Historical big data: integrated analysis of the past world by the workflow that bridges data structuring gaps

*Asanobu Kitamoto[1,2,3], Mika Ichino[1,2]

1. National Institute of Informatics, 2. ROIS-DS Center for Open Data in the Humanities, 3. SOKENDAI

1. Introduction

Reconstruct and analyze the past world through the extraction and integration of information from the past books and documents –we call this type of historical research as “historical big data.” This is because this approach has the same structure and purpose with “modern big data” on the current world, and it is meaningful to extend this approach into the past world. A big obstacle, however, is in data structuring. We start from data written in Kuzushiji (cursive scripts) on historical documents to high quality data usable for computaitonal analysis of the past world. For this purpose, we need information infrastructure to support this long data structuring workflow from digitzation to quality control. Because this is difficult to automate, we need efficient data structuring workflow based on human-machine collaboration. We therefore design a workflow to convert non-structured data (e.g. image and text) to structured data (analysis-ready data) and develop information platform for humanities dataset that is reusable and verifiable.

A similar but bigger project has started in EU called “Time Machine Flagship” involving more than 200 organizations. The project is building “time machine” to move freely within spatio-temporal space, especially within the big data of cities such as Venice in Italy and Amsterdam in the Netherlands. This trend has not been incorporated in Japan and Asia, so our platform may also be an entry point for this global initiative.

2. Issues

Many research projects for reconstructing the past have been carried out, but historical big data is a different approach in terms of the following.

First, the type of data is different. For example, paleoclimate research uses any data related to climate; not only books and documents, but also traces in the nature (proxy data such as ice core and growth ring). On the other hand, historical big data limits its scope on records written by human, and focuses on developing new data structuring workflow including the interpretation of textual records.

Second, the target of discipline is different. Research in one discipline had a tendency of focusing only on a limited type of phenomena, and did not pay attention to other phenomena. Even when researchers are studying the same diary, bibliographic data or extracted data were not shared across research groups, and many researchers repeated similar tasks again and again. To solve this problem, historical big data offers data structuring workflow that is effective across disciplines, and enables research that takes advantage of data sharing.

3. Data crossing and re-interpretation

To extend the modern big data into the past, relevant issues are not only the extension of technologies but also concepts and methodologies.

First, data crossing is a methodology to overlay different data to find unexpected relationship. A typical example is a map. By registering and overlaying data from different sources, we can take actions based on insights obtained from data. Here important challenges include the interoperability of API and sharing of terminology. To solve this challenge, we share future challenges within our research group, take advantage of the strength of each member without overlapping tasks, and develop information infrastructure maximizing the outcome within limited resources.

Second, data re-interpretation is a methodology to find effective reuse of data collected for one purpose into another purpose. A famous example in the modern big data is the case of reusing car probe data for the map showing passed routes after earthquake disasters. A same approach should be effective for historical big data. For example, is it possible to use the change of bureaucrat's directories for evaluating societal impact by climate change? Flexible imagination may lead to the creative usage of data.

4. Beyond “One Bed Multiple Dreams”

Historical big data is a multidisciplinary research. Of course, a big gap exists between humanities and science/technology, but difference within science and technology disciplines are not negligible. How to go beyond these gaps? Our basic idea is to recognize that we are in “one bed multiple dreams” situation in the sense that each member in the research group has a different dream. But at the same time, we recognize the advantages of being in one bed. For example, data and tools can be shared even if we have different dreams. After maximizing the advantages of sharing, an individual researcher tackles the reconstruction of a different part of the past world –this is the concept of historical big data research.

Keywords: Historical big data, Paleoclimate, Data structuring, Information infrastructure, Multi-disciplinary research , Historical documents