[3-2]

DATA CITATION ANALYSIS FRAMEWORK FOR OPEN SCIENCE DATA

Koji Zettsu*

**National Insititute of Information and Communications Technology, 3-5 Hikaridai, Seika-cho Soraku-gun, Kyoto, 619-0289, Japan*
Email: zettsu@nict.go.jp

Recent trend of open data encourages to publish and reuse science data across the networked data repositories. Data citation is the practices of providing a citation to data in the same way as a bibliographic reference to published articles (CODATA-ICSTI, 2013). Through data citation links are formed between articles and DOI-assigned data sets in repositories. The articles cite related data sets properly, and also the data sets provide citations of all articles related to the data set. As the result, the 'web of data citation' is being formed in emerging repositories like Pangaea, ICPSR and others.

Data citation is expected to create new ways of making research data actionable for reuse. Researchers can validate a conclusion of an article through the reanalysis of cited data. Moreover, the catalogue of cited data will encourage interdisciplinary collaboration and open up new user base for the data. In order to facilitate those benefits, it is the key to locate and discover appropriate data for reuse by analyzing the web of data citation. This talk introduces our research work on data citation mining for usage analysis of open science data.

Data citation mining process starts at collecting the citation metadata from the data repositories. It is done automatically by software programs which we developed in this work, extracting lists of citing articles for each data set through the standard API (OAI-PMH) or directly from the web pages. The collected metadata are uniformly formatted to create a "data citation graph", where a citation is represented by a directed link from an article node to a data node. More than 128,000 citations have been collected from Pangaea, ICPSR, ESDS, ADA, DRYAD, DataCite and others. Figure 1 shows examples of the data citation graphs.

The analysis process discovers following characteristics of data usage from the citation graph. The most intuitive one is the data linked by a large number of articles, which we call a reputable data here. The popularity is measured by number of citation links, while it can be further detailed by classifying the citing articles by subjects and/or authors. We found another type which is a cluster of datasets co-cited by a single or a very few articles, or a data collection community (Figure 1 (a)). A typical example can be seen in Pangaea, where each community corresponds to a research project collecting a lot of data. Those data are considered to be collected intentionally for a specific subject, thus the center article provides the index to many data sets. The community is more coherent if the fewer data are cited by those articles outside the community (independent), as well as the data are similar with each other in the community (consistent). In contrast, a cluster of articles citing a single or a small number of data sets with similar subjects can be called a data sharing community. For example, in Figure 1 (b), the center data, National Social Science Survey and its similarities, are shared by groups of articles related to 'working', 'inequality' and 'attitude' issues. In fact, the web of data citation is a composite of both kinds of communities. Moreover, a hub article can be found as an article citing many reputed data. For example, in Figure 1 (c), the hub article in the center, Australian Broadcasting Cooperation (Audience Research), cites radio survey data at different locations (Brisbane radio, Melbourne radio, Sydney radio, etc.), each of which is also cited by the local report articles. The hub article is considered to play a role of a catalogue of data sets from multiple communities.
A data can be characterized by the citing articles differently from its content. That is called a referential context of the data. For example, in Figure 1 (d), the population data and the income data can also be referred to as health insurance-related data through the citing articles, while those data originally have nothing to do with it. The referential context give a different viewpoint to the data, thus facilitate 'repurposing' of the data. A more common reference can be discovered by clustering the citing articles, which represents typical usage of that kind of data.

This talk introduces and demonstrates data citation mining for usage analysis of open science data with some analysis results. To the best of our knowledge, it is the first attempt of large scale and practical data citation analysis. This benefits both data repositories and researchers with improved discoverability of cited data, as well as increasing incentives to put more data citations. The data citation mining will enhance the analysis to more comprehensive networks of research relationships between scientific data, documents, authors and funding sources.

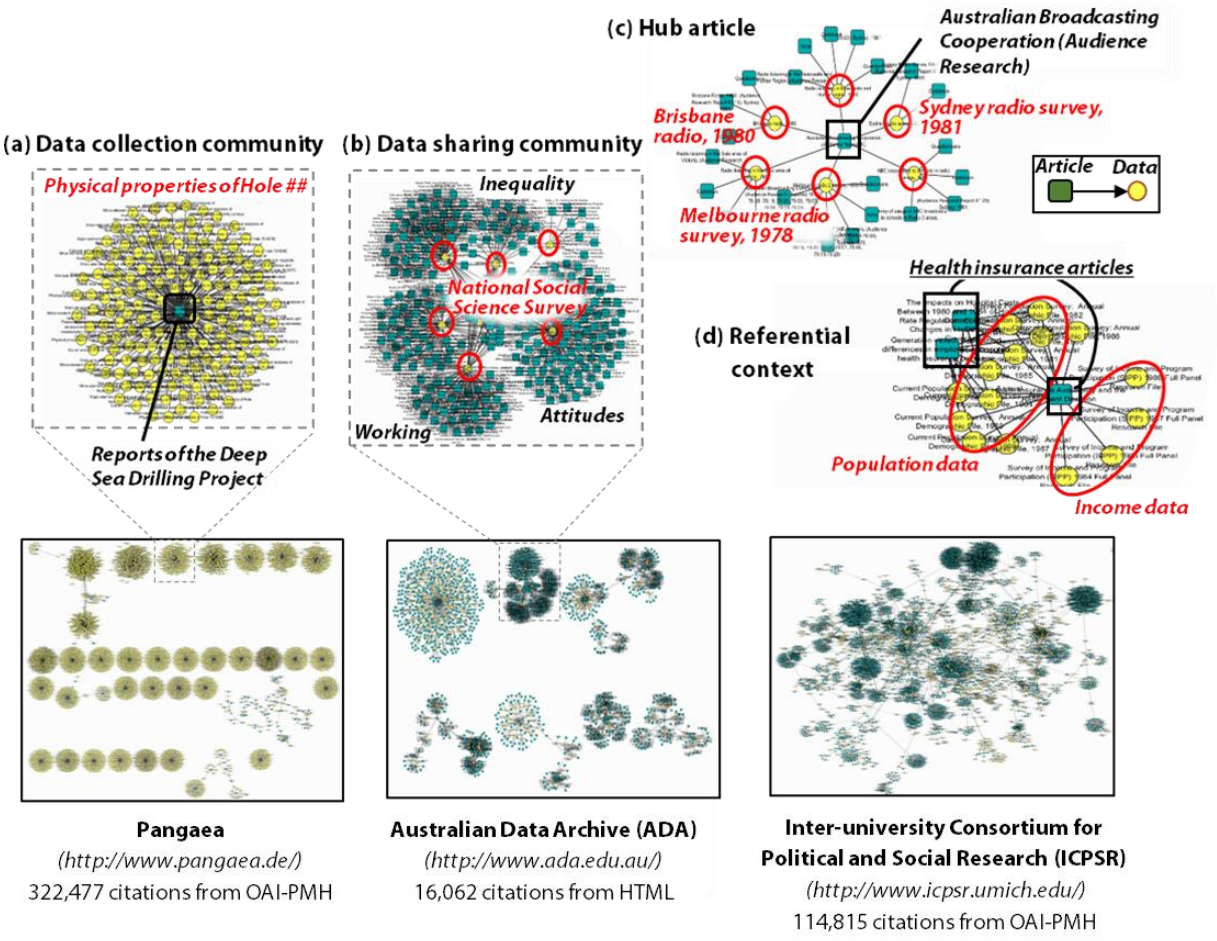

Figure 1: Examples of data citation mining

ACKNOWLEDGEMENTS

This research is partly supported by the data grant of Japan Link Center. I would like to acknowledge the valuable comments and supports of Dr. Yasuhiro Murayama, Prof. Takashi Watanabe, Mr. Yuhei Akahoshi and Dr. Takenari Kinoshita at NICT.

REFERENCES

CODATA-ICSTI Task Group on Data Citation Standards and Practices (2013) Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data. Data Science Journal 12, CIDCR1–CIDCR75.