HADatAc: A Framework for Scientific Data Integration using Ontologies

Paulo Pinheiro[1], Henrique Santos[1,2], Zhicheng Liang[1], Yue Liu[1], Sabbir M. Rashid[1], Deborah L. McGuinness[1], and Marcello P. Bax[1,3]

[1] Rensselaer Polytechnic Institute, Troy, NY, 12180, USA
[2] Universidade de Fortaleza, Fortaleza, CE, 60811-905, Brazil
[3] Universidade Federal de Minas Gerais, Belo Horizonte, MG, 31270-901, Brazil

Abstract. To investigate the cause and progression of a phenomenon, such as chronic disease, it is essential to collect a wide variety of data that together explains the complex interplay of different factors, e.g., genetic, lifestyle, environmental and social. Sharing information between studies is therefore of paramount importance. However, data that needs to be analyzed must be appropriately integrated, conceptually aligned, and harmonized. This implies that data collection must be done either in a sufficiently similar or a sufficiently transparent way in order to support meaningful synthesis from different studies. We will demonstrate[4,5] how the Human-Aware Data Acquisition (HADatAc) framework integrates and harmonizes data from multiple scientific studies and thus how to use it in interdisciplinary science investigations.

1 Introduction

The Human-Aware Data Acquisition (HADatAc) Framework is a schema-free, evolutionary, scalable and provenance-aware infrastructure for managing data and metadata content from multiple scientific studies. Three key goals of HADatAc are: (1) to extract relevant data value from instrument-generated files and to move these values into queryable content repositories, (2) to extract relevant metadata from scientist-generated documents and to move these values into queryable content repositories, and (3) to semantically annotate these values in a way that the entire content is logically linked and harmonized (i.e., unified representation) according to evolving collections of well-established scientific ontologies. HADatAc's core ontologies, that are fully integrated, aligned, and used in multiple scientific domains, include: W3C's Provenance Ontology (PROV), encoding provenance knowledge, Virtual Solar-Terrestrial Observatory (VSTO) [2], encoding knowledge about instruments and platforms, Human-Aware Science Ontology (HAScO)[6] [4], encoding knowledge about studies, study types, data elicitation from humans, and data simulation from computer models, and the Semantic science Integrated Ontology (SIO) [1], encoding knowledge about science-related entities and their characteristics.

[4] HADatAc's live demo is available at `http://bit.ly/HADatAc`
[5] Demo video: `http://bit.ly/HADatAc-iswc2018`
[6] `http://hadatac.org/ont/hasco`

Without any expansion, the framework and the collection of ontologies listed above are domain agnostic and ready for usage in scientific domains. Existing HADatAc deployments are built using these core ontologies, along with many other specialized ontologies, for the domain of interest. One specific domain ontology is often used to import specialized ontologies into a single document that HADatAc uses as a default namespace for a given domain of interest (e.g. the integrated exposure and health ontology CHEAR [3] that we use in our HADatAc backend for the NIEHS Child Health Exposure Analysis Resource implementation).

2 HADatAc Characteristics

Schema-free claim. HADatAc uses semantic technologies, ontologies, graph databases and non-relational databases to manage metadata and data from relevant scientific studies, e.g. CDC's National Health and Nutrition Survey (NHANES)[7] in the demonstration system. HADatAc is schema-free since the content from these study files is stored without a predefined and fixed structure. This allows HADatAc to include objects, e.g., subjects, samples, locations, into its repositories as they are presented, including attributes from objects regarded as relevant for the studies.

Evolutionary claim. Ontologies and the underlying graph, including the loaded data, are managed by one Apache SOLR repository and by one Blazegraph RDF graph database. The SOLR repository manages data values and a collection of URIs for each data value. URIs, in the collection of URIs of each stored data value, are links between the data value and semantic annotations in the Blazegraph. The Blazegraph repository is used to manage the HADatAc knowledge graph. HADatAc's knowledge graph evolves by either importing existing ontologies or by defining concepts and relations that are not available or not appropriate for reuse from existing ontologies.

Scalability claim. Scientific data management platforms must handle increasing data volumes. HADatAc's backend SOLR repository provides the required scalability for very large data repositories. As the metadata volume is significantly smaller than the data volume, it is easily managed by the Blazegraph triple-store. SOLR is also used to compute aggregate faceted values into the scientific study indicators, that would be costly to do with Blazegraph.

Provenance-Aware claim. HADatAc has been developed to work with a broad range of data sources, and the infrastructure captures and preserves provenance on how data values where acquired by a broad notion of instruments. HADatAc classifies data sources according to the instruments (and detectors) used to acquire the data. HADatAc understands if study data is the result of any of the data acquisition strategies: (a) *empirical measurement*, which is done using physical instruments like sensors; (b) *data and knowledge elicitation* from humans, which is done using questionnaires as instruments; (c) *computer data generation (simulation)*, which is done using simulation models as instruments.

[7] `https://www.cdc.gov/nchs/nhanes/index.htm`

3 HADatAc Architecture

Figure 1 shows a summarized description of HADatAc's architecture. The main box in the center of the figure represents HADatAc's core component, which is connected to six satellite subsystems shown as smaller boxes.

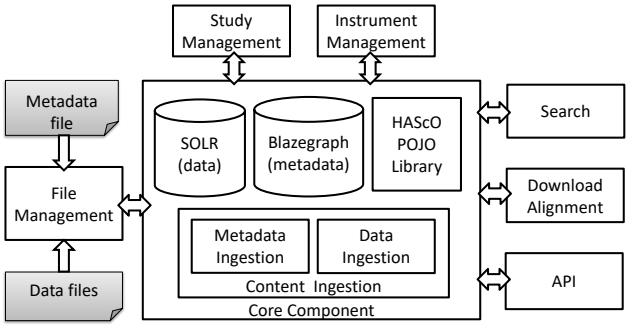

Fig. 1. HADatAc architecture, including content repositories and subsystems.

HADatAc is a framework and also a web application. Most of its web user interface is implemented as part of the six satellite subsystems. The `API Subsystem` is a special subsystem composed of a collection of RESTful services with programmatic access to HADatAc's content. The `Core Component` has the elements required to support the satellite subsystems including: the SOLR and Blazegraph content repositories, a Java API encoding the concepts of the Human-Aware Science Ontology (HAScO) as POJO Classes, and the subsystems responsible for extracting, annotating, and storing study content from data and metadata files into SOLR and Blazegraph. The HAScO POJO classes are used to build and maintain HADatAc's knowledge graph.

Content is added into HADatAc either through the parsing of uploaded files through the `File Management Subsystem` or on-line through user interaction with the `Study Management Subsystem` and the `Instrument Management Subsystem`. Content is directly presented to users through the `Search Subsystem` and downloaded through the `Object Alignment Subsystem` and `API Subsystem`.

4 HADatAc's Demonstration

Our demonstration includes the following six steps: (1) registration of CDC's NHANES as a new study in HADatAc including the generation of NHANES subjects as RDF instances, (2) registration of a semantic data dictionary (SDD) for NHANES that identifies how the content of NHANES data files are extracted, integrated, and harmonized, (3) upload and processing of NHANES data files that store NHANES data and metadata into databases, (4) automatic processing of uploaded data files, (5) display of harmonized content in a semantic faceted

search, as shown in Figure 2, and finally (6) download of datasets generated from the current selection of a search in the data faceted search tool. We have used a combination of techniques to identify and retrieve variables to be used in the demo, mainly the NHANES Variable Search[8] and, to generate datasets for ingestion into HADatAc, the RNHANES R package[9], which allows extraction of select named variables across NHANES datasets. Our claim is that HADatAc can play the roles of those multiple techniques for variable identification, retrieval, and dataset generation for diverse domains.

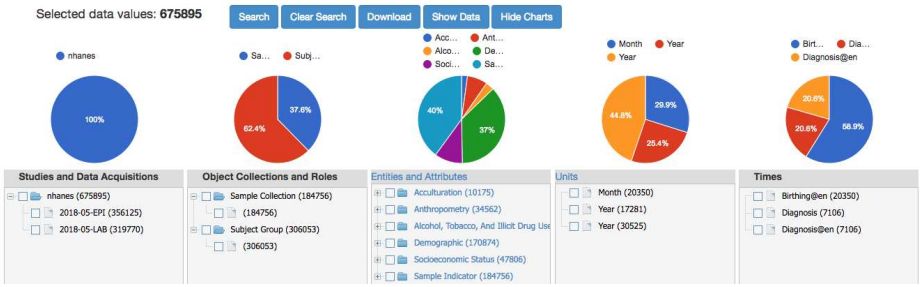

Fig. 2. HADatAc's faceted search for values with pie chart displays of distributions.

Acknowledgements This work was partially funded by the National Institute of Environmental Health Sciences (NIEHS) Award 0255-0236-4609 / 1U2CES02-6555-01 and CAPES Foundation Award 88881.120772 / 2016-01. It has been co-deployed with Mount Sinai School of Medicine and the Gates Foundation Healthy Birth, Growth, and Development Knowledge Integration program with collaborators from Yale's Center for Ecosystems in Architecture.

References

1. Dumontier, M., Baker, C.J., Baran, J., Callahan, A., Chepelev, L., Cruz-Toledo, J., Del Rio, N.R., Duck, G., Furlong, L.I., Keath, N., et al.: The Semanticscience Integrated Ontology (SIO) for Biomedical Research and Knowledge Discovery. Journal of Biomedical Semantics **5**(1), 14 (2014)
2. Fox, P., McGuinness, D.L., Cinquini, L., West, P., Garcia, J., Benedict, J.L., Middleton, D.: Ontology-supported Scientific Data Frameworks: The Virtual Solar-terrestrial Observatory Experience. Computers & Geosciences **35**(4), 724–738 (2009)
3. McCusker, J.P., Rashid, S.M., Liang, Z., Liu, Y., Chastain, K., Pinheiro, P., Stingone, J.A., McGuinness, D.L.: Broad, Interdisciplinary Science in Tela: An Exposure and Child Health Ontology. In: Proceedings of the 2017 ACM on Web Science Conference. pp. 349–357. ACM (2017)
4. Pinheiro, P., Bax, M., Santos, H., Rashid, S.M., Liang, Z., Liu, Y., McCusker, J.P., McGuinness, D.L.: Annotating Diverse Scientific Data with HAScO. In: Proceedings of the Seminar on Ontology Research in Brazil 2018 (ONTOBRAS 2018). São Paulo, SP, Brazil (2018)

[8] `https://wwwn.cdc.gov/nchs/nhanes/search/default.aspx`
[9] `https://CRAN.R-project.org/package=RNHANES`