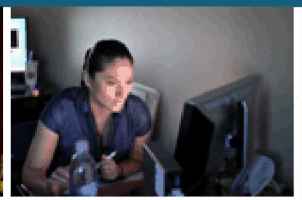
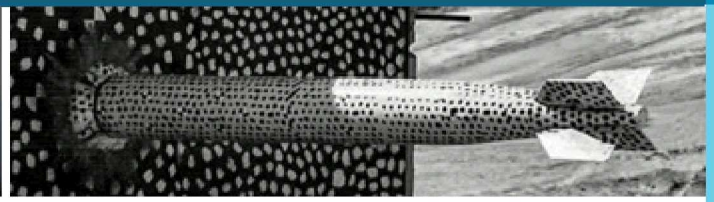
SAND2020-9921C

DCAT-eOS-AP: A Metadata Schema for Use in Nuclear Monitoring Projects with Applications to Other DOE Mission Areas





Kale Aur, Brian Young, Lauren Wheeler, Rose Borden & Russell Pate

DOE Data Day (D3), October 5-7, 2020

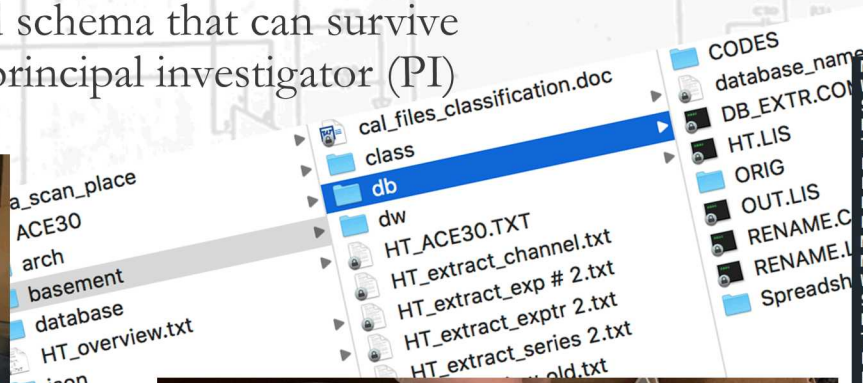
Sandia National Laboratories is a multimission laboratory managed and operated by National Technology & Engineering Solutions of Sandia, LLC, a wholly owned subsidiary of Honeywell International Inc., for the U.S. Department of Energy's National Nuclear Security Administration under contract DE-NA0003525.

Problem statement

Without appropriate metadata, data may become unidentifiable and unusable

Increasing understanding DOE needs:

- Standard metadata
- Understandable metadata
- Logical, documented schema that can survive beyond a project & principal investigator (PI)



Problem Solution

Approach: Evaluate metadata standards within cross-disciplinary projects of nuclear monitoring community, with focus on:

- Technical use
- Interoperability
- Discoverability & Accessibility
- Preservation, Reusability, & Version control
- Solicit Subject Matter Experts (SMEs) on domain specific metadata/data standards

Goal: Develop an extensible, modular, machine-readable metadata schema for domains relevant to nuclear monitoring

Impact:

- Formalized metadata standards/schema utilized across DNN R&D
- Extending usability of data
- Facilitating data accessibility, discovery, reuse, sharing, and preservation
- Widespread adoption to meet increasing government research data management needs

SME domains considered

- Atmospheric
- Geologic
- Geochemical
- GIS
- Geomechanics
- Geophysics
- Ground Truth
- LIDAR
- Modeling
- Radionuclide
- Derived Products
- Documents

Data Catalog Vocabulary (DCAT)

...the basis for DCAT-eOS-AP

DCAT – Introduction

- "…designed to facilitate interoperability between data catalogs published on the Web" (W3C, 2020)
- Extensive use of other standards/schemas, building off of Dublin Core
- Uses a standard model and vocabulary that facilitates consumption and aggregation of metadata from multiple catalogs
- Makes no assumptions about data format
- Developed in the context of government data catalogs
- Active user/development community, adopted by data.gov, Europe, others
- Extensible, modular, and can be presented in many forms including RDF, embedded in HTML, or serialized in XML, N3, Turtle, JSON, etc.

DCAT *classes* – What's in a name?

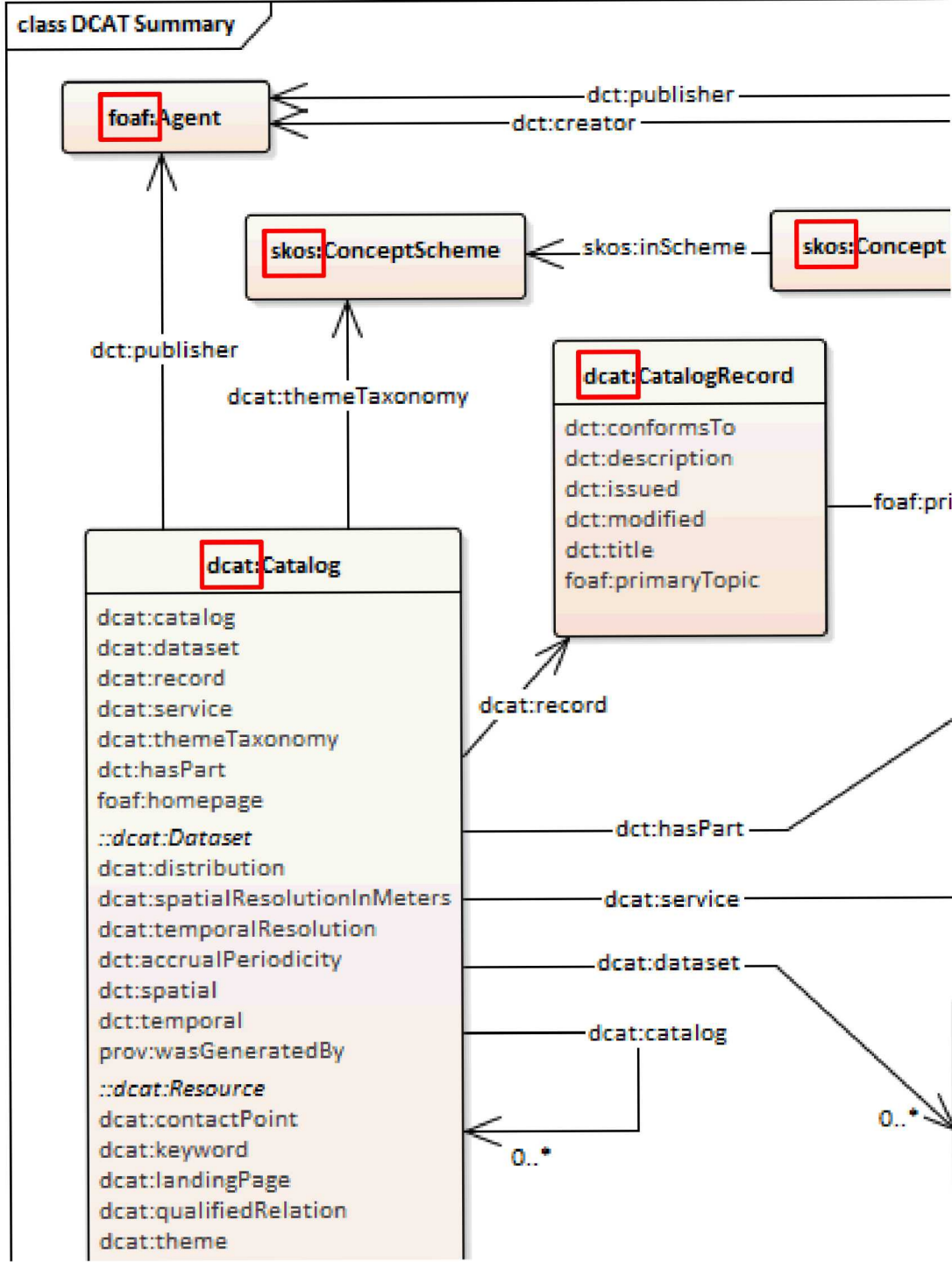
Prefixes define different *vocabularies*

Example *prefixes*

- dcat Data Catalog Vocabulary
- foaf Friend of a Friend
- skos Simple Knowledge Organization System

DCAT is extensible

- DCAT pulls in outside vocabularies using *prefixes*
- Why make new standards when you can use someone else's?



DCAT – Application Profiles

A **DCAT application profile** is a new standard that extends the core DCAT ***but still conforms with it***

Because DCAT is extensible, many organizations have added their own classes, properties, and constraints

Examples include:

- POD — Project Open Data
- DCAT-AP — DCAT Application Profile for Data Portals in Europe
- StatDCAT-AP — DCAT-AP for statistical data
- GeoDCAT-AP — DCAT-AP for geospatial data
- CiteDCAT-AP — DCAT-AP for dataset citation
- EPOS-DCAT-AP — European Plate Observing System
- CKANEXT-DCAT — Comprehensive Knowledge Archive Network Extension

- **DCAT-eOS-AP** — **Ecosystem for Open Science Application Profile**

DCAT-eOS-AP Core Classes

DCAT-eOS-AP

LEGEND: UML Diagram

DCAT-eOS-AP v0.2:
Data Catalog (Version 2)
Ecosystem for Open Science
Application Profile v0.2

Class Diagram:

vocab:Class
+ vocab:property
+ vocab:property
+ vocab:property
:: Inherited Properties

Blue boxes are demonstrative only (pending future development)

Arrows:

Association
Vocabulary which describes how one Class is associated with another Class.
ie: dct:publisher links doe:Project Class with dcat:Resource Class

Inheritance
One Class inherits properties from the other Class.
ie: dcat:Dataset inherits properties from dcat:Resource.

Junction
A junction dot indicates a connection between arrows. Cardinality constraints are placed at the junction dot.

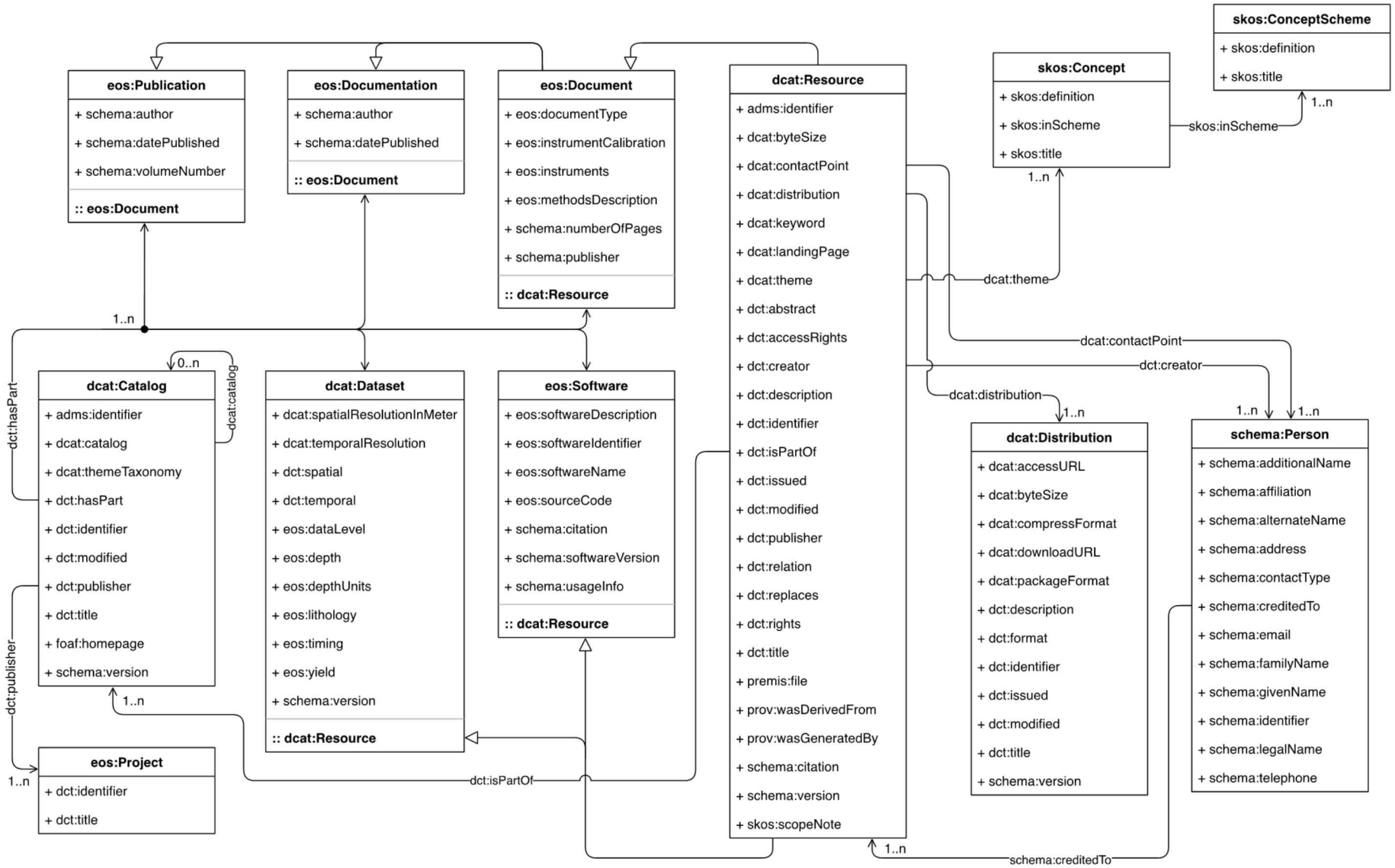
Line Jump
A line jump indicates no connection between arrows.

Cardinality Constraints:

1	Required, only 1 instance
1..n	Required, unlimited instances
0..1	Optional, only 1 instance
*	Optional, unlimited instances

Metadata Vocabularies:

adms	Asset Description Metadata Schema
dcat	Data Catalog Vocabulary
dct	Dublin Core
eos	eOS-specific Vocabulary
foaf	Friend of a Friend Vocabulary
format	eOS-specific Format
premis	PREMIS Vocabulary
prov	Provenance Ontology
rdf	Resource Description Framework Schema
schema	Schema.org Vocabulary
skos	Simple Knowledge Organization System

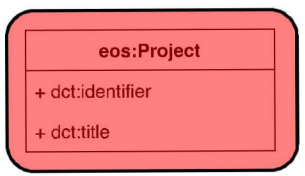


DCAT Ecosystem for Open Science Application Profile (DCAT-eOS-AP) Example

DCAT-eOS-AP Example

Each submission must start with a *Project:*

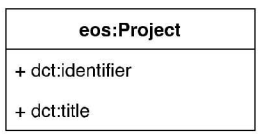
- NA-22 Project Name
- NA-22 Project Number



DCAT-eOS-AP Example

Then we add a *Dataset:*

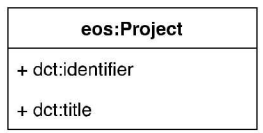
- NetCDF file with 24 hours of 3D atmospheric fields from a physical model



DCAT-eOS-AP Example

We add a second *Dataset:*

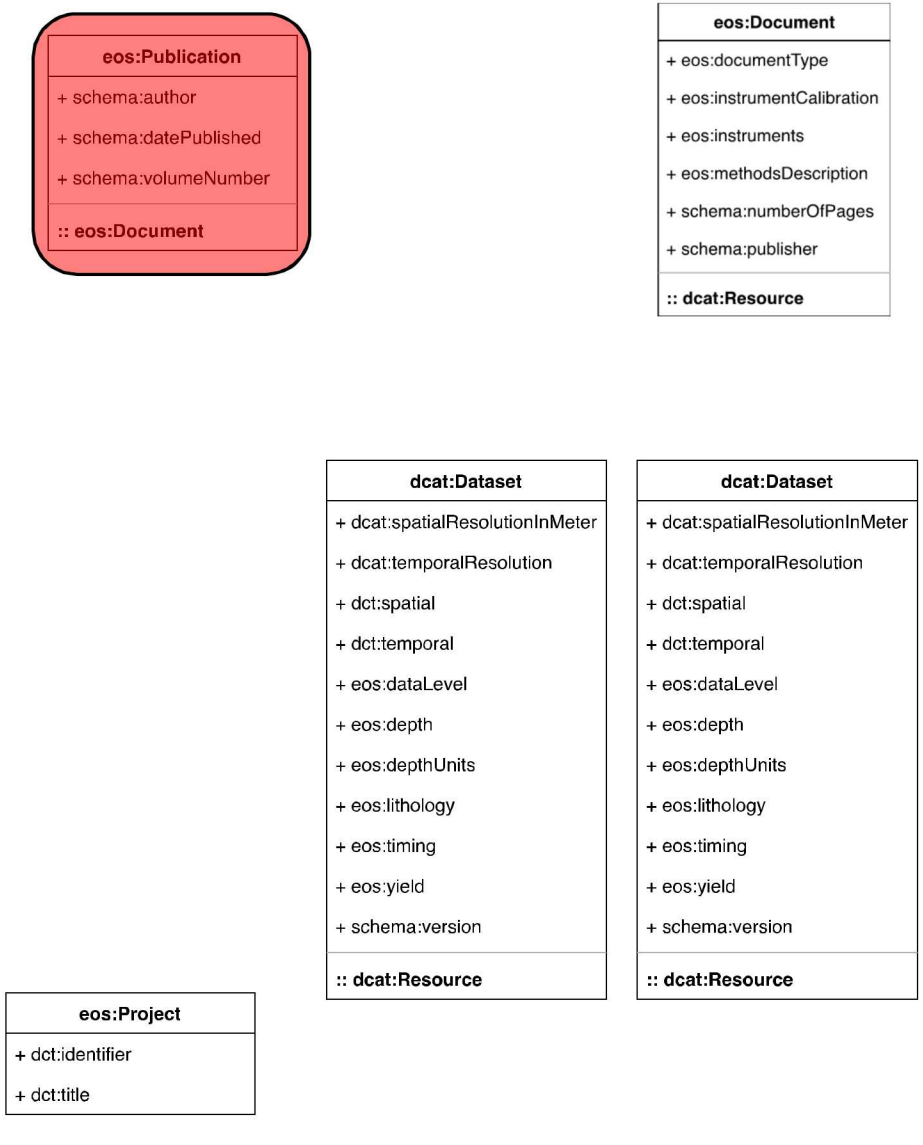
- ASCII file with a single atmospheric sounding



DCAT-eOS-AP Example

We also have a *Publication:*

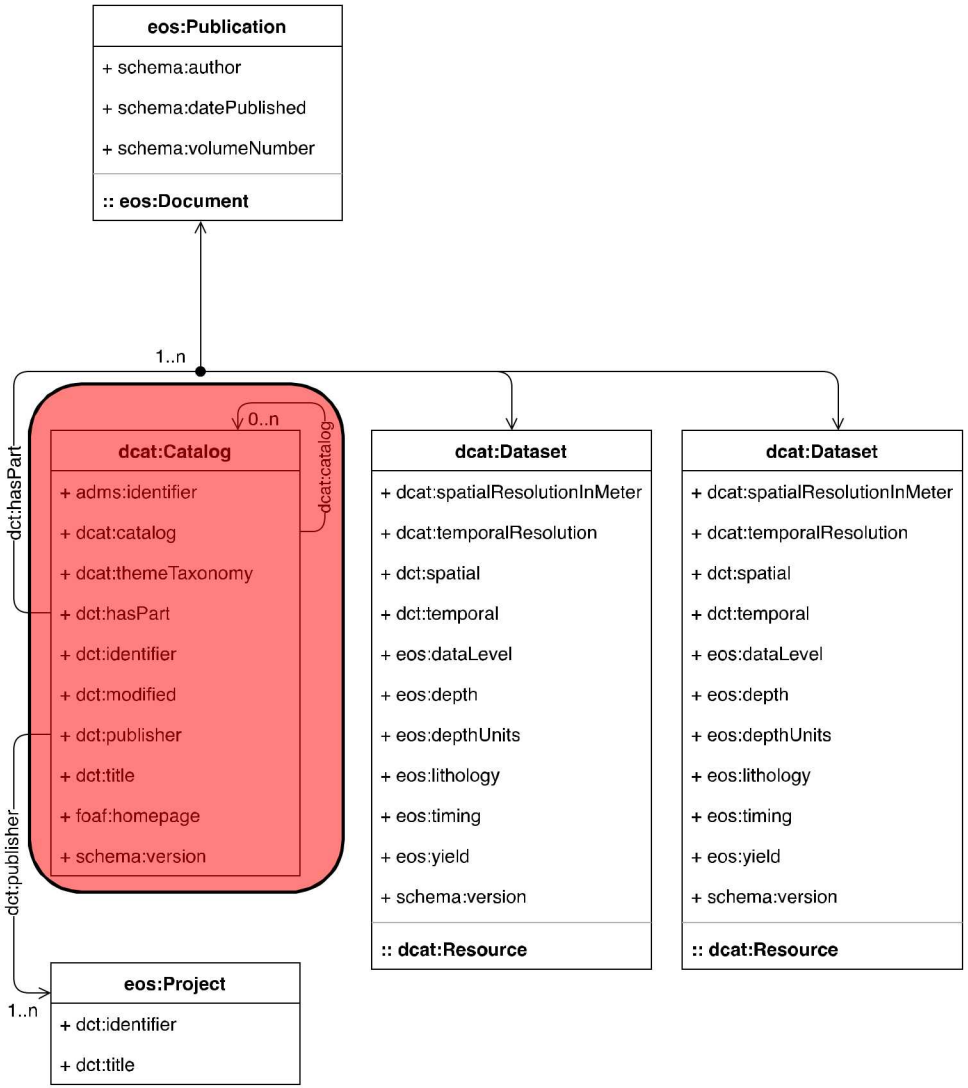
- PDF of SAND report



DCAT-eOS-AP Example

Datasets/*Publications* are contained within a *Catalog* via **hasPart**

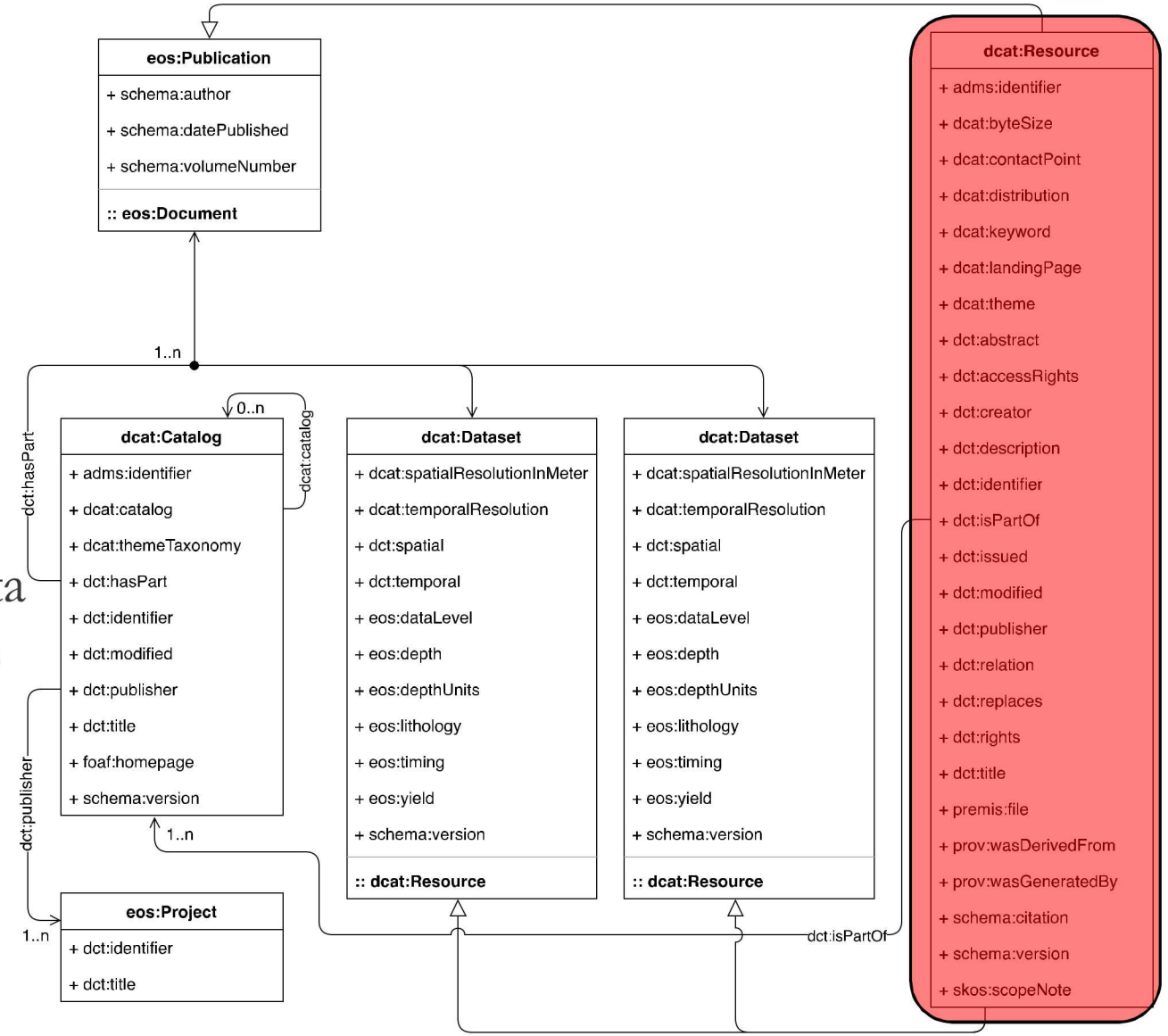
The *Catalog* is linked to the *Project* via **publisher**



DCAT-eOS-AP Example

Inheritance is represented by an open arrow

Common metadata are inherited from *Resource*



- **isPartOf** links a *Resource* to one or more *Catalogs*
- **relation** establishes relationships between *Resources*
- **replaces** used to denote which *Resource* is replaced or deprecated by this *Resource*
- **wasDerivedFrom** provides a link to another *Resource* this *Resource* may be derived from

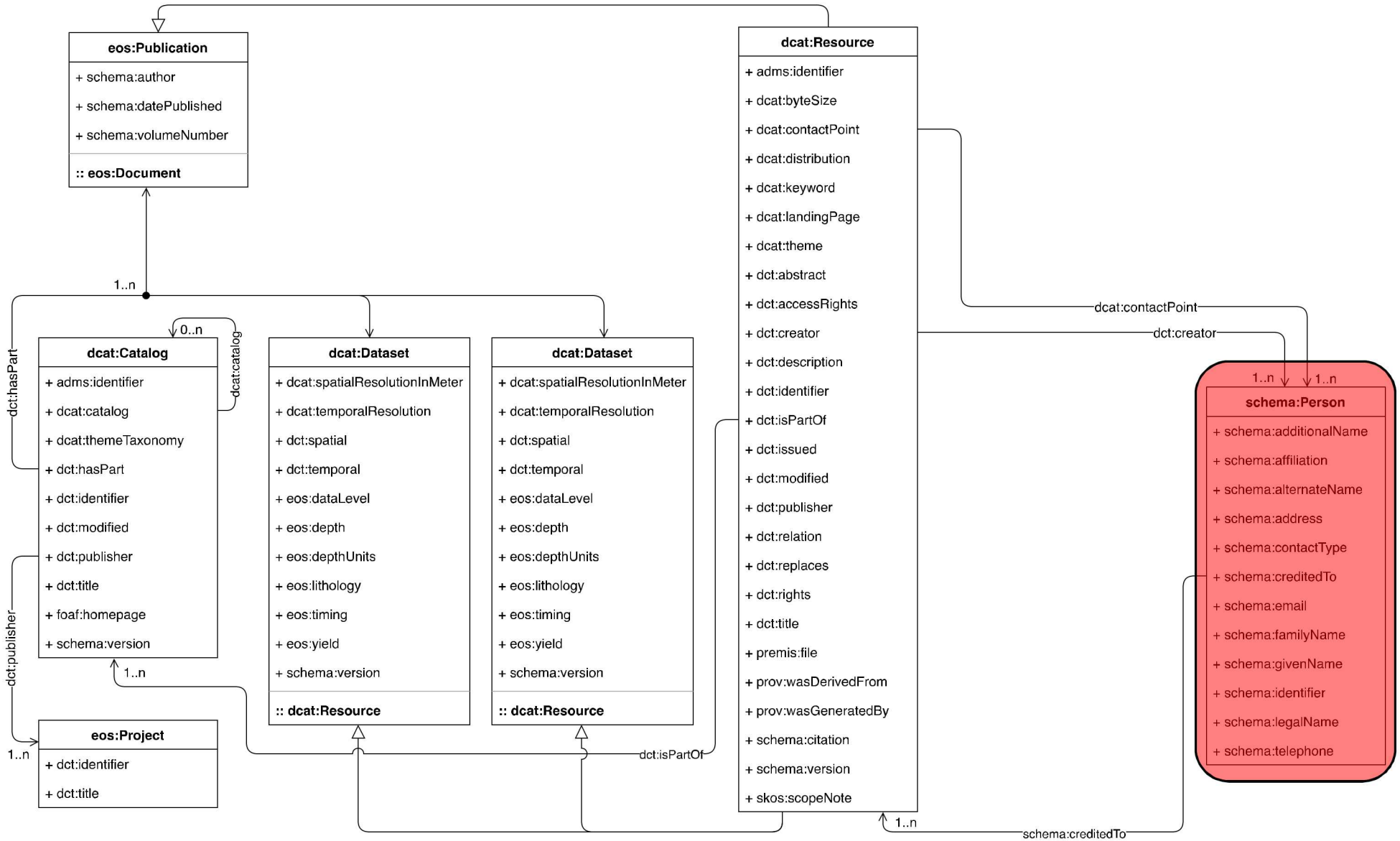
DCAT-eOS-AP Example

Each *Dataset* or *Publication* must have a **contactPoint**

Person relates to *Resource* via **contactPoint**

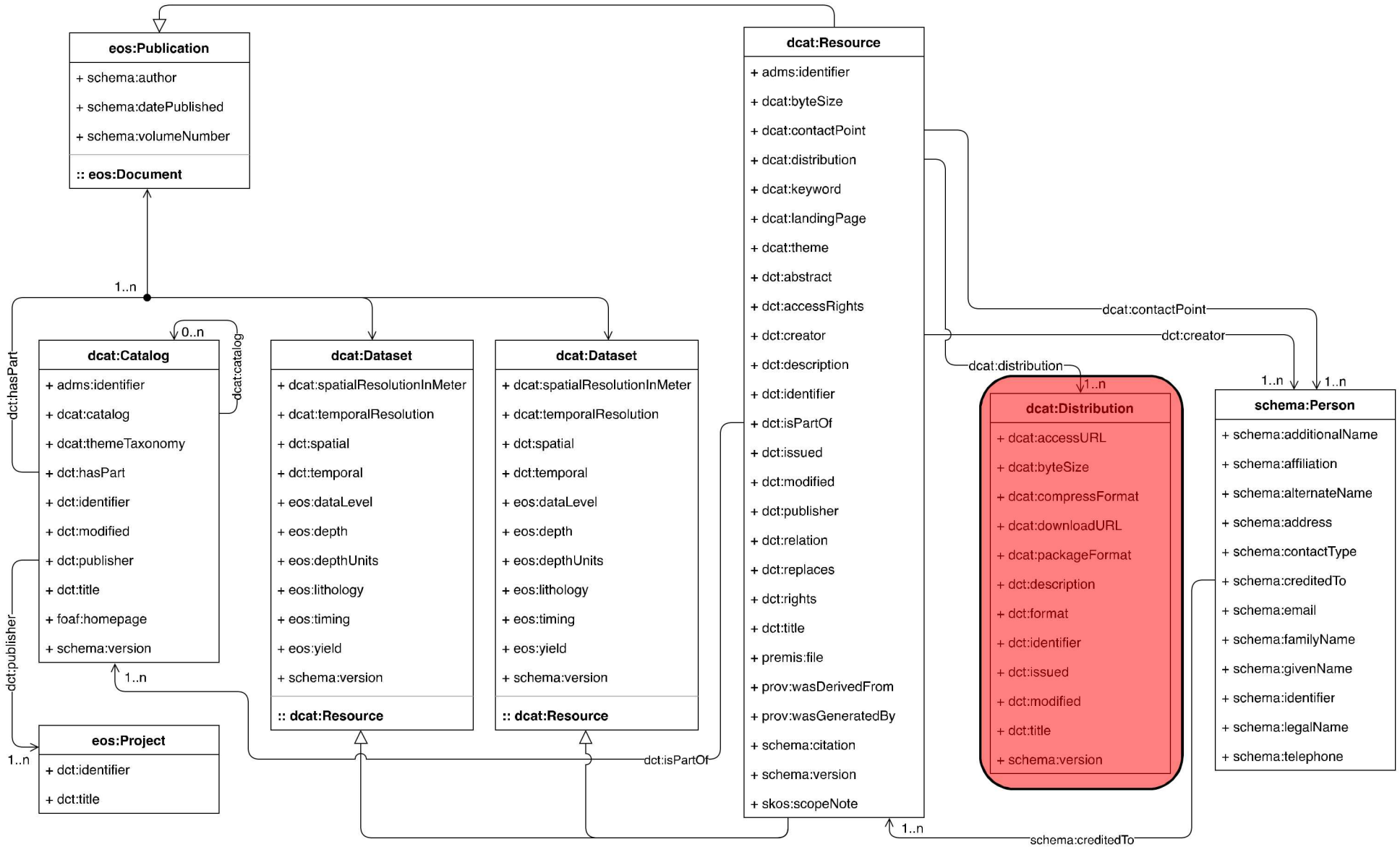
A *Person* is associated with multiple *Resources* using **creditedTo**

A *Person* also relates additional collaborators to *Resource* via **creator**



DCAT-eOS-AP Example

Distribution contains metadata about how the *Resource* is packaged and accessed



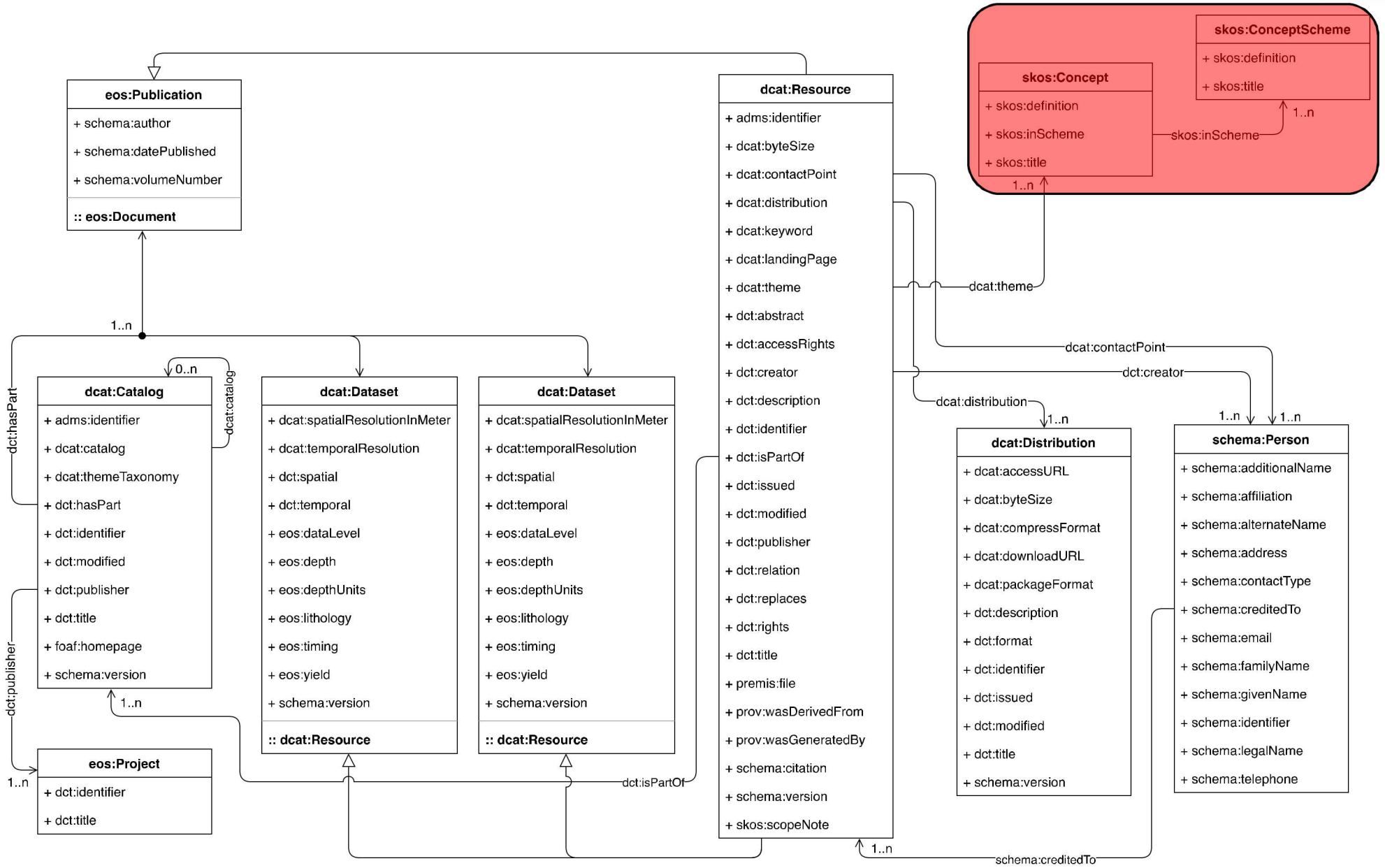
DCAT-eOS-AP Example

Concept

- Category or theme describing a *Resource* (e.g., waveform timeseries)

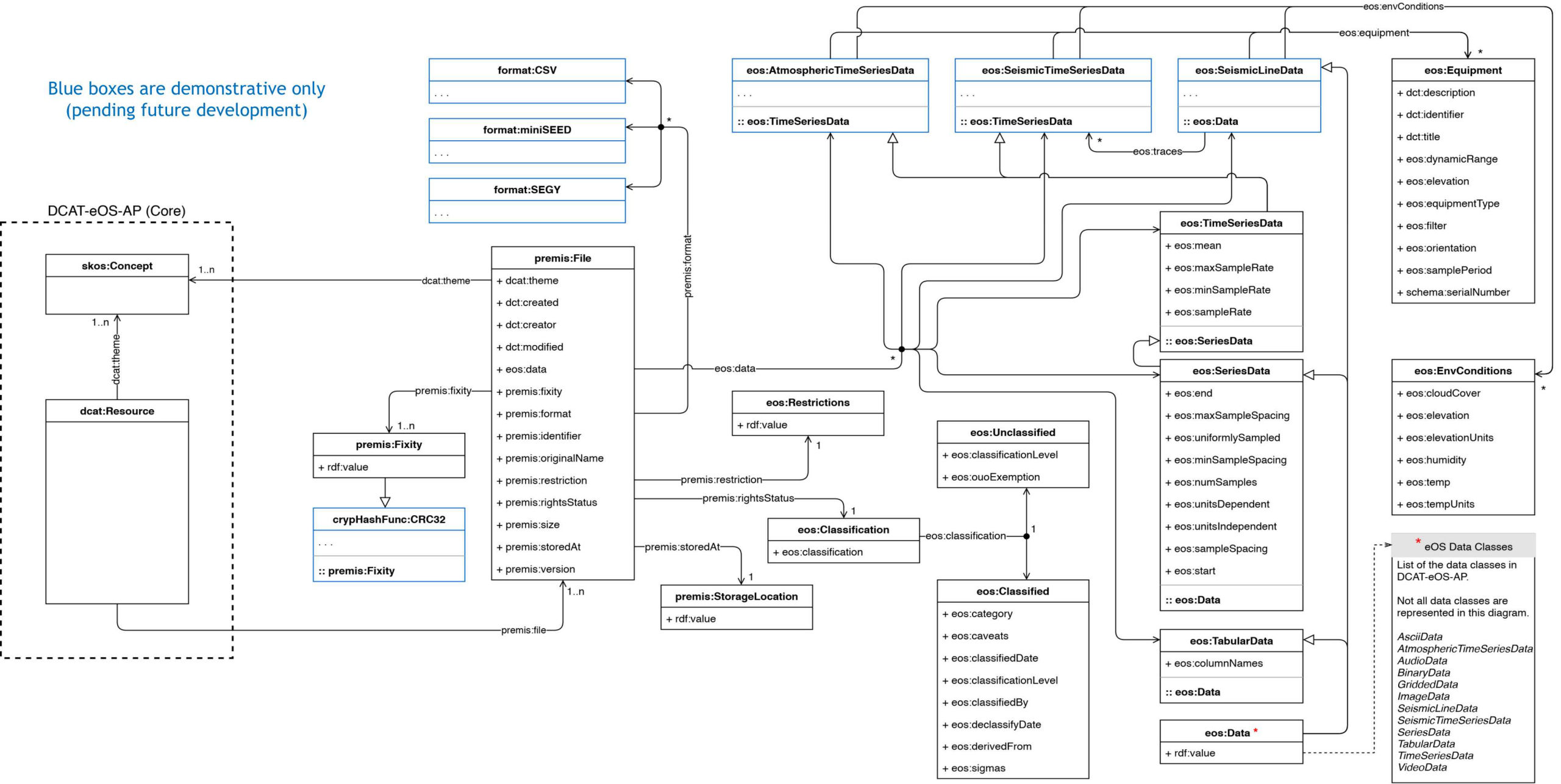
ConceptScheme

- A collection of related *Concepts* (e.g., seismology)



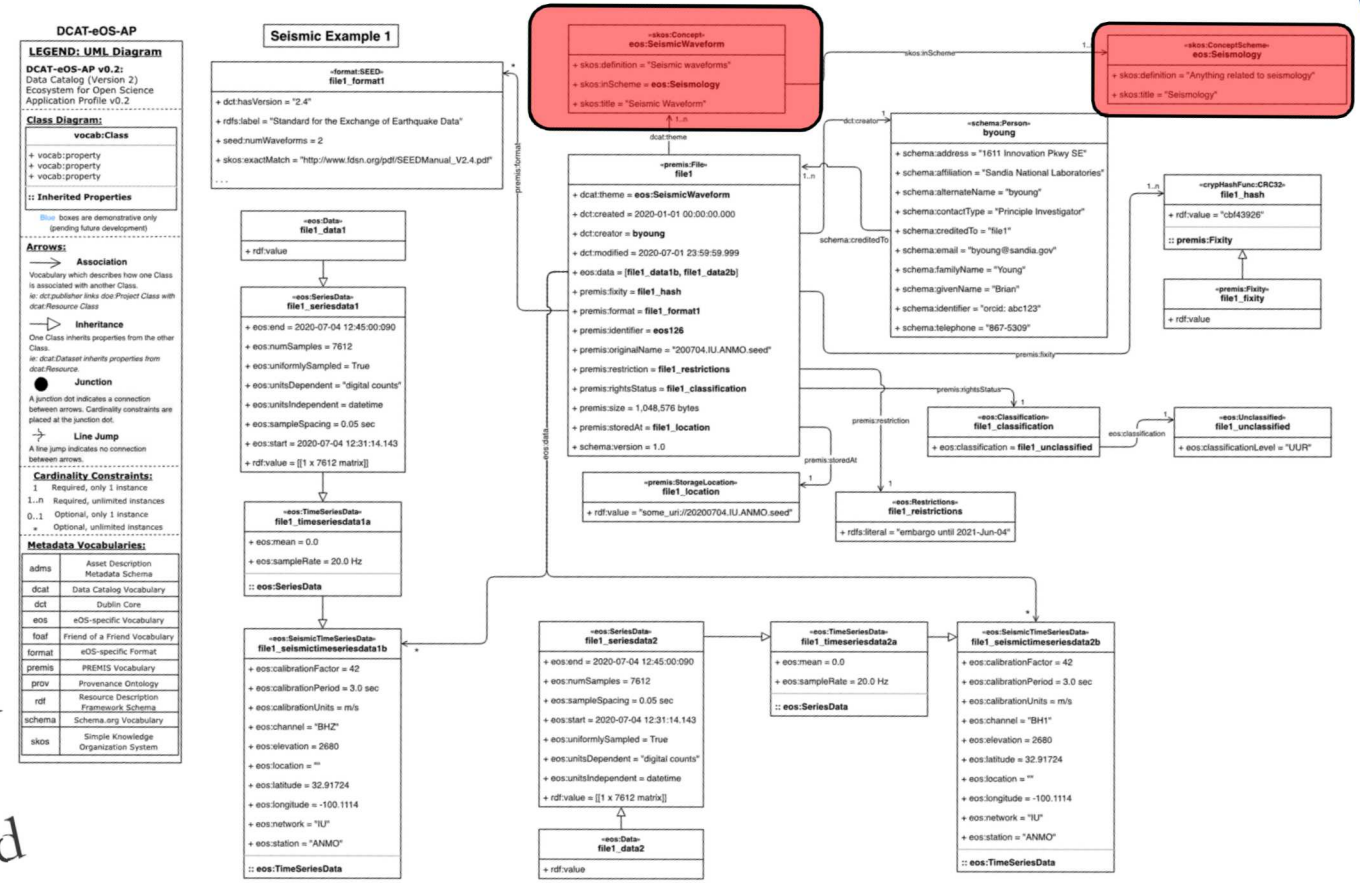
DCAT-eOS-AP Core + Granular Classes

Blue boxes are demonstrative only
(pending future development)



Summary

- DCAT-eOS-AP
  - **Extensible, modular, machine-readable** schema for domains relevant to nuclear monitoring
  - Domain specific granular-level metadata based on SME input
  - Scales with ever-increasing government data needs
    - Overall goal to gain widespread adoption within DNN R&D projects
    - Multiple phenomenologies represented within schema, anticipate extensible to other DOE mission areas



References

References

Office of Management and Budget (OMB), Office of Science and Technology Policy (OSTP), Executive Office of the President of the United States (EOP). (2020, March 10). *Project Open Data Metadata Schema v1.1*. Retrieved March 24, 2020, from Project Open Data: https://project-open-data.cio.gov/v1.1/schema/

World Wide Web Consortium (W3C). (2020, February 04). *Data Catalog Vocabulary (DCAT) - Version 2*, 20200204. Retrieved March 24, 2020, from World Wide Web Consortium: https://www.w3.org/TR/vocab-dcat-2/

Backup Slides

Metadata Standards Investigated

Dublin Core Metadata Initiative

- Originated in 1995 in Dublin, OH as an effort to describe digital resources, but also includes physical resources
- Many standards have roots in and/or are mapped to Dublin Core

Schema.org

- Joint effort focused on improving structured data markup schema on the web to make it easier to provide richer search results for users
- Metadata support for datasets added in 2018

Metadata Object Description Schema

- Maintained by Library of Congress, focus on bibliographic element sets

Project Open Data

- White House initiative to "unlock the potential of government data" by sharing open data (OMB & OSTP, 2020)
- Online public repository, collaborative open source project including collection of tools, code, case studies, etc.
- Schema based on DCAT, DCAT-US Schema v1.1

Data Catalog Vocabulary (DCAT)

- Hierarchical vocabulary (e.g., Resource Description Framework (RDF)) specific to datasets facilitating federated dataset search across catalogs

…And others