Knowledge Graph Enabled Scientific Data Repositories

Paulo Pinheiro[1], Henrique Santos[2,*], James Masters[3], Matthew Johnson[2], Jeanette A. Stingone[4], Sofia Bengoa[3], Marcello Bax[5] and Deborah L. McGuinness[2]

[1]*Instituto Piaget, Almada, Portugal*
[2]*Rensselaer Polytechnic Institute, Troy, NY, United States*
[3]*Icahn School of Medicine at Mount Sinai, New York, NY, United States*
[4]*Columbia University, New York, NY, United States*
[5]*Universidade Federal de Minas Gerais, Belo Horizonte, MG, Brazil*

Abstract

Most scientific data repositories have minimal capabilities for integrating data within studies and even less for supporting data harmonization across multiple studies. To prepare data for publication or analysis, it must be organized, normalized, and harmonized to allow the production of high-quality datasets for dissemination and reuse. The Findable, Accessible, Interoperable, Reusable (FAIR) principles have proven to be a key benchmark for scientific data, laying out the foundations to support a more straightforward way to accomplish these integration challenges in hybrid settings. The Human-Aware Data Acquisition Infrastructure (HADatAc) provides data repository software that uses FAIR principles to build and expose comprehensive knowledge, referred to as scientific knowledge graphs (SKG), using scientific data, data dictionaries, and study documentation. HADatAc employs metadata templates to capture the semantics of studies and systematically represents the scientific knowledge as RDF triples by annotating data points with community-built ontologies, providing users with features such as data browsing, faceted search, data summarization, and dataset generation. HADatAc has been used extensively in several National Institutes of Health and IBM-funded efforts, as well as across higher education institutions in the United States, Brazil, Portugal, and Canada, to support scientific data management and sharing.

1. Introduction

Scientific data integration and management pose challenges for scientists due to the large amount and diversity of data. New techniques and high-speed data generation tools have sparked an information revolution in scientific data management [1]. Scientists often face limitations when contributing to or retrieving data from repositories, and may need to adjust their approach based on what the repositories can offer [2]. For example, a repository may allow a scientist to select Alzheimer's studies from a collection of mental-related studies, but may not provide the option to restrict data download to specific variables.

Sci-K'24: 4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, November 12, 2024, Baltimore, MD
oliveh@rpi.edu (H. Santos)
0000-0001-8469-4043 (P. Pinheiro); 0000-0002-2110-6416 (H. Santos); 0000-0002-7975-4387 (J. Masters); 0000-0001-5212-8100 (M. Johnson); 0000-0003-3508-8260 (J. A. Stingone); 0009-0002-4952-2427 (S. Bengoa); 0000-0003-0503-3031 (M. Bax); 0000-0001-7037-4567 (D. L. McGuinness)

Existing semantic technologies, especially knowledge graphs (KGs) [3], are well-suited to bridge the gap between scientists' data requirements and data repositories' data operation capabilities. They can enable scientists to rely less on ad-hoc extract, transform, and load (ETL) tools/scripts and more on a data repository's semantic capability to acquire data tailored to their needs. The employment of such technologies has been focused on dataset annotation [4], with few approaches partially covering studies' and data semantics [5].

The Human-Aware Data Acquisition Infrastructure (HADatAc) is a semantic data repository for managing scientific data acquired through multiple sources including instruments, sensors, humans, and computer models. HADatAc builds a scientific knowledge graph (SKG) from studies' metadata, measurement data in datasets, and data dictionaries. HADatAc's contextual knowledge describes why data was acquired, how they came to be, and what are the many decisions that may have affected data quality during their acquisition. Contextual knowledge includes descriptions of study properties to support scientific activities, rationales for building sensing capabilities to support observations and experiments, and characterization of entities along with their quantities and qualities used to annotate acquired data.

We demonstrate HADatAc in the context of the National Institutes of Environmental Health Sciences (NIEHS)-funded Human Health Exposure Analysis Resource (HHEAR) program, in which it uses contextual knowledge to improve the way data are further analyzed through the use of analytics and visualization solutions. Further, we provide a brief assessment of HADatAc's adherence to the Findable, Accessible, Interoperable, Reusable (FAIR) [6] principles.

2. The Human-Aware Data Acquisition Infrastructure

HADatAc is a semantic scientific data repository based on the Resource Description Framework (RDF) that uses a web application to enable users to access and use its underlying SKG. Figure 1 shows a high-level view of the infrastructure's architecture. The *front-end* is composed of six web components (Section 2.1). The *back-end* (Section 2.2) is decomposed into *core components* that are responsible for storing the underlying SKG, a *content ingestion* component that is responsible for annotating and moving content from data sources into the SKG, and the *HAScO API* is responsible for providing a standardized way of manipulating the SKG. The top of the figure shows the types of instrument-provided content fed into HADatAc used to build the SKG, such as data files and data streams; supporting ontologies providing community-curated knowledge (Section 2.3); and scientist-generated metadata templates (Section 2.4).

2.1. HADatAc Web Interface

HADatAc's front end allows users to interact with the SKG and provides user interface (UI) capabilities to build, use, and share data. The UI can be described in terms of six subsystems. Three of these subsystems provide scientists with ways to search, find, select, and retrieve content from the SKG. The *Data Value Faceted Search* and the *Study Faceted Search* subsystems provide ways for scientists to find and select content from the SKG. The Dataset Generation subsystem provides the capability to build normalized data from faceted search selections. The *Data Source Management* subsystem is the main way for adding content into HADatAc: this subsystem is responsible for feeding content into the Content ingestion component and

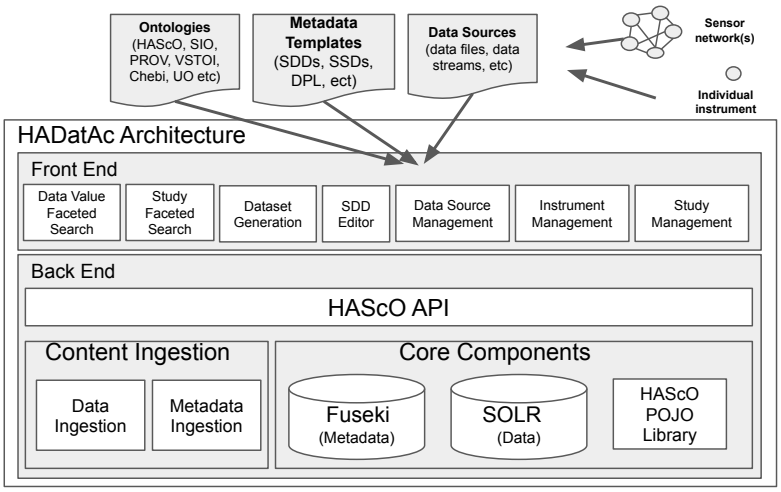

Figure 1: HADatAc architecture with data and metadata flow.

showing users whether content ingestion occurred successfully or not. In case of failure, a log is created for each data source to identify ingestion problems. The *Study Management* subsystem allows scientists to inspect cohort information and verify that study specifications are ingested correctly. Similarly, the *Instrument Management Subsystem* allows scientists to verify if instruments and supporting detectors are properly configured, if they are deployed or not, and on which platforms they are deployed. This is a key way of verifying if standard operating procedures are being properly followed when using instruments to collect scientific data. The *SDD-Editor* [7] is a feature that mimics the look and feel of a spreadsheet editor and loads in a Semantic Data Dictionary (SDD) [8] mapping file, enabling users to create or modify mappings between study data and ontology classes [9].

2.2. HAScO API, Core Components and Content Ingestion

The HAScO (Human-Aware Science Ontology) [10] API supports the storage of large volumes of scientific data and the comprehensive description of entities that compose a scientific study. The infrastructure provides core features in support of scientific data repositories, including a hybrid storage approach: a search engine/NoSQL database (Apache SOLR [11]) and an RDF triple-store (Apache Jena Fuseki [12]). The SKG created by HADatAc uses the RDF model and persists over this hybrid approach, relying on scalable NoSQL to store data and flexible triple stores to store metadata. Since all content follows the RDF model, the elements in both repositories are logically connected through the use of object properties. The union of these elements constitutes the overall knowledge graph.

2.3. Foundational Ontologies

For encoding knowledge about scientific studies, HADatAc's semantic data ingestion leverages several science ontologies including the Human-Aware Science Ontology (HAScO) [10], Se-

manticscience Integrated Ontology (SIO) [13], and Human Health Exposure Analysis Resource (HHEAR) [14]. HAScO is used for encoding knowledge about studies, study types, and data elicitation from human subjects. SIO is used for encoding knowledge about entities and their properties.

Scientific ontologies, like SIO (and others), tend to be domain-agnostic and used with the exclusive purpose of describing the data. They don't intend to support the process of the data throughout the scientific life cycle but are dependent on some data organizational needs and circumstances specified during the study design phase. HAScO supports scientific data organization by defining the notion of collections (SampleCollection and TimeCollection) and groups (SubjectGroup) of objects of interest and their relations.

Other ontologies generally used are W3C PROV Ontology [15] for encoding provenance knowledge and the Virtual Solar-Terrestrial Observatory (VSTO) [16] for encoding knowledge about instruments and platforms.

2.4. Metadata Templates for Knowledge Capture

HADatAc captures scientific knowledge through the use of metadata templates (MTs). MTs provide a framework for domain experts to identify and define the semantics of the study elements, including study metadata, study object collections (e.g., cohorts), roles that object collections play in studies (e.g., subjects, samples), object properties, and relationships among object collections. MTs are encoded in a tabular format, and when interpreted by HADatAc, each row will be translated to RDF resources in the SKG according to the purpose of the MT and the values in such row. Within each MT table, the column *hasUri* is used to inform the URI a row will be mapped to in the knowledge graph.

Study (STD) specification: Studies are where data acquisition activities designed by humans, i.e., scientists and engineers, are planned and executed. These activities acquire data that, once analyzed, should be able to answer scientific questions. STD specifications are used to capture and preserve knowledge from humans regarding their aims, scientific questions, and data acquisition activities. One important property of an STD is the nature (or type) of the study, which can be an observation, an empirical experiment, a computational experiment, or a combination of those.

Deployment (DPL) specification is used to comprehensively describe the measurement infrastructure of a study. A DPL has several tables to capture metadata about the data acquisition infrastructure of the study, including instruments, detectors attached to instruments, and platforms where instruments are deployed. In addition, the DPL allows scientists to define deployments to state the combinations of the aforementioned elements in which data has been acquired.

Semantic Study Design (SSD) describes study objects known at the time studies are designed. An SSD describes a scientific study in terms of its objects and object collections. Data in a study are values of an object's properties. In order to properly organize study data we need to be able to describe the study in terms of its objects.

Semantic Data Dictionary (SDD) describes the meaning of values in terms of objects and their properties. SDDs are composed of objects and attributes. Attributes are used to specify object properties including relationships among objects. In terms of spatial knowledge, objects

can be used to represent locations. For temporal knowledge, objects can be used to represent events and time instants.

Stream specification (STR) identifies if data are acquired as a stream of data files or messages. A data file stream can be composed of a single file. In addition to specifying the source of the data, streams also identify data ownership, data privacy, the deployed instrument used to collect the data, and, more importantly, from a data acquisition point of view, which SDD is used to ingest the data from files and messages into the KG.

When MTs and data files are ingested, HADatAc constructs an SKG with all the metadata, as seen in Figure 2. The SKG has five larger areas of knowledge representation: (A) Scientific activities as specified in STD, DPL, and STR MTs; (B) Instruments as specified in DPL MT; (C) Data schemas as specified in SSD MT; (D) Object collections as specified in SSD MT; (E) Semantically-annotated data repository.

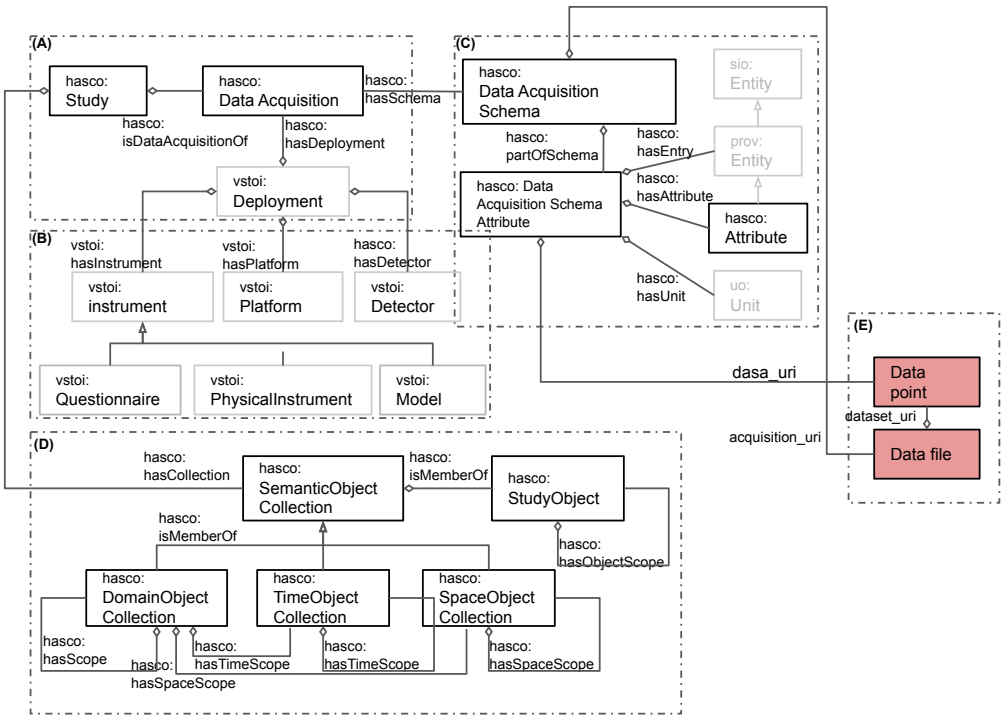

Figure 2: Overview of SKGs generated by HADatAc. [10]

3. HHEAR Data Center Use Case

The Human Health Exposure Analysis Resource (HHEAR) is an environmental health sciences research program established by the National Institute for Environmental Health Science [14]. HHEAR provides environmental epidemiology researchers with laboratory analysis of their environmental and biological specimens and to incorporate the laboratory results and statistical analyses with the original study data. The HHEAR Data Center makes the original study data,

the HHEAR laboratory results, and statistical analyses available to the public[1] as a means to improve our knowledge of the comprehensive effects of environmental exposures on human health throughout the life course and "to catalyze new scientific insight from the colocation, integration, and advanced statistical and data science analysis of multimodal data sets."

One challenge in HHEAR is integrating, normalizing, and harmonizing data from the several studies accepted by the program. For example, pregnancy cohorts provided urine samples across almost all studies, and several laboratory analyses were performed across many studies, such as measurement of environmental exposures including phthalate metabolites. Because many study variables are highly contextualized by both timing of measurement and study-specific characteristics, a solution that leveraged a semantic infrastructure was the most amenable to the program. The HHEAR Data Center uses the HADatAc infrastructure, Semantic Data Dictionaries, and well-established biomedical domain ontologies to model the metadata, collected data, and HHEAR laboratory analyses for the studies accepted by HHEAR and to build a single harmonized knowledge graph from these components. Domain ontologies and SDDs are used to normalize the semantics of each study variable, ensuring that variables across studies that share common specifications are aligned using the same formal terminology. When involved concepts are found not to be covered in existing domain ontologies, the HHEAR application ontology fills gaps in coverage when there is no appropriate term in an established biomedical ontology. We publish the HHEAR Ontology on BioPortal and release new versions whenever we add completed studies to the knowledge graph.

The HHEAR Harmonized Data Repository comprises a production instance of HADatAc, together with all of the loaded study metadata, measurement data, and ontology content. The HHEAR community can access the Harmonized Data Repository directly to search the data using HADatAc's built-in data and study search capabilities and generate normalized datasets from the search results. Custom facet search tools that leverage the SKGs built using Hadatac's infrastructure have also been integrated into the HHEAR Harmonized Data Repository and allow users to generate multi-study normalized datasets via HADatAc's APIs.

The ability to create normalized datasets across multiple studies is a significant tool for the research community because it enables data pooling across multiple HHEAR studies in which all of the concepts across studies share the same vocabulary. This ensures that when variables from different studies share the same context, the values from different studies appear in the same column. It also ensures that when categorical variables from different studies refer to the same entities, the values, and codes that appear in the dataset are globally unique and directly tied to the ontology terms that define the category value.

Table 1 provides some of the relevant overall statistics of the most recent release of the Harmonized Data Repository. Access to the HHEAR Portal and Harmonized Data Repository is available globally to any researcher who is affiliated with an academic or other institution with an Institutional Review Board, or its equivalent. Prospective users must also agree to the terms and conditions of the data use agreement.[2] A walkthrough of the HHEAR Data Center is available as an appendix (Section 7).

[1] https://hheardatacenter.mssm.edu/. Due to policies beyond the control of the authors, the HHEAR Data Center website is only accessible within the United States.

[2] https://hheardatacenter.mssm.edu/Register/Terms

Studies	Subjects	Variables	Active users	Data sets	Measurements
31	16,518	1,259	219	142	1,429,181

Table 1
HHEAR Data Center statistics.

4. FAIR Assessment

The FAIR guideline has four categories of principles: Findable, Accessible, Interoperable, and Reusable [6]. We have evaluated the HADatAc infrastructure under the lens of the FAIR principles to highlight how HADatAc can help scientists publish high-quality data repositories. We summarize the evaluation in Figure 3, listing the FAIR guidelines and associated principles and whether HADatAc meets each principle or not.

HADatAc meets the "Findable" guideline by employing the use of community-built RDF ontologies that use unique Internationalized Resource Identifiers (IRIs) to identify classes, relations, and entities within their domains, allowing the enrichment of the metadata representation that support data publishers to map measurements to the instrument level. The "Accessible" guideline is met by the use of SPARQL [17], allowing users to provide precise requests based on any possible collection of variables, studies, data sources, and time restrictions so that derived datasets can be generated from HADatAc. HADatAc meets the "Interoperable" guideline by employing the SKG to represent all study metadata and annotating every data point with elements in the SKG. The "Reusable" guideline is partially met as HADatAc delegates the metadata usage license to the repository owner, not currently encoding license metadata in the graph.

HADatAc supports the HHEAR data FAIR adherence by providing the infrastructure to create custom faceted searchers, which improves findability and accessibility. Importantly, because it leverages standardized terminology in ontologies, rather than original study variables, it promotes reusability as it clarifies the meaning of measurements and assessments reported in the datasets.

	Findable				Accessible				Interoperable			Reusable			
FAIR Principle	**F1**	**F2**	**F3**	**F4**	**A1**	**A1.1**	**A1.2**	**A2**	**I1**	**I2**	**I3**	**R1**	**R1.1**	**R1.2**	**R1.3**
HADatAc Meets Principle	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✗	✓	✗
Score	**4/4**				**4/4**				**3/3**			**2/4**			

Figure 3: HADatAc's FAIR assessment. Checks indicate principles met, while X's principles missed.

5. Community

HADatAc is an open-source project[3] that has been under development since 2014 [18]. Its infrastructure's source code was moved into a public GitHub Repository [4] on July 5th, 2015.

[3]https://hadatac.org
[4]https://github.com/hadatac/hadatac

HADatAc.org comprises fifteen organizations from four countries: from the United States (5 universities and 3 research groups); Portugal (2 universities and 1 private company); Brazil (2 universities and 1 national research organization); and Canada (1 university). The infrastructure includes comprehensive online documentation[5] with instructions on how to install and use the infrastructure. The HAScO, VSTO, and SIO ontologies that build HADatAc's foundation are all available at BioPortal.

6. Related Work & Conclusion

According to [19], scientific research data should ideally be shared through domain-specific repositories that use data types widely employed in a field. These repositories are like data warehouses, providing long-term access to data by assigning persistent IDs such as digital object identifiers (DOI). ImmPort is a platform that collects and curates immunological data, which is then shared through a public component [20]. ImmPort's domain is strictly for immunology [21], and the data model is not schema-free. The platform has some level of data lifting, although its significance is unclear. The NIMH National Data Archive (NDA) uses Global Unique Identifiers (GUIDs) to identify data from unique individuals [22]. Tools allow data download and search. Users must pass validation against the dictionary when uploading any data. Domain-agnostic repositories are generally chosen by investigators to deposit scientific data, such as Figshare [23] or Zenodo [24], along with metadata that accurately describes the included files and their format. However, these approaches do not provide any further data integration or harmonization, storing the data "as-is." European Data Spaces [25] is a policy that proposes a foundation for the "data economy" in Europe, which has fostered the development of several data frameworks [26].

The availability of entity characterization, along with logical linkage between data and scientific study knowledge, is one of HADatAc's benefits that scientists may immediately observe and recognize when processing scientific data. HADatAc has proven to be a useful tool for integrating data from multiple domains. Through the use of ontologies as shared metadata standards, data are annotated, integrated, and stored into a knowledge base. The metadata can then be used to query the knowledge base to retrieve relevant datasets without the domain expert having detailed knowledge of the original structures of these datasets.

Beyond the HHEAR program, HADatAc has become an important tool for the SKG community across multiple projects and scientific domains, supporting, for example, research efforts within Rensselaer Polytechnic Institute that use the National Health and Nutrition Examination Surveys (NHANES)[6] [27], promoting semantically-enabled data analysis [28, 29]. HADatAc is also being used to support projects involving the Internet of Things (IoT) in Europe, where we observe that usage scenarios can be far more complex since objects like cars and buildings can have thousands of sensors organized in many subsystems.

Acknowledgments

The HHEAR Data Center is funded by the National Institute of Environmental Health Studies

[5] https://github.com/paulopinheiro1234/hadatac/wiki
[6] https://www.cdc.gov/nchs/nhanes/index.htm

grant U2CES026555. Publicly available data used in this study was generated through grants supported by the National Institutes of Health as part of the Human Health Exposure Analysis Resource (HHEAR). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

References

[1] S. Abbasian Dehkordi, K. Farajzadeh, J. Rezazadeh, R. Farahbakhsh, K. Sandrasegaran, M. Abbasian Dehkordi, A survey on data aggregation techniques in IoT sensor networks, Wireless Networks 26 (2020) 1243–1263.

[2] E. Ramalli, B. Pernici, Challenges of a Data Ecosystem for scientific data, Data & Knowledge Engineering 148 (2023) 102236.

[3] V. Chaudhri, C. Baru, N. Chittar, X. Dong, M. Genesereth, J. Hendler, A. Kalyanpur, D. Lenat, J. Sequeda, D. Vrandečić, et al., Knowledge graphs: introduction, history and, perspectives, AI Magazine 43 (2022) 17–29.

[4] P. Manghi, A. Mannocci, F. Osborne, D. Sacharidis, A. Salatino, T. Vergoulis, New trends in scientific knowledge graphs and research impact assessment, Quantitative Science Studies 2 (2021) 1296–1300.

[5] S. J. Chalk, SciData: a data model and ontology for semantic representation of scientific data, Journal of Cheminformatics 8 (2016) 54.

[6] M. D. Wilkinson, M. Dumontier, I. J. Aalbersberg, G. Appleton, M. Axton, A. Baak, N. Blomberg, J.-W. Boiten, L. B. da Silva Santos, P. E. Bourne, et al., The fair guiding principles for scientific data management and stewardship, Scientific data 3 (2016) 1–9.

[7] M. Johnson, M. Ravi, P. Pinheiro, J. Stingone, D. McGuinness, A semi-automated approach to data harmonization across environmental health studies, in: ISEE Conference Abstracts, volume 2020, 2020.

[8] S. M. Rashid, J. P. McCusker, P. Pinheiro, M. P. Bax, H. Santos, J. A. Stingone, A. K. Das, D. L. McGuinness, The semantic data dictionary–an approach for describing and annotating data, Data Intelligence (2020) 443–486.

[9] M. Johnson, J. A. Stingone, S. Bengoa, J. Masters, D. L. McGuinness, Complex semantic tabular interpretation using sdd-gen, in: 2024 IEEE 18th International Conference on Semantic Computing (ICSC), IEEE, 2024, pp. 317–322.

[10] P. Pinheiro, M. Bax, H. Santos, S. M. Rashid, Z. Liang, Y. Liu, J. P. McCusker, D. L. McGuinness, Annotating Diverse Scientific Data with HAScO, in: Proceedings of the Seminar on Ontology Research in Brazil 2018 (ONTOBRAS 2018). São Paulo, SP, Brazil, 2018.

[11] Apache Software Foundation, Apache SOLR, http://solr.apache.org, 2006.

[12] Apache Software Foundation, Apache Jena, https://jena.apache.org, 2021.

[13] M. Dumontier, et al., The Semanticscience Integrated Ontology (SIO) for biomedical research and knowledge discovery, Journal of Biomedical Semantics 5 (2014) 14.

[14] S. M. Viet, J. C. Falman, L. S. Merrill, E. M. Faustman, D. A. Savitz, N. Mervish, D. B. Barr, L. A. Peterson, R. Wright, D. Balshaw, et al., Human health exposure analysis resource (hhear): A model for incorporating the exposome into health studies, International journal of hygiene and environmental health 235 (2021) 113768.

[15] T. Lebo, S. Sahoo, D. L. McGuinness, PROV-O: The PROV Ontology, W3C Recommendation, W3C, 2013. URL: https://www.w3.org/TR/2013/REC-prov-o-20130430/.
[16] P. Fox, D. L. McGuinness, L. Cinquini, P. West, J. Garcia, J. L. Benedict, D. Middleton, Ontology-supported scientific data frameworks: The virtual solar-terrestrial observatory experience, Computers & Geosciences 35 (2009) 724–738.
[17] S. Harris, A. Seaborne, Sparql 1.1 query language, 2013. URL: https://www.w3.org/TR/sparql11-query/.
[18] D. L. McGuinness, P. Pinheiro, H. Santos, M. Klawonn, K. Chastain, Semantic Support for Complex Ecosystem Research Environments, AGU Fall Meeting Abstracts 33 (2015).
[19] J. B. Byrd, A. C. Greene, D. V. Prasad, X. Jiang, C. S. Greene, Responsible, practical genomic data sharing that accelerates research, Nature Reviews Genetics 21 (2020) 615–629. Publisher: Nature Publishing Group.
[20] S. Bhattacharya, P. Dunn, C. G. Thomas, B. Smith, H. Schaefer, J. Chen, Z. Hu, K. A. Zalocusky, R. D. Shankar, S. S. Shen-Orr, et al., Immport, toward repurposing of open access immunological assay data for translational and clinical research, Scientific data 5 (2018) 180015.
[21] S.-A. Sansone, P. Cruse, M. Thorley, High-quality science requires high-quality open data infrastructure, Scientific data 5 (2018).
[22] D. Hall, M. F. Huerta, M. J. McAuliffe, G. K. Farber, Sharing heterogeneous data: the national database for autism research, Neuroinformatics 10 (2012) 331–339.
[23] M. Thelwall, K. Kousha, Figshare: a universal repository for academic resource sharing?, Online Information Review 40 (2016) 333–346.
[24] M.-A. Sicilia, E. García-Barriocanal, S. Sánchez-Alonso, Community curation in open dataset repositories: insights from zenodo, Procedia Computer Science 106 (2017) 54–60.
[25] A European strategy for data, 2020. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX%3A52020DC0066.
[26] B. Otto, A federated infrastructure for European data spaces, Communications of the ACM 65 (2022) 44–45.
[27] H. Santos, P. Pinheiro, D. L. McGuinness, Knowledge Graph Construction from Data, Data Dictionaries, and Codebooks: the National Health and Nutrition Examination Surveys Use Case, 2022. URL: https://us2ts.org.
[28] M. Qi, H. Santos, P. Pinheiro, D. L. McGuinness, K. P. Bennett, Demographic and socioeconomic determinants of access to care: A subgroup disparity analysis using new equity-focused measurements, PLOS ONE 18 (2023) e0290692. Publisher: Public Library of Science.
[29] P. Pinheiro, H. Santos, M. Qi, K. P. Bennett, D. L. McGuinness, Towards Machine-Assisted Biomedical Data Preparation: A Use Case on Disparity in Access to Health Care, in: Proceedings of the 6th Workshop on Semantic Web Solutions for Large-Scale Biomedical Data Analytics, volume 3466 of *CEUR Workshop Proceedings*, CEUR, Hersonissos, Greece, 2023. ISSN: 1613-0073.

7. Appendices

HHEAR Data Center walkthrough

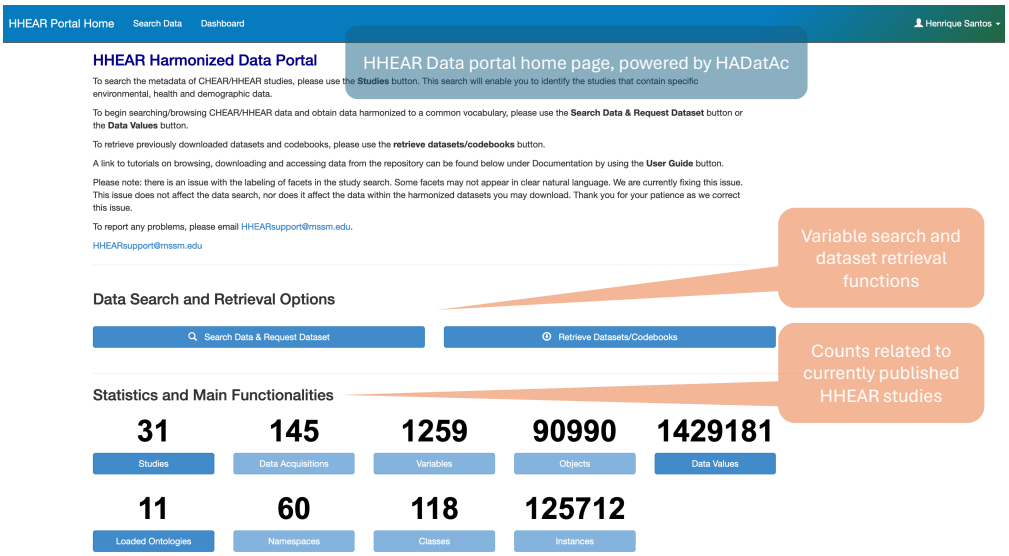

Figure 4: HHEAR Harmonized Data portal homepage.

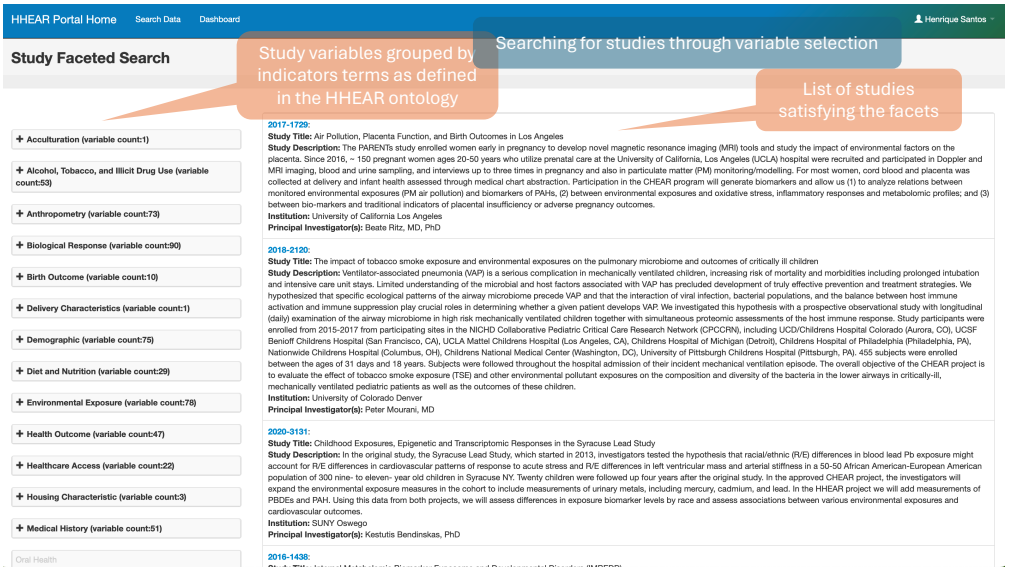

Figure 5: Faceted-search of available HHEAR studies. Indicators on the left side can be expanded to allow search by specific study variables. All contents on this page are dynamic and retrieved from a HADatAc-built SKG using SPARQL.

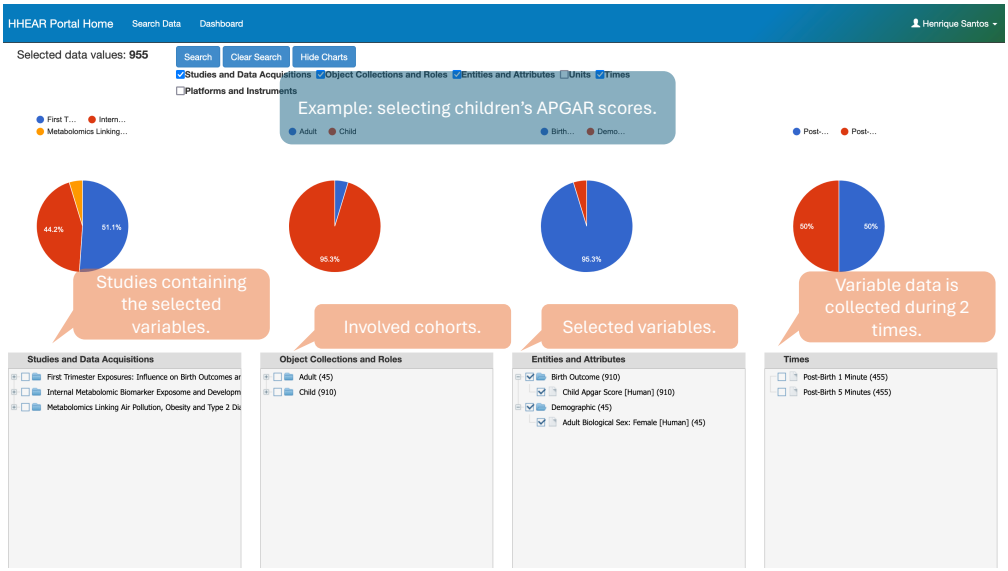

Figure 6: Data search view, leveraging the semantics in the SKG. In this example, we have specifically selected the APGAR score variable, which has 955 data values. We can further see that this variable is present in three studies and that it is measured at two different time points: 1 and 5 minutes post-birth.

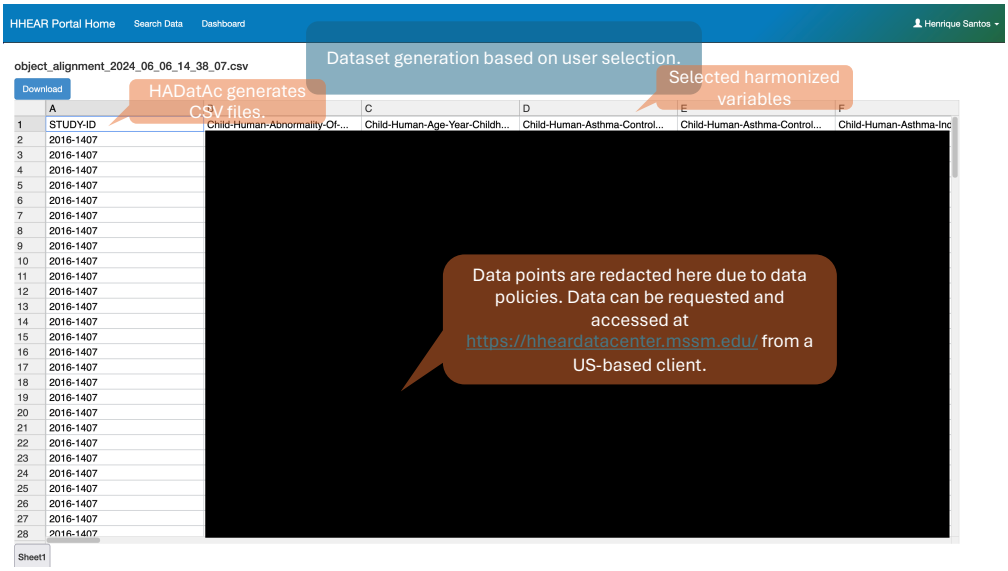

Figure 7: The dataset generation after variable selection. HADatAc generated CSV files with harmonized variables from potentially several studies. The first column indicates which study is a data point coming from.