



Semantic & Syntactic Consistency - A Critical Enabler for Big Data Analytics

***Mr. Victor Rohr – Aerospace
Mr. Ron Rudnicki – CUBRC***

GSAW Working Group - 4 March 2020

Agenda



- Semantic Consistency – The objective: assuring meaning
 - *Mr. Victor Rohr, Aerospace*
- Semantic Consistency – The value and path forward
 - *Mr. Ron Rudnicki, CUBRC*
- Semantic Consistency – The means for representing knowledge
 - *Dr. David Limbaugh, SUNY Buffalo*
- Semantic & Syntactic Consistency – The prudent use and development of standards
 - *Mr. Scott Houchin, Aerospace Corporation*

Objectives

- Is Semantic / Syntactic consistency a necessary or worthwhile objective?
- What are the obstacles?
- How can we overcome them?
- Is it possible to accurately capture forming knowledge in a data layer?

Semantic Consistency

The Object: Assuring Meaning

- Semantic consistency – The means for assuring information meaning through:
 - *Clear & precise term definitions*
 - *Consistent use of terms*
 - *Precise interpretation*
 - *Accurate & complete information capture*
 - ...
- The working group will first explore the issue of semantic consistency
 - *Stories from participants*
 - *Issues encountered*
 - Mission impact?
 - How handled?

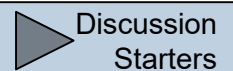


Photo by Steven McDole, used with permission from GANT News

- **Fire Truck supports ladder**
- **Fire Truck supports Fire fighting activity**
- **Fired Department supports *Fire fighting***

How can a machine interpret the meaning of the word *supports*?

Is semantic consistency required?



Semantic Consistency

Value and Path Forward

- The working group will next explore pathways toward semantic consistency, soliciting from the group opinions on:
 - *Best practices for defining data labels*
 - *The value of ontologies*
 - *Formal ontologies vs other approaches*
 - *The value of standardized upper-level ontologies*
 - ISO/IEC 21838: Top level ontologies
 - *Governance, participation and acceptance*
 - Balancing governance with operational implementation
 - Value of a repository of terms
 - *What is required for the repository*
 - *Examples of historical success (e.g. OBO Foundry)*

Is semantic consistency an achievable objective?



Discussion
Starters

Semantic Consistency

A Means for Representing Knowledge

- Big Data Analytics & Artificial Intelligence
 - *How does semantic consistency contribute?*
 - *How does it make diverse methods of machine reasoning more interoperable?*
 - *How does it make machine reasoning more accessible?*
 - *How can instructions be structured for increasingly complex machines, to best –*
 - Query the data they process
 - understand the provenance of that data
- The working group will discuss these challenges with an eye toward an increase in explainability – and thus of controllability – of the sorts of things our machines and networks of machines are doing in the field of intelligence collection and analysis.

Can we accurately capture knowledge without it?






Discussion
Starters



Semantic and Syntactic Consistency

Prudent Use and Development of Standards

- The need for Semantic and Syntactic Consistency for reliable data interpretability 
 - *A significant impedance to optimized ground-based processing and exploitation systems if standards are not developed at the right level of granularity and subsequently followed.*
 - *XML, GML, JSON, JPEG 2000, NITF, ...*
 - Building blocks for syntactic and semantic consistency
 - Enforceable agreements on using building blocks consistently
- War stories from participants: 
 - *Lessons learned*
 - *Successes and failures of using broad, generic standards to meet the needs of specialized systems*
- Success stories 
 - *Development and documentation of broad-standards-based solutions*
 - *The need for custom solutions or custom extensions*

Is there value in data format community agreement?



Thank you

Technical Operations Report (TOR) 2016-03027

Data Sharing and Semantic Understanding across the IC



AEROSPACE REPORT NO.
TOR-2016-03027

Ontology?

- *Why?*
- *The impact of poor choices*

Seven true short stories...

Data Sharing and Semantic Understanding Across the Intelligence Community (IC)

August 30, 2016

Victor S. Rohr¹ and Sara R. Miller²

¹GEOINT Systems Architectures and Strategy, Advanced IC Programs

²System Architecture Engineering and Cost Department, Acquisition Analysis and Planning
Subdivision

Prepared for:

Space and Missile Systems Center
Air Force Space Command
483 N. Aviation Blvd.
El Segundo, CA 90245-2808

Contract No. FA8802-14-C-0001

Authorized by: National Systems Group

Distribution Statement A: Approved for public release; distribution unlimited.





1. Conflicting & Ambiguous Vocabularies

My terms are best

- Term selection left to developers and analysts
 - *Missing or poor definitions*
 - *No need to define – everyone will know what I mean*
- In reality
 - *Individuals entering data must “best guess” the meanings of terms*
 - *Result:*
 - Vocabulary that does not accurately represent concepts
 - Chosen vocabulary is simply incorrect for the concept being expressed
- Example of terms leading to inaccurate or confusing meaning:
 - ***Mary Associated With St. Patrick’s Day Parade***
 - ***XYZ Squadron allotted Aircraft, TN 123***
 - ***Aircraft, TN 123 assigned to XYZ Squadron***
 - ***XYZ Security Group assigned to 2016 Olympic Games***
 - ***2016 Olympic Games allotted XYZ Security Group***

Return






2. Inconsistent Information Capture

Everyone will understand my terms

- The terms I've chosen will be used in a consistent manner
- In reality
 - *Multiple vocabulary choices used for the same meaning*
 - *Multiple meanings for the same vocabulary*
- Examples (within a single database!) to express a relation between an actor and an event:
 - **Actor Associated With Activity**
 - **Actor Affiliated With Activity**
 - **Actor Supports Activity**
 - **Actor Assigned To Activity**
 - **Actor Involved With Activity**
 - **Activity Affiliated Personnel** <Unconstrained text property>
 - **Actor Participant in Activity**
- Other term meanings
 - **Tower structure supports equipment**
 - **Actor assigned to Role**

More than one term per concept / more than one concept per term


Return 



3. *Inaccurate or Incomplete Information Capture*

These terms are sufficiently broad

- The terms provided are sufficiently expressive
- In reality
 - *Terms that seems reasonable when conceived, quickly prove too limiting*
- Example:
 - **Wile E Coyote** employed by **ACME Corporation**
- With this term, how is it possible to express:
 - *The job held by Wille Coyote*
 - Over what time period
 - *The division or group under which this position was held*
 - *Current and past managers for this position*
 - *The physical location for this position*
 - Office location
 - Performance location
 - *Equipment needed or used for this position*
 - *Other positions held by Wile E Coyote*

Return 



4-1. Data Interoperability

Because my way is best, all will understand

As an example of this problem space, consider the statements:

- **James Kirk** is the commanding officer of the **Enterprise**
- **James Kirk** is an instance of a **person**
- **Enterprise** is an instance of a **starship**

Program 1

- **Enterprise** *Commanding Officer* “CAPT James Kirk”
- **James Kirk** *Position* “Commander of the Enterprise”
- Alternatively:
- **Enterprise** *Commanded_By* **James Kirk**
- **James Kirk** *Commands* **Enterprise**

Program 2

- When program 2 began, it adapted and was aligned with the language in use by program 1. The possible approaches to expressing the above statements are nearly identical.
- **James Kirk** *Title* “Commander of the Enterprise”
- Alternatively:
- **Enterprise** *Commanded By* **James Kirk**
- **James Kirk** *Commands* **Enterprise**

Program 3

- **James Kirk** *In Command Of* **Enterprise**
- No reciprocal term

Program 4

- **James Kirk** *Title* “Commander of the Enterprise”
- Alternatively:
- **Enterprise** *Unit Commanded By* **James Kirk** ← implies a military unit, not a starship
- **James Kirk** *Commander of* **Enterprise**

Program 5

- **James Kirk** *Third Party Reference Title* “Commander of the Enterprise”
- Alternatively:
- **Enterprise** *Led By* **James Kirk**
- **James Kirk** *Leader of* **Enterprise**

All are limiting in scope!



4-2. Data Interoperability

Expanded Scope

James Kirk *instance of* **Person**

Star Fleet Command *instance of* **Organisation**

Star Fleet Command HQ *instance of* **Facility**

Enterprise *instance of* **Starship**

Enterprise Crew *instance of* **Organisation**

- **Commander of Enterprise Crew** *instance of* **Organisation Internal Role Specification**
 - *Internal Position Type* : Head of Organisation
 - *Internal Role Defined By* **Star Fleet Command**
 - *Internal Role Assignment Location* **Star Fleet Command HQ**
 - *Role Description* : “To command a crew to go where no one has gone before using the Starship Enterprise”
- *Internal Role Performance* [**James Kirk : Commander of: Enterprise Crew**]
- ... Other properties

[James Kirk: Commander of: Enterprise Crew] *instance of* **Organization Internal Role**

- *Internal Role Provider* **Star Fleet Command**
- *Internal Role Holder* **James Kirk**
- *Internal Role Performed* **Commander of Enterprise Crew**
- *Internal Role Performance Time Period* : Star date A to star date B

Example using terminology from the NSG Application Schema (NAS)

Return






5. Data Interconnectivity

Sharing knowledge

- Object based production is a key enabler
 - *Observations need only reside in my local system*
- In reality:
 - *Maximum benefit comes from integrating knowledge*
 - Information in one analytic cell may be of immediate value to another
 - Methods for rapid discovery and integration is required
 - *The requirement for linked data remains under-defined and inconsistently understood*
- Innovative technologies and methodologies for dynamic integration of enterprise-wide information is required

UUID: 123e4567-e89b-12d3-a456-426655440000

Goal: Tap into the full corpus of IC-wide knowledge


Return 



6. Analytic Workflow Capture and Information Provenance

Why did you say that?

- I need only provide the resultant conclusions of my analytic process
 - *Resultant facts are all that matter*
- In reality:
 - *Maximum value if results of analysis are provide with linkage to the analytic processes, algorithms, and data used*
 - *Required to advance machine learning*

Return 

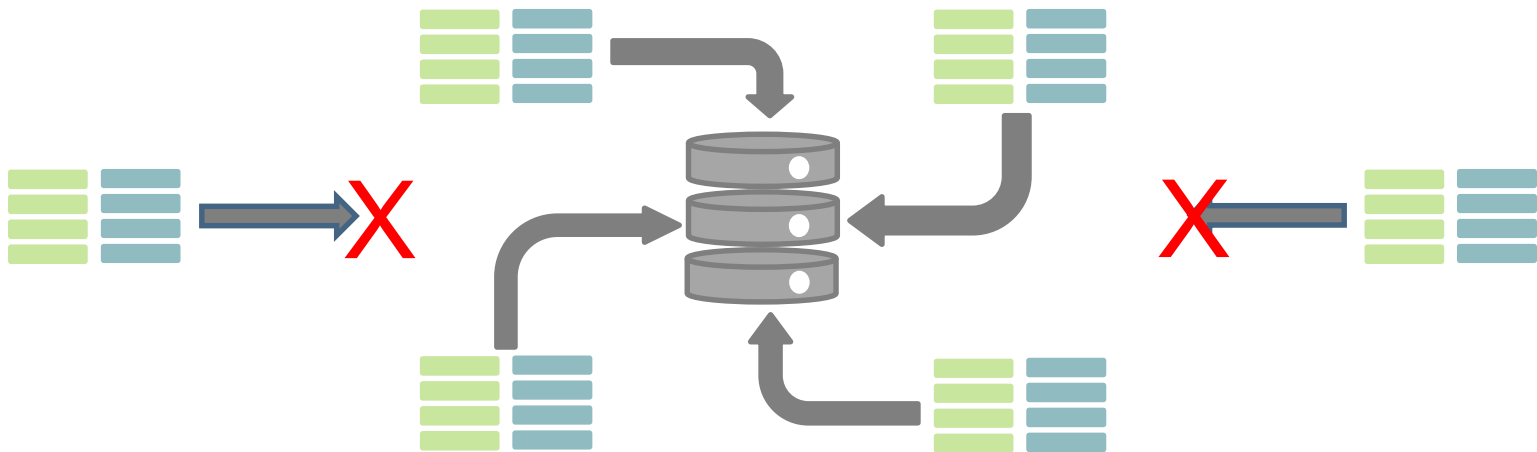
Semantic Consistency – the Value and Path Forward

GSAW Working Group - March 4, 2020

The Value - Delivering on the Promise of Big Data

The promise of Big Data is based partly on the thought that any data may prove to be valuable to someone or some algorithm at some point in time

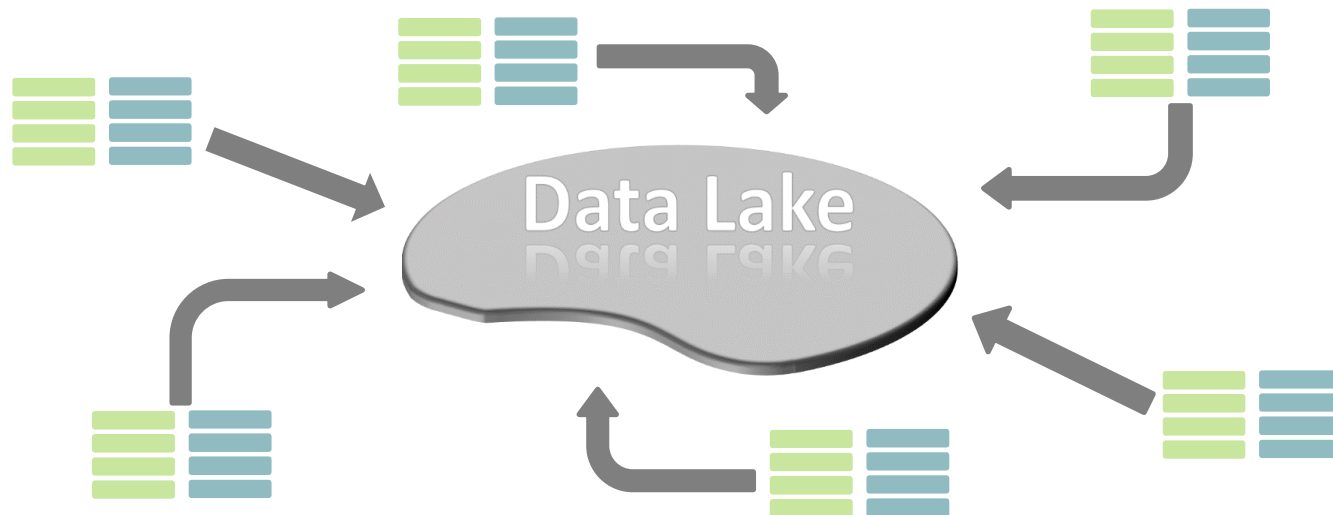
Architectures such as the data warehouse that use a “schema at write” run counter to this by determining at design which data is important enough to be available for analysis by being included in the warehouse



The Value - Delivering on the Promise of Big Data

An uncurated data lake uses a “schema at read” approach by being a repository of all data in native format

Experiences with data swamps lead to the view* that a curated data lake better enables analysis of big data



*Terrizzano, Ignacio G., et al. "Data Wrangling: The Challenging Journey from the Wild to the Lake." *CIDR*. 2015.

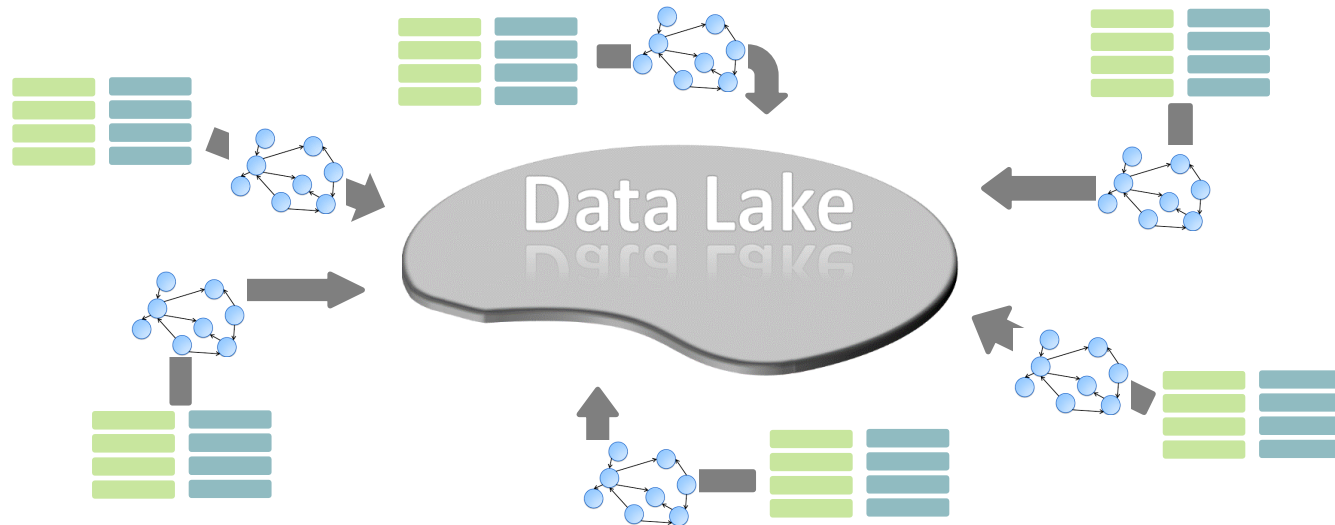
Curation of a data lake is a complex process comprising the subtasks of:

- *Procuring data: Identifying data sources for inclusion*
 - *Vetting data: Understanding transaction schedules, legal use and security*
 - *Obtaining data*
 - *Describing data*
 - *Grooming data: Standardizing data formats, entity resolution*
 - *Provisioning data: policies and process for data retrieval*
 - *Preserving data: maintenance and archival tasks*
- } traditional extract, transform and load processes

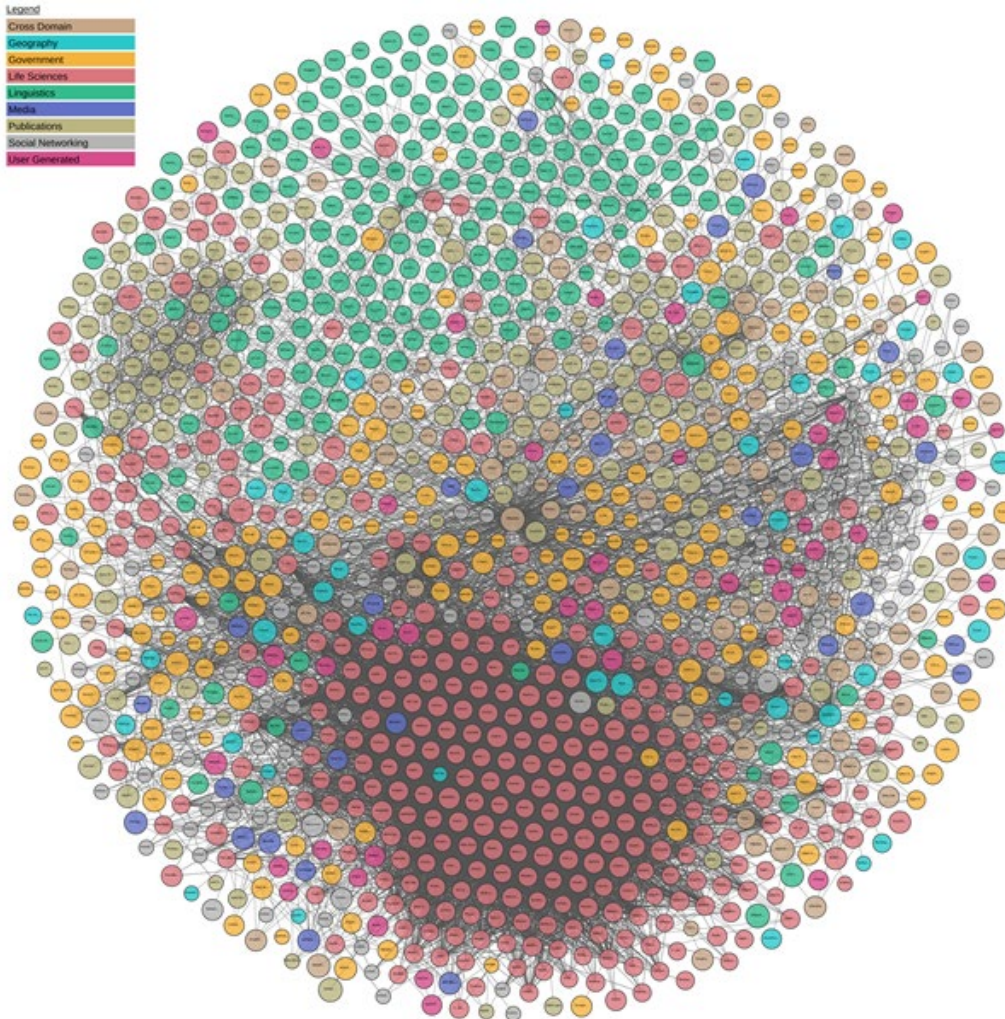
The optimal curated data lake would be one in which all data used and produced by all of these tasks was standardized and linked

The Path – Enabling Standardization and Linking

Ontologies can provide a standard semantics and linkage of data but there are different development methodologies from which to choose



The Path - Linked Data Method of Development



One method of building ontologies is proposed by the Linked Data Community*

Ontologies are created for domain of interests and data sets are linked to others via mappings

[*http://linkeddata.org/home](http://linkeddata.org/home)

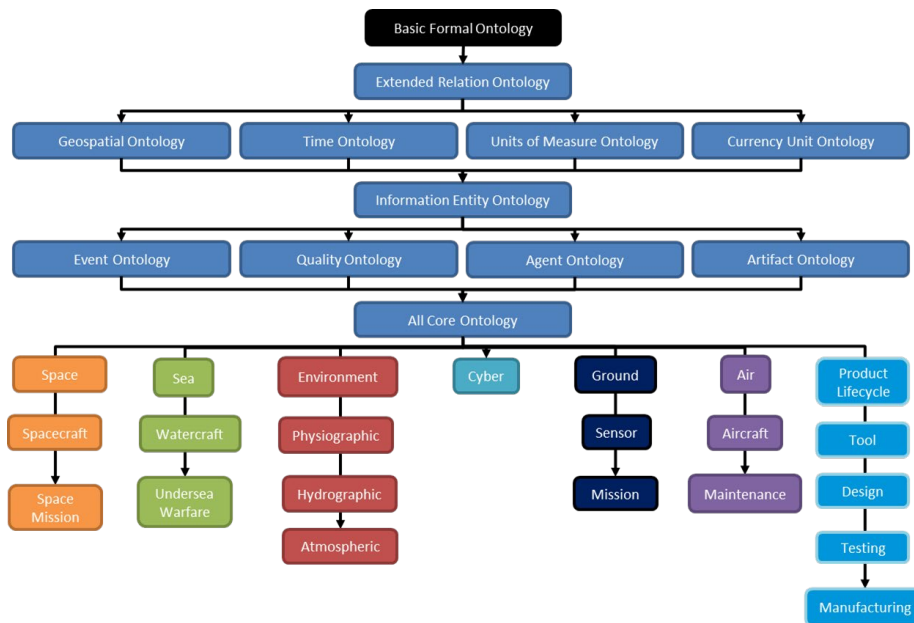
Pros:

- *Agile, low overhead of constraints on ontology development*
- *Active community, 1,239 data sets (of 1000 or more elements) and 16,147 links as of March 2019*
- *High usability, method is straightforward*

Cons:

- *High maintenance costs, changes cause ripple affect across mappings*
- *Error prone, mappings created by humans tend to ignore context of terms (e.g. inheritance of semantics from ancestor classes)*
- *Precision and recall of queries depend on completeness and accuracy of mappings*

The Path – Foundry Method of Development



A method of building ontologies proposed by the Open Biological and Biomedical (OBO) Foundry* and the National Center for Ontological Research**

Ontologies are created by extending from a hierarchy of ontologies comprised of upper-level, mid-level and domain level ontologies. Data sets linked via mappings to these ontologies

*<http://obofoundry.org/>

**<https://ubwp.buffalo.edu/ncor/>

Pros:

- *Active community, primarily in biomedical domains but growing within IC and DoD*
- *Semantic consistency between ontologies promotes re-use of queries and algorithms*
- *Singular terms and hierarchies improve precision and recall of queries*

Cons:

- *Rigor of method extends development time*
- *Higher learning curve than Linked Data*
- *Quality of data lake content dependent on quality of mappings*

Semantic Consistency

Governance

1. Establish and resource a standing working group focused on IC & DoD wide semantic integration and collaboration
2. Establish a DoD and IC ontology repository to serve as standard for information system semantics
3. Establish rules and best practices for developing ontologies and submitting them to the repository
4. Establish processes for the continual review and vetting of discrepancies and issues
5. Establish artifacts to foster understanding of ontologies, best practices, and other required topics

Return



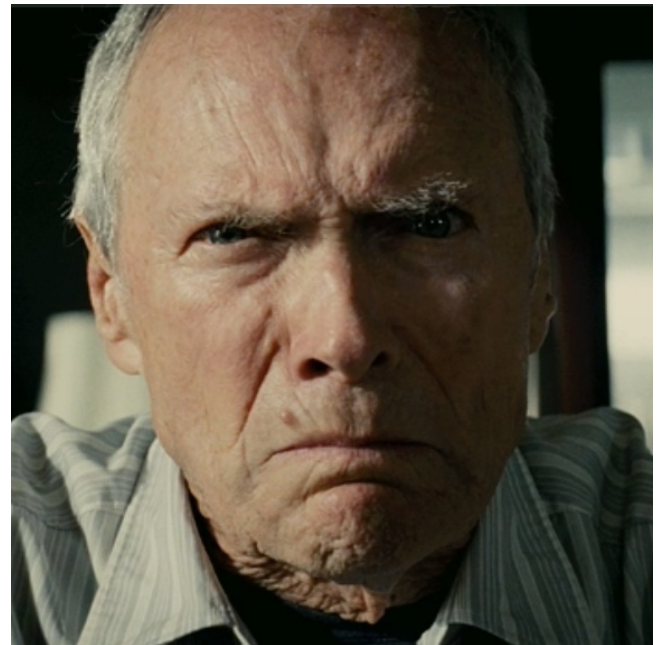
Semantic Consistency – The means for representing knowledge

**David Gordon Limbaugh
Intelligence Community Postdoc
University at Buffalo
GSAW Working Group - 03/04/2020**

Referent Tracking and Portions of Reality

- A Portion of Reality is literally anything: a fleet of ships, a flash of insight, a risk of cyberattack, and so on.
- A benefit, for example, is this allows us to construe ‘indicator’ much more broadly.
- An *indicator* is a portion of reality that, if it exists, affects our estimation that some other portion of reality exists.

Indicators and Portions of Reality



Building Datasets That Mirror Reality

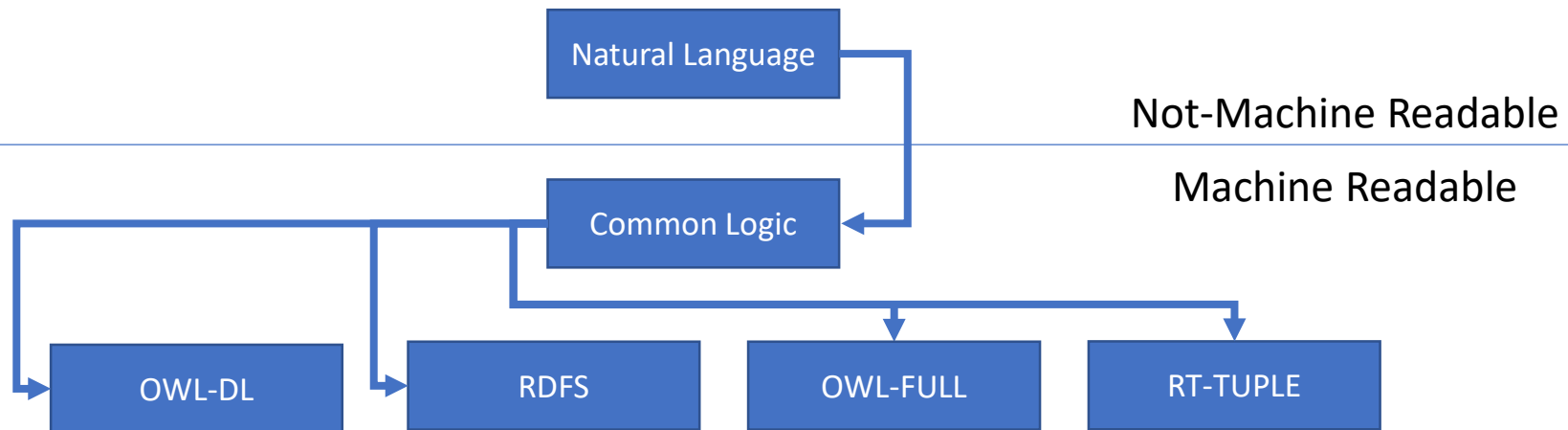
- Reality is made of unique entities with shared features and relationships indexed to locations and times.
- An Referent Tracking System (RTS) mirrors reality by using
 1. unique identifiers to refer to unique entities,
 2. terms from a controlled vocabulary to represent features and relationships, and
 3. time-indexed, first-order logic expressible, assertions to represent when an entity has some feature or some relationship to other entities.

Terms and Controlled Vocabularies

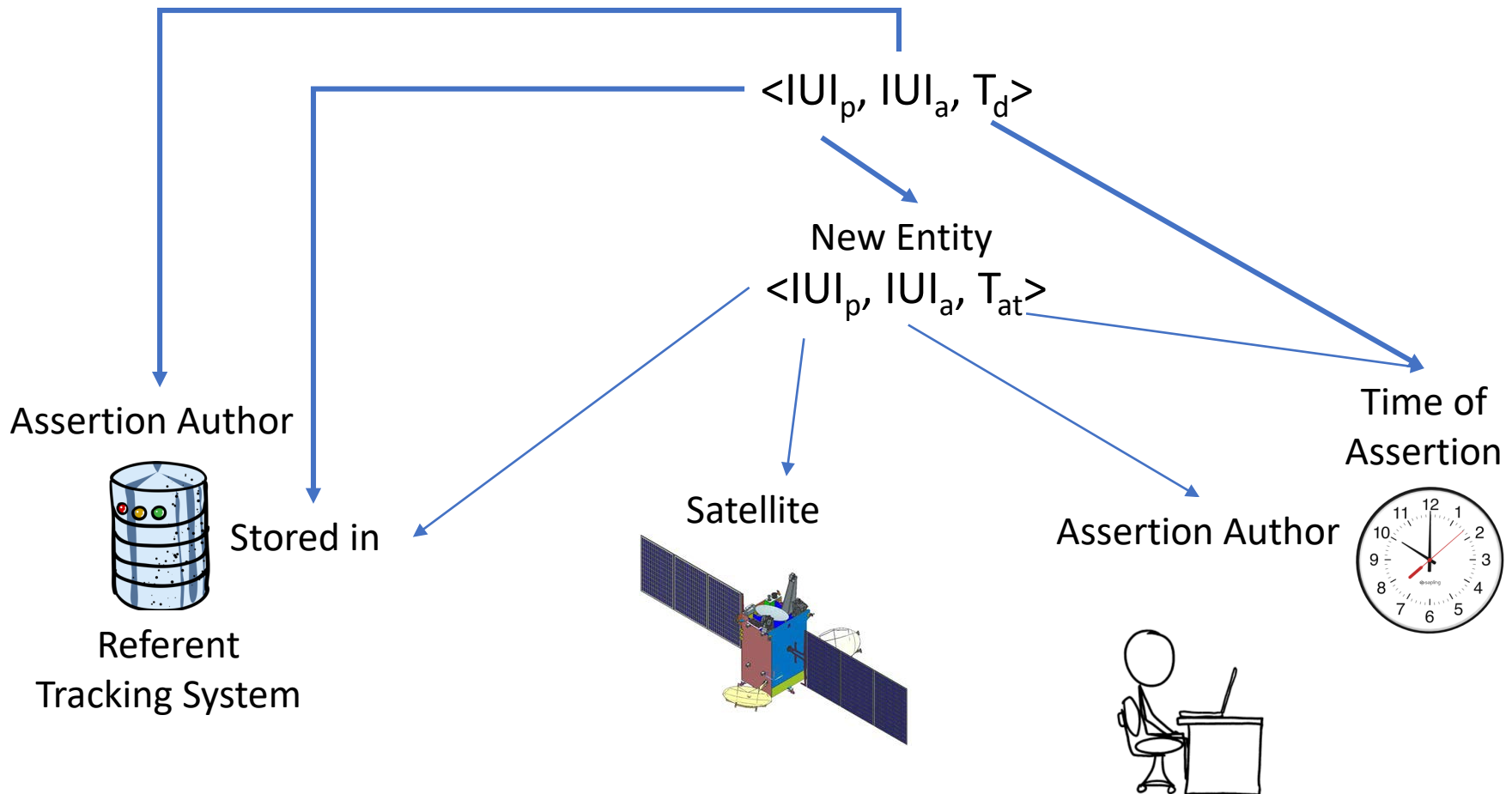
- Controlled vocabularies are organized in a modular fashion
- Basic Formal Ontology (BFO: ISO/IEC 28138-2) as top-level hub
- The Common Core Ontologies (CCO) as mid-level spokes
- The metric system of ontology.

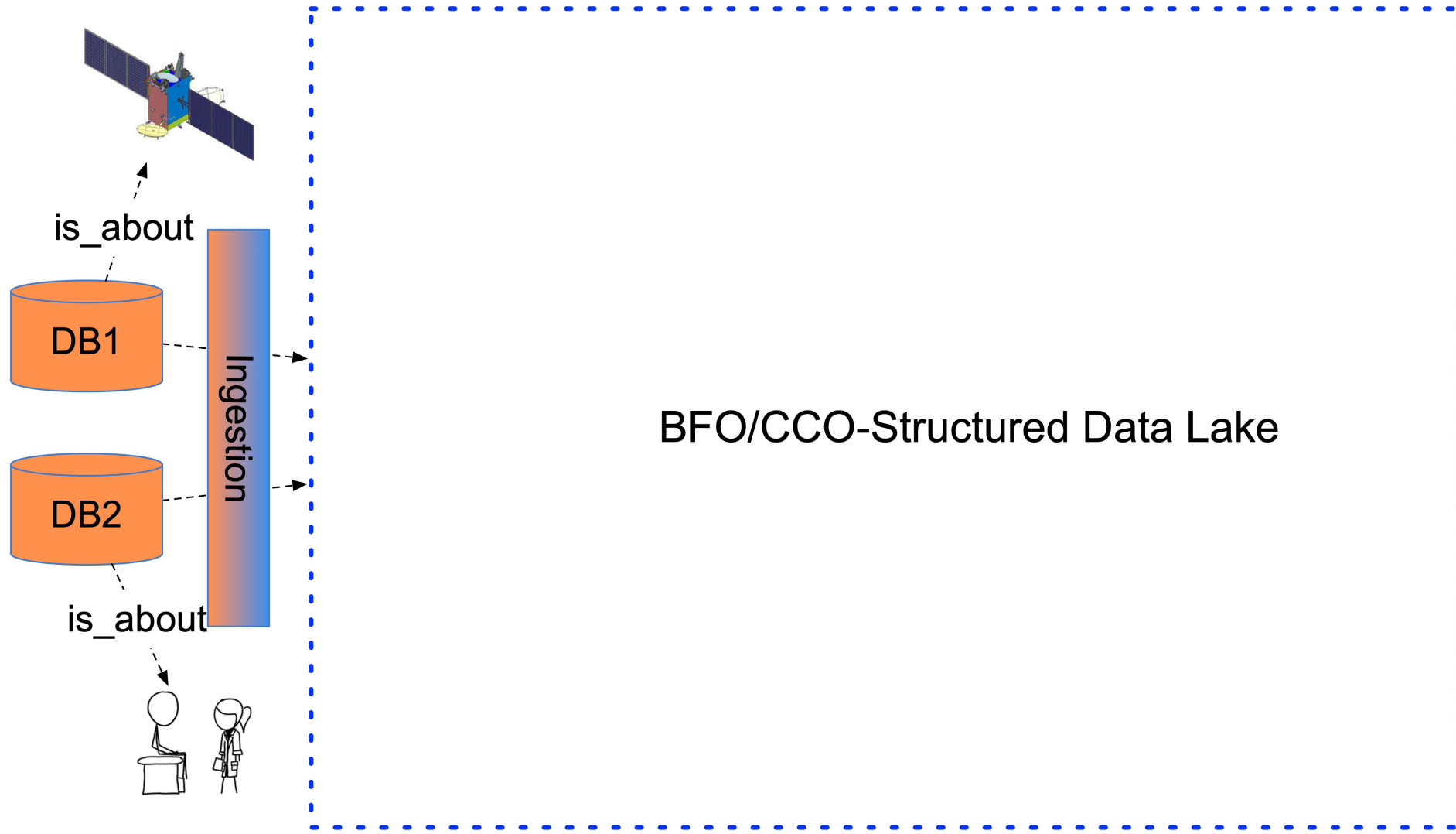
Assertions Consistent with Common Logic

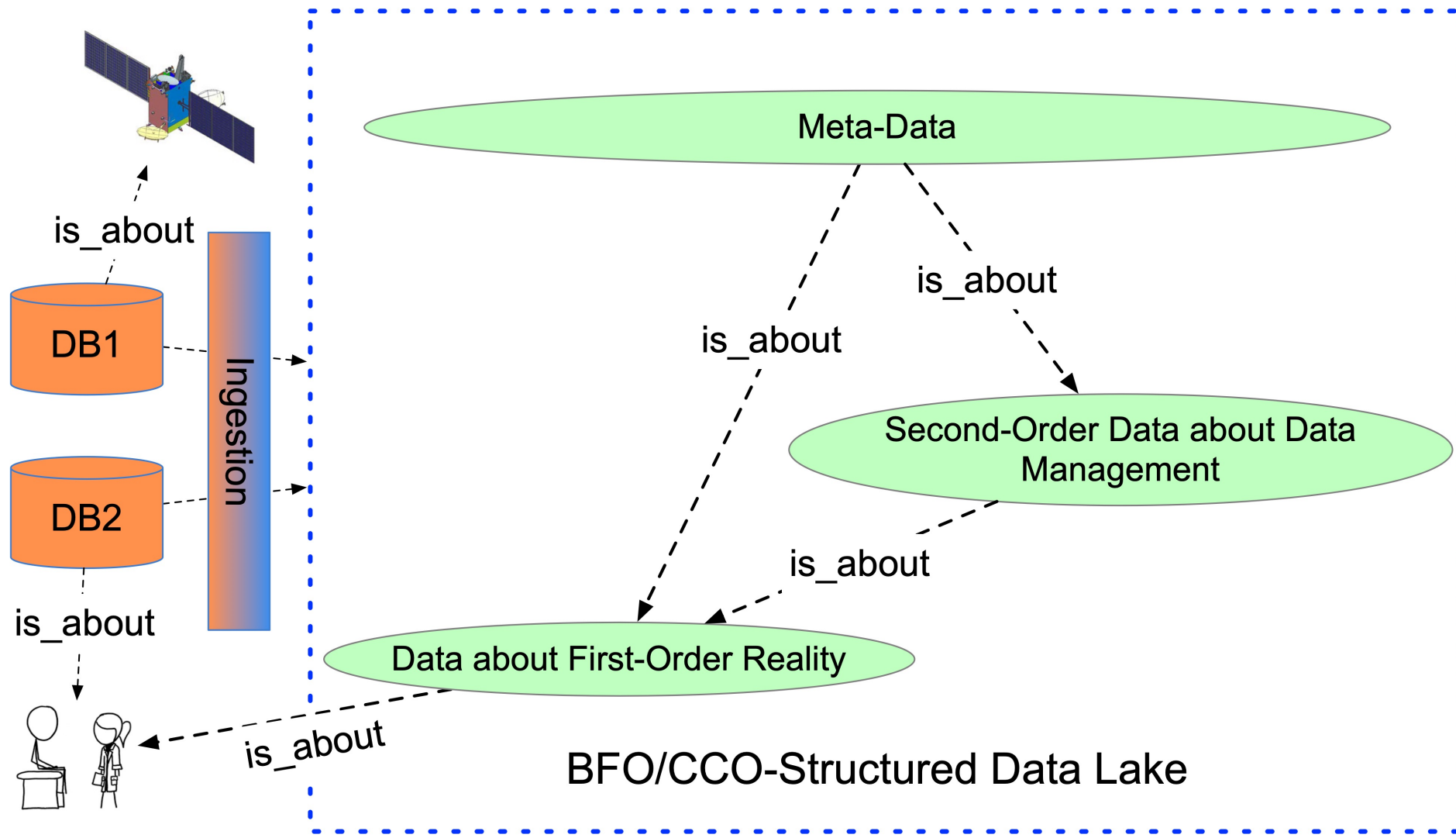
Common Logic (CL) is a framework for a family of logic languages, based on first-order logic, intended to facilitate the exchange and transmission of knowledge in computer-based systems.



Multiple Levels of Data







Meta-Data

**Second-Order
Data**

**First-Order
Data**

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Meta-Data

**Second-Order
Data**

**First-Order
Data**

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Meta-Data

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Second-Order Data

First-Order Data

Tuple3:<Doctor1 examines Patient1 at 9:00Pst>

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Meta-Data

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Tuple6:<Tuple3 asserted_by System at 12:53Est>

Second-Order Data

First-Order Data

Tuple3:<Doctor1 examines Patient1 at 9:00Pst>

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Meta-Data

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Tuple6:<Tuple3 asserted_by System at 12:53Est>

Second-Order Data

Tuple2:<Ingestion1 outputs Tuple3 at 12:54Est>

First-Order Data

Tuple3:<Doctor1 examines Patient1 at 9:00Pst>

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Meta-Data

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Tuple6:<Tuple3 asserted_by System at 12:53Est>

Tuple5:<Tuple2 asserted_by System at 12:55Est>

Second-Order Data

Tuple2:<Ingestion1 outputs Tuple3 at 12:54Est>

First-Order Data

Tuple3:<Doctor1 examines Patient1 at 9:00Pst>

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Meta-Data

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Tuple6:<Tuple3 asserted_by System at 12:53Est>

Tuple5:<Tuple2 asserted_by System at 12:55Est>

Second-Order Data

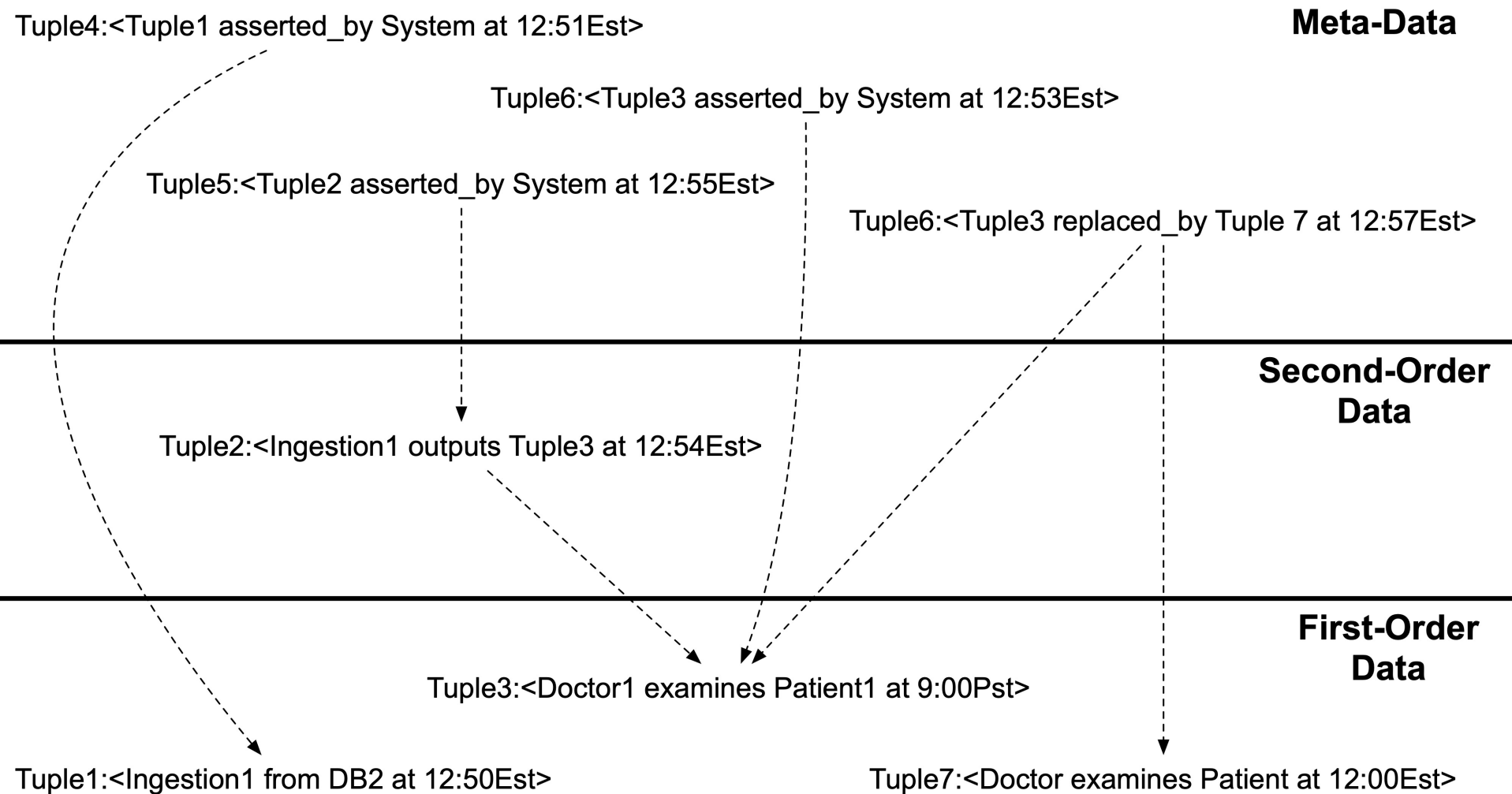
Tuple2:<Ingestion1 outputs Tuple3 at 12:54Est>

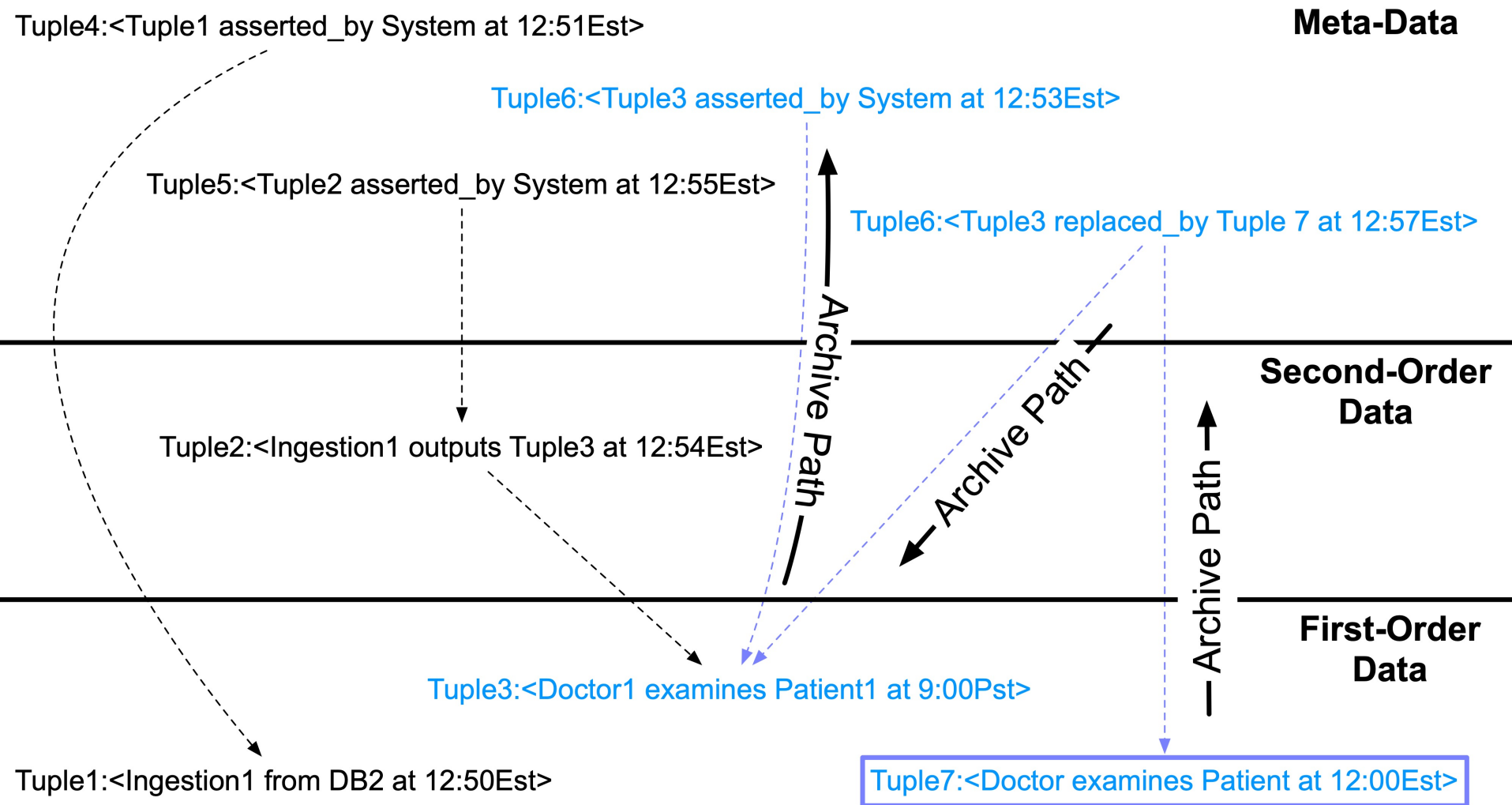
First-Order Data

Tuple3:<Doctor1 examines Patient1 at 9:00Pst>

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Tuple7:<Doctor examines Patient at 12:00Est>





Meta-Data

Second-Order Data

First-Order Data

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Tuple5:<Tuple2 asserted_by System at 12:55Est>

Tuple2:<Ingestion1 outputs {T,T,T, ...} at 12:54Est>

Tuple 8

Tuple 9

Tuple 13

Tuple 10

Tuple 12

Tuple 11

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Meta-Data

Second-Order Data

First-Order Data

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Tuple5:<Tuple2 asserted_by System at 12:55Est>

Tuple2:<Ingestion1 outputs {T,T,T, ...} at 12:54Est>

Tuple 8

Tuple 9

Tuple 13

Tuple 10

Tuple 12

Tuple 11

Tuple1:<Ingestion1 from DB2 at 12:50Est>

Meta-Data

Tuple4:<Tuple1 asserted_by System at 12:51Est>

Tuple5:<Tuple2 asserted_by System at 12:55Est>

Second-Order Data

Tuple2:<Ingestion1 outputs {T,T,T, ...} at 12:54Est>

First-Order Data

Tuple 8

Tuple 9

Tuple 13

Tuple 10

Tuple 12

Tuple 11

Tuple1:<Ingestion1 from DB2 at 12:50Est>

This research was supported by an appointment to the Intelligence Community Postdoctoral Research Fellowship Program at the University at Buffalo, administered by Oak Ridge Institute for Science and Education through an interagency agreement between the U.S. Department of Energy and the Office of the Director of National Intelligence.

These and similar images reproduced from xkcd.com and with thanks to Randall Munroe.





The need

Semantic and Syntactic Consistency

- How much commonality do we really need when considering data from disparate but similar sources
 - e.g., JPSS and GOES both provide spectral image data of the earth, but in
 - Different data formats (generic HDF vs netCDF)
 - Different mechanisms to georeference individual pixels
 - Different metadata and different metadata formats
 - *If I want to use GOES data in a JPSS-centric toolset, I've got to start from scratch*
 - *So where should the line be? At what point do I start losing what's special about GOES if I make it's data look just like JPSS?*
- How forward leaning do we need to be in our data formats? Designing in forward compatibility from the start is great, but at what point do we wind up with something that can't be implemented without profiling it down?
- How do we encourage developers (and contracts) to focus on the broader standards and not just on the narrow profile immediately at hand?



War stories

Semantic and Syntactic Consistency

- Have you run into product for formats (or documentation of said formats) that promised to be reusable, promote interoperability, minimize downstream impact to data consumers, but ultimately failed
- How did your programs respond and redirect?
- Have you fought against the “my system is a one-of-a-kind thing, and thus there’s no reason to try to make my data just like everyone else” monster?
 - *Promoting common standards and data modeling*
 - *Spending more on the provider in order to spend less on the consumer*

Success stories

Semantic and Syntactic Consistency



- Have you developed or found data/product format standards that
 - *Could be easily profiled for use by disparate systems*
 - *Had enough flexibility to handle data from complex one-of-a-kind systems*
 - e.g., not just a simple image but a data product that includes both collected and supplemental images (cloud grid, pixel quality, ...) within the same file
 - *Could be profiled/documented without having expert-level skill in that standard*
 - *Did not require any/significant retooling by existing data consumers (e.g., image viewers)*
 - *Image standards, Signal standards, Metadata/analysis output*
- Have you found format documentation and design methods that
 - *Allowed significant level of reuse*
 - *Minimized redundancy across documents, minimized errors stemming from redundancy*
 - *Encouraged reuse and building upon existing standards*
 - *Facilitated later reuse and extension*