IJDC | *Conference Paper*

Synchronic Curation for Assessing Reuse and Integration Fitness of Multiple Data Collections

Maria Esteva
Texas Advanced Computing Center

Weijia Xu
Texas Advanced Computing Center

Nevan Simone
Oden Institute

Kartik Nagpal
Oden Institute

Amit Gupta
Texas Advanced Computing Center

Moriba Jah
Oden Institute

Abstract

Data driven applications often require using data integrated from different, large, and continuously updated collections. Each of these collections may present gaps, overlapping data, have conflicting information, or complement each other. Thus, a curation need is to continuously assess if data from multiple collections are fit for integration and reuse. To assess different large data collections at the same time, we present the Synchronic Curation (SC) framework. SC involves processing steps to map the different collections to a unifying data model that represents research problems in a scientific area. The data model, which includes the collections' provenance and a data dictionary, is implemented in a graph database where collections are continuously ingested and can be queried. SC has a collection analysis and comparison module to track updates, and to identify gaps, changes, and irregularities within and across collections. Assessment results can be accessed interactively through a web-based interactive graph. In this paper we introduce SC as an interdisciplinary enterprise, and illustrate its capabilities through its implementation in ASTRIAGraph, a space sustainability knowledge system.

Submitted date 2022 ~ *Accepted* date 2022

Correspondence should be addressed to Maria Esteva, 3206 South Oak Drive, Austin, TX, 78704 USA. Email: maria@tacc.utexas.edu

The *International Journal of Digital Curation* is an international journal committed to scholarly excellence and dedicated to the advancement of digital curation across a wide range of sectors. The IJDC is published by the University of Edinburgh on behalf of the Digital Curation Centre. ISSN: 1746-8256. URL: http://www.ijdc.net/

Introduction

As data curation becomes a habitual practice for researchers, there are challenges that need to be resolved in generalizable ways to support data-driven applications. To answer research questions using AI and knowledge systems, researchers may need to integrate data from multiple collections. Each of these datasets may have a unique structure, missing or incorrect information, and have different frequency of updates. While there is often conflicting information between collections that have to be integrated, they can also be complementary, or may have similar data points but labelled differently. Typically, prior to integration, users identify these issues in each individual dataset through time-consuming analyses. Therefore, an outstanding challenge is to support the curation of large datasets from multiple collections (Freitas & Curry, 2016). Instead of a case-by-case solution we developed Synchronic Curation (SC), a general framework with protocols to assess fitness for integration and reuse of many collections simultaneously and across time.

SC involves:

1. A data modelling process;
2. implementation and infrastructure, and;
3. an analysis and comparison module.

It can be used to identify gaps, overlaps, consistency, or contradictions in a single, large, and evolving data collection, or to assess multiple different ones. Researchers and curators can build this framework to diagnose and identify relevant and reliable data for reuse and integration within an area of knowledge.

SC's central component is the data model, which bridges the researchers' questions to curated data across different collections. Because building a data model for an entire area of knowledge can be daunting, we propose modelling data progressively. For efficiency, SC is conducted in relation to problems that researchers are investigating, which leads to selecting and integrating smaller chunks of data at a time. As new research problems are incorporated, the model grows to represent an area of knowledge more comprehensively. For consistency and normalization, the data model is developed using controlled vocabularies.

The analysis and comparison module is the diagnostic component of SC through which curators can monitor collections' trends including updates, growth, and completeness over time. In addition, by identifying gaps, changes, and irregularities in the data, SC allows for the assessment of the quality of the collections. Researchers can use the module to gauge and contrast collections, and to identify which ones have the information that they need and are more reliable for reuse. The data model and the analysis and comparison module can be made available as interactive graphs.

We developed SC within the Texas Advanced Computing Center (TACC) research infrastructure.[1] Originally developed for ASTRIAGraph[2], a knowledge system to explore space sustainability (Slavin et al., 2021), SC has also been applied to the AI Soil Dataset[3]. In this work we demonstrate its capabilities using ASTRIAGraph as a case study.

[1] Texas Advanced Computing Center: http://www.tacc.utexas.edu

[2] ASTRIAGraph: http://astria.tacc.utexas.edu

[3] DIVE Domain Informational Vocabulary Extraction. AI-SOIL Dataset Visualization: http://er.tacc.utexas.edu/datasets/aisoil/aisoil_visualization_lab_data_model

Related Work

At large, curation remains a one-time process for mostly single and static datasets for purposes of publication in open data repositories. Such practice, typically manual and time consuming, does not scale (Lee & Stvilia, 2017). Numerous initiatives are addressing the problem of scaling curation through tooling extensions to data repository software (Weber et al., 2020) and leveraging machine learning applications (Lafia et al., 2021).

Curation of large domain science data collections that are continuously aggregated in databases involve many of the processes that we use in this work; including significant up-front planning, a data model or schema to organize the data, infrastructure where tasks can be automated, and regular revisiting of processes to adjust to data changes and growth (Martin et al., 2021; Pouchard et al., 2019). Our approach differs in that it involves assessment of aggregated data from different collections at the same time. Data quality is fundamental to transparency and balance in AI applications. Our work is inspired by some modules of The Data Nutrition Project, which develops evaluative tools and practices to produce unbiased AI models (Holland et al., 2018).

The focus of many projects to improve data accessibility and usability for access to databases is to map data collections to an ontology (Konstantinou et al., 2008; Rodriguez-Muro et al., 2008; Sequeda & Miranker, 2013). Often, the results are systems extending data query interfaces with selected ontology semantics. Rather than mapping to a fixed set of semantics, SC aims to integrate multiple data collections and mapping them to an evolving semantic data model.

Knowledge graphs have been used to support integrative analysis from multiple data collections in health (Hasan et al., 2020), manufacturing (Buchgeher et al., 2021), and smart city applications (Hryhoruk et al., 2021), linking multiple datasets in a meaningful way so that specific research use cases can be supported. Our work shares this same high-level vision, except that it is not limited to a specific use case. In SC we designed an extendable framework that can be used to connect multiple data collections to multiple research problems.

Data curation is often perceived as a burden both by the scientists that create data, and by data curators (Wallis et al., 2013; Trelogan et al., 2010). In addition, depending on the scientific domain, curation can be a very specialized and contextual process (Borgman et al., 2015). Therefore, interdisciplinary collaboration between scientists and curators is crucial for developing understandable and reusable datasets (Chen & Chen, 2019; Borgman et al., 2015). To develop a data model for curation and publication of simulations in DesignSafe, curators and scientists worked together in the design of categories and metadata tags to organize and describe the resultant datasets (Esteva et al., 2019). A step forward from the former, SC incorporates curation as part of the research process. Researchers, curators and computer scientists develop the SC data model fueled by research questions that are contributed progressively and continuously (Esteva et al., 2020).

The SC Framework

SC requires infrastructure comprising:

1. Storage where raw data collections are preserved;
2. a database where data is ingested and organized according to the data model, and;
3. a computing environment which enables automation and scalability of tasks.

At TACC SC uses both cloud and storage resources. To process the data, we use a NEO4J database.[4] A graph database was chosen because of its flexibility in capturing overlapping

[4] NEO4J Graph Data Platform: https://neo4j.com/

relationships between different entities, as well as different types of relationships between two entities. One of our goals was to identify data changes over time, which in a relational database is more difficult to implement (See Figure 4.).

An example of how capturing complex relationships is facilitated by graph databases is shown in Figure 1. A property of space objects, captured in many data collections in ASTRIAGraph, is "country name". However, across these different sources, the values are not standardized nor reflect continent dependencies. Using an open-source NLP tool that converts country names and provides continent information, we wrote a script that maintains a mapping between the non-standard country names in the NEO4J database and the standard country names provided by the NLP tool. The query is formulated to avoid repetitive information from the multiple sources. This function can also be used by curators to normalize country names across collections and improve data quality. The graph database was also chosen for ease of use. Iterating through the data model by adding classes or properties, or by changing their relationships is easy to do. Additionally, graph databases avoid repetitive data storage which improves performance.

The SC framework has three components:

1. Development of the data model;
2. data processing based on the data model, and;
3. data analysis and comparison module.

2019-072B	SSES-1	SDN
2021-059BT	AYAN-21	RWA
None	Alsat-1B	Algeria
None	AlSat-1N	Algeria
None	Alsat-2B	Algeria
None	Alsat-2A	Algeria
None	NigeriaSat-2	Nigeria
None	Nigeriasat-X	Nigeria
None	ETRSS-1	Ethiopia
None	ZACUBE-1	South Africa
None	ZACUBE-2	South Africa
None	Alcomsat	Algeria
None	TIBA-1	Egypt
2000-046B	Nilesat 102	Egypt
1998-024A	Nilesat 101	Egypt
2007-018A	NigComSat 1	Nigeria
2007-063A	Rascom-QAF 1	Mauritius
2010-037A	Nilesat 201	Egypt
2010-037B	RASCOM-QAF 1R	Mauritius
2011-016A	Intelsat 28 (New Dawn)	Mauritius
2011-077A	NigComSat 1R	Nigeria
2017-070A	MOHAMMED VI-A	MA
2017-086A	ANGOSAT 1	AGO
2018-095A	MOHAMMED VI-B	MA
2021-022AA	CHALLENGEONE	TUN

Figure 1. Results of a query in ASTRIAGraph showing all the satellites that belong to Africa

Development of the Data Model

The data model is a representation of a research space. It is a co-development between the data needs of the users to solve a research problem, and the contents of the different collections that are integrated in the system. It provides a common language to organize and access data from multiple collections within an area of study based on identifying and describing entities as units of analysis, phenomena to observe, and problems to solve. Entities are modelled as classes with

properties and relations between them. The data model is a dictionary of sorts, built through the expertise of researchers that identify units of analysis, observations, and problems. Both classes and properties are labelled using, as much as possible, domain-specific vocabularies.[5] All the terms in the data model are described and maintained in a master dictionary, and their entries are used to normalize field labels across the different collections.

Developing the data model is an interdisciplinary process. It includes domain researchers, information scientists, and computer scientists. As researchers describe their investigation, the information scientist organizes the narrative as classes and properties. Classes are units of analysis, problems or observations, and properties correspond to data fields in the collections that characterize the classes. For example, the class space object (e.g., a satellite), has properties such as country, launch date, model, name, and identifier. To maintain data provenance, we included a collection's description class with properties from the Dublin Core metadata schema. Computer scientists in the team implement the data model in the graph database, devise ingest scripts, connect raw data to storage, and maintain all the infrastructure components.

We converted the data model into a web-based interactive graph that users can consult to learn what datasets are available in ASTRIAGraph and how field labels within those collections map to the terms in the master data dictionary.[6] The graph displays each of the collection's sources, the classes and the properties, and the relations between all (See Figure 2). Gradually, the data model can become a schema or an ontology about an area of study.

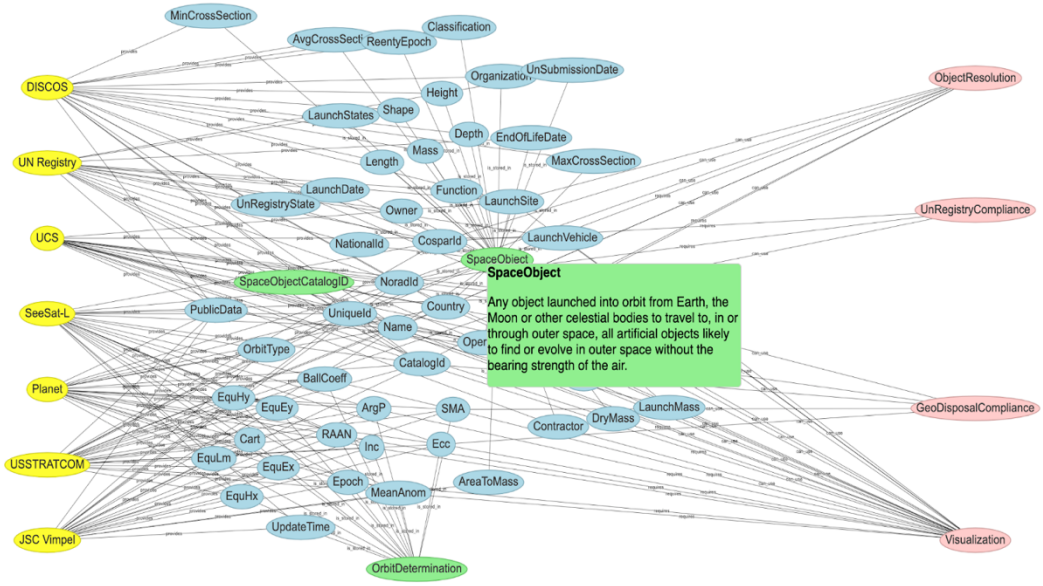


Figure 2. ASTRIAGraph data model highlighting the class SpaceObject (green node) with corresponding definition and related properties (blue nodes). Research problems are also classes (pink nodes), and we can see the relationships between classes and properties across seven data collections (yellow nodes).

Data Processing Based on the Data Model

The process to integrate and store data from each collection in a graph database is accomplished through the following steps:

Extraction

Fields from the files of each data collection are scraped and coalesced as field-value pairs. Additionally, the field labels are gathered in a list. The curator and the domain researcher

[5] Unified Astronomy Thesaurus, American Astronomical Society: https://astrothesaurus.org/

[6] Synchronic Curation Data Model: http://astriaservices.tacc.utexas.edu/liveschema

review them against the terms in the data model to match each label to a corresponding term. This is codified into a set of rules and developed into a script used in the translation step.

Translation

The process addresses the variety of file structures of the different collections by establishing the formatting, and by normalizing the labels to the terms defined by the data model and maintained in the master data dictionary. This process is completed for each data collection that is ingested to the system.

Ingest

All translation functions are invoked from a central ingest script which is run daily to check for and process newly received files. This schedule prompts the recording of the times of data gathering and of ingestion to SC so that data can be compared across time. Data is then ingested to the graph database where each collection is represented by a class node. Data integration happens through the data model, as fields from different collections are mapped to its properties.

Collections Analysis and Comparison Module

Developing the analysis and comparison module involves both automated and human review steps. These steps are conducted on a continuous basis, as data from some collections are frequently updated, and new collections are ingested in the framework. The goal is to achieve a protocol that allows comparing multiple data collections consistently.

Analysis

During this step we gather statistics to evaluate the quality of the collections. Data points corresponding to each field in each collection are counted to determine how often they are populated, and to estimate their relevance (i.e., if a field is seldom populated it can be ignored). A report listing data points by field is also generated to record characteristics of their values, such as if they are numerical, textual, or alphanumeric. Data is also processed to identify misspellings and to determine whether existing problems can be corrected. The translation step is automated, and the results are reviewed by the curator. At this step, curators can identify the need for a solution like the one featured in Figure 1. to normalize and group terms using NLP tools to improve data quality.

Comparison

Data can be compared at any time between different ingests or versions of one collection, and between different collections. A scripted comparison tool generates a 3D matrix from which 2D slices can be analyzed. These are used to do pairwise comparisons to identify gaps, differences, and irregularities between collections. Reports are generated through queries to the graph database. Results of predetermined queries can be displayed in the interactive graph where users can select properties to compare (See Figure 3).

SC Assessments

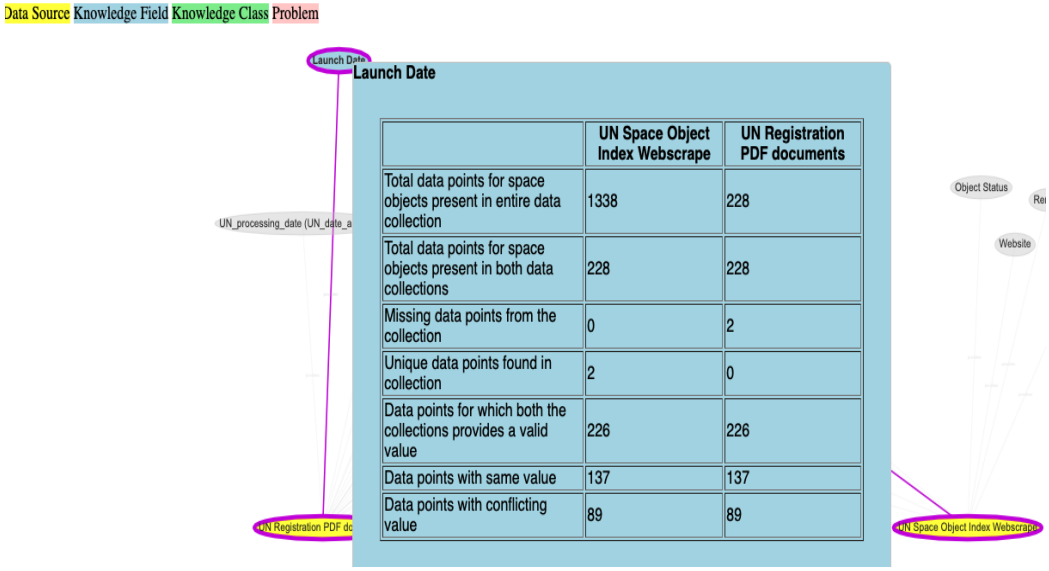
We use the SC implementation in ASTRIAGraph to illustrate how curators can identify gaps, differences, and irregularities within and across data collections.[7] Below are examples of SC assessments.

[7] ASTRIAGraph (http://astria.tacc.utexas.edu/AstriaGraph/) ingests data from multiple private and public data providers. It has agreements with private data providers that allow some data to be made public and reusable. For this work we queried the publicly available data ingested to ASTRIAGraph whose sources are clearly identified.

A curator may need to evaluate consistency between two versions of a collection to decide which one is more reliable for reuse. Figure 3. shows the interactive graph[8] with the comparisons of two versions of a collection about space objects registered with the United Nations Office for Outer Space Affairs (UNOOSA).[9] The versions are:

- The Outer Space Objects Index [10] and;
- the Space Object Registry.[11]

Each version has the same temporal coverage and source of information, but is processed and displayed differently, and has different levels of detail. In one version, data is displayed in a table form and in the other as individual pdf. files.



	UN Space Object Index Webscrape	UN Registration PDF documents
Total data points for space objects present in entire data collection	1338	228
Total data points for space objects present in both data collections	228	228
Missing data points from the collection	0	2
Unique data points found in collection	2	0
Data points for which both the collections provides a valid value	226	226
Data points with same value	137	137
Data points with conflicting value	89	89

Figure 3. Comparison of the data property "Launch Date" between two versions of a data collection with information about space objects registered with the United Nations Office of Outer Space Affairs.

Using the analysis and comparison module in the interactive graph, a curator can select any property and review the completeness of the corresponding values in comparison to another version or to another collection. In this case one version has more information about Launch Date, a property of the class Space Objects than the other, and there are 89 instances in which the values for Launch Date do not coincide between the two versions. The data resulting from this analysis is publicly available in the ASTRIAGraph Dataverse at the Texas Data Repository (Esteva et al., 2022).

In another case, a curator may need to know if data values are recorded consistently in a frequently updated collection. For this, the ASTRIAGraph database was queried to obtain a report about how the Cospar ID field of the Space Object class was populated over time across the USSPACECOM collection with ~18,000 reported space objects (See Figure 4). The Cospar ID is a unique identifier for space objects and thus should not change. The collection was

[8] Synchronic Curation: Data Analysis and Comparison Module: http://astriaservices.tacc.utexas.edu/liveschema/dataset_comparison

[9] United Nations Register of Objects Launched into Outer Space (UNOOSA): https://www.unoosa.org/oosa/en/spaceobjectregister/index.html

[10] UNOOSA Outer Space Objects Index: https://www.unoosa.org/oosa/osoindex/index.jspx?lf_id=

[11] UNOOSA National Space Objects Registry: https://www.unoosa.org/oosa/en/spaceobjectregister/national-registries/index.html

updated 600 times during 2019 and cases of gaps in reporting, changing values, and removal of IDs during updates were identified.

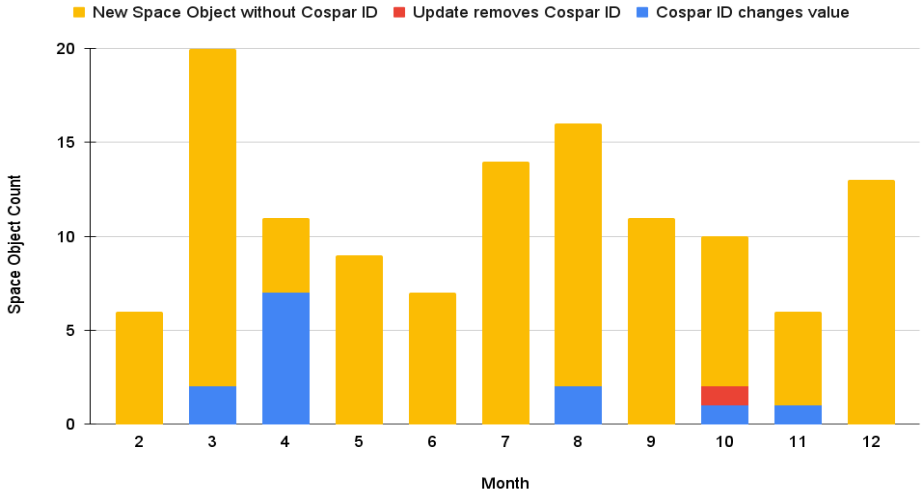


Figure 4. Gaps and irregularities in the property "COSPAR ID" reported by the USSPACECOM collection that was regularly updated during 2019.

Comparisons between collections that are frequently updated are useful both for curators and researchers to contrast their scope and content, and identify their differences. Querying the graph database, we obtained results about three collections that report similar fields about space objects during 2019. In Table 1. below we observe that collection 1 reports about a large quantity of space objects, while collections 2 and 3 only report about a subset.

Table 1. Assessment of three collections showing coverage of space objects data in 2019.

Collections during the year 2019	Number of space objects reported	Number of space objects not always reported	Number of days without collection updates
1.USSPACECOM	18.168	760	2
2.Planet Labs [12]	215	0	3
3. SeeSat-L[13]	191	191	5

Collection 2 consistently reports on the same 215 objects, and did not receive updates for only 3 days. Collection 3 is the least reliable, it has gaps in reporting for each space object at some point and the most days without collection updates. Mapping field names to a common property in the data model allows comparisons across multiple different collections.

A useful assessment to decide which collection to use or how collections complement each other is to compare the completeness of values across similar fields at a given period. Figure 5. shows such a comparison between two collections. The results of a query for similar fields for each collection is expressed as probabilities ranging from 0 - never reported values - to 1 - always reported values. In this case collection 0 (USSPACECOM) is more complete and thus

12 Planet: https://www.planet.com/

13 SeeSat-L HomePage: http://www.satobs.org/seesat/seesatindex.html

more reliable than 1 (UNOOSA), as it always reports values for the compared fields. Again, the comparison is possible because the fields' labels have been normalized as properties in the data model. The comparison can be extended to multiple data collections.

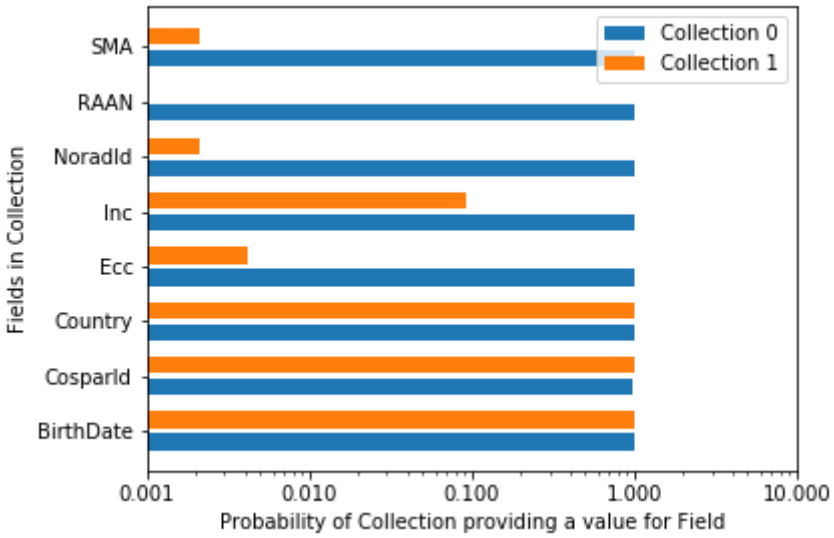


Figure 5. Collection 0 is more reliable than collection 1 because it always reports values for the compared fields.

Conclusions

SC is a framework that can be implemented to curate data collections to solve multiple research use cases in different scientific fields. SC fills an urgent need in data driven research that requires usage of large and diverse data collections. To reuse data, the first step is to assess its quality and its fitness to address the research use case at hand. SC proposes modelling data collections to research questions to enable targeted analyses and comparisons that can help users identify which collections are more reliable and adequate to solve them. Importantly, SC enables curators and researchers to assess multiple datasets at the same time.

SC is flexible and scalable. As more data is integrated in a system, it is possible to tackle more research questions, and to design more assessments about the curation status of the ingested collections without losing sight of their provenance. The framework is not designed as a turn-key solution for a specific use case. By merging the data model and semantics, the framework can be extended with additional data and research use cases. New research cases may demand adding classes and properties to the data model, expanding the dictionary, and mapping to fields from existing or new collections depending on the user's data needs. As existing collections grow and new ones are added, conflicting and missing information across them can be identified through the analysis and comparison module. Such information is essential for curators to evaluate the quality of the data and for researchers to decide if data is reliable for reuse. Once issues of gaps, changes and irregularities are diagnosed, curators and researchers can further investigate what happened and decide how to address them. SC connects data curators and domain users so that data management is coupled with data usage in targeted research projects.

References

Borgman, C. L., Darch, P. T., Sands, A. E., Pasquetto, I. V., Golshan, M. S., Wallis, J. C., & Traweek, S. (2015). Knowledge Infrastructures in Science: Data, Diversity, and Digital libraries. *International Journal on Digital Libraries*, *16*(3–4), 207–227. https://doi.org/10.1007/s00799-015-0157-z

Buchgeher, G., Gabauer, D., Martinez-Gil, J., & Ehrlinger, L. (2021). Knowledge Graphs in Manufacturing and Production: A Systematic Literature Review. *IEEE Access*, 9. https://doi.org/10.1109/ACCESS.2021.3070395

Chen, S., & Chen, B. (2019). Practices, Challenges, and Prospects of Big Data curation: A Case Study in Geoscience. *International Journal of Digital Curation* 14(1). https://doi.org/10.2218/ijdc.v14i1.669

Esteva, M., Jansen, C., Arduino, P., Sharifi-Mood, M., Dawson, C. N., & Balandrano-Coronel, J. (2019). Curation and Publication of Simulation Data in DesignSafe, a Natural Hazards Engineering Open Platform and Repository, *Publications,* 7(3), 51. https://doi.org/10.3390/publications7030051

Esteva, M., Xu, W., Simone, N., Gupta, A., & Jah, M. (2020). *Modeling Data Curation to Scientific Inquiry: A Case Study for Multimodal Data Integration*. Proceedings of the JCDL'20 ACM/IEEE Joint Conference on Digital Libraries (pp. 235-242). https://doi.org/10.1145/3383583.3398539

Esteva, M., Xu W., Simone, N., Nagpal, K., Gupta, A., Jah, M. (2022). *Replication Data for: Synchronic Curation for Assessing Reuse and Integration Fitness of Multiple Data Collections*. Texas Data Repository. https://doi.org/10.18738/T8/OOTALX V1

Freitas, A, & Curry, E. (2016). Big Data Curation. In, Cavanillas, J. M et al. (Eds.), New Horizons for a Data-Driven Economy: A Roadmap for Usage and Exploitation of Big Data in Europe, *Springer International Publishing* (pp. 87-118). https://doi.org/10.1007/978-3-319-21569-3_6

Hasan, S. S., Rivera, D., Wu, X. C., Durbin, E. B., Christian, J. B., & Tourassi, G. (2020). Knowledge Graph-Enabled Cancer Data Analytics. *IEEE Journal of Biomedical and Health Informatics,* 24(7), 1952-1967. https://doi.org/10.1109/JBHI.2020.2990797

Holland, S., Hosny, A., Newman, S., Joseph, J., & Chmielinski, K. (2018). The Dataset Nutrition Label. Data Protection and Privacy, *Volume 12: Data Protection and Democracy, 12, 1.* http://arxiv.org/abs/1805.03677

Hryhoruk, C. C., Leung, C. K., Wen, Y., & Zheng, H. (2021). *Smart City Transportation* Data *Analytics with Conceptual Models and Knowledge Graphs*. Proceedings of the IEEE Smart World, Ubiquitous Intelligence & Computing, Advanced & Trusted Computing, Scalable Computing & Communications, Internet of People and Smart City Innovation (pp. 455-462). https://doi.org//10.1109/SWC50871.2021.00068

Konstantinou, N., Spanos, D. E., & Mitrou, N. (2008). Ontology and Database Mapping: A Survey Of Current Implementations And Future Directions. *Journal of Web Engineering,* 001-024. https://dl.acm.org/doi/10.5555/2011262.2011263

Lafia, S., Thomer, A., Bleckley, D., Akmon, D., & Hemphill, L. (2021). Leveraging Machine Learning to Detect Data Curation Activities. Proceedings of the IEEE 17th International Conference on eScience (pp. 149-158). https://doi.org/10.48550/arXiv.2105.00030

Lee, D.J.,& Stvilia, B. (2017). Practices of Research Data Curation In Institutional Repositories: A Qualitative View from Repository Staff. *PLoS ONE* 12(3): e0173987. https://doi-org/10.1371/journal.pone.0173987

Martin, A., Chazeau, C., Gasco, N., Duhamel, G., & Pruvost, P. (2021). Data Curation, Fisheries, and Ecosystem-Based Management: The Case Study of the Pecheker Database. *International Journal of Digital Curation,* 16(1). https://doi.org/10.2218/ijdc.v16i1.674

Pouchard, L., Kleese van Dam, K., & Campbell, S. I. (2019). Experimental Data Curation at Large Instrument Facilities with Open Source Software. *International Journal of Digital Curation*, 14(1). https://doi.org/10.2218/ijdc.v14i1.637

Rodriguez-Muro, M., Lubyte, L., & Calvanese, D. (2008). *Realizing Ontology Based Data Access: A Plug-in for Protégé.* Proceedings of the IEEE 24th International Conference on Data Engineering Workshop (pp. 286-289). https://doi.org/10.1109/ICDEW.2008.4498333

Sequeda, J. F., & Miranker, D. P. (2013). Ultrawrap: SPARQL Execution on Relational Data. *Journal of Web Semantics,* 22, 19-39. https://doi.org/10.1016/j.websem.2013.08.002

Slavin, M., Wood, D., Jah, M. (2021). *Use of ASTRIAGraph to Inform Detectability, Identifiability, and Trackability Metrics for Space Sustainability.* Proceedings of ASCEND 2021. American Institute of Aeronautics and Astronautics. https://doi.org/10.2514/6.2021-4088

Trelogan, J., Rabinowitz, A., Esteva, M. & Pipkin, S. (2010). *What Do We Do with the Mess? Managing and Preserving Process History in Evolving Digital Archaeological Archives*. In Contreras, F., and Melero, F.J. (eds.), Proceedings of the 38th Conference on Computer Applications and Quantitative Methods in Archaeology, Granada, Spain, April 6 – 9.

Wallis, J. C., Rolando, E., & Borgman, C. L. (2013). If We Share Data, Will Anyone Use Them? Data Sharing and Reuse in the Long Tail of Science and Technology. *PLOS ONE*, *8*, e67332. https://doi.org/10.1371/journal.pone.0067332

Weber, J., Karcher, S., & Myers, J. (2020). Open Source Tools for Scaling Data Curation at QDR. *The Code4lib Journal*, (49). https://journal.code4lib.org/articles/15436