Annotating Diverse Scientific Data with HAScO

Paulo Pinheiro[1], Henrique Santos[12], Marcello Bax[13] , Zhicheng Liang[1], Sabbir M. Rashid[1], Yue Liu[1], James P. McCusker[1], Deborah L. McGuinness[1]

[1]Rensselaer Polytechnic Institute, Troy, NY – USA

[2]Universidade de Fortaleza, Fortaleza, CE – Brazil

[3]Universidade Federal de Minas Gerais, Belo Horizonte, MG – Brazil

{pinhep,oliveh,baxm2,liangz4,sabbir,liuy30,mccusker,dlm}@rpi.edu

Acknowledgement

The work presented here has been developed in collaboration with a large number of contributors from **hadatac.org** including the following:

HADatAc Framework

Motivations

Scientific Knowledge Representation in Support of Data Analysis

HAScO provides the foundation for developing domain ontologies used to semantically annotate scientific data.

"*The repeatability of data and computational intensive experiments is a key aspect in making their results more trustworthy, and thus also of higher relevance to other researchers in the community*"

– MAYER; MIKSA; RAUBER, 2014

Motivations

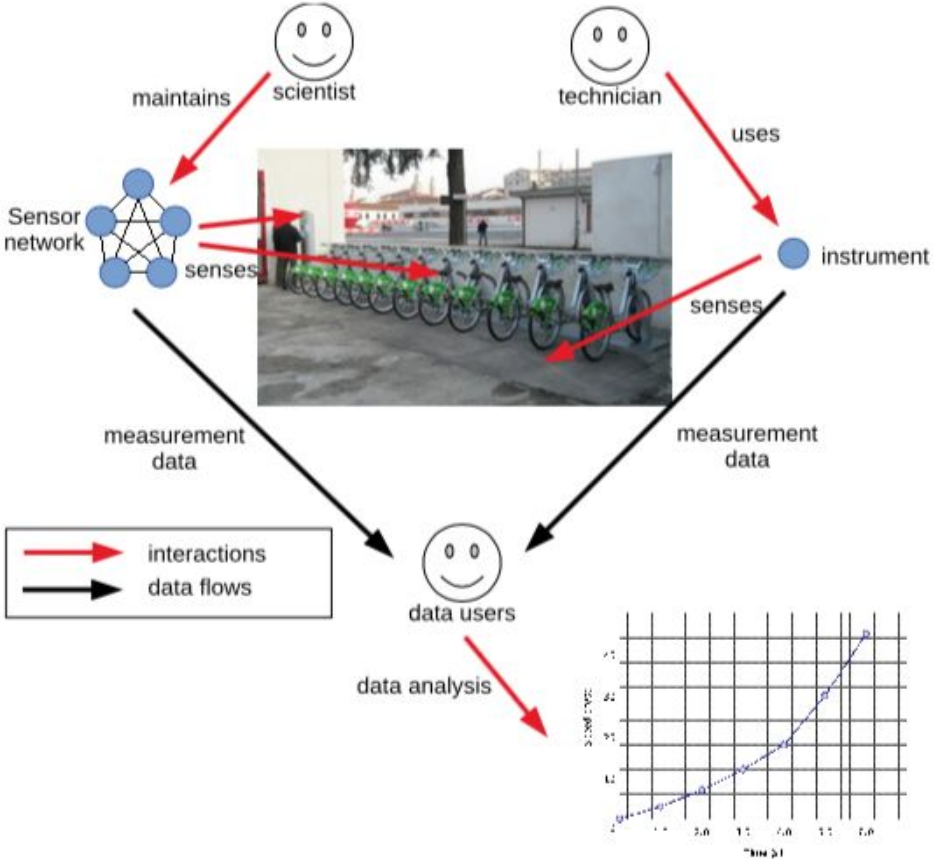
Scientific Knowledge Representation in Support of Data Analysis

- Ontologies are being widely used in science activities, most notably in roles related to **acquiring**, **preparing**, **integrating** and managing data resources.

- Data acquisition and preparation are often difficult to reuse since they tend to be **domain dependent**, as well as dependent on how data is acquired: through **measurement**, **subject-elicitation**, and/or **model-generation** activities.

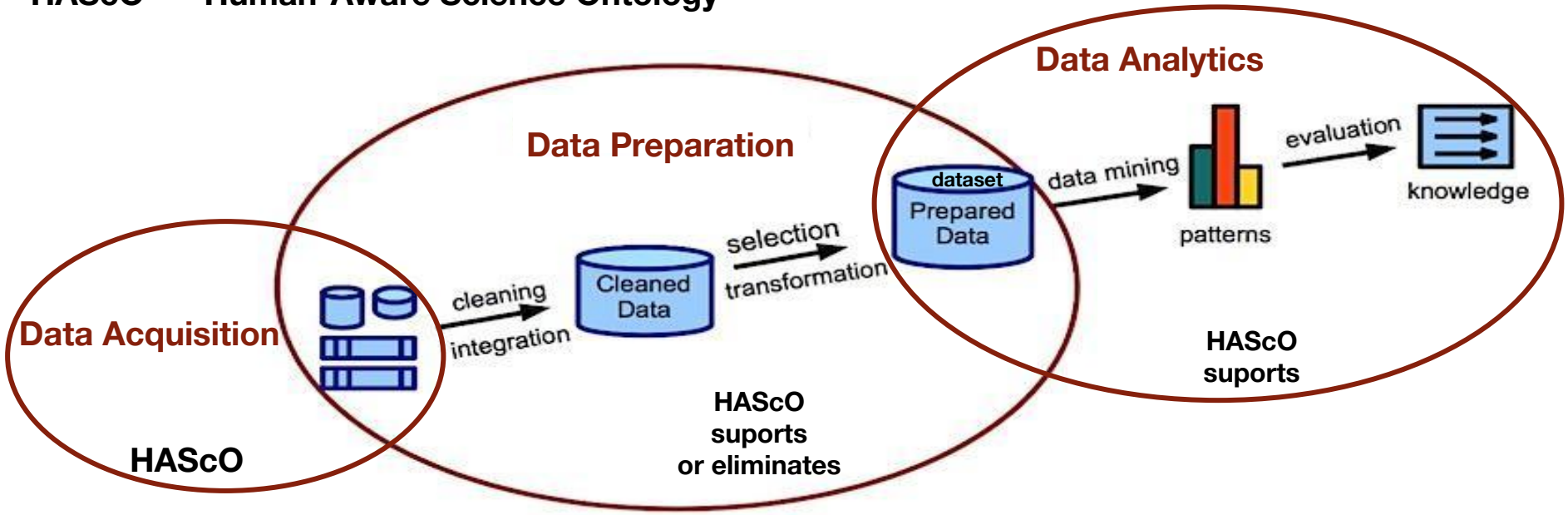
- Therefore, tools developed for preparing data from one scientific activity often cannot be adapted to prepare data from other scientific activities.

Typical data acquisition scenario (Santos, 2018)



Source: Santos, Henrique. *An Indicator-based Approach for Variable Alignment based on Knowledge Graphs*, 2018.

HAScO - Human-Aware Science Ontology

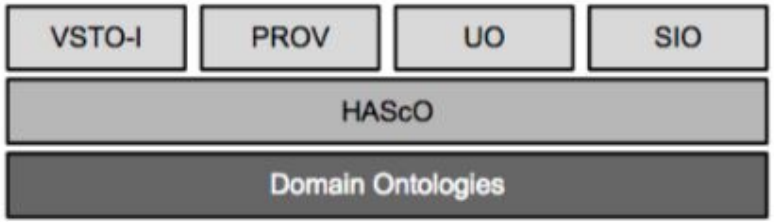


- It is a process that involves many different tasks and which **cannot be fully automated**.
- Many of the data preparation activities are **routine, tedious, and time consuming**.
- It has been estimated that data preparation **accounts for 60%-80% of the time** spent on a data mining project.
- **HAScO** is an ontology for **conceptually describing** the tasks of data preparation.

[1] https://www.datapreparator.com/what_is_data_preparation.html with adaptations

HAScO: Human-Aware Science Ontology

- HAScO integrates a collection of well-established science-related ontologies.



- Designed for encoding **metadata of scientific studies** in **large data ecosystem**, where data can come from diversified data sources including sensors, lab results, and questionnaires.

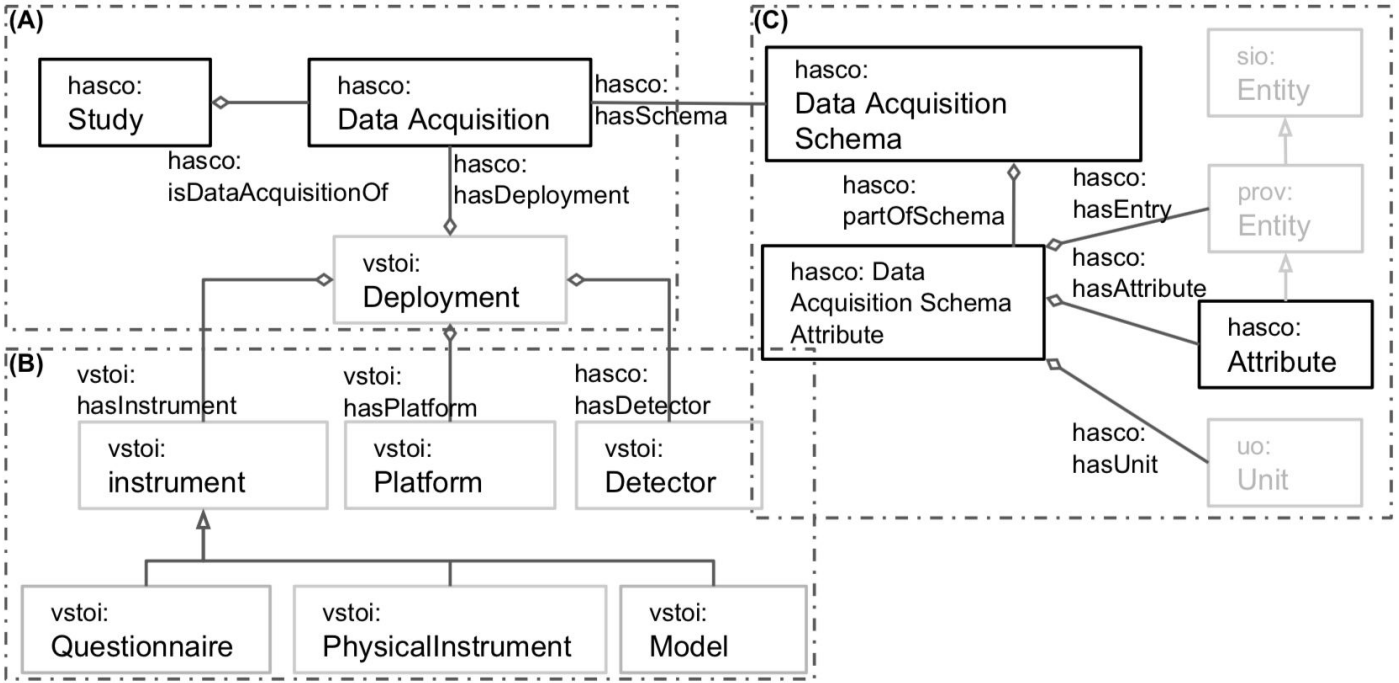
- The paper presents HAScO based on our experience applying it to annotate data, facilitating its exploration and analysis.

Using HAScO

- Data files produced by scientific studies are processed to identify and **annotate the objects** (a gene, for instance) with the appropriate ontological terms.
- Study > Object Collections > Objects > Attributes > Values
- Objects, Attributes, Values are typed based on HAScO
- HAScO is used for harmonizing data across studies since the meaning of and interrelationships between the data is explicit. Harmonization is achieved through the use of **data annotation**, semantically **rich query support**, and **data driven views** for specific user groups.
- One benefit is that software platforms can support scientists in their data **acquisition**, preparation and exploration activities for analysis .
- HAScO supports the design and implementation of the Human-Aware Data Acquisition Framework (HADatAc) [Pinheiro et al. 2018]
  => a platform to **support broad scientific data acquisition activities**.

Overview of HAScO

(A) scientific activities;
(B) instruments of data acquisition;
(C) scientific data schema.



Requirements for Scientific Data Representation

Data Acquisition Activity Metadata				HAScO's Use Cases (Research Projects)		
				Environmental Studies	Human Health Studies	Building Sciences Studies
Study Metadata	Study Type	Observations		X		X
		Experiments		X	X	X
	Study Description	Identification of objects and their inter-relations			X	X
		Data quality management at study level		X		X
		Temporal support for study description			X	X
		Spatial support for study description		X		X
Sensing Infrastructure Metadata	Instruments (Data Acquisition Methods)	Measurement Data	sensor networks	X		X
			lab controlled		X	
		Elicited Data	questionnaires		X	X
			documents as source			X
		Model Generated Data	simulation	X		
	Activities (Quality Control)	Uncertainty Provenance	lab managed		X	
			deployment managed	X		X

Conclusion

- To encode scientific findings in a structured, knowledge-enhanced way using ontologies, can support research exploration and potentially identify novel connections, thereby increasing the overall research impact [Brodaric/Gahegan 2010].

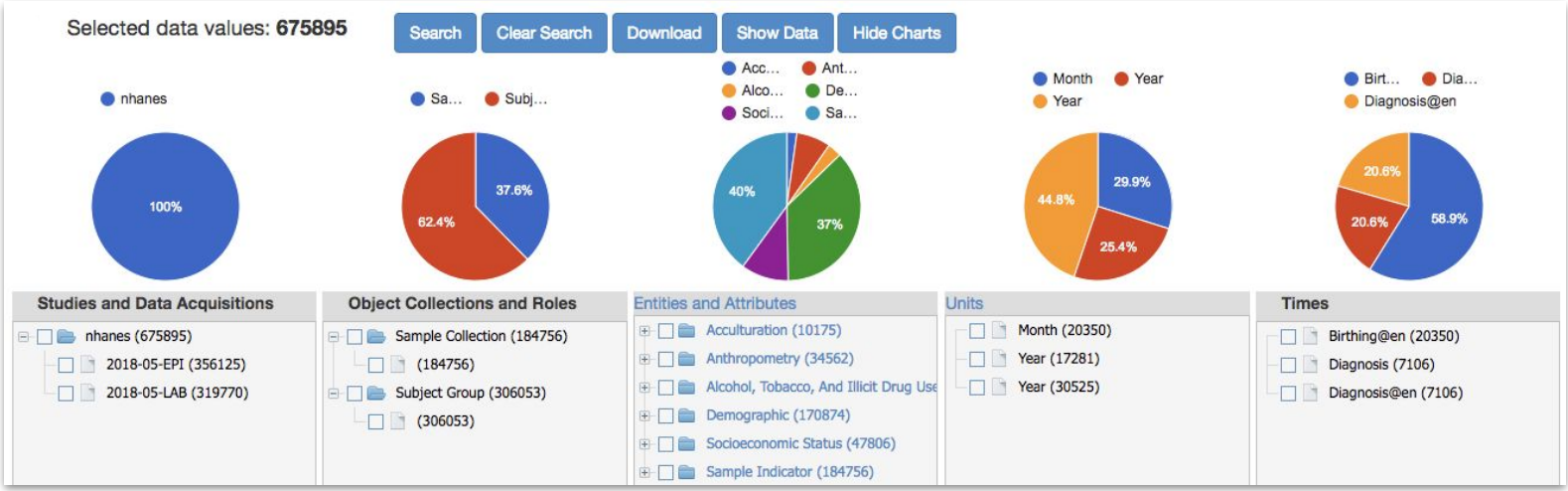
- HAScO was developed and applied to major scientific projects with the goal of helping scientists with data preparation (in support of data analysis).

- HAScO is domain-agnostic, and leverages a combination of well-established foundational ontologies including SIO, OBO Foundry's UO, W3C's PROV, and VSTO-I.

- In addition to supporting some high-level scientific concepts such as Studies, Subjects, Samples and others, HAScO provides a quality dimension of data based on a new generalizable concept called Data-Acquisition.

HAScO and HaDatAc

- The ontology is available under MIT license http://hadatac.org/ont/hasco/
- We are maintaining and evolving the ontology through its use in the HaDatAc infrastructure.
- To the extent that the HaDatAc framework is used as a basis for the implementation of new research projects, HAScO will evolve accordingly, guaranteeing the necessary support for the progress of the ontology.



Acknowledgements

This work was partially supported by

- the National Institute of Environmental Health Sciences (NIEHS) Award 0255-0236-4609 / 1U2CES026555-01,

- National Science Foundations Award DBI 1625044, the Gates Foundation through the Healthy Birth, Growth, and Development knowledge integration (HBGDki),

- the RPI Tetherless World Constellation, and

- CAPES Award 88881.120772 / 2016-01.

Main References

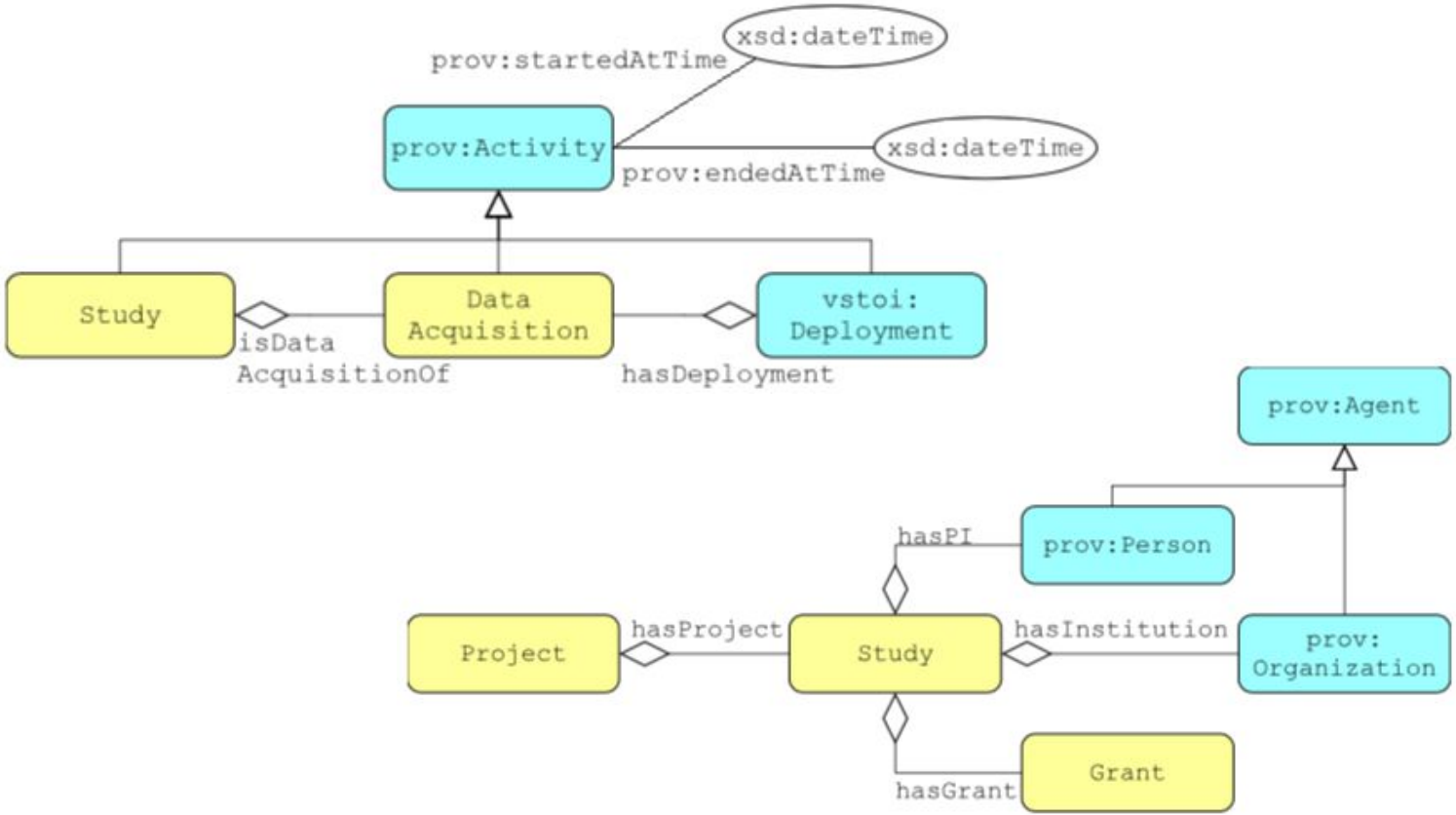
- Pinheiro, P., Santos, H., Liang, Z., Liu, Y., Rashid, S., McGuinness, D., and Bax, M. (2018). HADatAc: A Framework for Scientific Data Integration using Ontologies. In Proceedings of the ISWC 2018 Posters & Demonstrations Track.

- Santos, Henrique. An Indicator-based Approach for Variable Alignment based on Knowledge Graphs, 2018 (doctoral thesis).

- McGuinness, D., Pinheiro, P., Patton, E., and Chastain, K. (2014). Semantic escience for ecosystem understanding and monitoring: The jefferson project case study. In AGU Fall Meeting Abstracts, volume 1, page 3712.

- Brodaric, B. and Gahegan, M. (2010). Ontology use for semantic e-science. Semantic Web, 1(1, 2):149–153.

Studies and Data Acquisition activities



Source: Santos, Henrique. *An Indicator-based Approach for Variable Alignment based on Knowledge Graphs*, 2018.

Instruments