

Ontologies for Scientific Data Transformation

Leonardo Salayandia



Outline

- Problem
- Background
- Research Question
- Related Work
- Framework
- Evaluation
- Conclusions

Problem

- Scientists Collect and Transform Data

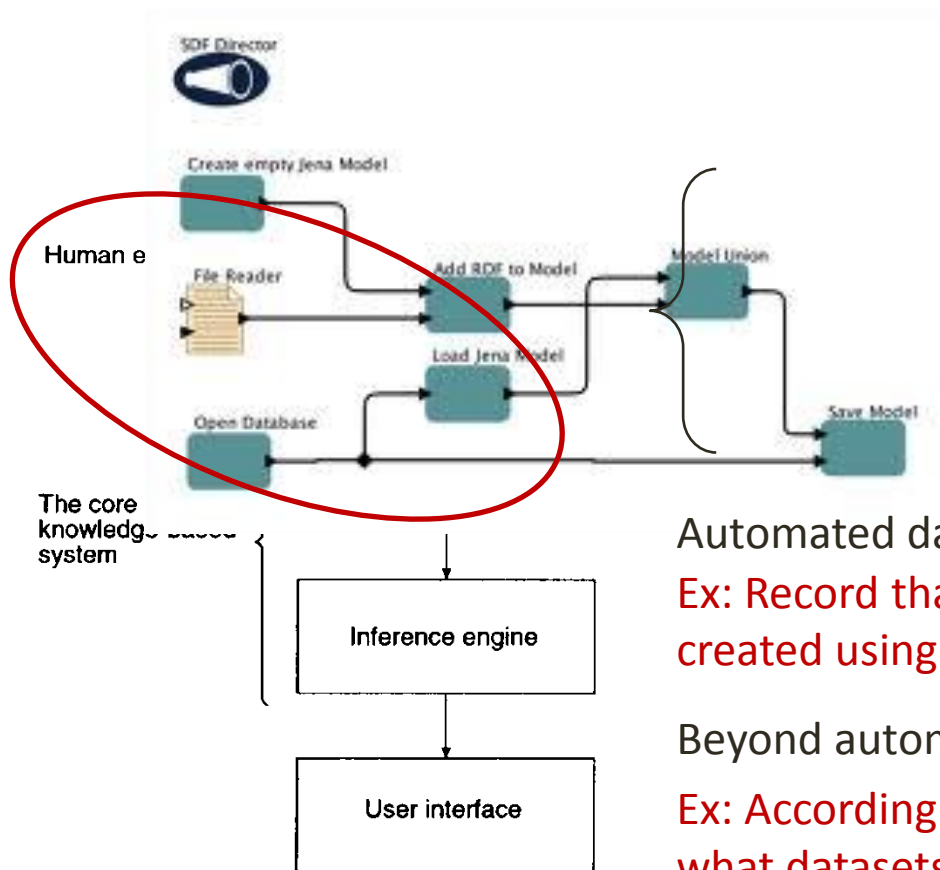


- Datasets are diverse and voluminous
- Computers are needed to handle data

Scientists need to represent their knowledge about data in a computer

Background

Knowledge-Based System Management Systems

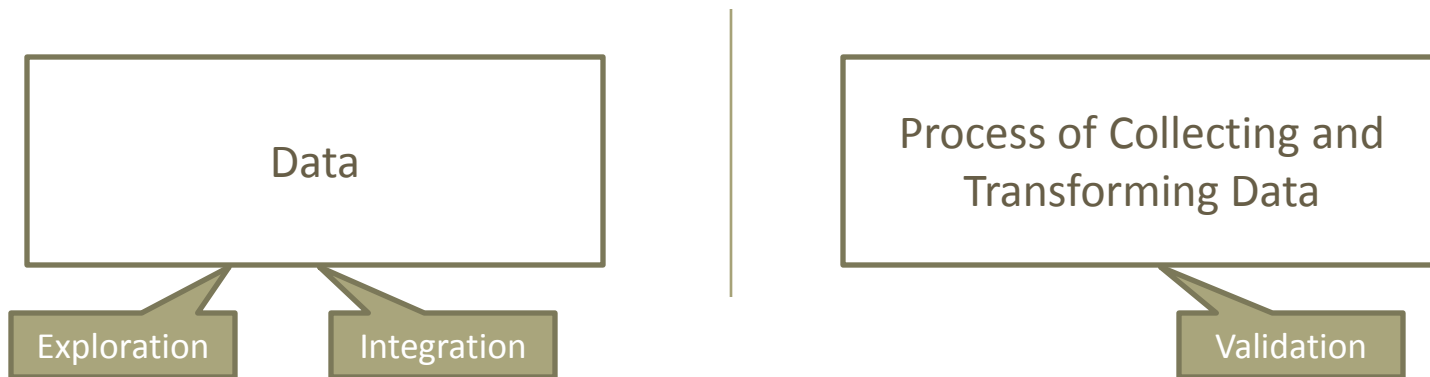


Automated data book-keeping
Ex: Record that Dataset X was created using Method Y.

Beyond automated data book-keeping
Ex: According to my project requirements, what datasets can I use?

Research Question

- Ontologies are used to represent knowledge in a computer



- Ontologies are not trivial to produce
(Brachman and Levesque, 1984)

How can scientists create ontologies about the process of data collection and transformation?

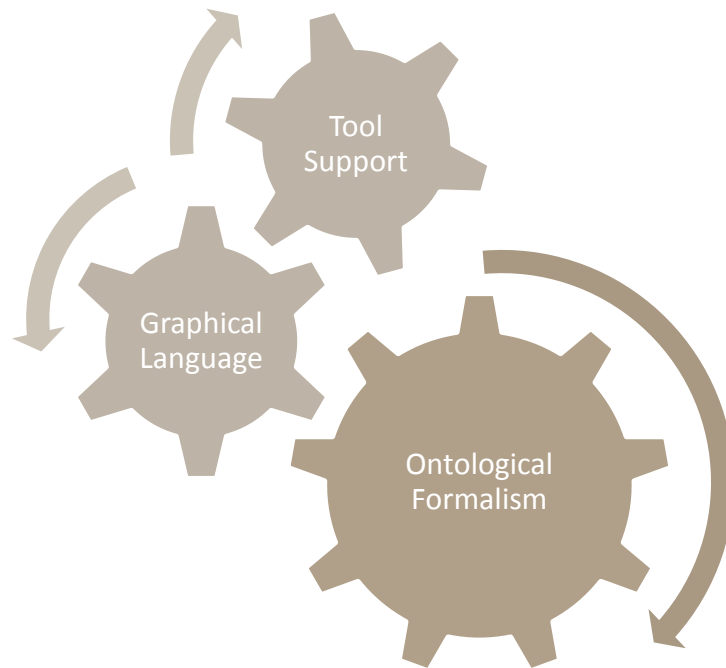
Related Work

- General purpose ontology editors/frameworks, e.g., Protegé
 - High flexibility
- Process-specific ontologies, e.g., PSL
 - Wide audience

High learning curve for non-technical person

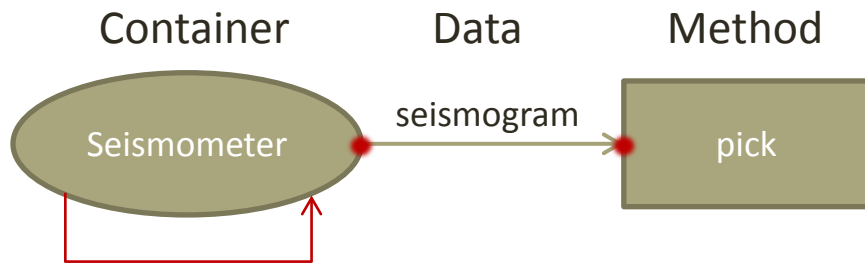
Framework (1/2)

Goal: Create a framework for scientists to formally represent their knowledge about data collection and transformation



Framework (2/2)

Semantic Abstract Workflow



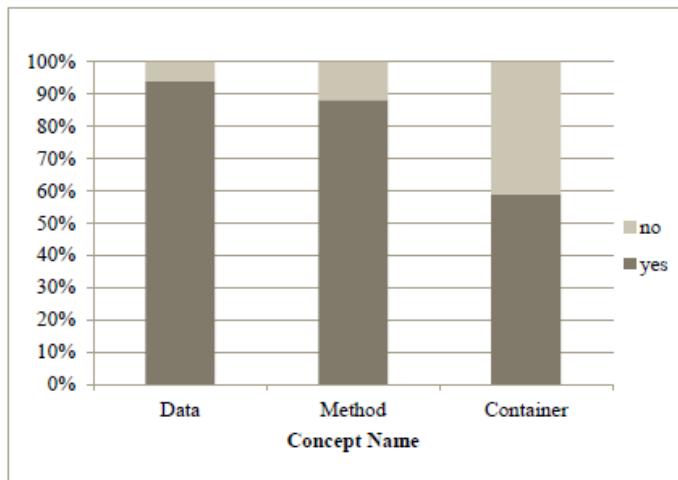
Framework prevents invalid formalisms to be specified

Workflow-Driven Ontology

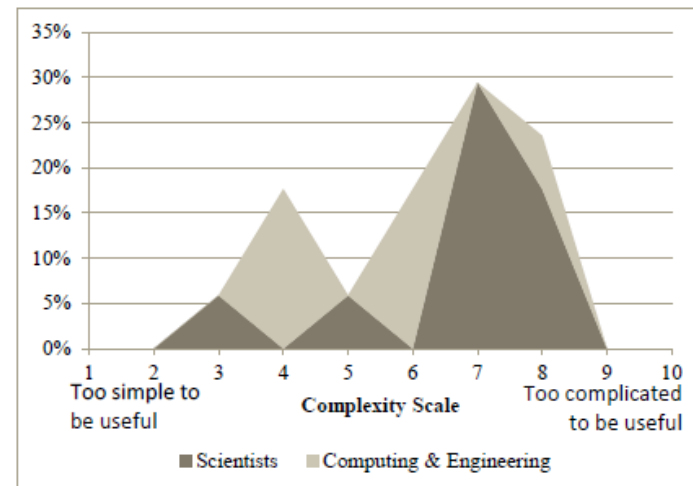
- Seismometer *subsumed by* Container
- Seismogram *subsumed by* Data
- Seismometer *has output* Seismogram
- Pick *subsumed by* Method
- Seismogram *is input to* Pick

Evaluation

- Graphical language favors ease of use
 - (Davis, 1990), (Petre, 1995)
- Pilot survey results
 - 18 subjects from engineering and science
 - Undergraduates, graduates, faculty, researchers, professionals



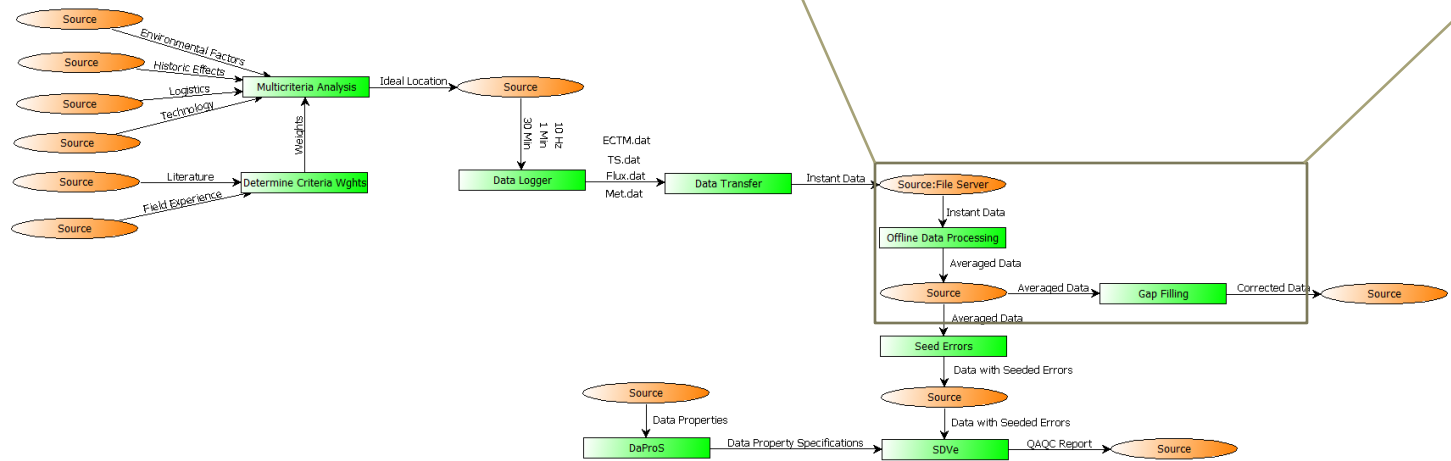
Intuitiveness of terminology



Usefulness

Case Study: Eddy Covariance

- Framework is useful for scientists to:
 - plan their processes
 - represent knowledge
- Scientists want to:
 - specify more details
 - execute



Conclusions

- Framework useful for scientists
- Usefulness validated in Cyber-ShARE's interdisciplinary environment
- Next steps:
 - Extend evaluation efforts beyond Cyber-ShARE
 - PNNL
 - Use represented knowledge (from Cyber-ShARE) to
 - Support tool development efforts
 - Support acquisition of data book-keeping information
 - Improve framework's tool support