





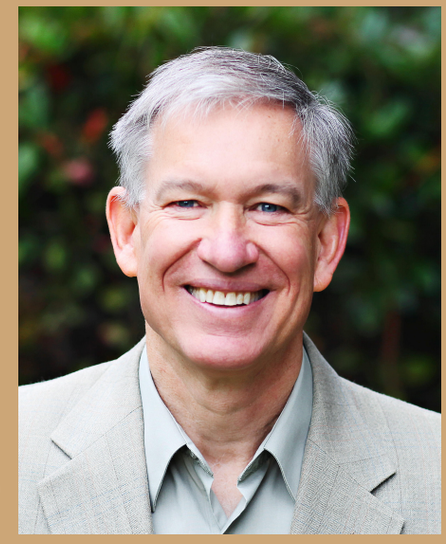
Protege, BioPortal, CEDAR: Building Complete Pipelines for Engineering Metadata

Introduction

John Graybeal
Technical Program Manager
CEDAR & BioPortal

Stanford Center for
Biomedical Informatics Research (BMIR)

- Science computing support years
  - 19: astronomy
  - 12: marine
  - 5: biomedicine
- Discipline years
  - 12: operations
  - 10: cyberinfrastructure
  - **13: metadata**
  - **15: semantics**
  - lots: project lead
- Project Lead, Marine Metadata Interoperability Project and MMI Ontology Registry and Repository



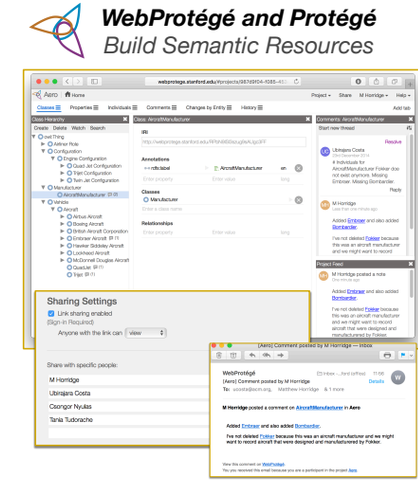
In this talk

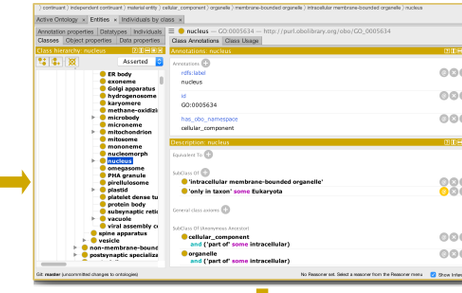
1. Make the case for the importance of standardized semantics in metadata.
2. Describe how to build and make available standardized semantics (e.g., vocabularies and concepts).
3. Show you how to easily build most metadata models *but with standardized semantics*

Stanford Center for Biomedical Informatics Research
Open Source Solutions for Semantic Knowledge Management

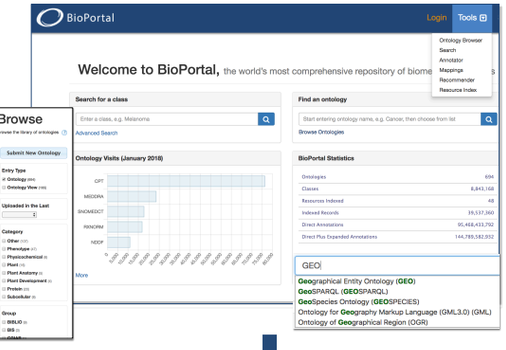
John Graybeal, Marcos Martinez-Romero, Matthew Horridge, Mark Musen
Stanford University

WebProtégé and Protégé
Build Semantic Resources

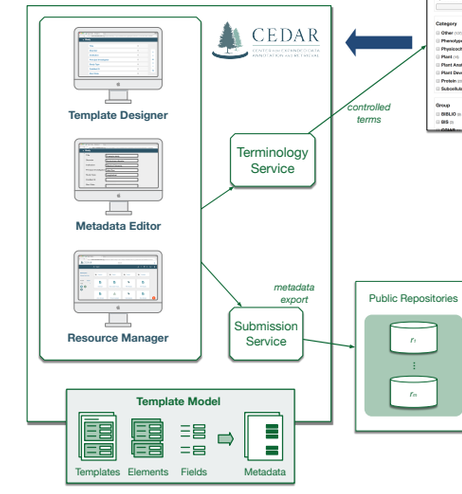


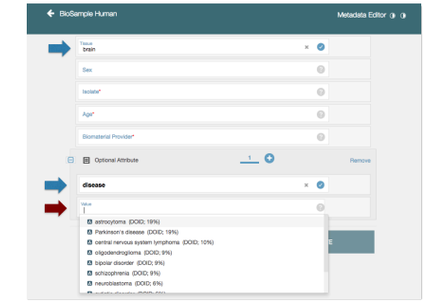


BioPortal
Store and Access Semantic Resources



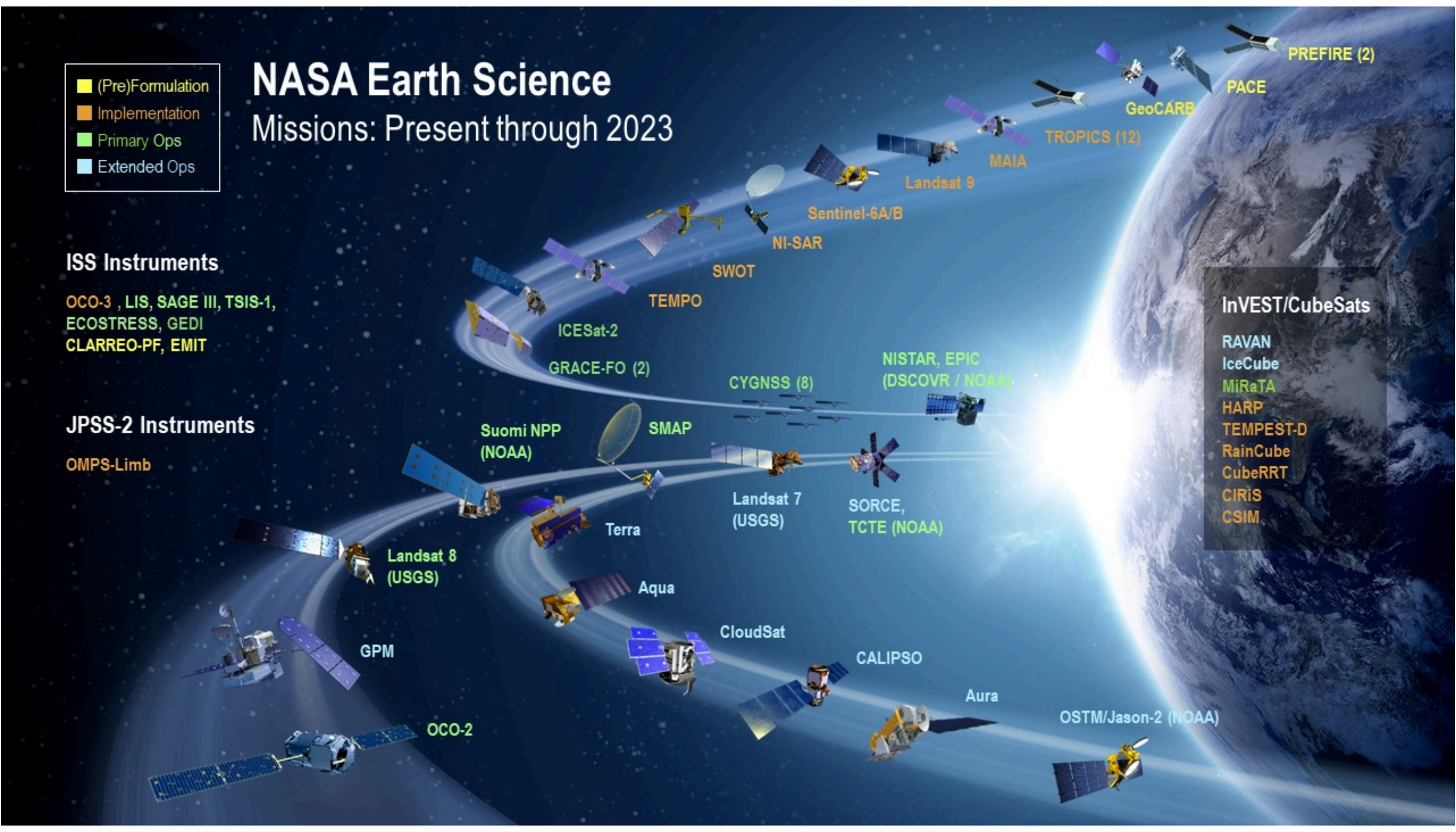
CEDAR
Manage Semantic Metadata

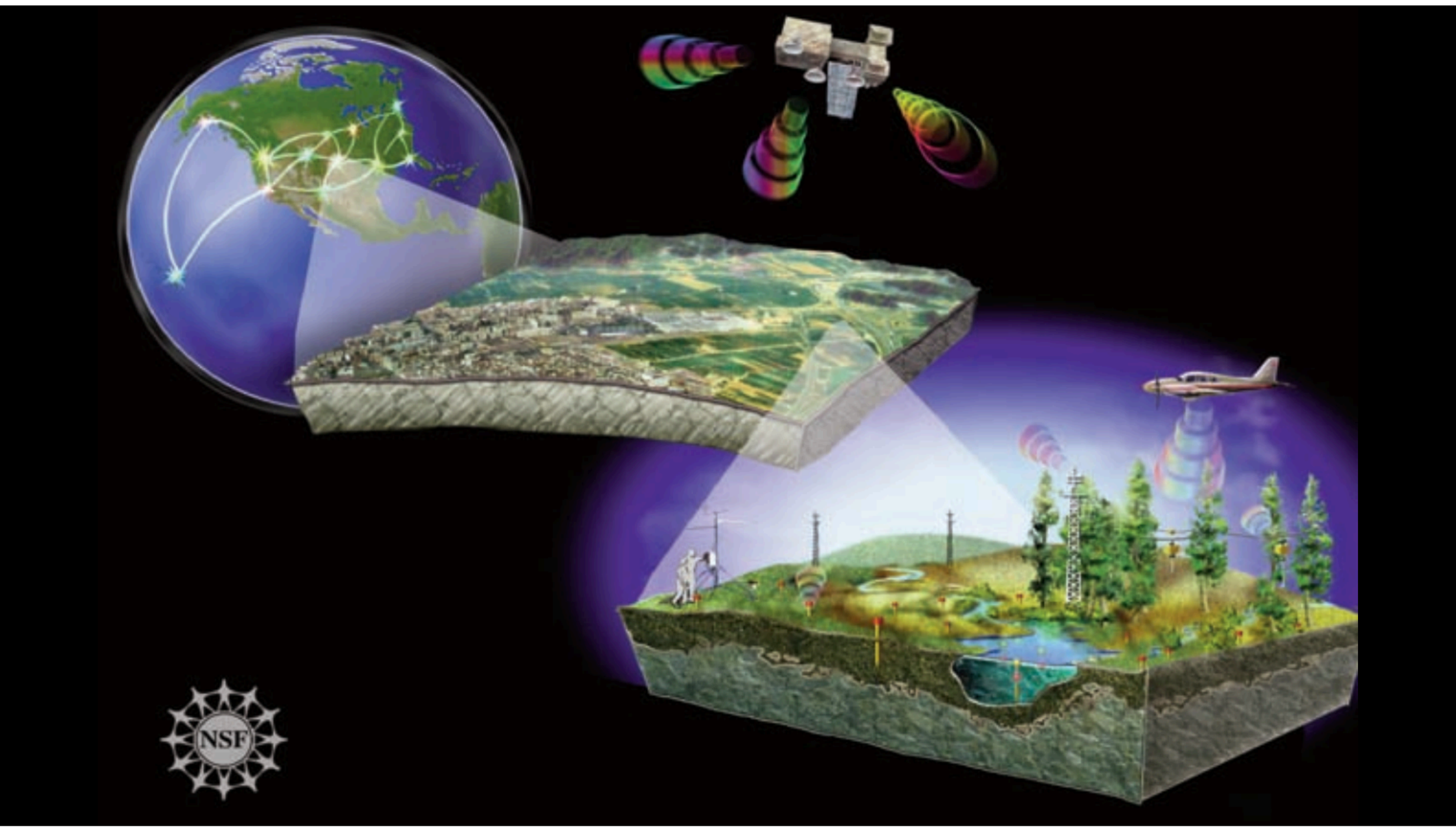




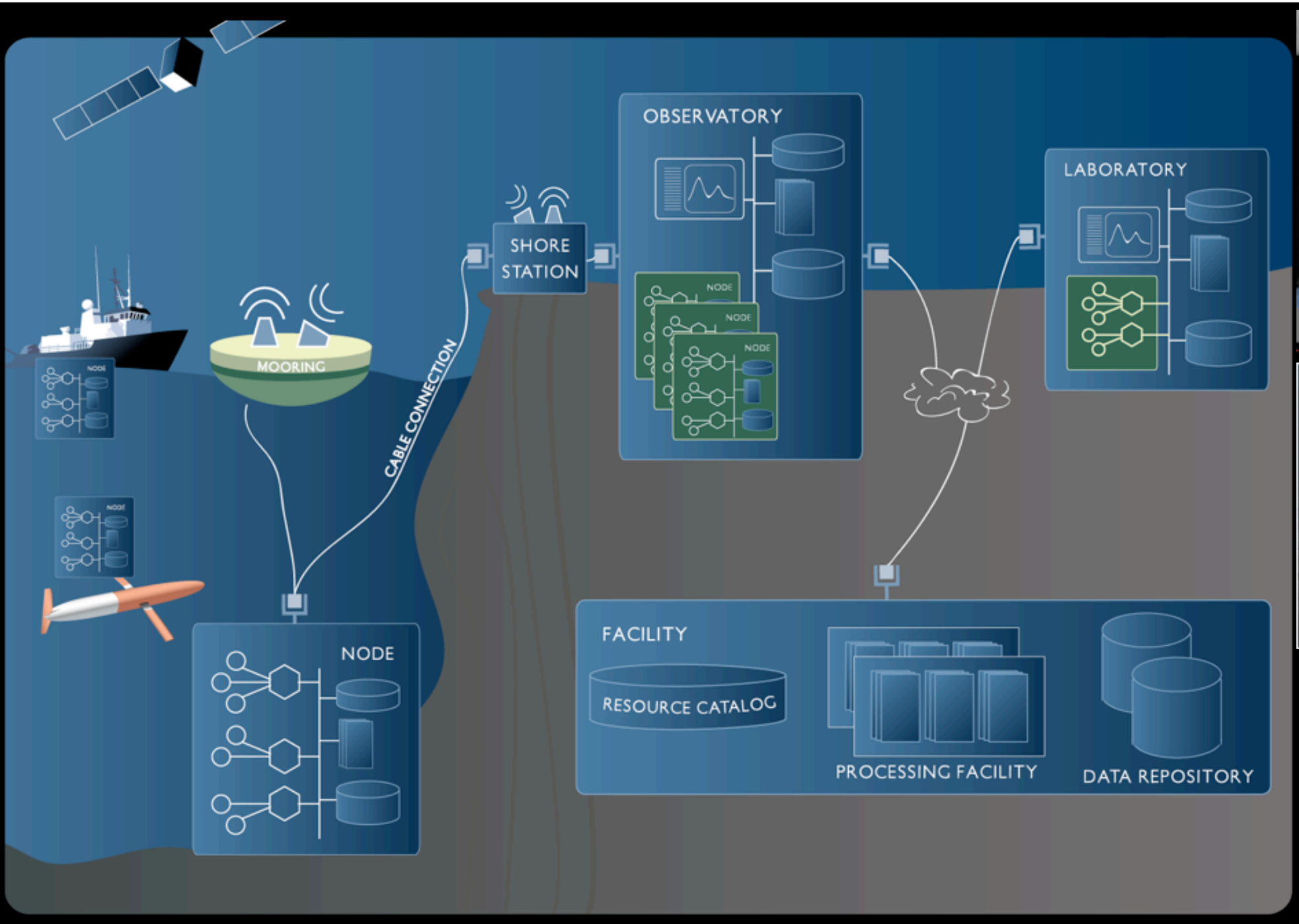
Part 1:
Why are standardized semantics important?

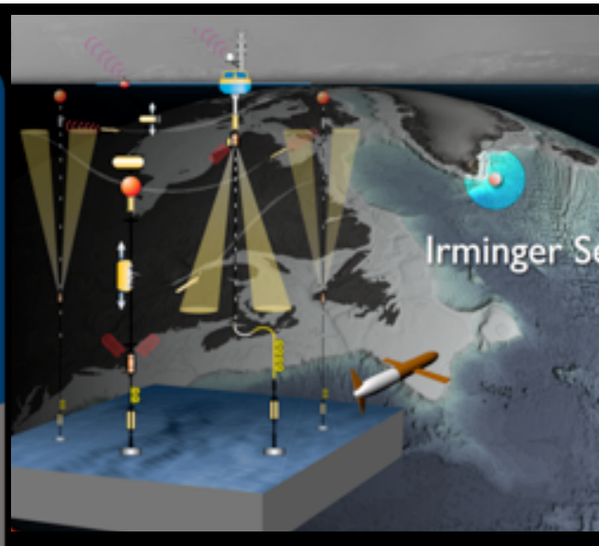
Why?


(Pre)Formulation
Implementation
Primary Ops
Extended Ops
NASA Earth Science
Missions: Present through 2023
ISS Instruments
OCO-3 , LIS, SAGE III, TSIS-1,
ECOSTRESS, GEDI
CLARREO-PF, EMIT
JPSS-2 Instruments
OMPS-Limb
PREFIRE (2)
PACE
GeoCARB
TROPICS (12)
MAIA
Landsat 9
Sentinel-6A/B
NI-SAR
SWOT
TEMPO
ICESat-2
GRACE-FO (2)
CYGNSS (8)
NISTAR, EPIC
(DSCOVR / NOAA)
SMAP
Suomi NPP
(NOAA)
Landsat 7
(USGS)
SORCE,
TCTE (NOAA)
Terra
Landsat 8
(USGS)
Aqua
CloudSat
GPM
CALIPSO
Aura
OCO-2
OSTM/Jason-2 (NOAA)
InVEST/CubeSats
RAVAN
IceCube
MiRaTA
HARP
TEMPEST-D
RainCube
CubeRRT
CIRiS
CSIM

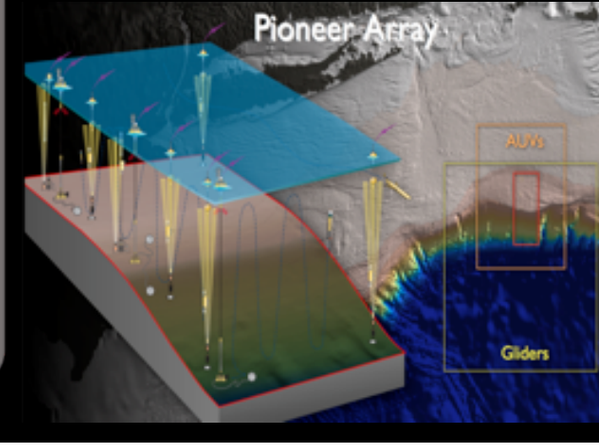

NSF



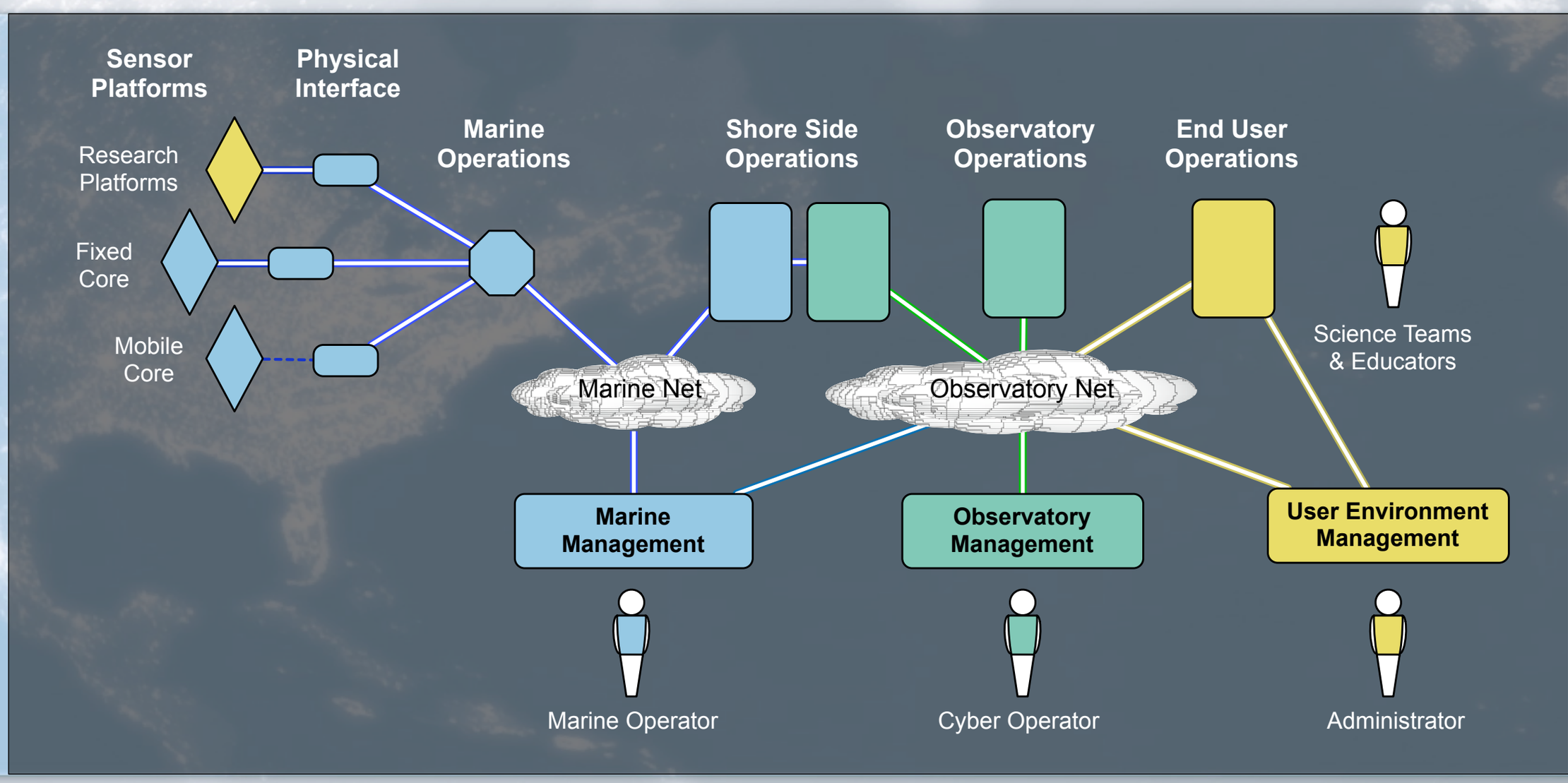

OBSERVATORY
NODE
NODE
NODE
SHORE STATION
LABORATORY
MOORING
CABLE CONNECTION
NODE
NODE
FACILITY
RESOURCE CATALOG
PROCESSING FACILITY
DATA REPOSITORY


Irminger Se

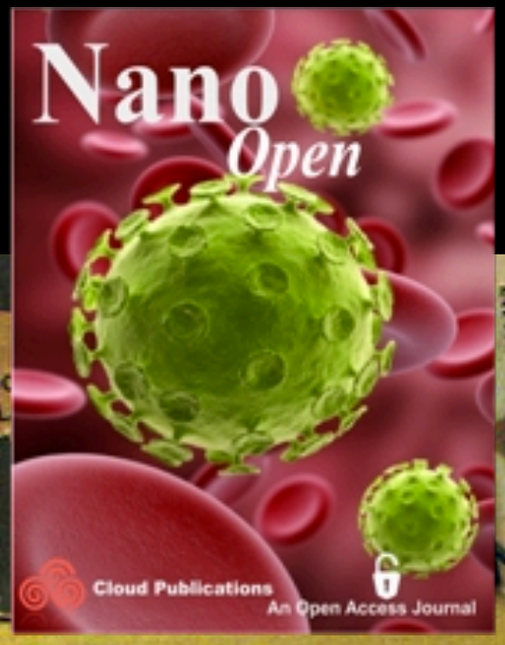


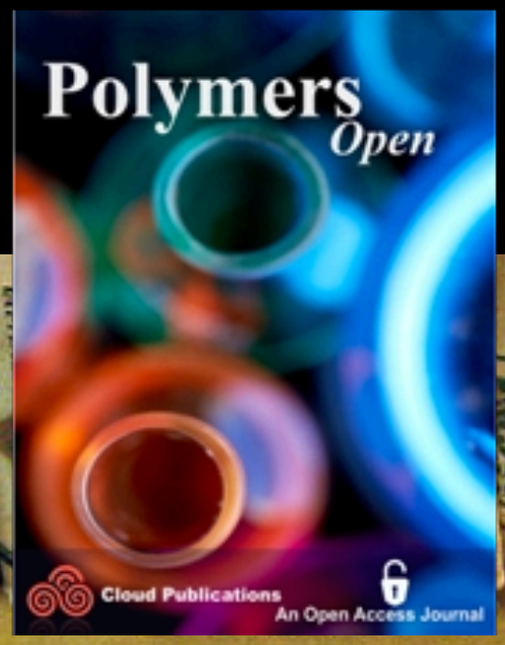

Pioneer Array
AUVs
Gliders

Observatory Topology









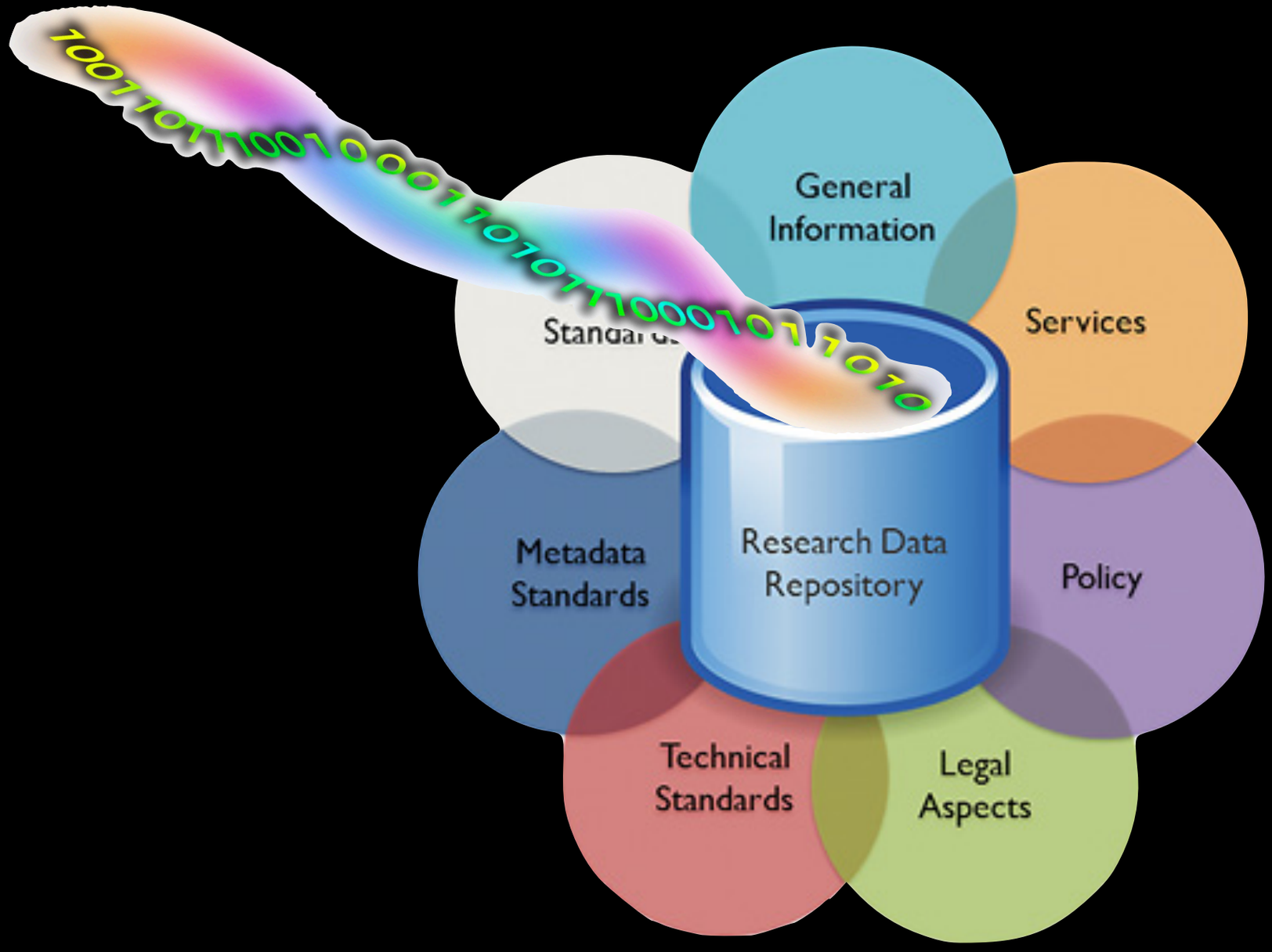


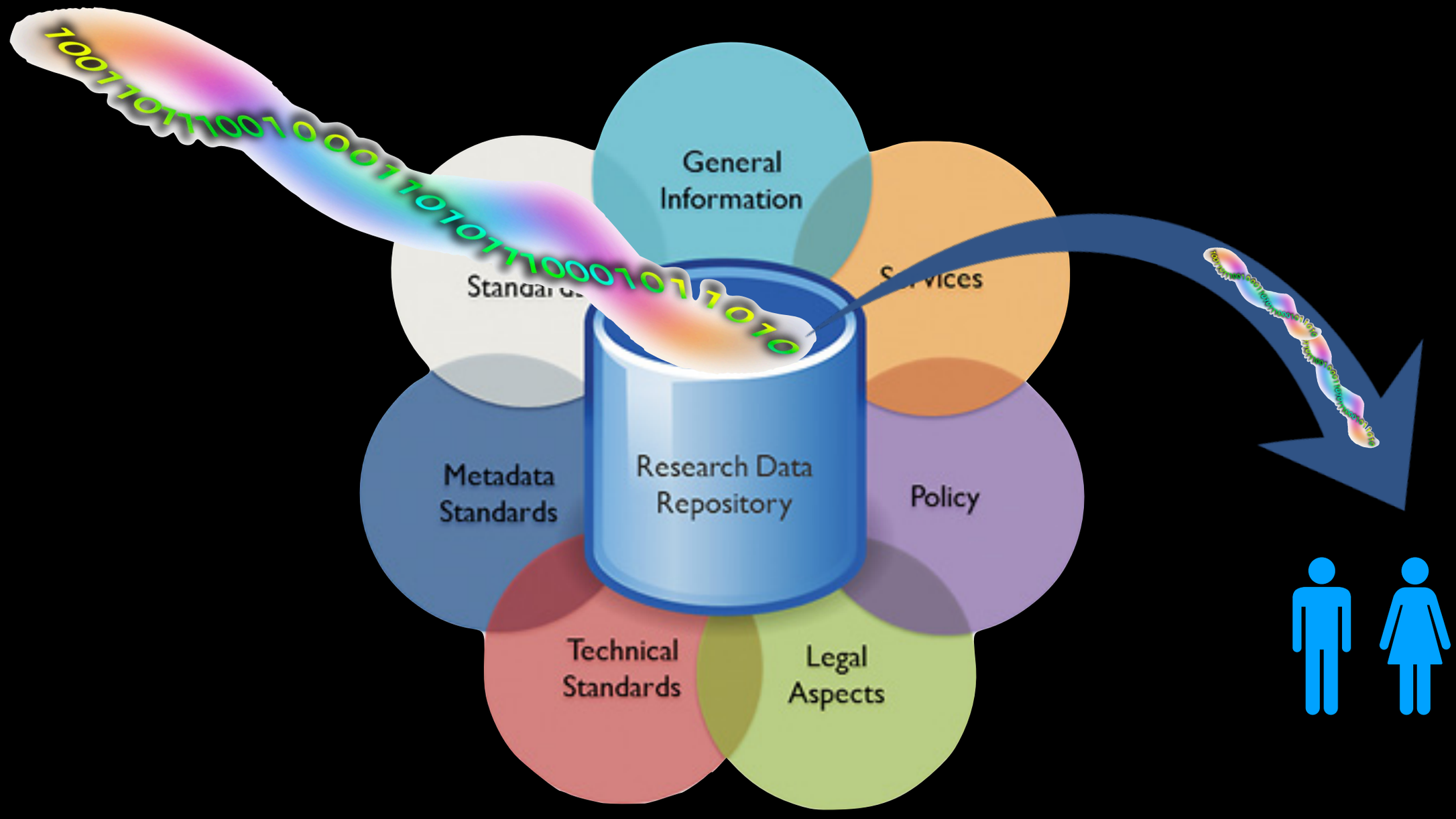


PUBLICATIONS

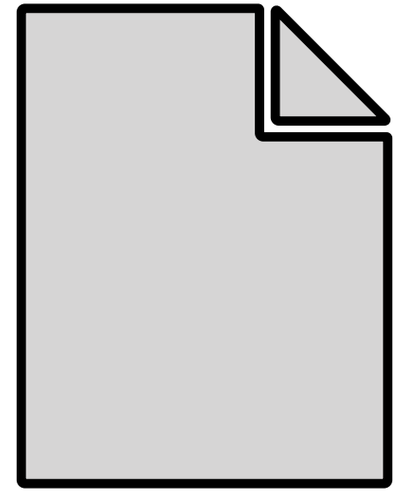


from
Research Data Repositories: The What, When, Why, and How
by Ray Uzwyshyn
https://www.infotoday.com/cilmag/apr16/Uzwyshyn--Research-Data-Repositories.shtml

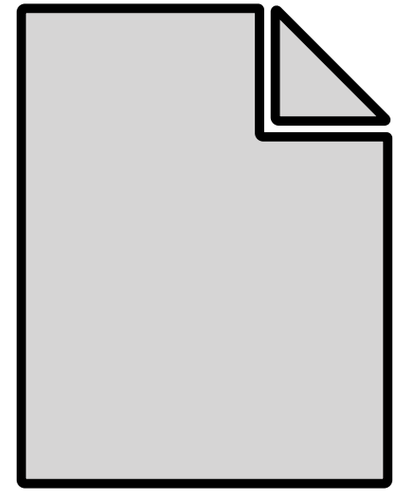

General Information
Services
Metadata Standards
Research Data Repository
Policy
Technical Standards
Legal Aspects


General Information
Metadata Standards
Research Data Repository
Policy
Technical Standards
Legal Aspects

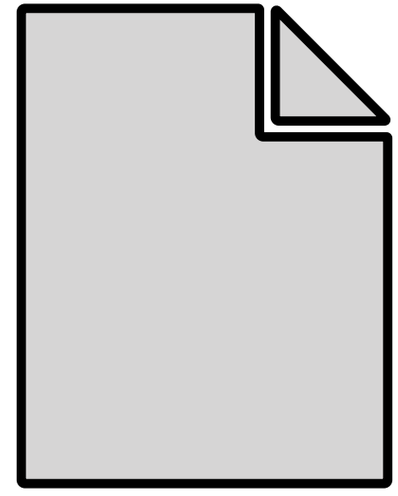
What do we want that data to look like?



Attachment



untitled



D3222944.pdf

We want research data to be self-describing.

“We need to develop one universal [metadata] standard to cover everyone’s use cases.”

–XKCD #927

HOW STANDARDS PROLIFERATE:

(SEE: A/C CHARGERS, CHARACTER ENCODINGS, INSTANT MESSAGING, ETC)

SITUATION:
THERE ARE
14 COMPETING
STANDARDS.

14?! RIDICULOUS!
WE NEED TO DEVELOP
ONE UNIVERSAL STANDARD
THAT COVERS EVERYONE'S
USE CASES.



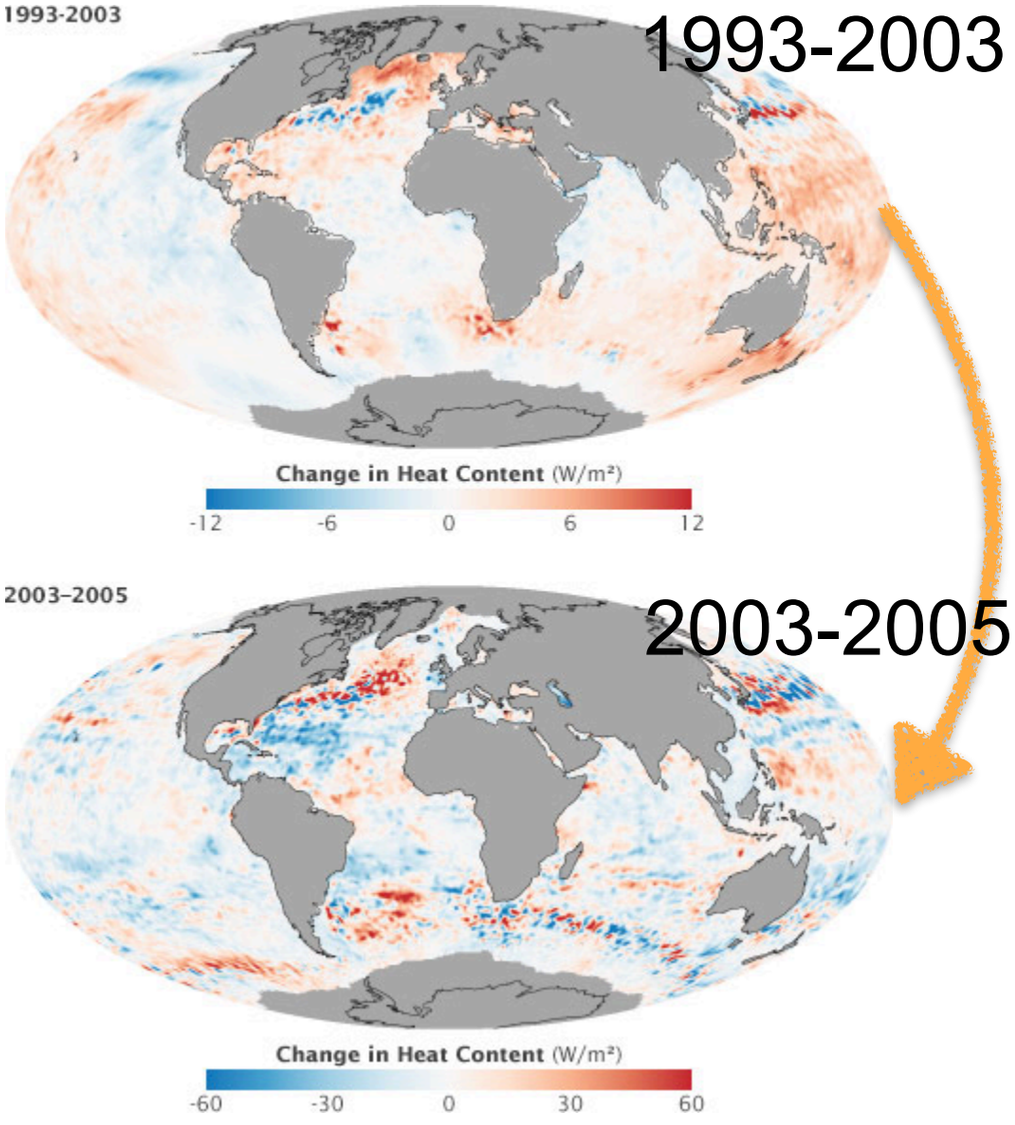
SOON:

SITUATION:
THERE ARE
15 COMPETING
STANDARDS.

What is ‘metadata’ again?

- It’s not just ‘data about data’—e.g., “here are my definitions of what my values mean” (though that helps)
- It’s about origins / provenance / context of the data
- This can include things like…
  - a precise description of how to parse data streams
  - the reason we collected the data streams (and who paid)
  - basically, any context related to the metadata

A story about ocean temperatures…

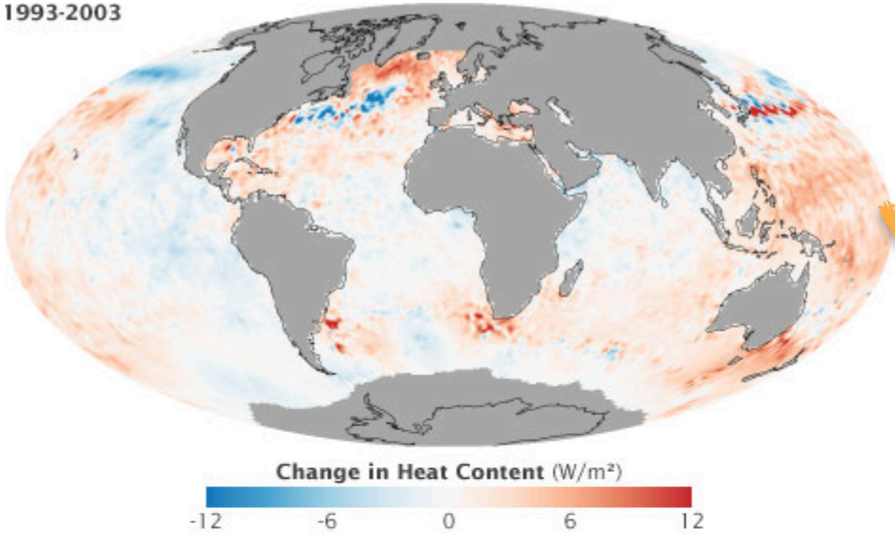


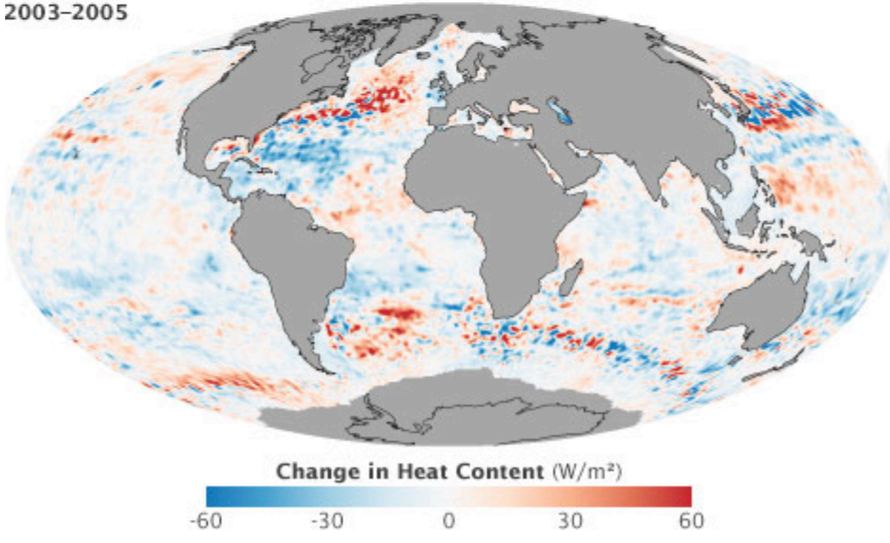
1993-2003

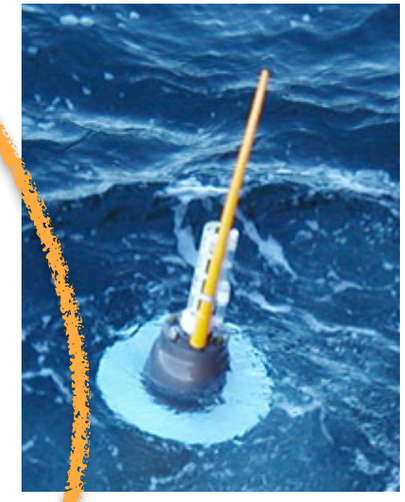
2003-2005

from Correcting Ocean Cooling,
a NASA Earth Observatory blog post
(https://earthobservatory.nasa.gov/features/OceanCooling)

A story about ocean temperatures…





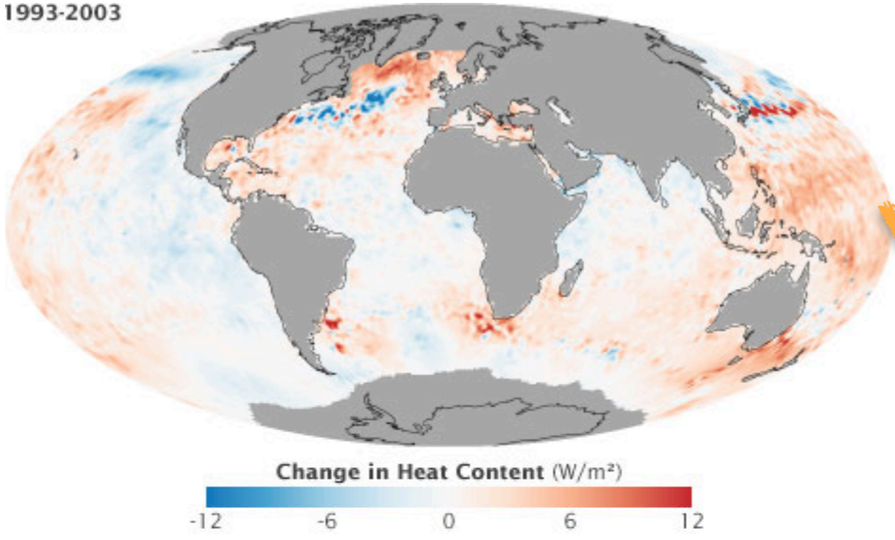


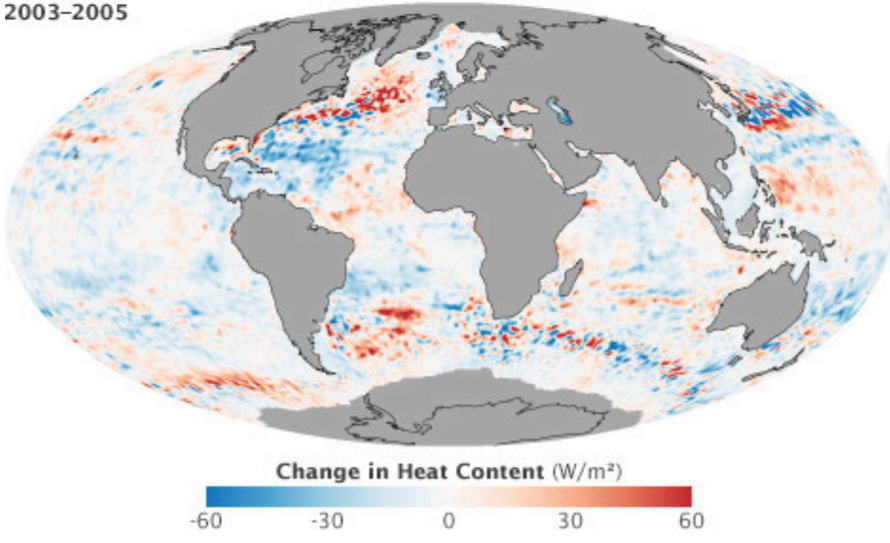
Argo Float

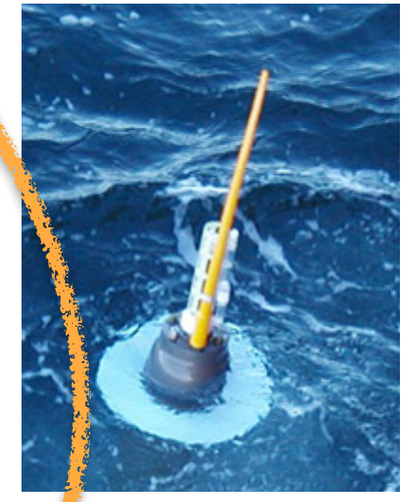


XBT Instrument

A story about ocean temperatures…







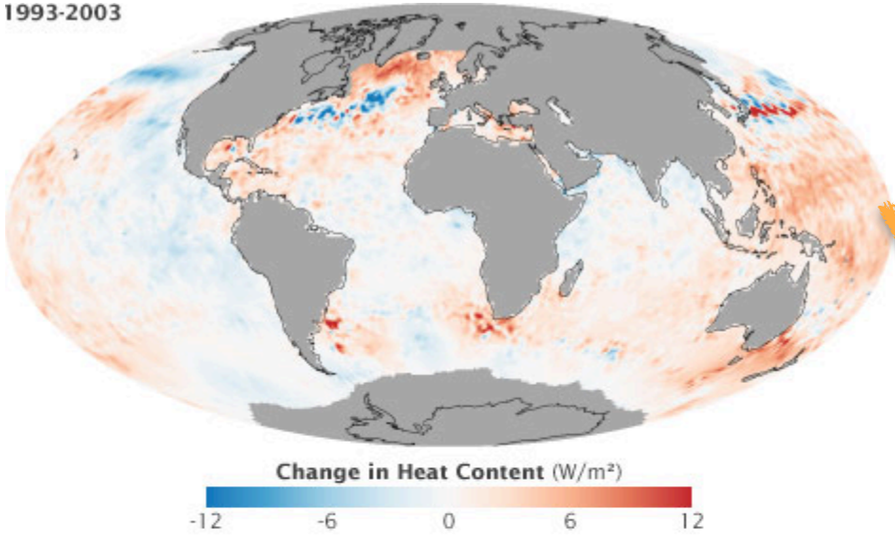
Argo Float

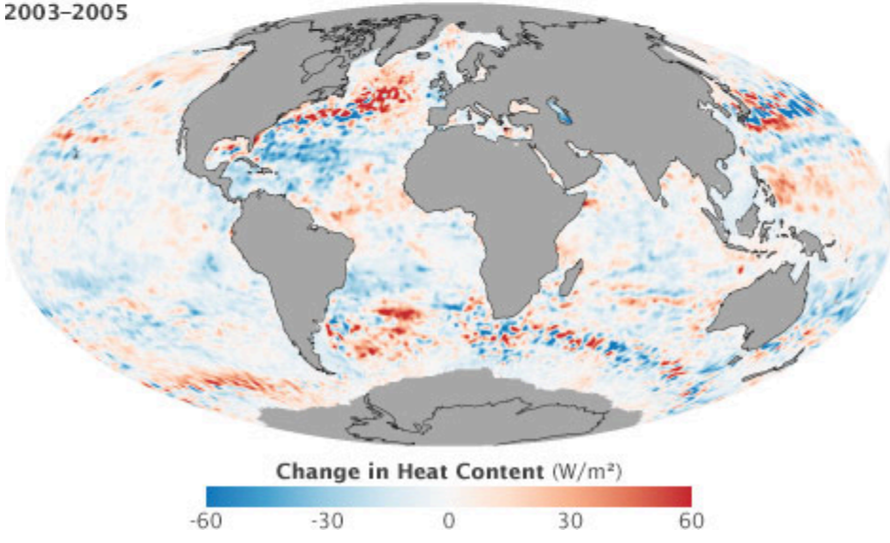


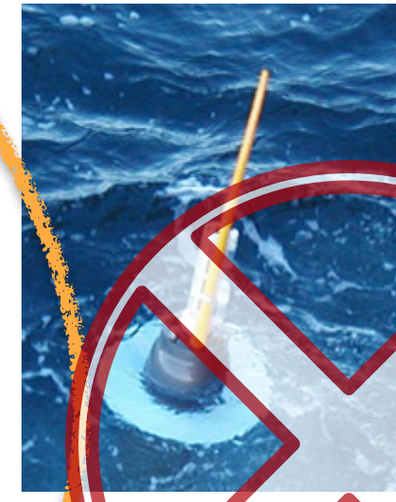
XBT Instrument

Biased Measurements

A story about ocean temperatures…





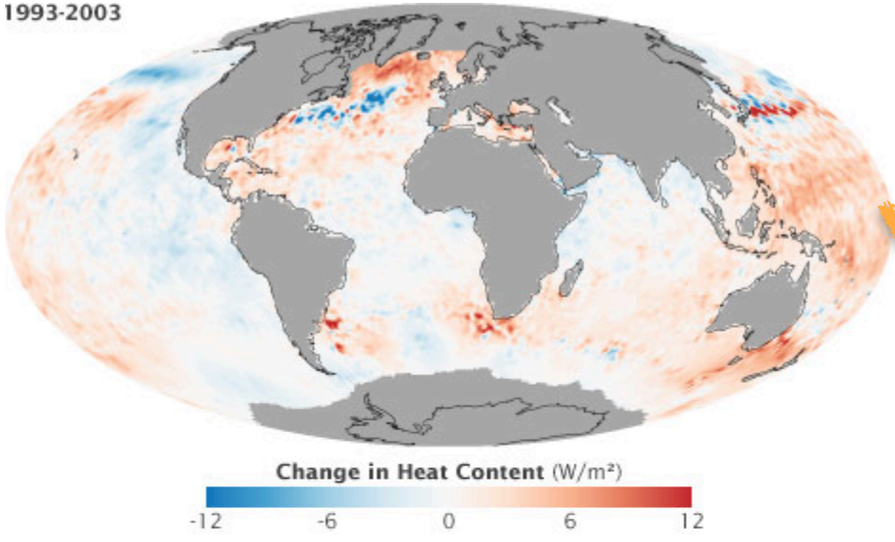


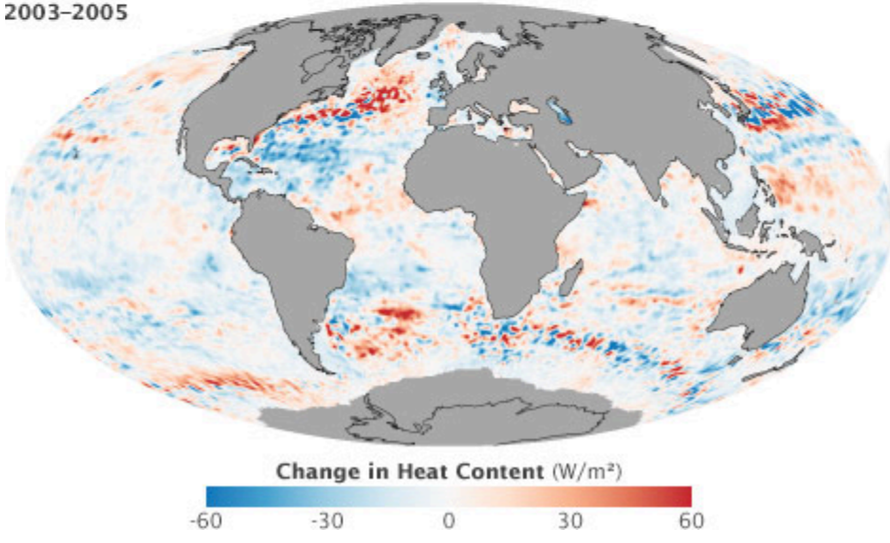


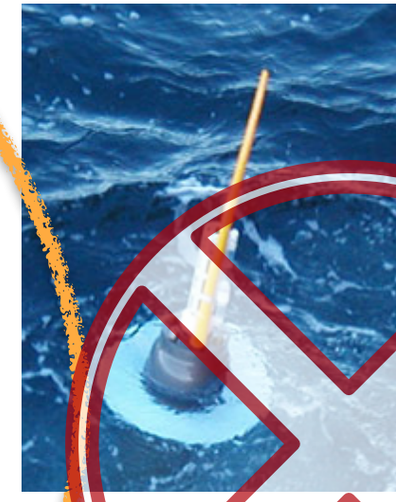
Argo Float XBT Instrument

Biased Measurements

A story about ocean temperatures…



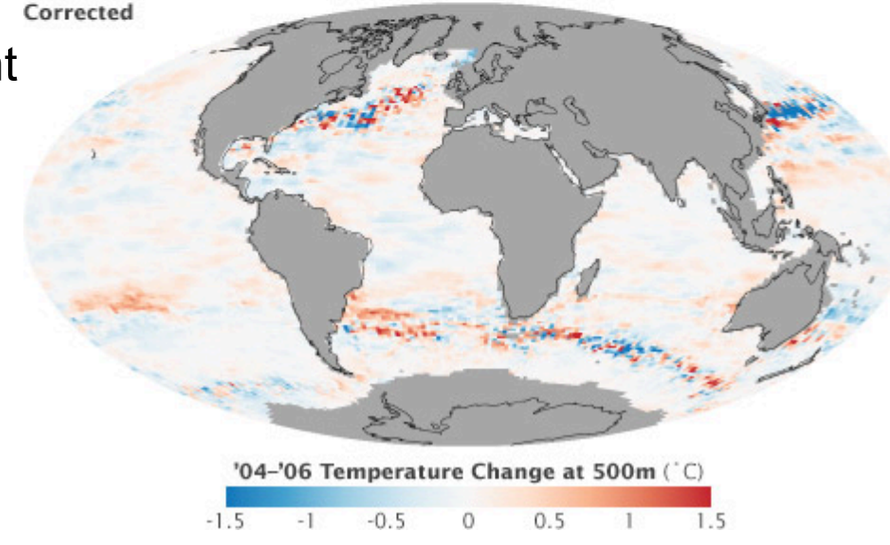




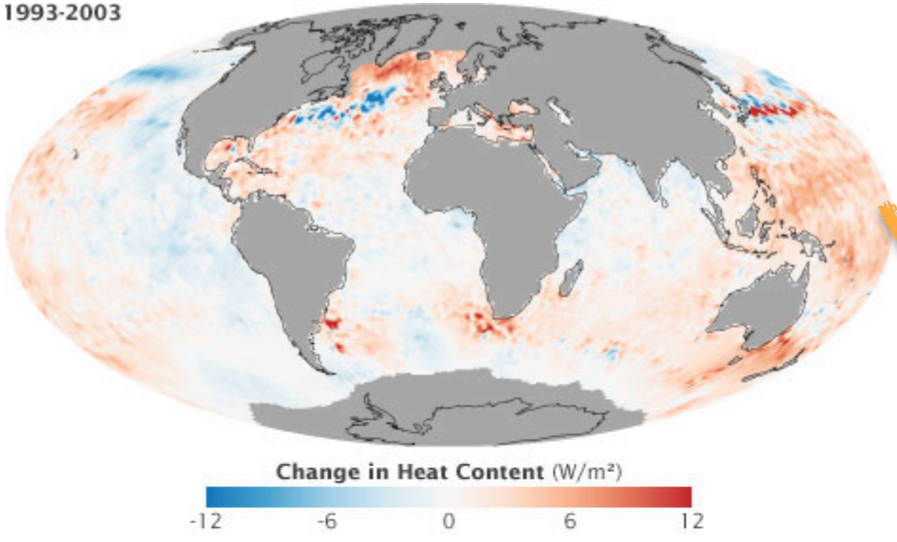


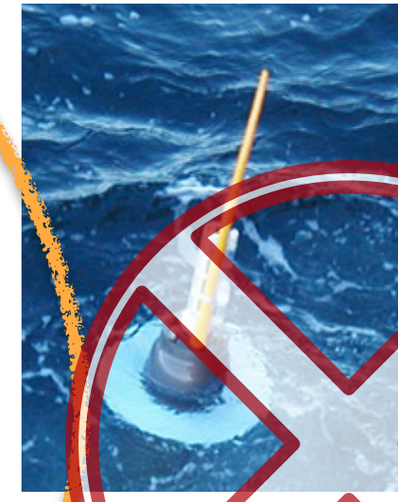
Argo Float XBT Instrument

Biased Measurements



A story about ocean temperatures…

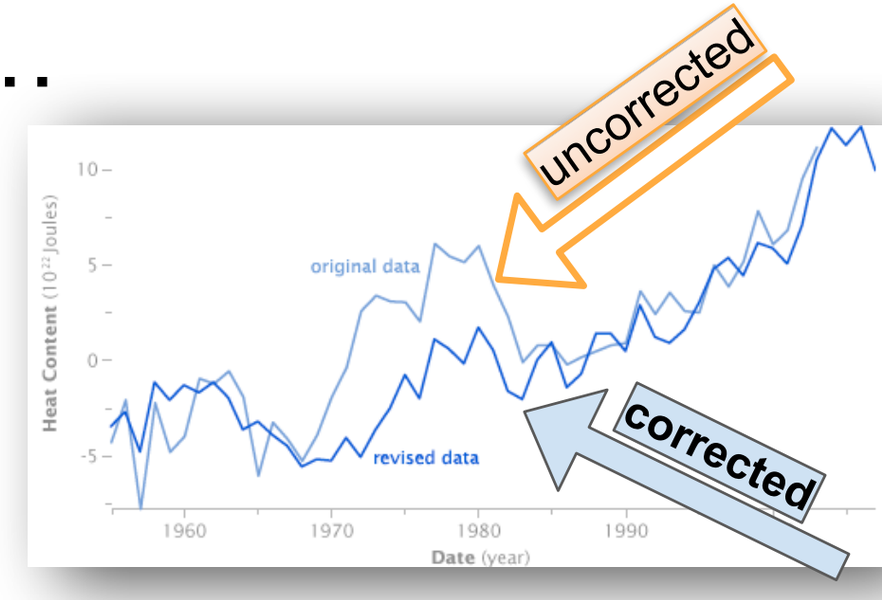


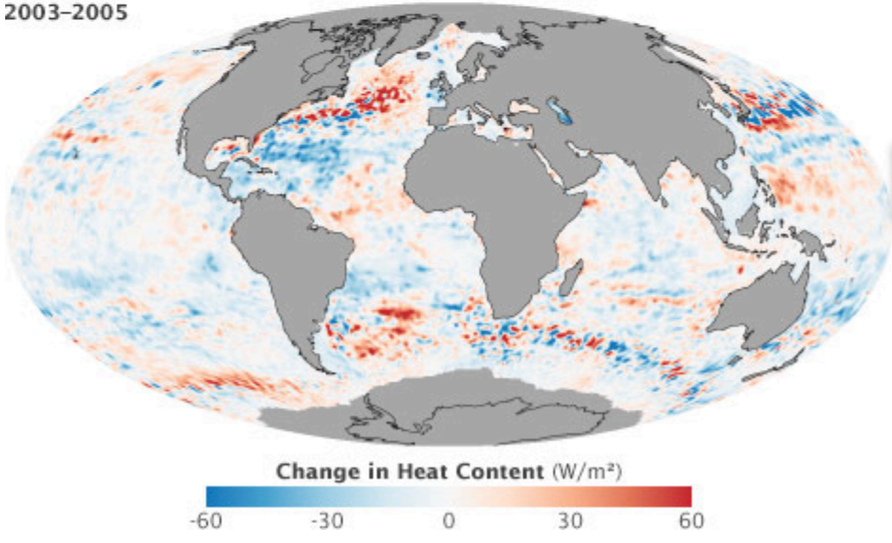


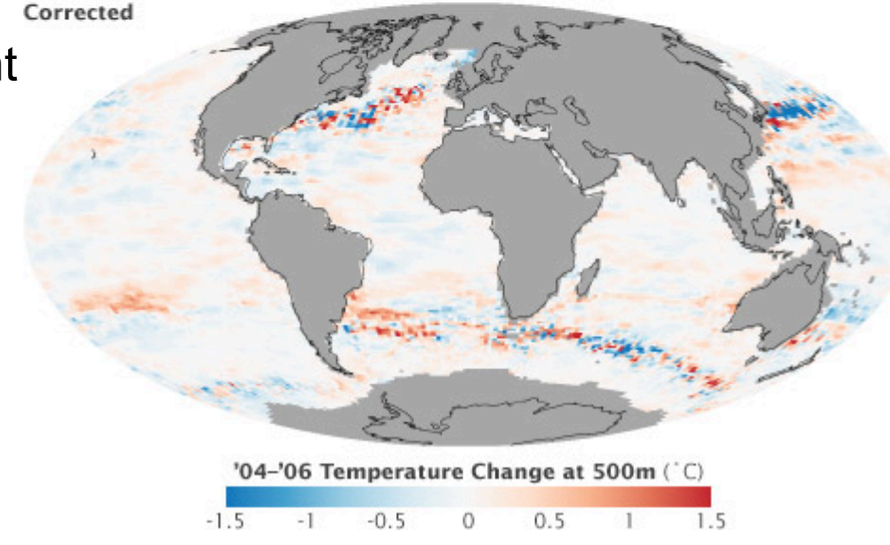


Argo Float XBT Instrument

Biased Measurements







A story about ocean temperatures…

The method gives detailed information on biases for specific vintages of XBTs, but **this information cannot be used with confidence to correct the majority of XBT profiles archived that are lacking critically important metadata**

A review of global ocean temperature observations: Implications for ocean heat content estimates and climate change.
J. P. Abraham et al. August 2013.
doi://10.1002/rog.20022

depth-dependent corrections were introduced for each XBT temperature observation (the corrections were obtained by comparing XBT and CTD/Nansen bottle anomalies for overlapping 2° × 4° boxes). **Though a cruise-wise correction of the XBT is preferable, it remains beyond the scope of this study because of the absence of collocated high quality data and of the respective metadata (manufacturer, probe and acquisition system type).**

How much is the ocean really warming?
Viktor Gouretski et al
January 2007.
doi://10.1029/2006GL027834

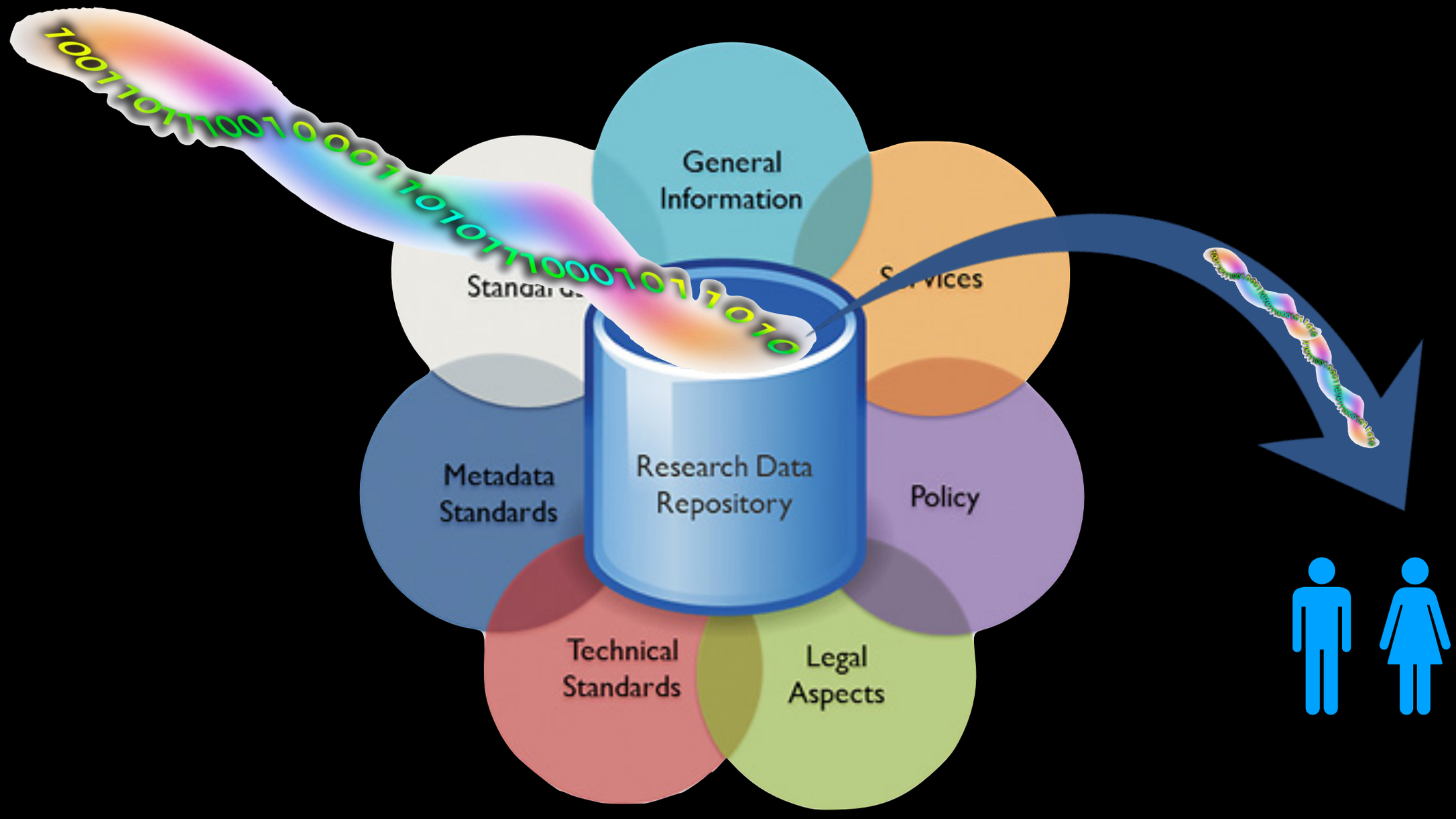
How much metadata do we have to keep?

Do we forever follow the “Provenance Trail” to keep any data-related context?



Metadata (redefined):

“Information useful to work with data”


General Information
Metadata Standards
Research Data Repository
Policy
Technical Standards
Legal Aspects

How do we decide what metadata to provide?

- Given our metadata definition: “Information useful to work with data”
  - The question becomes: useful to **whom**?
- To compensate for human beings’ lack of imagination, let’s ask *What if…*
  - *Someone else has to understand this data?*
  - *And they aren’t working in this domain?*
  - *And don’t speak our language fluently (or at all)?*
  - *And it’s 10 years in the future?*
  - *And the meaning of some of these terms have changed?*

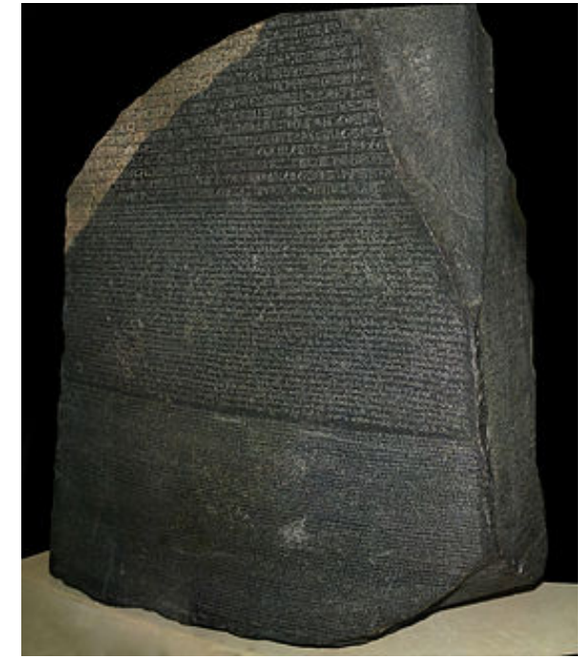
The choice is yours, because the end goal is yours

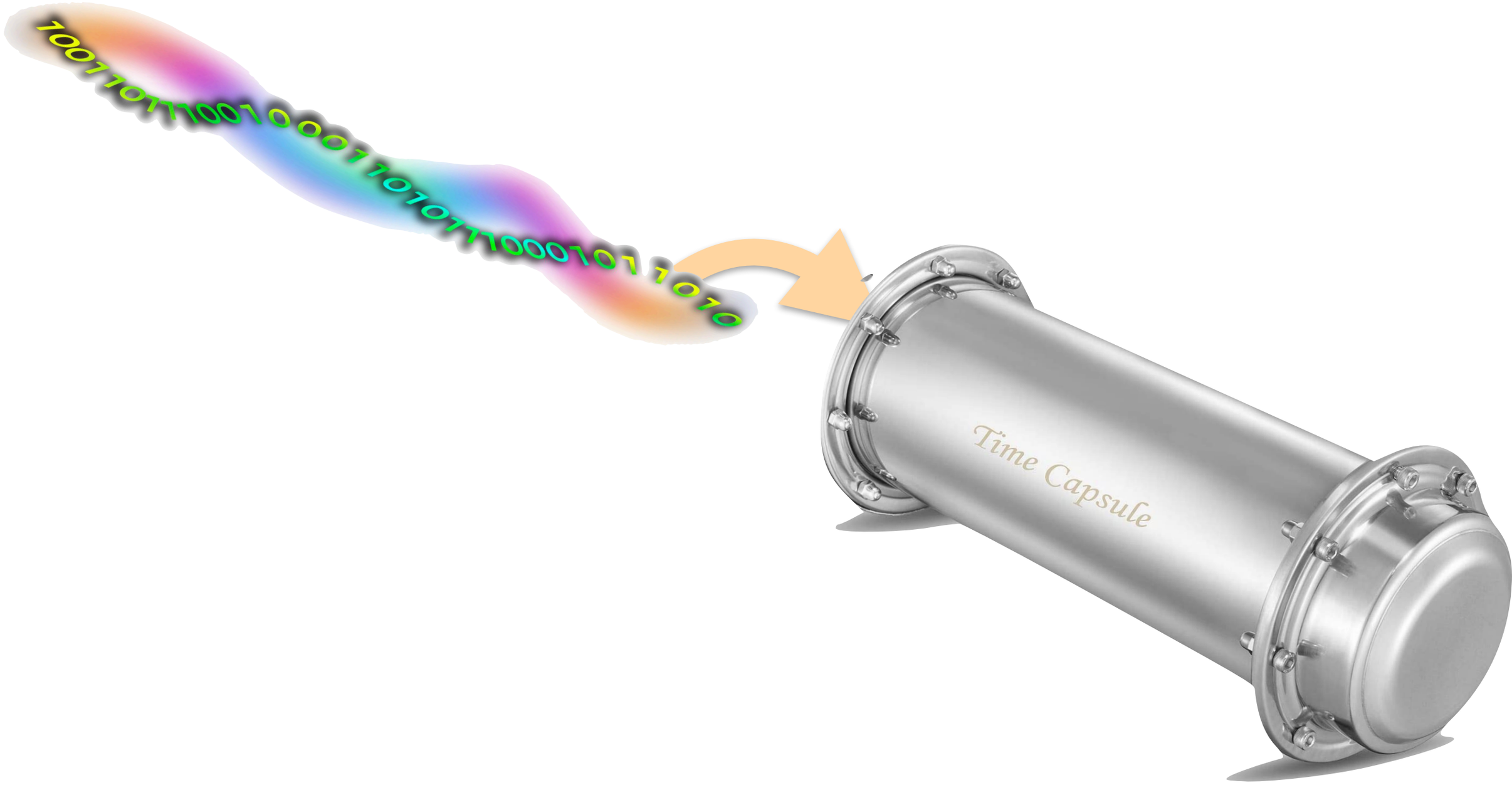
- The premise of FAIR data: We want this data to be findable, understandable, interoperable, and reusable across all factors that may interfere
  - Factors we want to consider include individuals, languages, domains, time frames, evolving scientific understanding…
  - Well, as many as is reasonable, anyway
- Use engineering (and financial) judgment to decide how far to go to satisfy the above

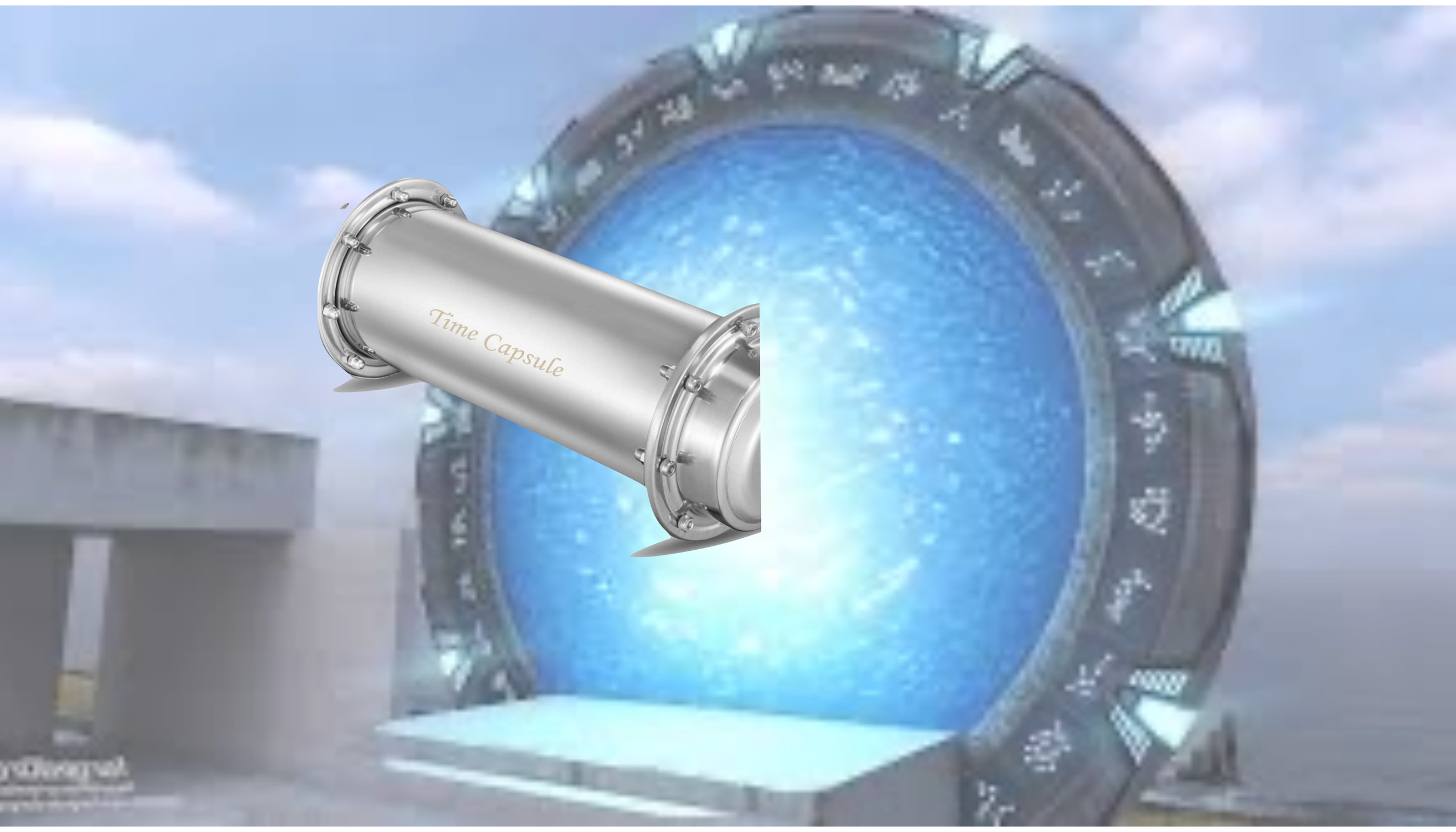
And why are semantics important?

- *The premise of FAIR data: We want this data to be **findable**, **understandable**, **interoperable**, and **reusable** across all factors that may interfere*
- *Factors include **individuals**, **languages**, **domains**, **time frames**, **evolving scientific understanding**…*

- For these goals, and these factors that interfere with them, **meaning** is critical.
- So you want an unambiguous ‘Rosetta Stone’ that helps interpret the data.




1001101110010 0011101011000101 1010
Time Capsule


Time Capsule

Part 2:
How do we create (and make available) standardized semantics?

Semantic Introduction (Good news!)

In semantics, there IS one standard to rule them all!



Primer: Resource Description Framework (RDF)

An Astonishingly Simple Guide

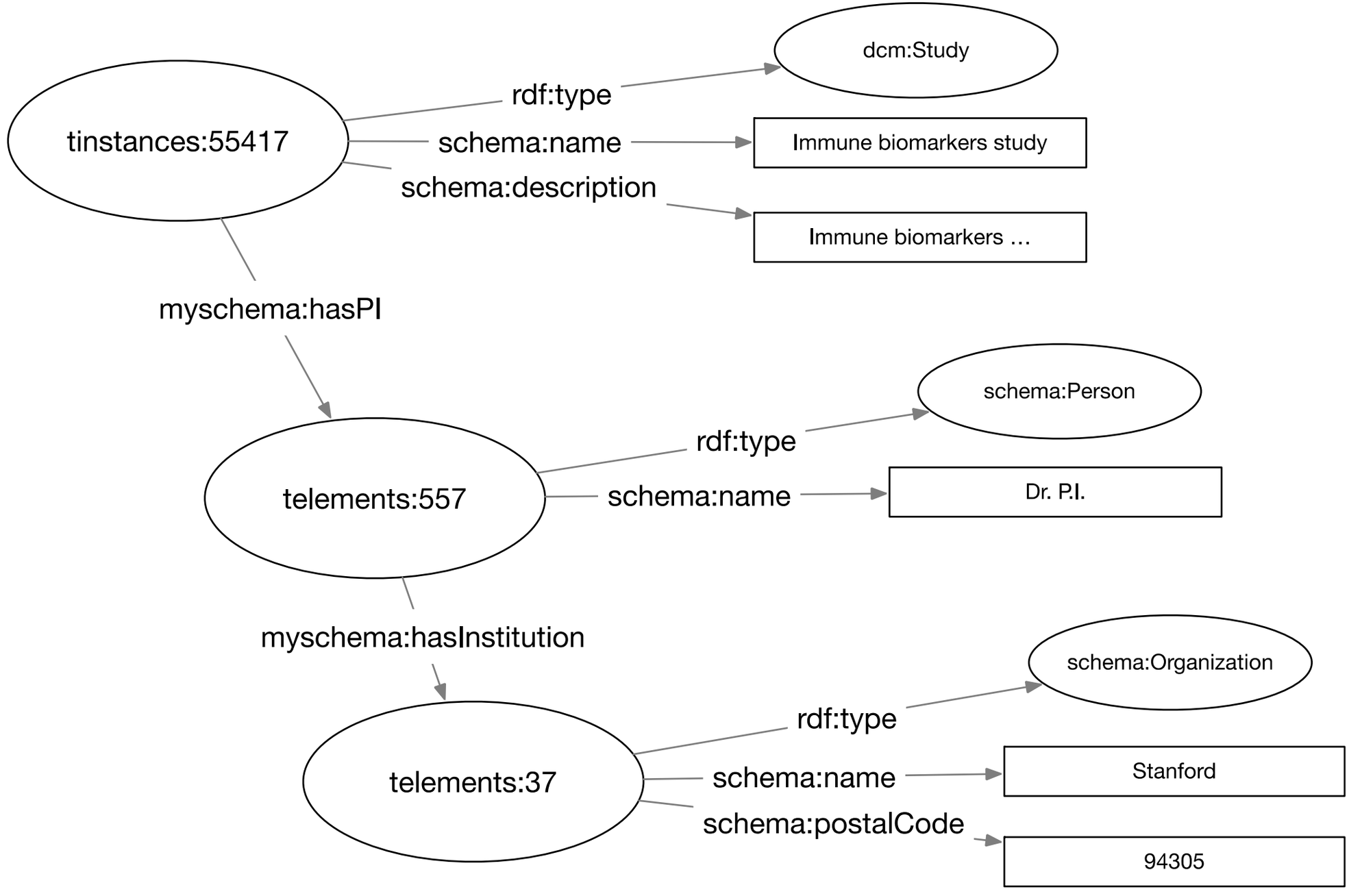
RDF: Making Statements with Triples

Subject	Predicate	Object
(a class)	(aka **Property**)	data or class
identifier (IRI)	*identifier (IRI)*	*“string” or identifier (IRI)*
https://ex.org/dataset	https://ex.org/createdOn	"20190226T145903Z"
ex:dataset	ex:createdOn	"20190226T145903Z"



The power of RDF comes from (a) declaring logical attributes of properties, like transitivity; and (b) creating shared rules about certain classes and properties. RDFS, SKOS, and OWL add more of these shared classes and properties (and describe rigorously how they work).

Metadata as RDF Graph for a study



Here's how simple RDF is (* = nice on specific topic)

(I was putting together this list and my browser died…)

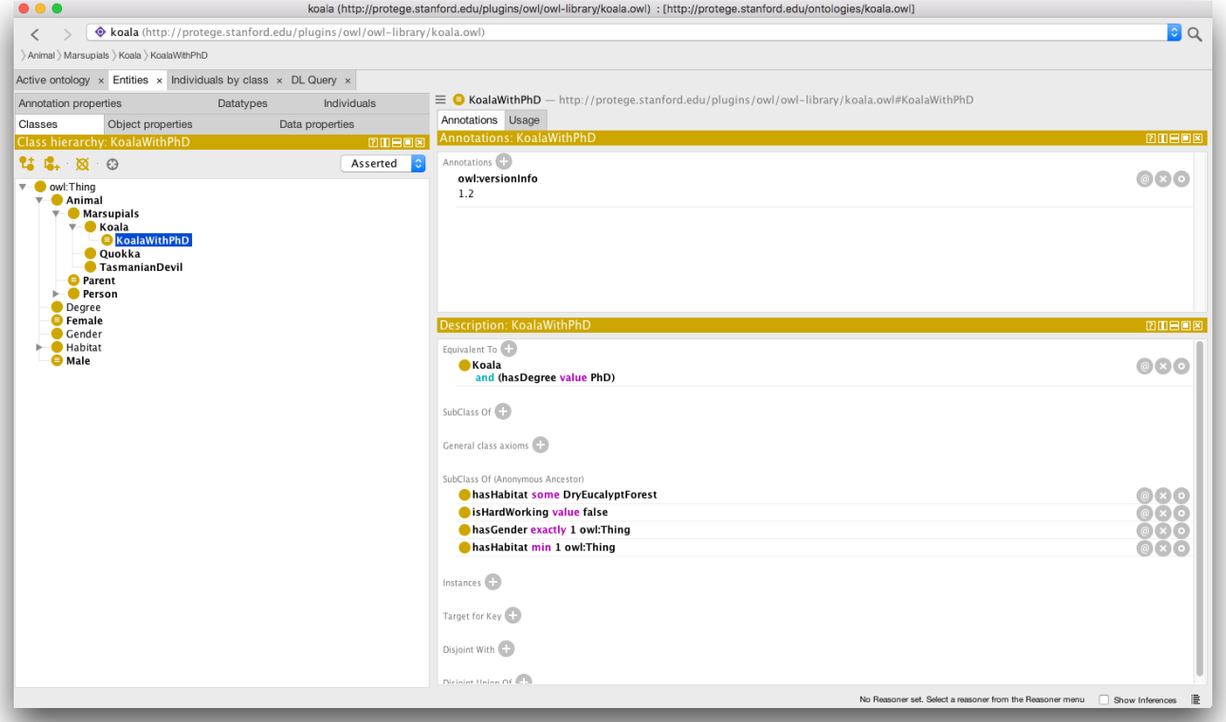
- https://inqlab.net/2019-11-19-a-primer-on-the-semantic-web-and-linked-data.html
- https://programminghistorian.org/en/lessons/intro-to-linked-data
- http://dev.iptc.org/Introduction-To-RDF
- https://www.cambridgesemantics.com/blog/semantic-university/learn-rdf/
- RDF and RDFS (slides)
- Understanding Linked Data Formats *
- Data Modeling with RDF(S) *
- https://www.w3.org/TR/rdf11-concepts/ (technical but precise)
- https://patterns.dataincubator.org/book/ (book)
- http://www.learningsparql.com/ (book)
- https://twitter.com/WorkingOntology (book)
- What is Linked Data (12-minute video, totally basic for your non-tech friends)

2A:
Create

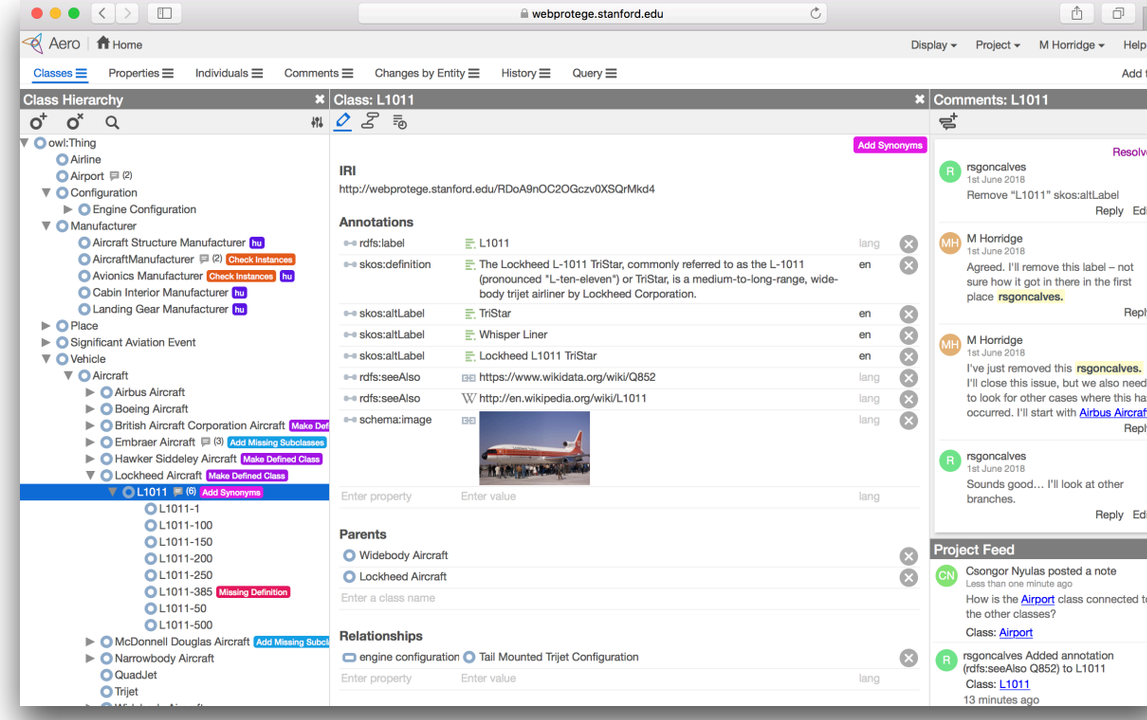
How do I create a file with RDF statements?

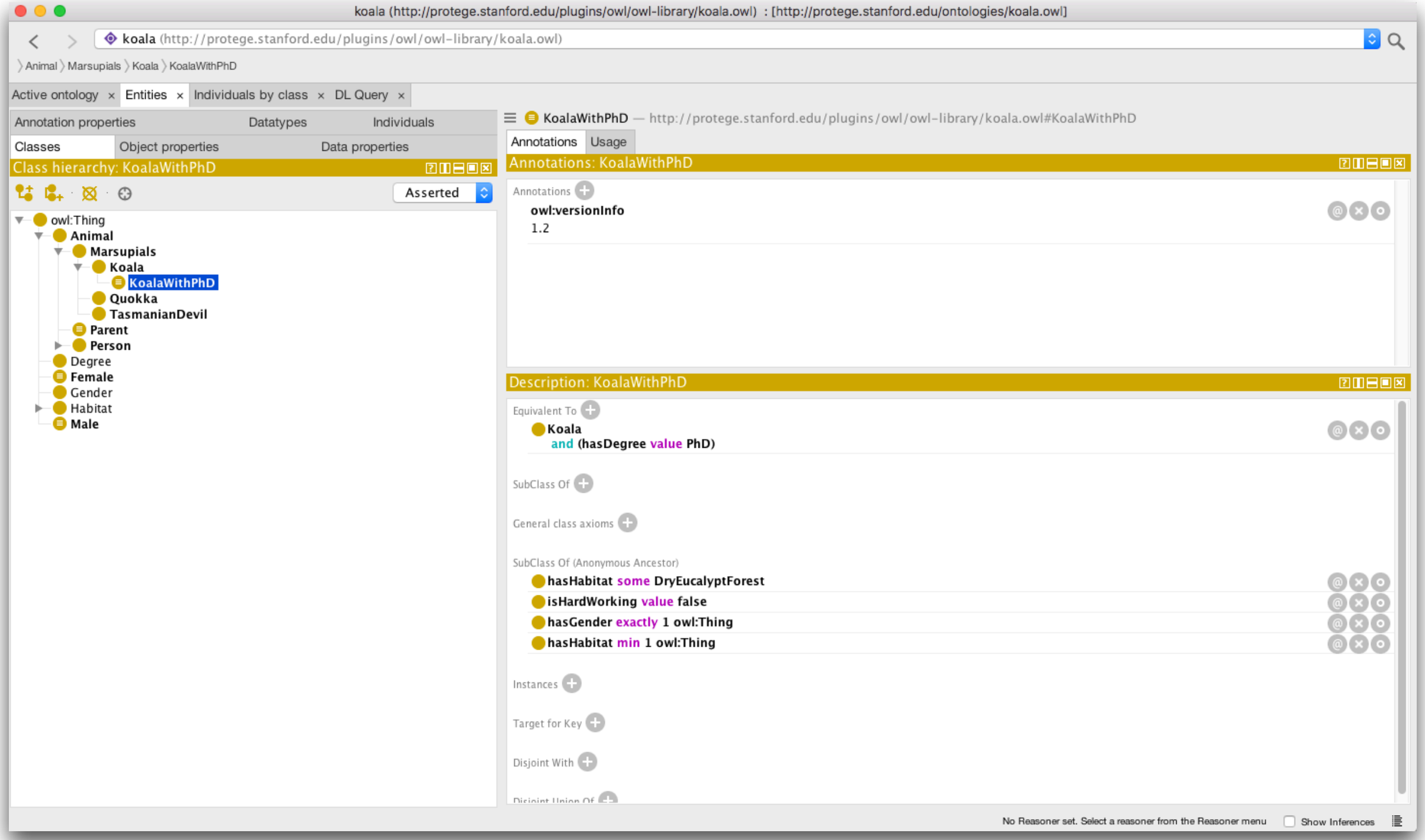
- A lot of options, from hand-editing the text (I still do this) to free to expensive
- We will briefly show one free option (pair), the one we created 30+ years ago

Protégé Desktop



WebProtégé




koala (http://protege.stanford.edu/plugins/owl/owl-library/koala.owl) : [http://protege.stanford.edu/ontologies/koala.owl]
koala (http://protege.stanford.edu/plugins/owl/owl-library/koala.owl)
Animal Marsupials Koala KoalaWithPhD
Active ontology
Entities
Individuals by class
DL Query
Annotation properties
Datatypes
Individuals
Classes
Object properties
Data properties
Class hierarchy: KoalaWithPhD
Asserted
owl:Thing
Animal
Marsupials
Koala
KoalaWithPhD
Quokka
TasmanianDevil
Parent
Person
Degree
Female
Gender
Habitat
Male
KoalaWithPhD — http://protege.stanford.edu/plugins/owl/owl-library/koala.owl#KoalaWithPhD
Annotations
Usage
Annotations: KoalaWithPhD
Annotations
owl:versionInfo
1.2
Description: KoalaWithPhD
Equivalent To
Koala
and (hasDegree value PhD)
SubClass Of
General class axioms
SubClass Of (Anonymous Ancestor)
hasHabitat some DryEucalyptForest
isHardWorking value false
hasGender exactly 1 owl:Thing
hasHabitat min 1 owl:Thing
Instances
Target for Key
Disjoint With
Disjoint Union Of
No Reasoner set. Select a reasoner from the Reasoner menu
Show Inferences

webprotege.stanford.edu

Aero | Home

Display | Project | M Horridge | Help

Classes | Properties | Individuals | Comments | Changes by Entity | History | Query | Add tab

Class Hierarchy

- owl:Thing
  - Airline
  - Airport (2)
  - Configuration
    - Engine Configuration
  - Manufacturer
    - Aircraft Structure Manufacturer hu
    - AircraftManufacturer (2) Check Instances
    - Avionics Manufacturer Check Instances hu
    - Cabin Interior Manufacturer hu
    - Landing Gear Manufacturer hu
  - Place
  - Significant Aviation Event
  - Vehicle
    - Aircraft
      - Airbus Aircraft
      - Boeing Aircraft
      - British Aircraft Corporation Aircraft Make Def
      - Embraer Aircraft (3) Add Missing Subclasses
      - Hawker Siddeley Aircraft Make Defined Class
      - Lockheed Aircraft Make Defined Class
        - L1011 (6) Add Synonyms
          - L1011-1
          - L1011-100
          - L1011-150
          - L1011-200
          - L1011-250
          - L1011-385 Missing Definition
          - L1011-50
          - L1011-500
      - McDonnell Douglas Aircraft Add Missing Subcl
      - Narrowbody Aircraft
        - QuadJet
        - Trijet

Class: L1011

Add Synonyms

IRI

http://webprotege.stanford.edu/RDoA9nOC2OGczv0XSQrMkd4

Annotations

rdfs:label	L1011	lang
skos:definition	The Lockheed L-1011 TriStar, commonly referred to as the L-1011 (pronounced "L-ten-eleven") or TriStar, is a medium-to-long-range, wide-body trijet airliner by Lockheed Corporation.	en
skos:altLabel	TriStar	en
skos:altLabel	Whisper Liner	en
skos:altLabel	Lockheed L1011 TriStar	en
rdfs:seeAlso	https://www.wikidata.org/wiki/Q852	lang
rdfs:seeAlso	http://en.wikipedia.org/wiki/L1011	lang
schema:image		lang
Enter property	Enter value	lang

Parents

- Widebody Aircraft
- Lockheed Aircraft

Enter a class name

Relationships

engine configuration	Tail Mounted Trijet Configuration	
Enter property	Enter value	lang

Comments: L1011

Resolve

R rsgoncalves
1st June 2018
Remove “L1011” skos:altLabel
Reply Edit

MH M Horridge
1st June 2018
Agreed. I'll remove this label – not sure how it got in there in the first place **rsgoncalves.**
Reply

MH M Horridge
1st June 2018
I've just removed this **rsgoncalves.** I'll close this issue, but we also need to look for other cases where this has occurred. I'll start with Airbus Aircraft.
Reply

R rsgoncalves
1st June 2018
Sounds good… I'll look at other branches.
Reply Edit

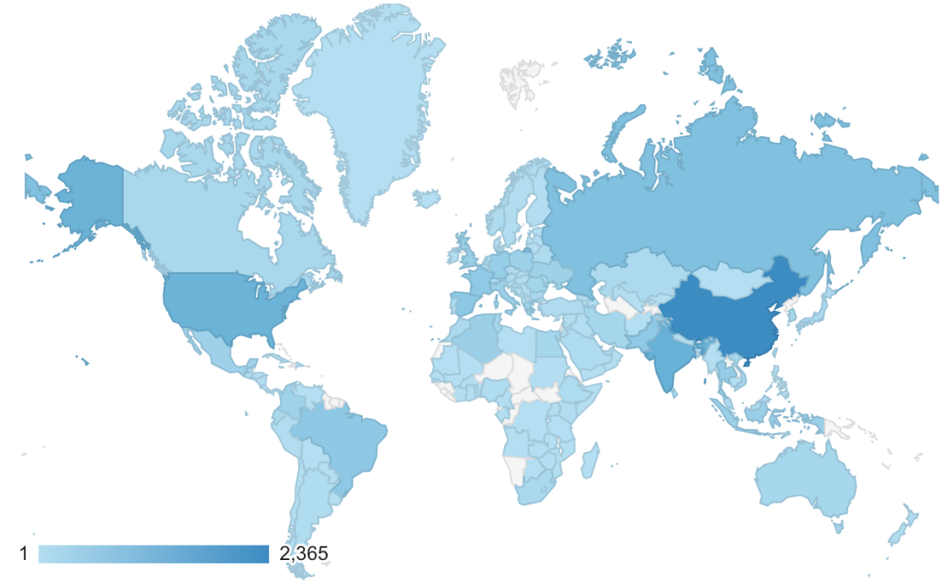
Project Feed

CN Csongor Nyulas posted a note
Less than one minute ago
How is the Airport class connected to the other classes?
Class: Airport

R rsgoncalves Added annotation (rdfs:seeAlso Q852) to L1011
Class: L1011
13 minutes ago

Protégé Community

- Over 365,000 registered users
- Over 21,000 mailing list subscribers
- Academia, industry, government
- Over 76,000 downloads of Protégé v5.5.0



- Over 141,000 downloads of 5.5.0 beta series in 2019
- 55,000+ projects; 70,000+ users on https://webprotege.stanford.edu

Who uses Protégé?

- World Health Organization (WHO) – ICD-11
- National Cancer Institute (NCI) – NCI Thesaurus
- Gene Ontology Consortium – Gene Ontology
- Object Mgmt. Group & Enterprise Data Mgmt. Council - FIBO
- The Open Biological and Biomedical Ontologies (OBO Foundry)
- Foundational Model of Anatomy (FMA)

Protégé Web Site and Wiki

URL: http://protege.stanford.edu

- Download new versions of Protégé
- Subscribe to mailing lists
- Learn about support options
- Links to Protégé wiki, Facebook, and Twitter

URL: http://protegewiki.stanford.edu

- WebProtégé user documentation
- Protégé Desktop user documentation
- Developer documentation
- Plug-in and ontology libraries

MAILING LISTS

Get free support for all of your Protégé questions via our mailing lists. The lists are actively monitored by Stanford's Protégé team, as well as many experienced Protégé users from the community at large.

Click the Subscribe button next to the list you wish to join. Please note that you must be subscribed to a list before you can post messages.

Protégé User Support

User support for all versions of WebProtégé and Protégé Desktop.

SUBSCRIBE

ARCHIVE »

Protégé Developer Support

Developer support for all versions of WebProtégé and Protégé Desktop.

SUBSCRIBE

ARCHIVE »

Protégé Announcements

Low traffic, announcement-only list for new releases, availability of short courses, and information regarding Protégé Conferences.

SUBSCRIBE

ARCHIVE »

Protégé on GitHub

All software (Protégé, WebProtégé, plugins, etc.) is on GitHub

https://github.com/protegeproject

Protégé: https://github.com/protegeproject/protege

WebProtégé: https://github.com/protegeproject/webprotege

Make feature requests & bug reports in our GitHub issue trackers

Social Media

Twitter (@protegeproject)



Facebook (Protege Project)



2B:
Make Available

BioPortal Ontologies Search Annotator Recommender Mappings Login Support

Welcome to BioPortal, the world's most comprehensive repository of biomedical ontologies

Search for a class

Enter a class, e.g. Melanoma

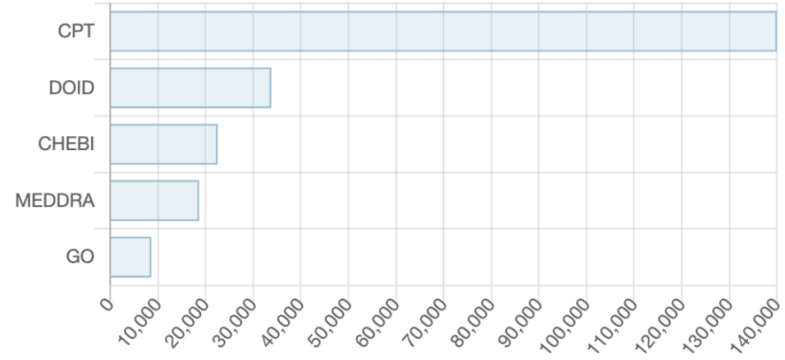
Advanced Search

Find an ontology

sweet

Browse Ontologies

Ontology Visits (August 2020)



More

BioPortal Statistics

Ontologies	890
Classes	12,070,303

PRODUCTS
BioPortal

SUPPORT
Contact Us

ABOUT
About Us

CONNECT

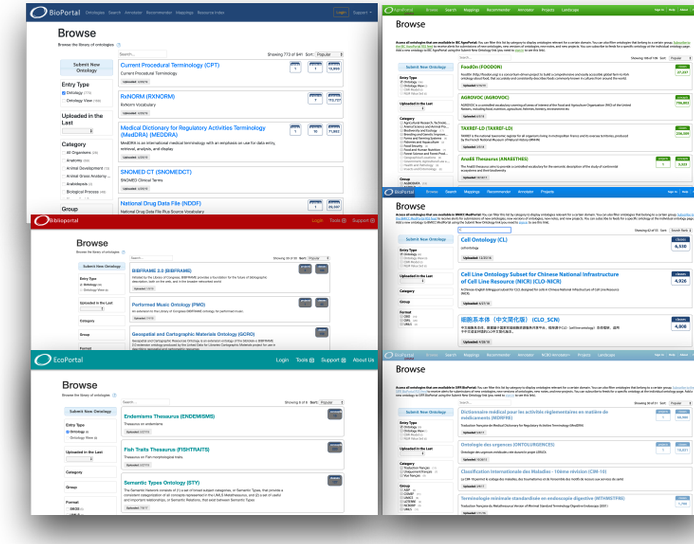
BioPortal (BMIR)

Storing Semantic Content

Biomedical and more!

- NCBO BioPortal (900+ community ontologies)
  - Started in 2004/2005 with National Institutes of Health grant
  - Goal: represent biomedical knowledge (concepts) in one place
  - Emphasized practicality over semantic fidelity
- Significant funding over first 12 years, largely by NIH
  - Developed large number of features, deployable ‘appliance’
  - Started as relational system, migrated to triple store backend…
  - While originally focused on biomedicine, now branching out

Where are we now?



- BioPortal (900+ community ontologies)
  - Extremely large user base (>12K)
  - Extremely large usage (408K hits/mo, 1.4M unique users/yr)
- Large number of deployments => OntoPortal Alliance
  - Community center of support at https://ontoportal.org
  - Opens up new avenues for pursuing support
- Significant ongoing support (>99.5% uptime, new triple store)

OntoPortal Deployment Software

- OntoPortal Virtual Appliance
  - Download as VMWare Virtual Appliance OVF (44 requests, 11 active))
  - Runnable as Amazon Web Service AMI (30)
  - Installable (and even maintainable) by us for you (for a fee)
- Significant upgrades in recent 3.0 release
  - Improved ability to deploy systems and ‘call home’ re upgrades
  - Latest UIs, APIs, and bug fixes
  - Allegrograph triple store (actually quad store) support
  - Self-installable in full preview mode; free license for most users

https://ontoportal.github.io/administration

BioPortal Features

- Submission/Ingest
  - Format Conversions
  - Versioning
  - Differencing
  - Metrics re Vocab.
  - Interpretation/ Property Mapping
- Search
  - Ontology Name
  - Terms
  - Best Terms
  - Attribute/Value
  - Metadata
  - Triples
- Subselection
  - Views (partial voc.)
  - Slices (partial repo)
  - Preferred Vocabs
- Terms
  - Browsing
  - Create ID
  - Resolve ID
  - Term Pages
  - Mappings
    - Auto–syntax
    - Auto–identifiers
    - Manual–user
- Vocab. Recommend
  - Keyword-based
  - Text-based
  - Multi-vocabulary
  - Controllable criteria
  - Selectable vocabs
- Annotation
  - Keyword/text based
  - Selectable vocabs

Widgets
- Term completion
- Term browsing
- Visualization

Social
- Reviews
- Notes/Comments
  - Ontology
  - Term

Usage Tracking
- # Requests
- # Downloads
- # Page hits
- # API requests

API Features

Resource Index

Visualization

The Plan So Far

- Define your concepts
  - Find concepts and definitions in existing terminologies/ontologies, OR
  - Transform your existing dictionaries into standard such as SKOS, OR
  - Create a controlled terminology or ontology based on your needs (using Protégé, WebProtégé, or other tools), OR
  - Mix and match the above
- Put your concepts into BioPortal so they can be seen and used
  - (You can keep them private if you must, but CEDAR users will see them)
  - BioPortal can read any RDF, SKOS, or OWL file (good practices help!)
  - BioPortal will make your ontology and its concepts easy to find

Part 3:
How do we make metadata using standardized semantics?

The CEDAR Vision

A simple 'life of metadata' for users

- Compatibility with best known practices and standards
- Use of existing semantic and data structure resources
- Interoperability with existing workflows/systems
- End-to-end improvement in user-supplied metadata



CEDAR

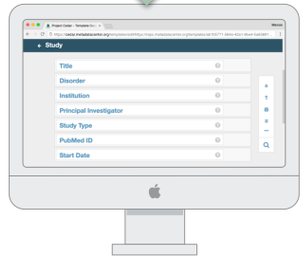
metadatacenter.org



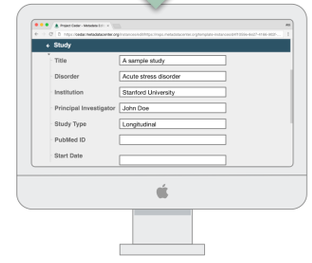
Template Specifiers



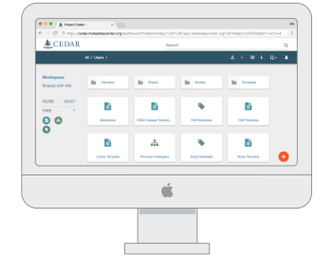
Researchers/Operators



Template Designer



Metadata Editor



Resource Manager

Open Services

- Authentication Service (Keycloak)
- User Service

user management

- Resource Service

resource management

- Value Recommender Service

intelligent authoring

- Submission Service

metadata export

- Terminology Service

controlled terms

NCBO Open Services

BioPortal ontologies

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

Public Repositories

r_1

r_m

= CEDAR components

= Third-party components

CEDAR

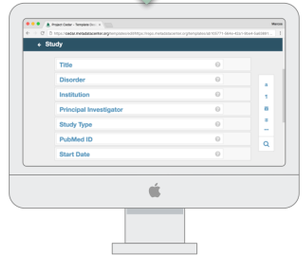
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL



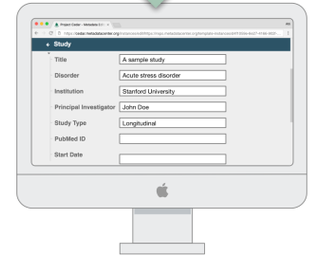
Template Specifiers



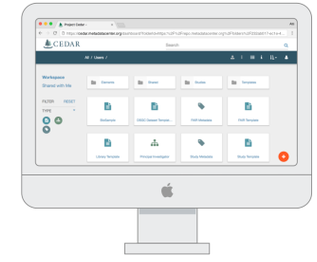
Researchers/Operators



Template Designer



Metadata Editor



Resource Manager

Open Services

- Authentication Service (Keycloak)
- User Service

user management

- Resource Service

resource management

- Value Recommender Service

intelligent authoring

- Submission Service

metadata export

- Terminology Service

controlled terms

NCBO Open Services

BioPortal ontologies

THE NATIONAL CENTER FOR BIOMEDICAL ONTOLOGY

Public Repositories

r_1

r_m

CEDAR

CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL

= CEDAR components

= Third-party components

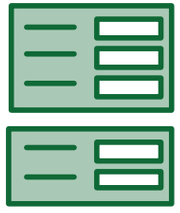
CEDAR Workspace elements







Metadata templates



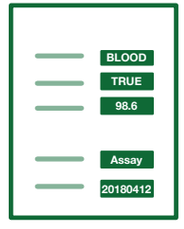


Template elements
(groups of fields and/or elements)





Template fields





Instantiated metadata (metadata instances)

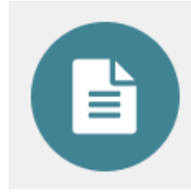


Folders

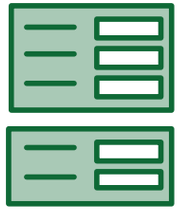
CEDAR Workspace elements







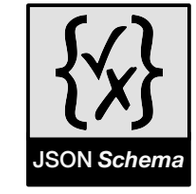
Metadata templates





Template elements
(groups of fields and/or elements)



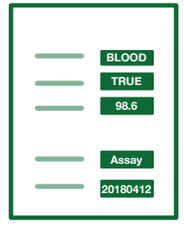








Template fields



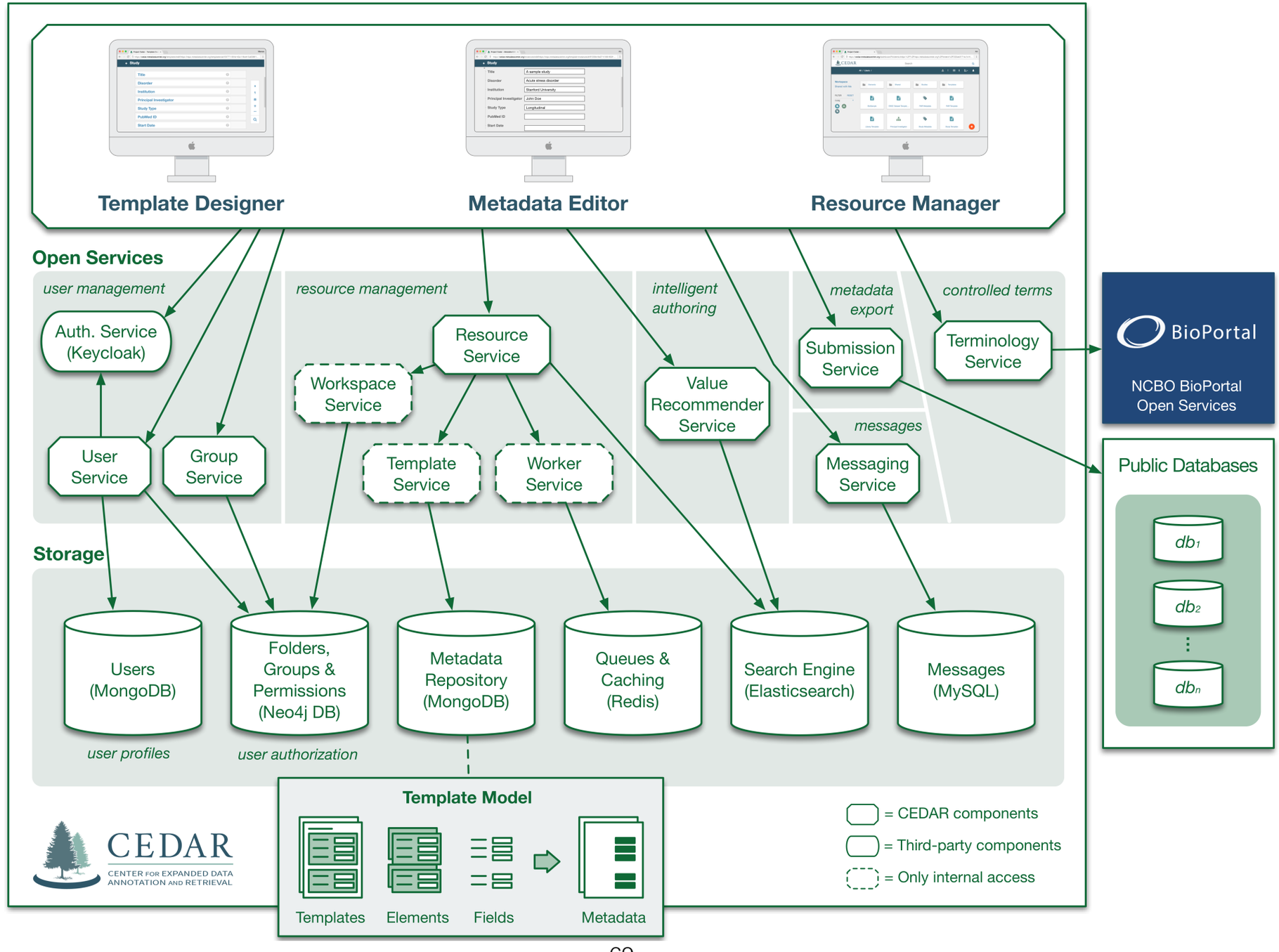


Instantiated metadata (metadata instances)

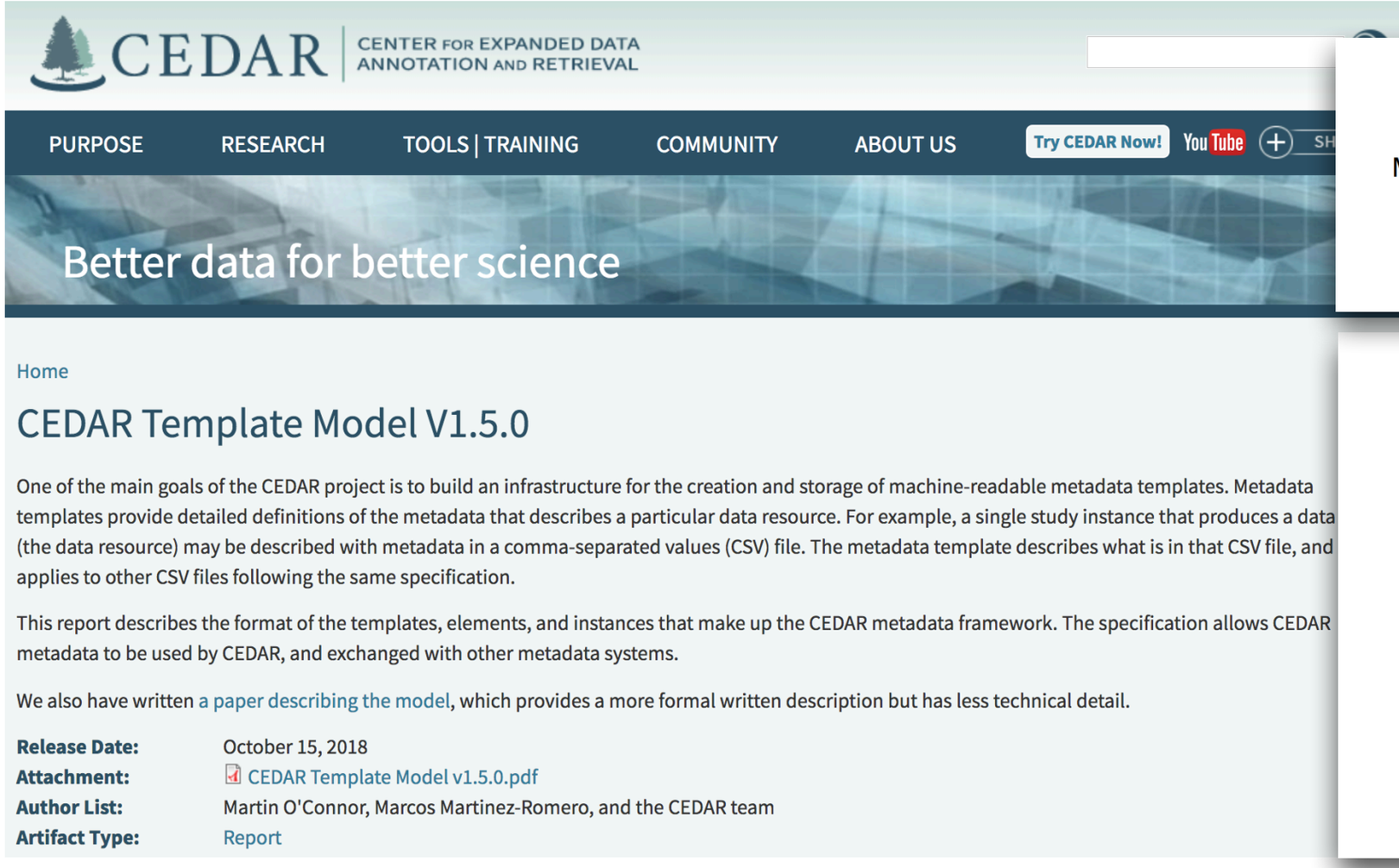


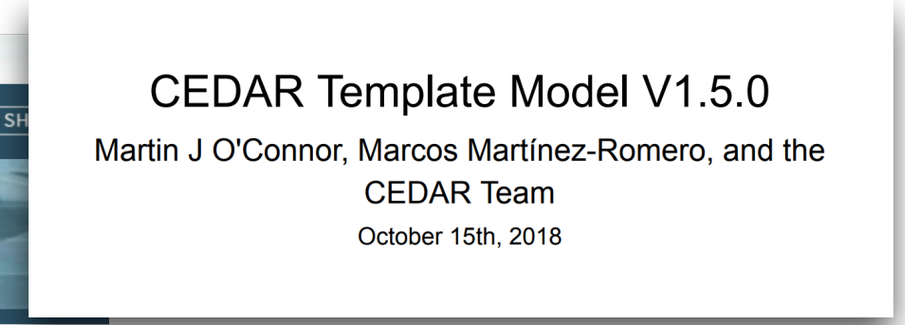


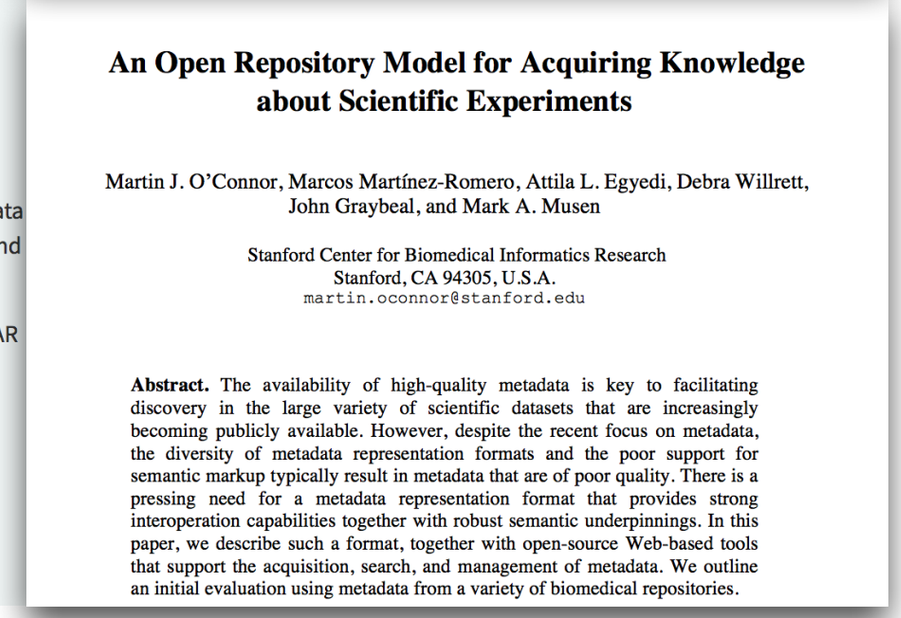
Folders


Template Designer
Metadata Editor
Resource Manager
Open Services
user management
Auth. Service (Keycloak)
User Service
Group Service
resource management
Resource Service
Workspace Service
Template Service
Worker Service
intelligent authoring
Value Recommender Service
metadata export
Submission Service
controlled terms
Terminology Service
messages
Messaging Service
BioPortal
NCBO BioPortal Open Services
Public Databases
db1
db2
dbn
Storage
Users (MongoDB)
Folders, Groups & Permissions (Neo4j DB)
Metadata Repository (MongoDB)
Queues & Caching (Redis)
Search Engine (Elasticsearch)
Messages (MySQL)
user profiles
user authorization
Template Model
Templates
Elements
Fields
Metadata
CEDAR
CENTER FOR EXPANDED DATA ANNOTATION AND RETRIEVAL
= CEDAR components
= Third-party components
= Only internal access

Model Specifications







https://metadatacenter.org/cedar-template-model

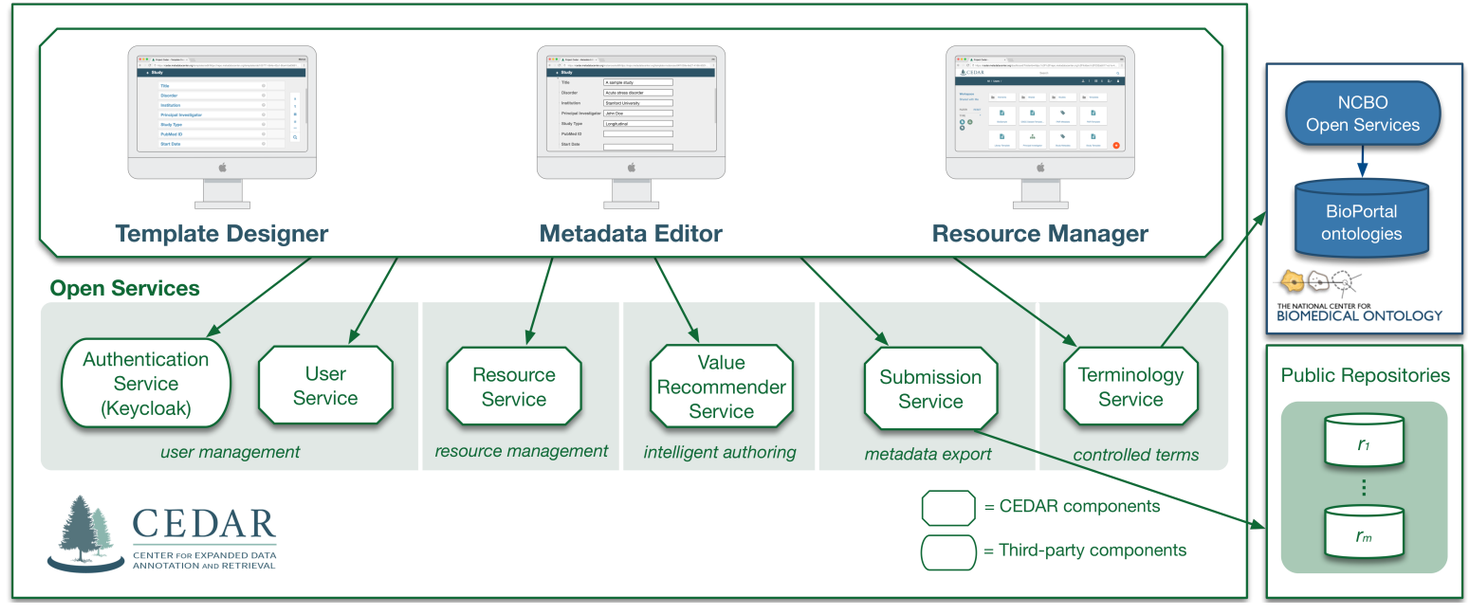
Demonstration

To see more demonstrations visit YouTube

- Playlists: https://www.youtube.com/c/MetadataCenterOrg/playlists
- Short videos of CEDAR features: GO FAIR M4M DeiC Workshop

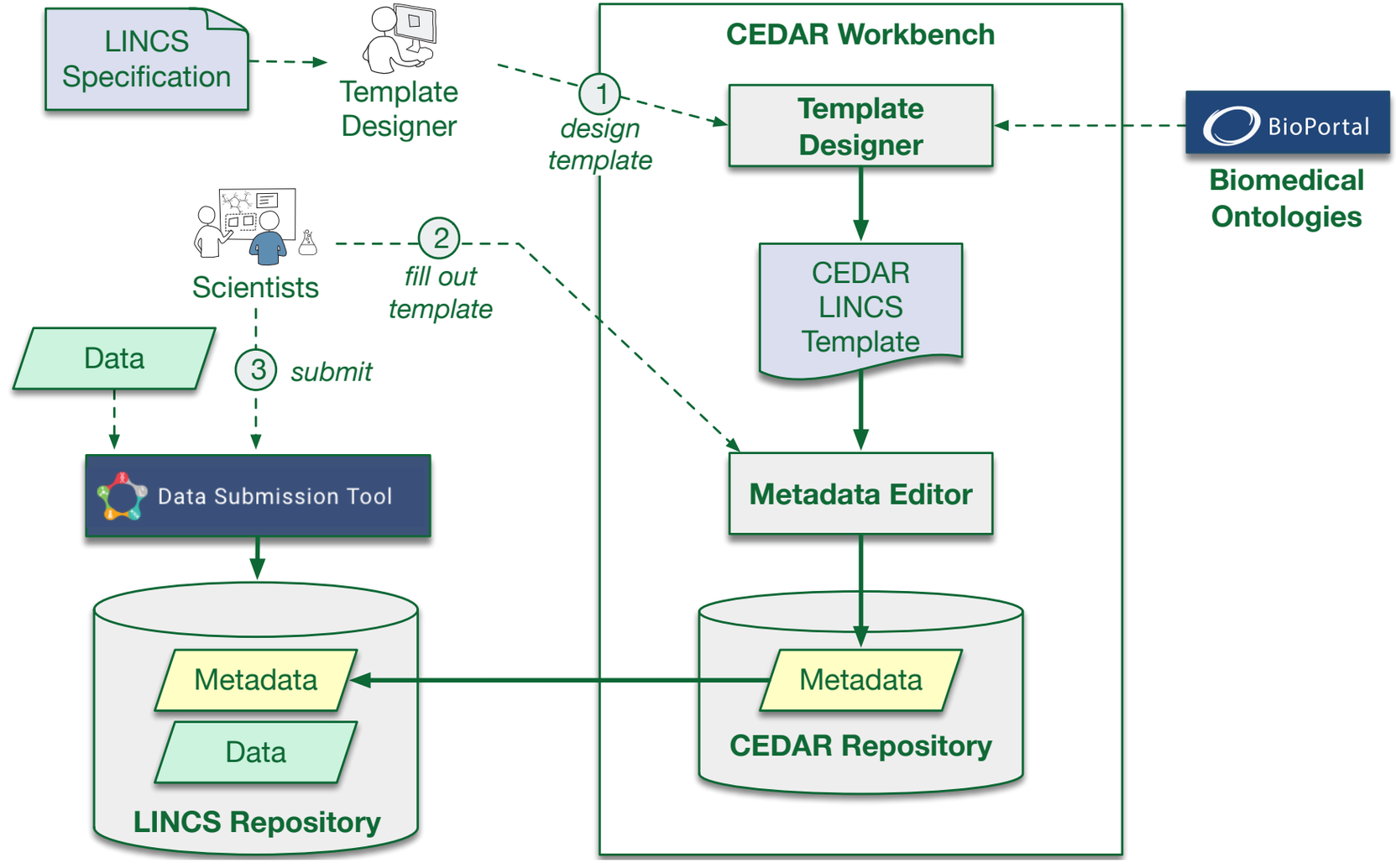
Using Principled Approaches

- All knowledge artifacts managed as first-class identifiable entities
- All software components have well-defined APIs
- All content represented in JSON and JSON-LD
- Domain-independent metadata specification
- On-the-fly validation of metadata templates and content



Deployable for Any Occasion

LINCS uses it to **get** metadata for their data repository.



Deployable for Any Occasion

The AIRR community uses it to **submit** metadata to the National Center for Biotechnology Information (NCBI) repositories.

