Data Curation Process Evaluation

Loic Niragire

TIM-8130 v4: Data Curation

Dr. Ulrich Vouama

February 4, 2022

Table of contents

Data Curation

Transforming raw data into curated data is a process that involves techniques and algorithms for extracting, classifying, linking, merging, enriching, sampling, and summarization of data and knowledge (Beheshti, Tabebordbar, Benatallah, & Nouri, 2017). This process is aimed at providing contextualized data and knowledge to end-users for the purpose of further analysis. Figure one illustrates common curation tasks involved in transforming raw data into a curated form.

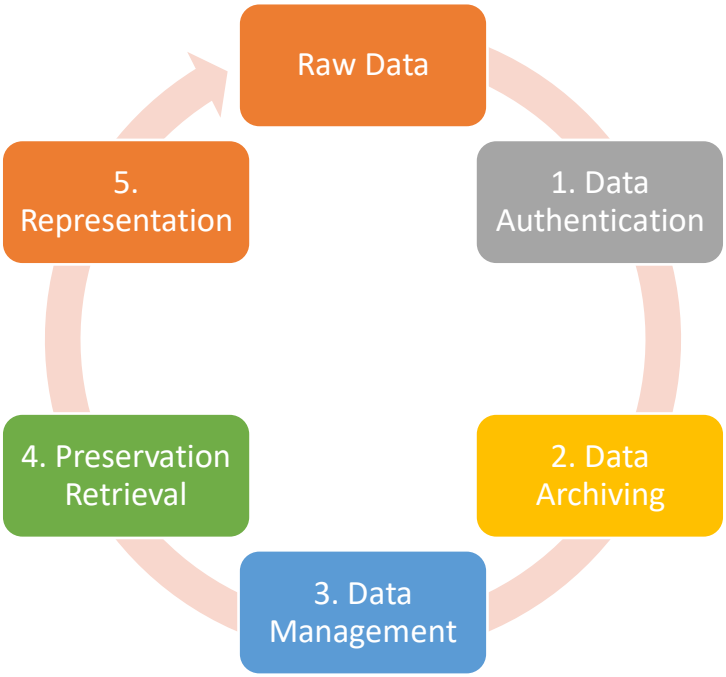

Figure 1 Curation Tasks

Data Authentication

Data authentication and authorization process ensures that resources are accessed by the users in a secure way and that the proper rights have been granted accordingly. To ensure data authenticity and originality, it is necessary to have a mechanism to identify data's original source

(Meng, Meng, & Qiao, 2020). Moreover, data source identification is an effective means of clustering data from the same source.

Data Archiving

Digital archive is a component of the knowledge infrastructure responsible for mitigating access to research data. Archives may be community specific, such as focusing on natural science, or may data type specific such as weather or genome data. Knowledge infrastructure is comprised of research institutions, various artifacts, and a network of people. The process of archiving data, depicted by 'P' in figure two, involves data selection, data tagging, and the work to add contextual information necessary for data interpretation (Borgman, Scharnhorst, & Golsham, 2019).

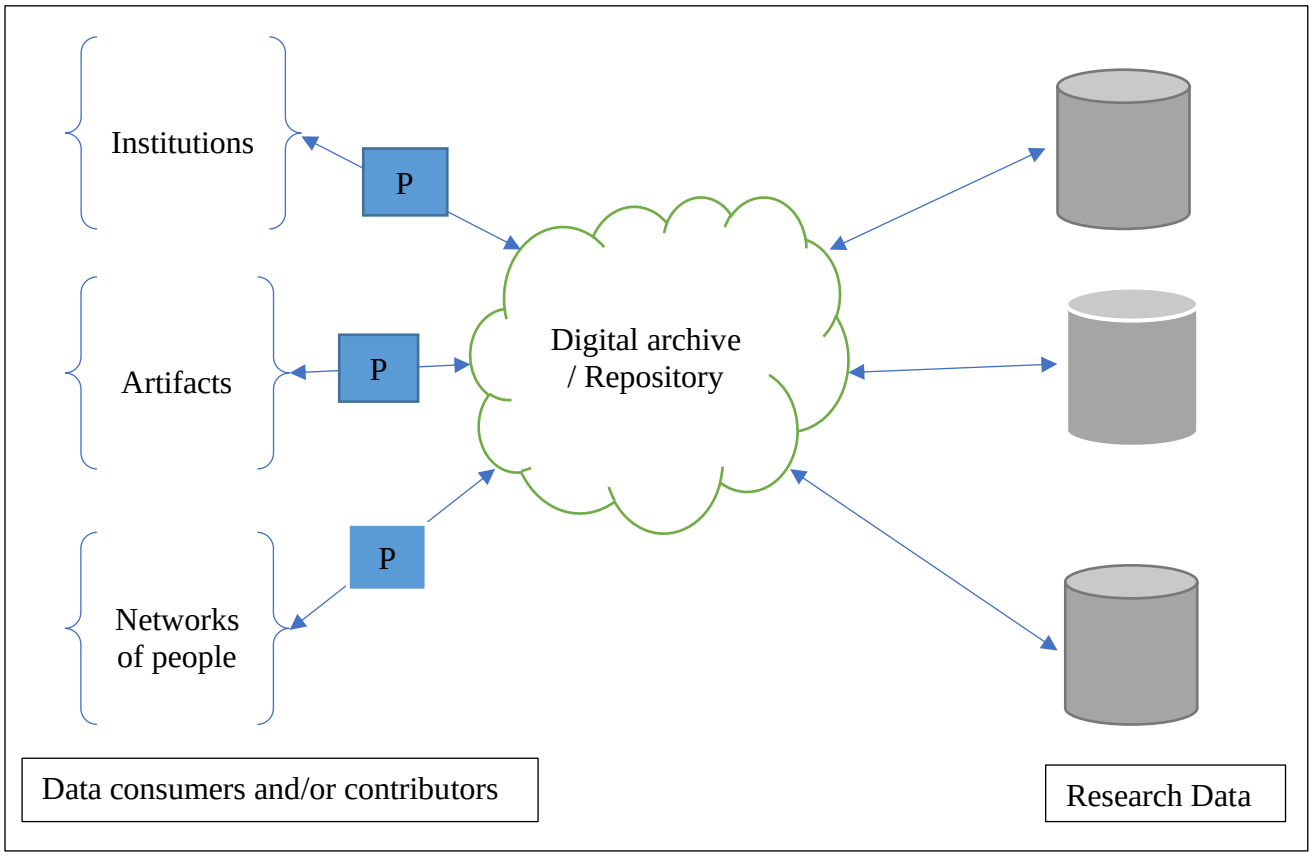

Figure 2 Knowledge infrastructure

Data Management

Data management refers to the adoption of data repositories to facilitate data sharing, discovery, and preservation (Zhang & Chen, 2015). Some of the well-known repositories include *Dryad*, *Dataverse*, *figshare*, *GigaScience*, *OSF* and *Zenodo*, each of which is distinguished by supported data type and format. Another characteristic of data repositories is supported data accessibility and provided integration analysis tools. Most open-public repositories provide Web APIs for data access or support dataset direct download.

Preservation Retrieval

Preservation retrieval is concerned with data reusability over a long period. It is the ability to store and continuously improve data quality for ease to usability in the future. It answers the question of data availability and usability over changes in information technology. This entails more than just the preservation of data but also all software and tools used to process and analyze the data. Thus, metadata collection becomes a vital aspect for better preservation, as they provide sufficient information for data discovery, contextualization, and action.

Representation

Data representation refers to the data layout, or format, of the curated data such as xml, json, or others. Representation plays a crucial role in long-term data preservation as technology changes may turn certain formats unusable. Therefore, it is important to rely on proven formats that are well documented. Additionally, the choice of data format is a consequence of the type of data being preserved and influences data usability.

Importance of Data Curation

Data curation ensures that data is organized, structured and accessible. It is an important aspect of any data driven research or study. Most notably, it gives researchers access to reusable datasets. The process of data curation led to the creation of large national, and international, data repositories to facilitate collaboration on major research issues in various research communities. Curation is a continuous effort to improve and maintain available datasets for ease of discoverability and usability. Moreover, curation allows researchers the ability to reproduce previous study results.

Data Collection

Data repositories serve as a centralized location for all data related access and analysis. They provide researchers easy access to collected data and analysis tools by aggregating data from multiple sources. Moreover, aggregating data into a centralized location help research communities and companies to reduce duplication of effort and related costs. The term *Data Warehouse* is often used to reference large data repositories. Moreover, repositories can also store unstructured, or semi-structured, data tagged with metadata - we refer to such repositories as *Data Lakes*.

Sample Work Summary

This section analyzes the data curation process utilized by *N. Tempini* and *S. Leonelli* in their paper on actionable data for precision oncology. In this paper, they explore how researchers make decisions about actionability of specific datasets and conditions that allow data to be trusted (Tempini & Leonelli, 2021). Moreover, they discuss the efforts involved in maintaining

the Catalogue of Somatic Mutations in Cancer (COSMIC). Figure three below highlights curation process used in their paper and how each task is accomplished.

Curation Task	*How the task is accomplished*
Data Authentication	• Post-doctoral researchers with a background in biology aim to curate all key papers that are published in genes they specialize in. • Papers deemed to be of high quality or promise are prioritized.
Data Management	• Curators receive a daily automated report of all publications about any genes. They read and extract data from papers and input them into COSMIC through web APIs. • Post-doctoral researchers with programming skills download data from repositories and integrated them into COSMIC through scripting code.
Preservation Retrieval	• Updates are released every three months with a new version of the database and the team publishes a summary of its curatorial processes on COSMIC website. • Web browsing • File downloads
Representation	• Genomic data: xml, tsv, json

Figure 3 Curation methods usage

References

Beheshti, S.-M.-R., Tabebordbar, A., Benatallah, B., & Nouri, R. (2017). On Automating Basic Data Curation Tasks. *2017 International World Wide Web Conference Committee (IW3C2)* (pp. 165-169). Perth, Australia: ACM.

Borgman, C., Scharnhorst, A., & Golsham, M. (2019). Digital Data Archives as Knowledge Infrastructures: Mediating Data Sharing and Reuse. *Journal of the association for information science and technology*, 888-904.

Meng, X., Meng, K., & Qiao, W. (2020). A Survey of Research on Image Data Sources Forensics. *AIPR* (pp. 174-179). Xiamen, China: ACM.

Tempini, N., & Leonelli, S. (2021). Actionable data for precision oncology: Framing trustworthy evidence for exploratory research and clinical diagnostics. *Social Science & Medicine*, 1-10.

Zhang, Y., & Chen, H.-l. (2015). Data Management and Curation Practices: The Case of Using DSpace and Implications. *ASIST*, 6-10.