Korean Diaspora Digital Archive Project:
AI-based integrated database system development and translocality reinterpretation

-- Korean Diaspora Digital Archive System --

1. Overview of Korean Diaspora Digital Archive System (KoDDAS)

The purpose of Korean Diaspora Digital Archive System (KoDDAS) is to provide a digital data management infrastructure for aggregating, organizing, and disseminating Korean Diaspora data for research communities as well as the general public. In that regard, KoDDAS extends the conventional archival functions of collection and preservation to focus on information dissemination and data reuse. To promote data reuse, KoDDAS will implement an integrated online database system that contains not only the raw data (i.e., interview transcripts, photos, videos), but also the contextual (e.g., historical, regional) and structural metadata (e.g., object relationships) that can facilitate the understanding and analysis of Korean Diaspora data. For information dissemination, the project will construct a Korean Diaspora Digital Archive Center (KoDDAC) website where users can search, browse, and download Korean Diaspora data from the KoDDAS database.

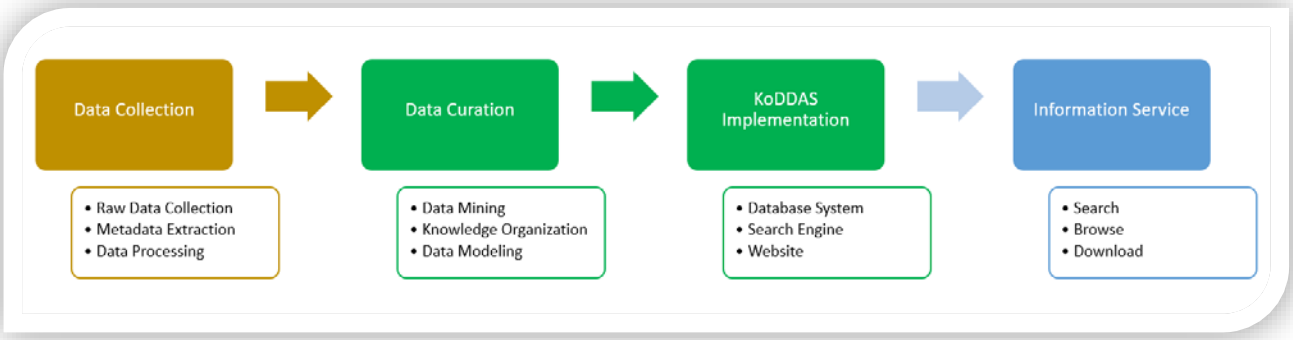

Figure 1. Overview of KoDDAS Construction Process

Figure 1 shows an overview of KoDDAS construction process that starts with the collection and curation of data followed by the implementation of database system, search engine and website to provide a Web-based information service of the Korean Diaspora Digital Archive. The data collection phase includes a data processing step, which entails such tasks as digitization, data archival, and indexing of keywords and metadata. The data curation phase involves an augmented intelligence process, where artificial intelligence (AI) tools will be employed to enhance the data mining and knowledge organization outcomes, both of which will facilitate the design of KoDDAS database and search engine. The third phase entails the implementation of KoDDAS database, search engine and website, integration of which will provide web-based services of search, browse, and download of archived data.

The construction of KoDDAS will not only enable the efficient and effective operation of the Korean Diaspora Digital Archive Center but also serve as a template for constructing other

diaspora digital archive systems by providing a roadmap for transforming unstructured raw data into a multi-faceted archival database system via AI-assisted curation (Rodriguez-Esteban, Iossifov & Rzhetsky, 2006; Yu, Beam & Kohane, 2018). Thus, KoDDAS will provide invaluable services to diaspora researchers, offer an easy access to diaspora data for the general public, and establish a generalizable mechanism to manage, disseminate, and utilize unstructured information. Furthermore, KoDDAS will serve as an example of how cutting-edge information and communication technology can be applied to organize and archive cultural contents for general dissemination and academic analysis.

2. Data Collection

The initial KoDDAS data consists of audio recordings of interviews, video recordings of events, and copies of documents and photographs provided by interviewees. Audio recordings will be transcribed into digital interview documents, after which metadata will be assigned to each content items. The exact metadata construct will require in-depth examination of raw data and careful consideration of KoDDAS objectives, but metadata will include descriptive information that describes the data item, contextual information that informs the context of data item such as location and historical background, and structural information that defines relationships between data items.

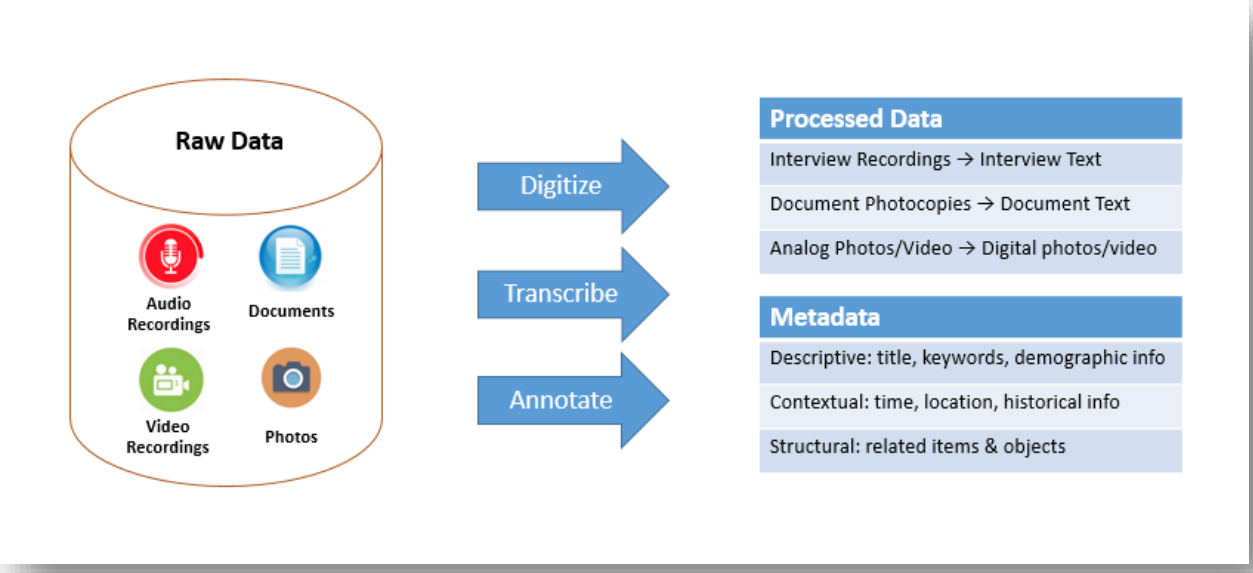

Figure 2. Data Collection Process

3. Data Curation

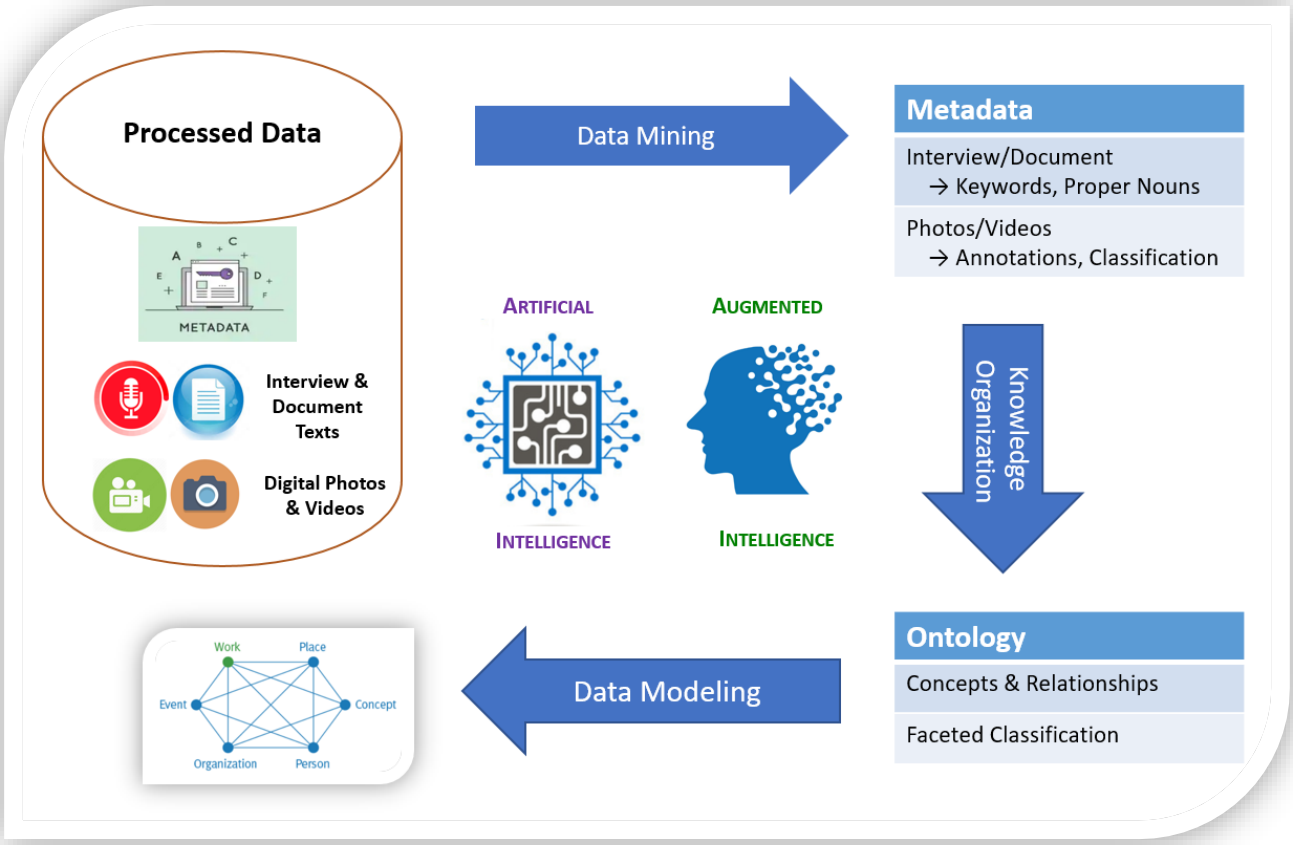

Figure 3. Data Curation Process

The first step of the data curation phase is the extraction of keywords and proper nouns from interview and document texts and annotation and classification of photos and videos via AI-driven data mining (Hofmann & Klinkenberg , 2013; Jungermann, 2009) and machine learning tools (Abadi et al., 2016; Géron, 2017) like Rapid Miner and TensorFlow. Machine-generated metadata will then be analyzed by a team of subject experts to identify a candidate pool of facets and isolates, which will serve as the basis for the faceted classification system (Broughton, 2006; Ranganathan, 1939; Vickery, 2008). In addition, all text data and metadata will be fed into an iterative cycles of semi-supervised learning followed by manual validation to construct a knowledge map of concepts and relationships (Tseng et al., 2007; Tseng et al., 2010).

Thus, data curation phase generates the metadata to enhance searching and ontology to enable browsing and knowledge discovery as well as facilitating the construction of the data model. The data curation phase will also serve as a working example of augmented intelligence application, where artificial intelligence facilitates and enhances the human task of knowledge organization for the purpose of information discovery (Oliver & Harvey, 2017; Lokshin, 2015; Yakel, 2007).

4. Design and Implementation of KoDDAS

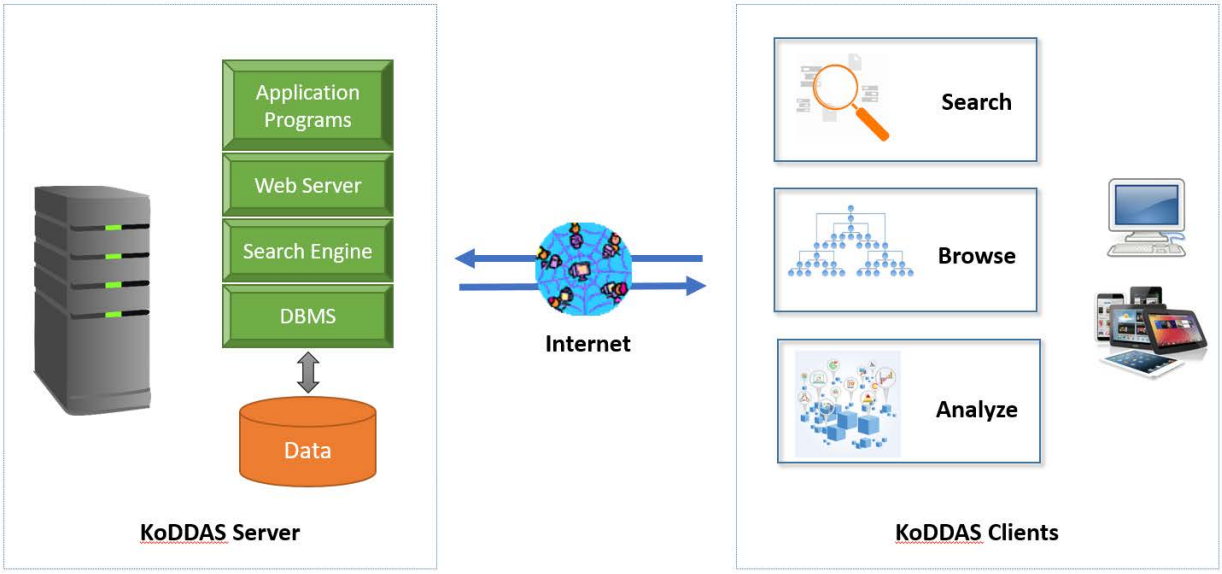

Figure 4. KoDDAS Client-Server Architecture

As shown in Figure 4, KoDDAS is to be implemented in a client-server architecture, where the KoDDAS server will consist of a database management system (DBMS), search engine, Web server, and database application programs to provide search, browse, and data analysis services to KoDDAS clients over the Internet. The data model, a draft of which is shown in Figure 5, is to be implemented in MySQL DBMS to encapsulate both the raw data and metadata in a structured format that will facilitate data storage, retrieval, and analysis. In addition to the conventional relational data model that captures the structural and contextual metadata, DBMS will also house the faceted classification data and concept-relationship knowledge map in the Resource Description Framework (RDF) format (Broekstra et al., 2002; Decker et al., 2000). The search engine will comprise ~~of~~ a database search of all data, a free-text search of interview transcripts and media annotations, and an integrated search that combines both. The database search will be implemented in an application program environment to query the DBMS, whereas the free-text search will index and search interview transcripts in a traditional vector space model framework. In addition to the database search, other application programs (e.g., PHP/MySQL, Node.js) will be constructed to provide the KoDDAS data browsing interface of faceted classification hierarchy and knowledge map in a Liked Open Data (LOD) format as well as the interactive components of the KoDDAS website such as data statistics display by region.

The KoDDAS server, to be housed in a 10-core Linux workstation with 64 gigabytes of RAM and 9 terabytes of hard disk space, will run MySQL (DBMS) and Apache (Web server) to provide an HTML/PHP interface via Web browsers (e.g., Internet Explorer, Chrome) to client devices such as personal computers, tablets and smart phones.

By aggregating data in a centralized server in the form of a relational database, proposed Korean Diaspora Digital Archive Center (KoDDAC) will ensure the preservation and utilization of Korean Diaspora information and become a comprehensive data source for diaspora researchers who seek structured high-quality information enriched with metadata conducive to such tasks as cultural and socio-economic analysis. The search engine, with its free-text search of interviews and documents to find specific persons, records, and historical facts, will supplement the structured search of the database that can pinpoint queries according to field-specific filters such as time, place, and people. For the general public, the KoDDAC website will offer a browse-by-facet (Smith et al., 2006) and knowledge map LOD interface (Aue, Bryl & Tramp, 2014; Karagiannis & Buchmann, 2016) of the Korean Diaspora data as well as a statistical information interface that displays overviews of descriptive statistics, such as diaspora population map, and customized statistics generated from user-specified queries.

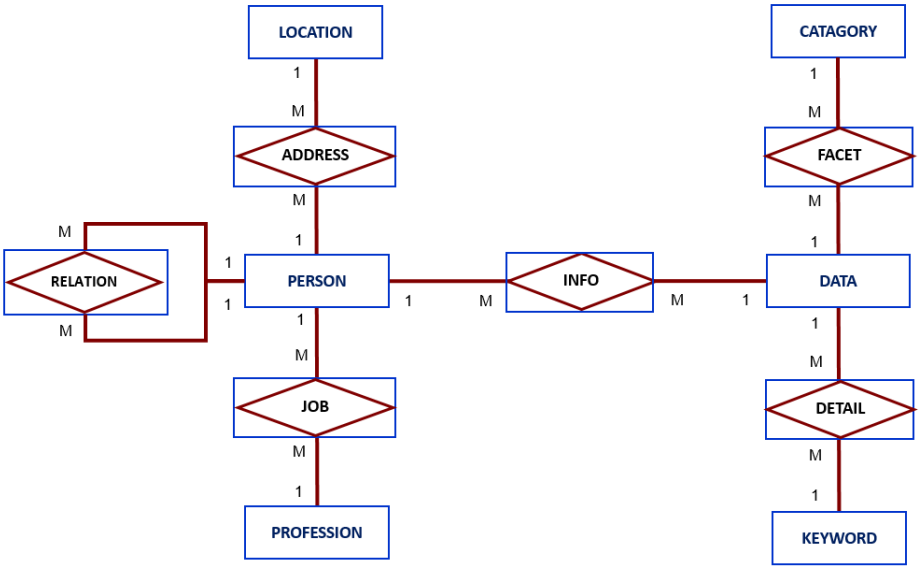

Figure 5. KoDDAS Data Model

Figure 5 displays a tentative data model for KoDDAS. Two key entities in the model are PERSON and DATA. The bridge entity of JOB will hold the profession information of a person over time, ADDRESS will hold living locations of a person over time, and RELATION will store family relationships between people. The DATA entity, which is to store the core data items such as photos, videos, and interview transcripts collected over the years, will be classified into categories in a faceted manner via the FACET entity and assigned a controlled vocabulary of keywords via the DETAIL entity. The INFO entity, which connects PERSON and DATA, will record who collected what data about whom at what date. The model as proposed will enable

searching and browsing by categories and keywords as well as data analysis queries to generate descriptive statistics.

The tentative data model, by representing the core items of Korean Diaspora data (e.g., data, person, profession, location), its metadata (e.g., category, keywords), and data relationships (e.g., relation, info, facet), encapsulates the basic building blocks of the envisioned KoDDAS database. A data archive without a proper data model is nothing but a disorganized data warehouse that hinders data reuse and analysis. Without a well-constructed data model, it will be quite difficult to leverage the Korean Diaspora data for efficient access let along meaningful analysis.

After the initial construction of a KoDDAS prototype, we envision multiple iterations of model refinement to validate and streamline the database component of KoDDAS and progressive implementation of search, browse, and analysis interfaces via PHP/HTML and underlying SQL queries. KoDDAS will also contain a data entry/update interface along with a data import mechanism for batch input processing and a data export mechanism for data backup and distribution in standard formats such as XML.

5. KoDDAS Services

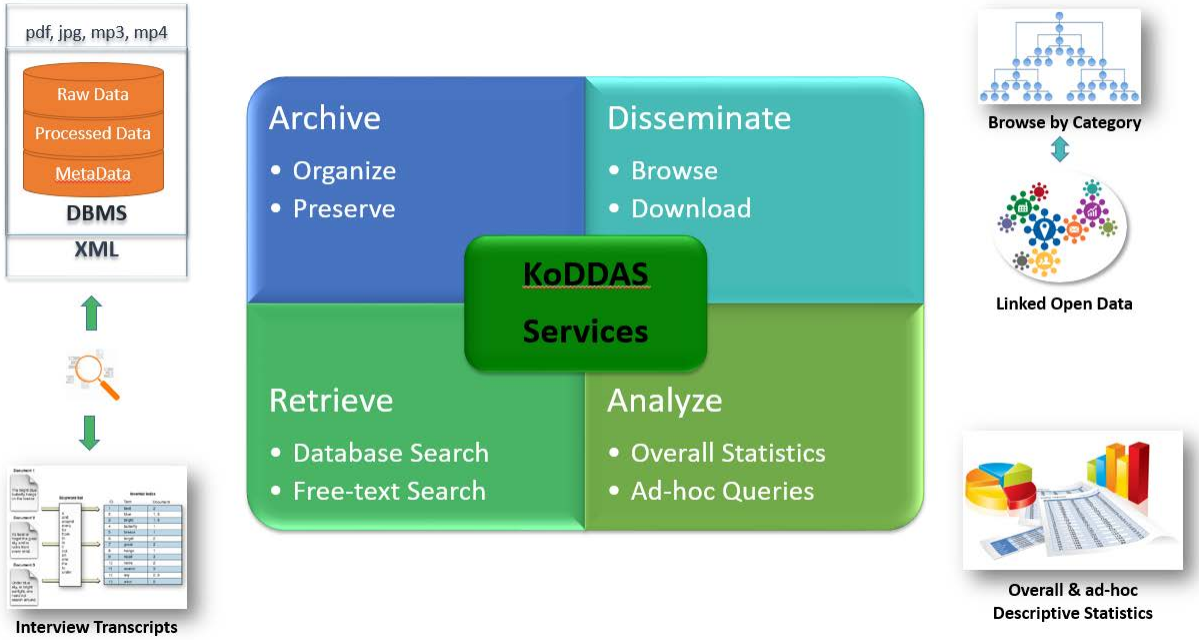

Figure 6. Overview of KoDDAS Services

The key services of KoDDAS can be categorized into data archival, data dissemination, information retrieval, and data analysis. Data archival service includes aggregation, organization, and preservation of data, where the original raw data in physical forms as well as the digitized raw data and metadata in application (e.g., DBMS) and preservation formats

(e.g., XML, PDF) will be stored and maintained in the Korean Diaspora Digital Archive Center. Data dissemination service will provide browsing of KoDDAS data by categories (e.g., time, region, keywords) in a liked open data format that can be downloaded when appropriate. Information retrieval service, consisting of database search and free-text search, will enable the search of interview content as well as field-specific search of the KoDDAS data. Last but not least, data analysis service will offer overall descriptive statistics computed from KoDDAS database as well as customized statistics compiled from ad-hoc queries posed by users.

The construction of KoDDAS as described will contribute to the successful operation of the proposed Korean Diaspora Digital Archive Center, which is to aggregate, organize, preserve, and provide online access to Korean diaspora data in a way that promotes data reuse and facilitate data analysis. The significance of KoDDAS lies in its innovative integration of information technology applications to enable a multitude of services that reach far beyond those provided by conventional archival centers. In that light, the proposed project embodies the cutting-edge vision of multi-disciplinary user-centric research where subject expertise in various areas of social science is pooled together to generate a wide range of information services to a multitude of users.

References

Abadi, M., Barham, P., Chen, J., Chen, Z., Davis, A., Dean, J., & Kudlur, M. (2016). Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)* (pp. 265-283).

Auer, S., Bryl, V., & Tramp, S. (Eds.). (2014). *Linked Open Data--Creating Knowledge Out of Interlinked Data: Results of the LOD2 Project* (Vol. 8661). Springer.

Broekstra, J., Klein, M., Decker, S., Fensel, D., Van Harmelen, F., & Horrocks, I. (2002). Enabling knowledge representation on the web by extending RDF schema. *Computer networks*, *39*(5), 609-634.

Broughton, V. (2006, January). The need for a faceted classification as the basis of all methods of information retrieval. In *Aslib proceedings* (Vol. 58, No. 1/2, pp. 49-72). Emerald Group Publishing Limited.

Decker, S., Melnik, S., Van Harmelen, F., Fensel, D., Klein, M., Broekstra, J., ... & Horrocks, I. (2000). The semantic web: The roles of XML and RDF. *IEEE Internet computing*, *4*(5), 63-73.

Géron, A. (2017). *Hands-on machine learning with Scikit-Learn and TensorFlow: concepts, tools, and techniques to build intelligent systems.*" O'Reilly Media, Inc.".

Hofmann, M., & Klinkenberg, R. (Eds.). (2013). *RapidMiner: Data mining use cases and business analytics applications*. CRC Press.

Jungermann, F. (2009, February). Information extraction with rapidminer. In *Proceedings of the GSCL Symposium'Sprachtechnologie und eHumanities* (pp. 50-61).

Karagiannis, D., & Buchmann, R. A. (2016). Linked open models: extending linked open data with conceptual model information. *Information Systems*, *56*, 174-197.

Lokshin, D. J. (2015). *U.S. Patent No. 8,929,709*. Washington, DC: U.S. Patent and Trademark Office.

Oliver, G., & Harvey, R. (2017). *Digital curation*. American Library Association.

Ranganathan, S. R. (1939). *Colon classification*. Madras Library Association, Madras.

Rodriguez-Esteban, R., Iossifov, I., & Rzhetsky, A. (2006). Imitating manual curation of text-mined facts in biomedicine. *PLoS computational biology*, *2*(9), e118.

Smith, G., Czerwinski, M., Meyers, B., Robbins, D., Robertson, G., & Tan, D. S. (2006). FacetMap: A scalable search and browse visualization. *IEEE Transactions on visualization and computer graphics*, *12*(5), 797-804.

Tseng, S. S., Sue, P. C., Su, J. M., Weng, J. F., & Tsai, W. N. (2007). A new approach for constructing the concept map. *Computers & Education*, *49*(3), 691-707.

Tseng, Y. H., Chang, C. Y., Rundgren, S. N. C., & Rundgren, C. J. (2010). Mining concept maps from news stories for measuring civic scientific literacy in media. *Computers & Education*, *55*(1), 165-177.

Yakel, E. (2007). Digital curation. *OCLC Systems & Services: International digital library perspectives*, *23*(4), 335-340.

Yu, K. H., Beam, A. L., & Kohane, I. S. (2018). Artificial intelligence in healthcare. *Nature biomedical engineering*, *2*(10), 719.

Vickery, B. (2008). Faceted classification for the web. *Axiomathes*, *18*(2), 145-160.