Developing Ontologies and Their Role for Engineering Information Fusion Systems

David G Limbaugh, Ron Rudnicki, Barry Smith

Intelligence Community Postdoc

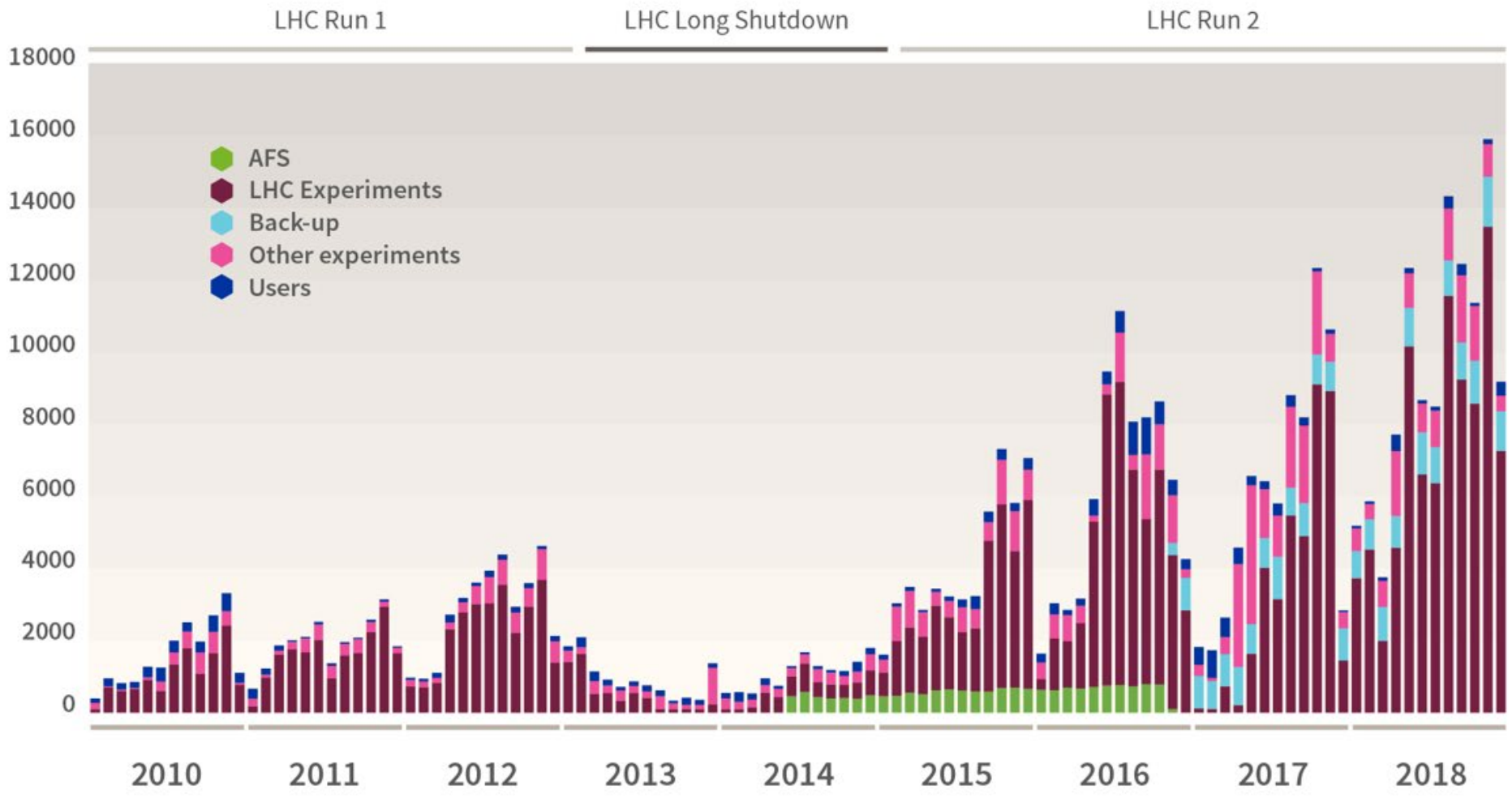
University at Buffalo

GSAW 2020 – 03/04/2020

Big Data (Data Hoarding)

- Any item of data may prove to be valuable to someone or some algorithm at some point in time
- Thus, collect everything and, if possible, delete nothing
- Data Lakes are a way of centralizing data for future exploitation

CERN (terabytes-to-tape per month)

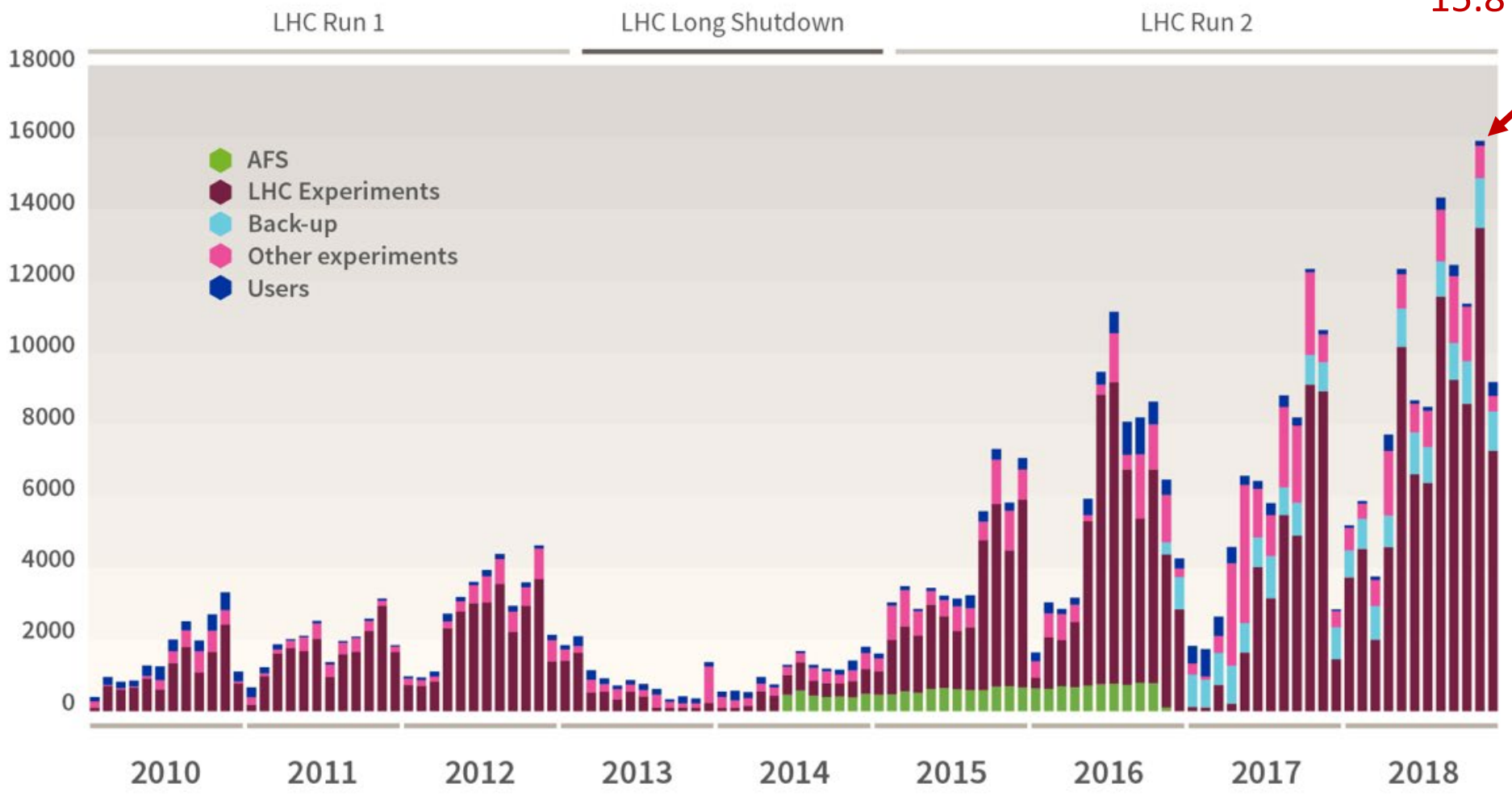


Esma Mobs/CERN https://cds.cern.ch/images/CERN-HOMEWEB-PHO-2019-003-1 (accessed 02/03/2020)

1 PB = 1000 TB

CERN (terabytes-to-tape per month)

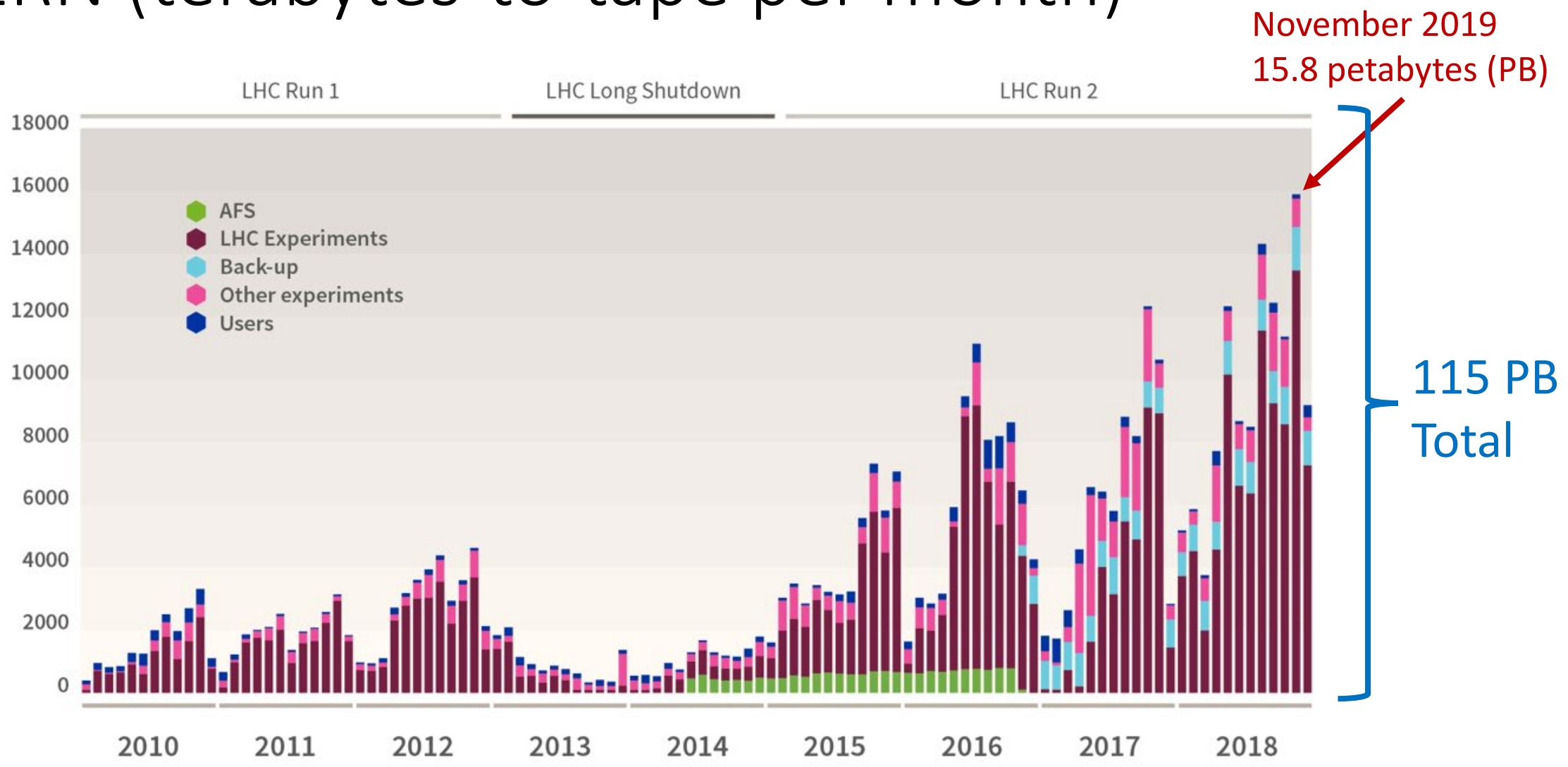
November 2019
15.8 petabytes (PB)



Esma Mobs/CERN https://cds.cern.ch/images/CERN-HOMEWEB-PHO-2019-003-1 (accessed 02/03/2020)

1 PB = 1000 TB

CERN (terabytes-to-tape per month)



Esma Mobs/CERN https://cds.cern.ch/images/CERN-HOMEWEB-PHO-2019-003-1 (accessed 02/03/2020)

Data Lake

- Un-curated (unstructured) Data Lake:
  - Schema-at-read
  - Difficult to find data without a known schema

- Curated (structured) Data Lake:
  - Schema-at-write
  - Requires a schema appropriate for all data types

Curated Data Lake

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

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

Tracking Provenance of Curation

- Where did the data come from?
- When did it become available?
- Who (or what) procured the data?
- What format was it in previously?
- Are there restrictions on use?
- What errors have been corrected?

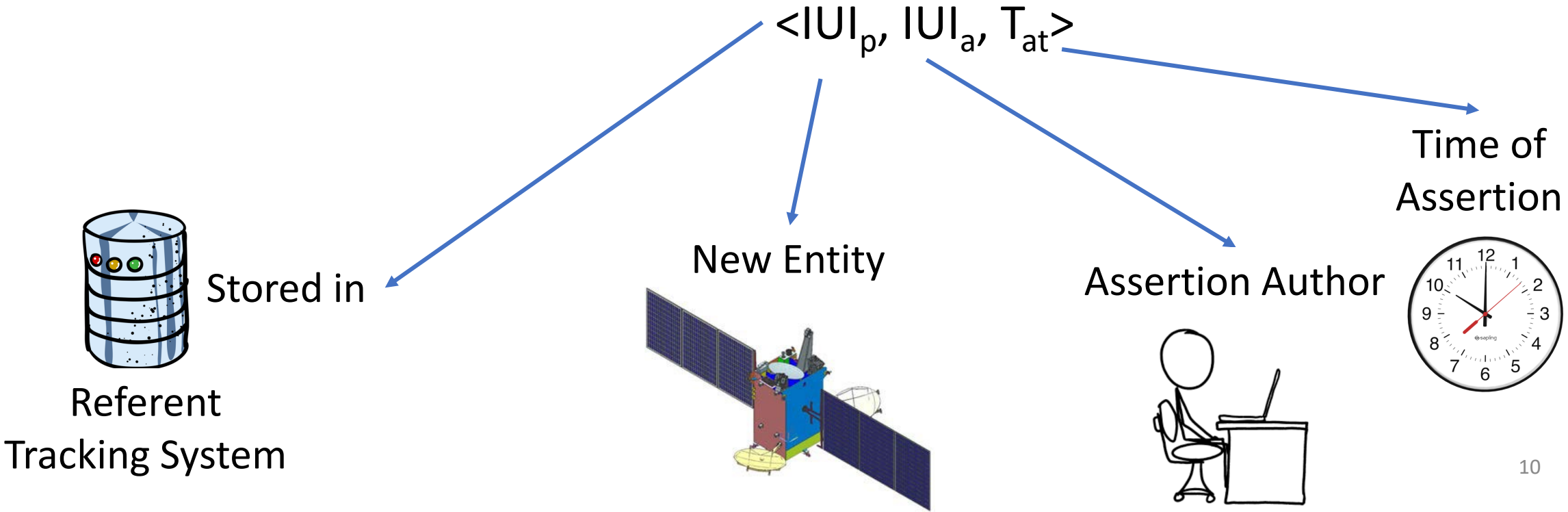
Tracking Provenance of Curation

- Reservation Assertion reserves UUID (or IUI) of new entity
- Particular to Portion of Reality assertion puts the new entity into relationship with other entities
- Particular not-to Portion of Reality assertion explicitly excludes some relationship between an entity and a portion of reality
- Particular to Name assertion assigns non-unique human-readable labels as alternatives to UUIDs
- Meta-Data Assertion captures the author, time, integrity status, of assertions

IUI = Instance Unique Identifier

Reservation Tuple

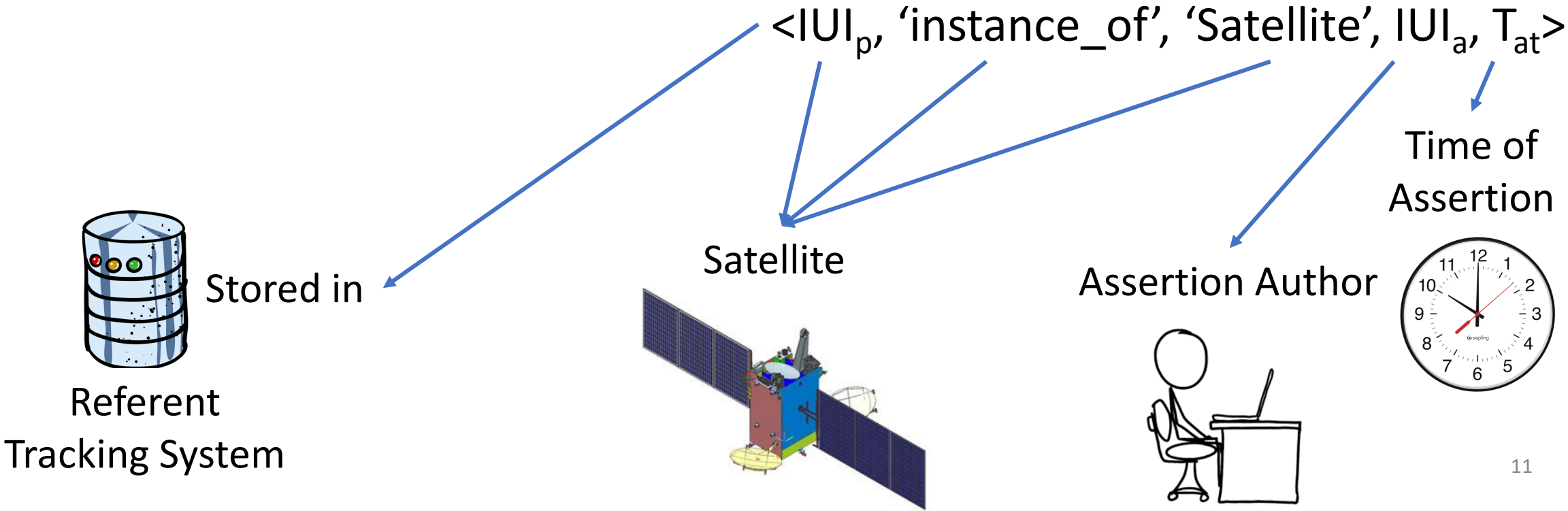
- Reserves an IUI and asserts that it refers to some portion of reality.



IUI = Instance Unique Identifier

Particular to Portion of Reality Tuple

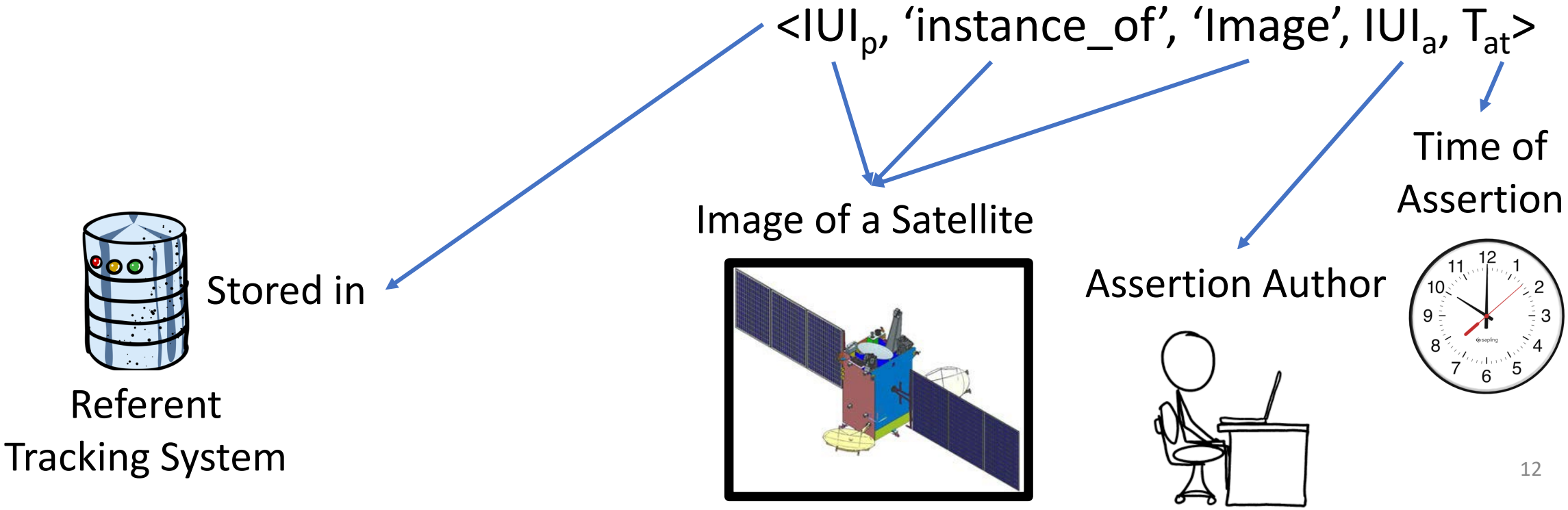
- Asserts that the referent IUI is in some relationship with some other portion of reality.



IUI = Instance Unique Identifier

Particular to Portion of Reality Tuple

- Asserts that the referent IUI is in some relationship with some other portion of reality.



IUI = Instance Unique Identifier

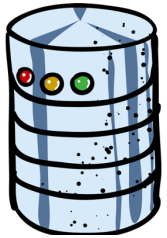
Meta-Data Tuple

<IUI$_p$, IUI$_a$, T$_d$, E, S>

New Entity

<IUI$_p$, IUI$_a$, T$_{at}$>

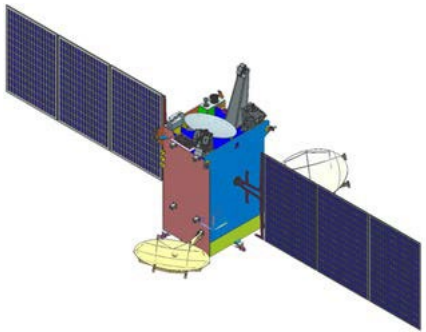
Assertion Author



Stored in

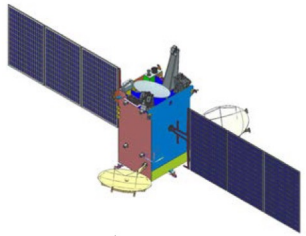
Referent Tracking System

New Entity



Assertion Author

Time of Assertion



is_about

DB1

DB2

Ingestion

is_about

is_about

Meta-Data

is_about

is_about

Second-Order Data about Data Management

is_about

Data about First-Order Reality

BFO/CCO-Structured Data Lake

<IUI#123, instance_of, BUS, t1>

is_about

Bus

<IUI#Tuple1, IUI#Assigner1, t2, Good, Null>

is_about

<IUI#123, instance_of, BUS, t1>

is_about

Bus

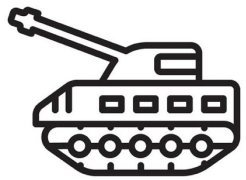
<IUI#Tuple1, IUI#Assigner1, t2, Good, Null>

is_about

<IUI#123, instance_of, BUS, t1>

is_about

Tank



<IUI#Tuple1, IUI#Assigner1, t2, Good, Null>

is_about

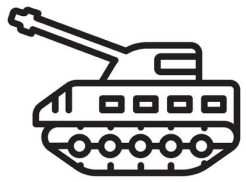
<IUI#Tuple1, IUI#Assigner1, t3, Bad, IUI#Tuple2>

is_about

<IUI#123, instance_of, BUS, t1>

is_about

Tank



<IUI#Tuple1, IUI#Assigner1, t2, Good, Null>

is_about

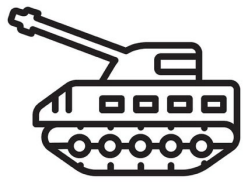
<IUI#Tuple1, IUI#Assigner1, t3, Bad, IUI#Tuple2>

is_about

<IUI#123, instance_of, BUS, t1>

is_about

Tank



<IUI#Tuple1, IUI#Assigner1, t2, Good, Null>

is_about

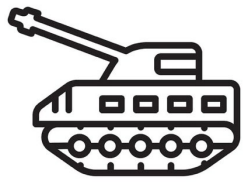
<IUI#Tuple1, IUI#Assigner1, t3, Bad, IUI#Tuple2>

is_about

<IUI#123, instance_of, BUS, t1>

is_about

Tank



<IUI#Tuple2, IUI#Assigner, t5, P+1, Null>

<IUI#Tuple1, IUI#Assigner1, t2, Good, Null>

is_about

is_about

<IUI#Tuple1, IUI#Assigner1, t3, Bad, IUI#Tuple2>

is_about

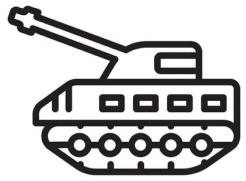
<IUI#123, instance_of, BUS, t1>

<IUI#123, instance_of, TANK, t4>

is_about

is_about

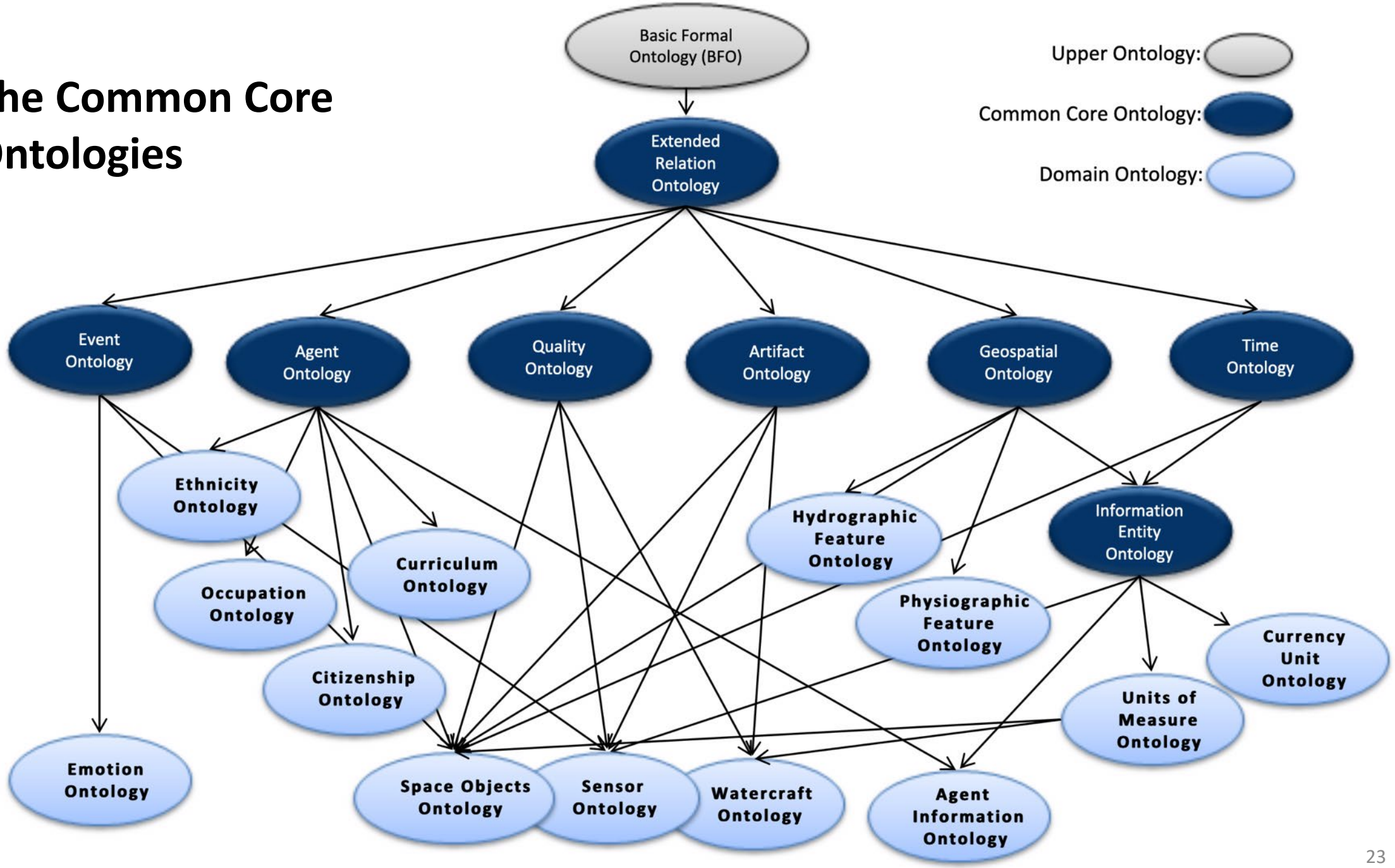
Tank



Realist Curation

- Data and Meta-data should be tagged with Realist Curation Ontologies, which means terms should
  - ➢ make clear reference to scientific reality and
  - ➢ have easy to grasp definitions.

The Common Core Ontologies



Realist Curation

- Data and Meta-data should be tagged with realist curation ontologies meaning terms
  - ➢ make clear reference to scientific reality and
  - ➢ have easy to grasp definitions.

- Curation ontologies should fall under a top level ontology conformant with ISO-21838 Information Technology – Top-Level Ontologies (TLO) – Part 1: Requirements

ICS › 35 › 35.060

ISO/IEC PRF 21838-1 [ISO/IEC DIS 21838-1]

Information technology — Top-level ontologies (TLO) — Part 1: Requirements

GENERAL INFORMATION

Status : ⊙ Under development **Publication date :** 2020-03

Edition : 1

ICS › 35 › 35.060

ISO/IEC PRF 21838-2 [ISO/IEC DIS 21838-2]

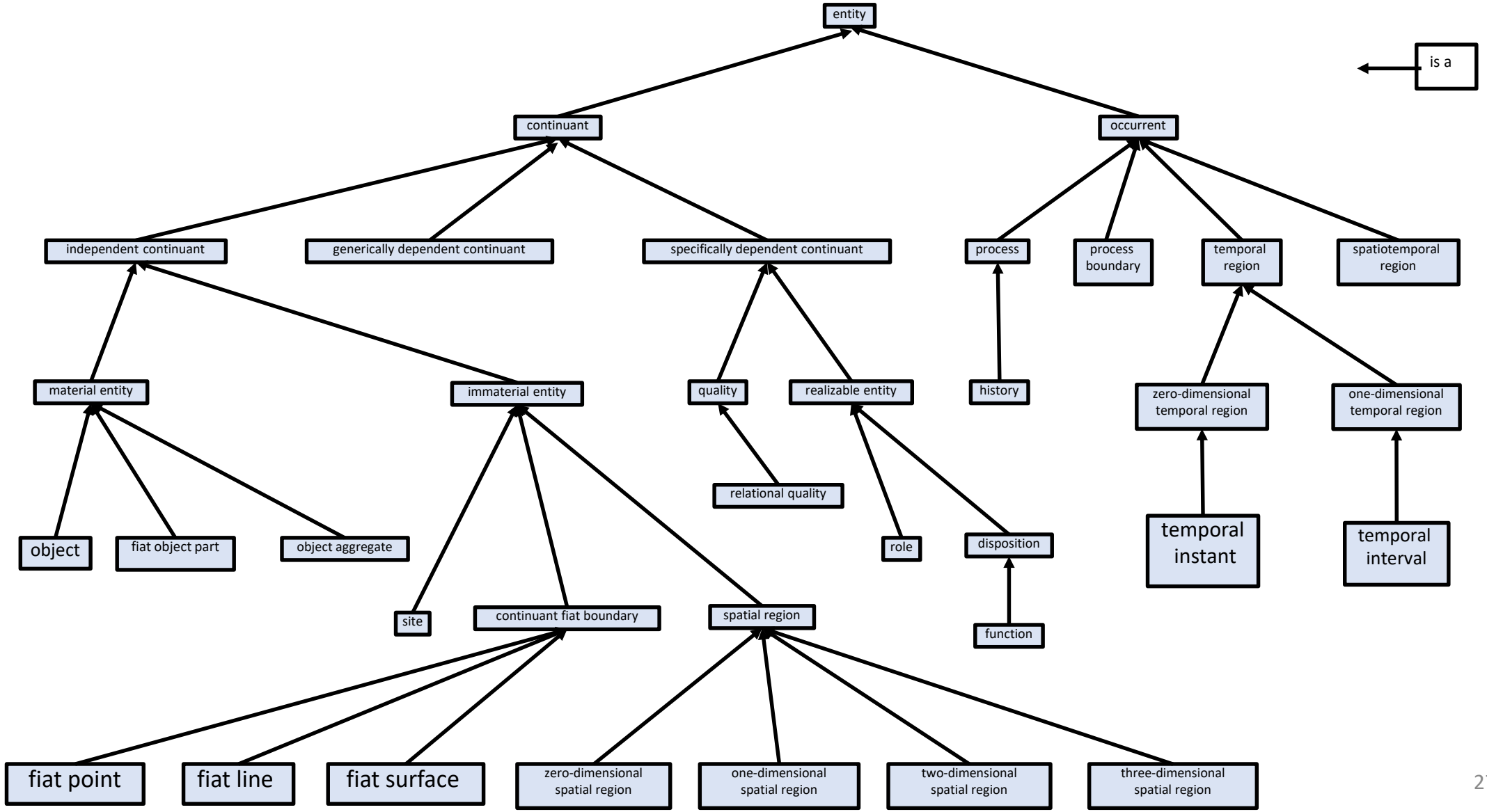
Information technology — Top-level ontologies (TLO) — Part 2: Basic Formal Ontology (BFO)

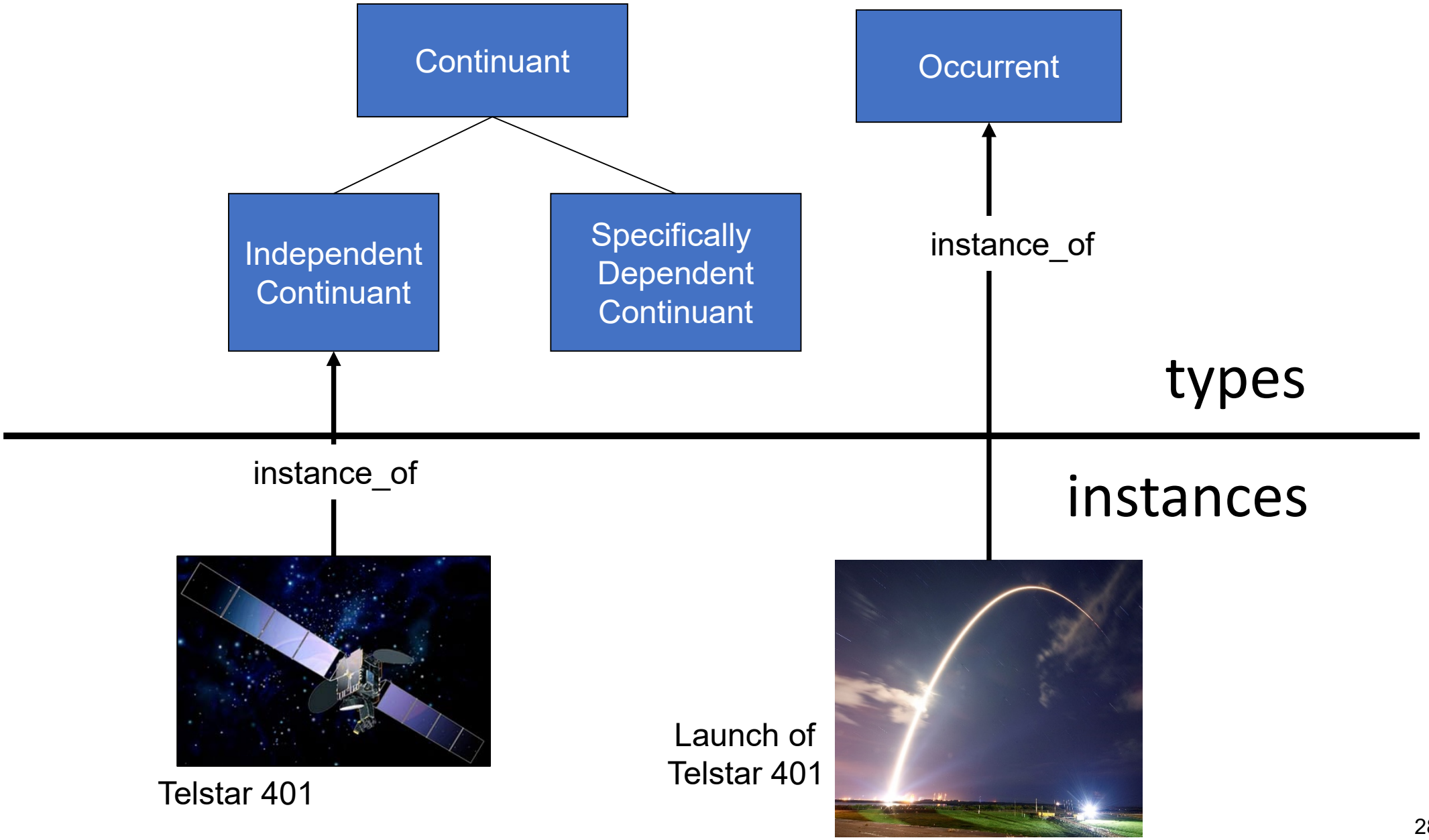
GENERAL INFORMATION

Status : ⊙ Under development **Publication date :** 2020-03

Edition : 1

BFO-2020 (ISO/IEC 21838-2)



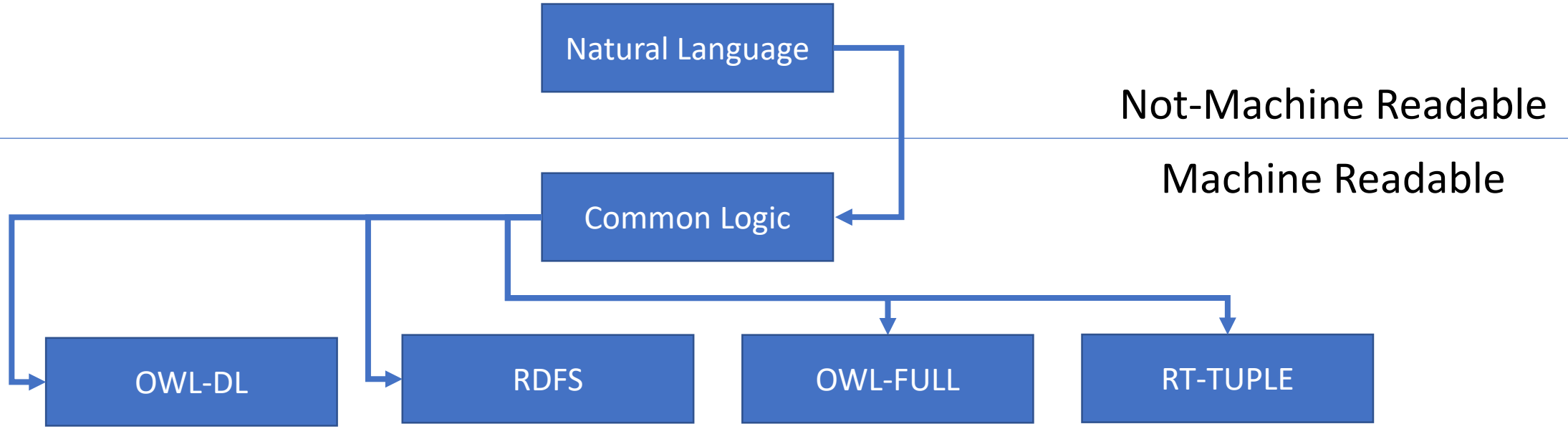

Continuant
Occurrent
Independent Continuant
Specifically Dependent Continuant
instance_of
types
instance_of
instances
Telstar 401
Launch of Telstar 401

Realist Curation

- Data and Meta-data should be tagged with realist curation ontologies meaning terms
  - ➢ make clear reference to scientific reality and
  - ➢ have easy to grasp definitions.
- Curation ontologies should fall under a top level ontology conformant with ISO-21838 Information Technology – Top-Level Ontologies (TLO) – Part 1: Requirements
- All definitions and database assertions should be expressible in Common Logic

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.



Realist Curation

- Terms that clearly refer to reality
- With Definitions that are easily graspable
- Structured in a way that allows for terms representing new knowledge and new domains to be added in modular fashion

Tracking Provenance of Curation

- Where did the data come from?
- When did it become available?
- Who (or what) procured the data?
- What format was it in previously?
- Are there restrictions on use?
- What errors have been corrected?

Answers encoded in format that is

- highly interoperable with other data under the same top-level ontology,
- without having to be under the same mid- or domain-level ontology.
- Relationship between items of data is explicit without e.g., 'join table' and
- Adding new types doesn't require a new data-base schema.

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.