool to discover hints and/or create new knowledge.

To overcome these drawbacks, a new project has been launched to develop an innovative information platform for open humanities data. This platform comprises three sublayers. The first layer is "Open Data Layer" which accumulates heterogeneous metadata. This layer uses RDF to describe data of different structures. The second layer is "Data Link Layer." This layer uses ontology techniques such as RDFS and OWL to link ambiguous (uncontrolled) vocabularies and emerge "humanities big data." The third layer is "Application Layer." As humanities big data is too huge and complicated to retrieve, categorize, and analyze by hands. This layer provides utilities to process big data. This platform will prepare for APIs to help mashup applications. We expect the platform to reconstruct knowledgebase from heterogeneous databases, which is used to construct meaningful chunks from scattered data.

As a pilot study to determine the validity of the platform, we prepared a dataset of "Japanese Journal of Southeast Asian Studies" as a core dataset (the first layer) for trying to link words or documents in the core dataset to external resources such as KURENAI or DBpedia (the first layer also). Then, the relationship between heterogeneous internal and external databases will be described in the second layer, so that whole data is structured for clean API and exploits that data in an annotated paper viewer (the third layer).

Primary author: Prof. SHOICHIRO, HARA (Center for Integrated Area Studies, Kyoto University)

Co-author: Prof. AKIHIRO, KAMEDA (Center for Integrated Area Studies, Kyoto University)

Presenter: Prof. SHOICHIRO, HARA (Center for Integrated Area Studies, Kyoto University)

Session Classification: Humanities, Arts & Social Sciences I

Track Classification: Humanities, Arts, and Social Sciences (HASS) Applications