ODM2: Developing a Community Information Model and Supporting Software to Extend Interoperability of Sensor and Sample Based Earth Observations

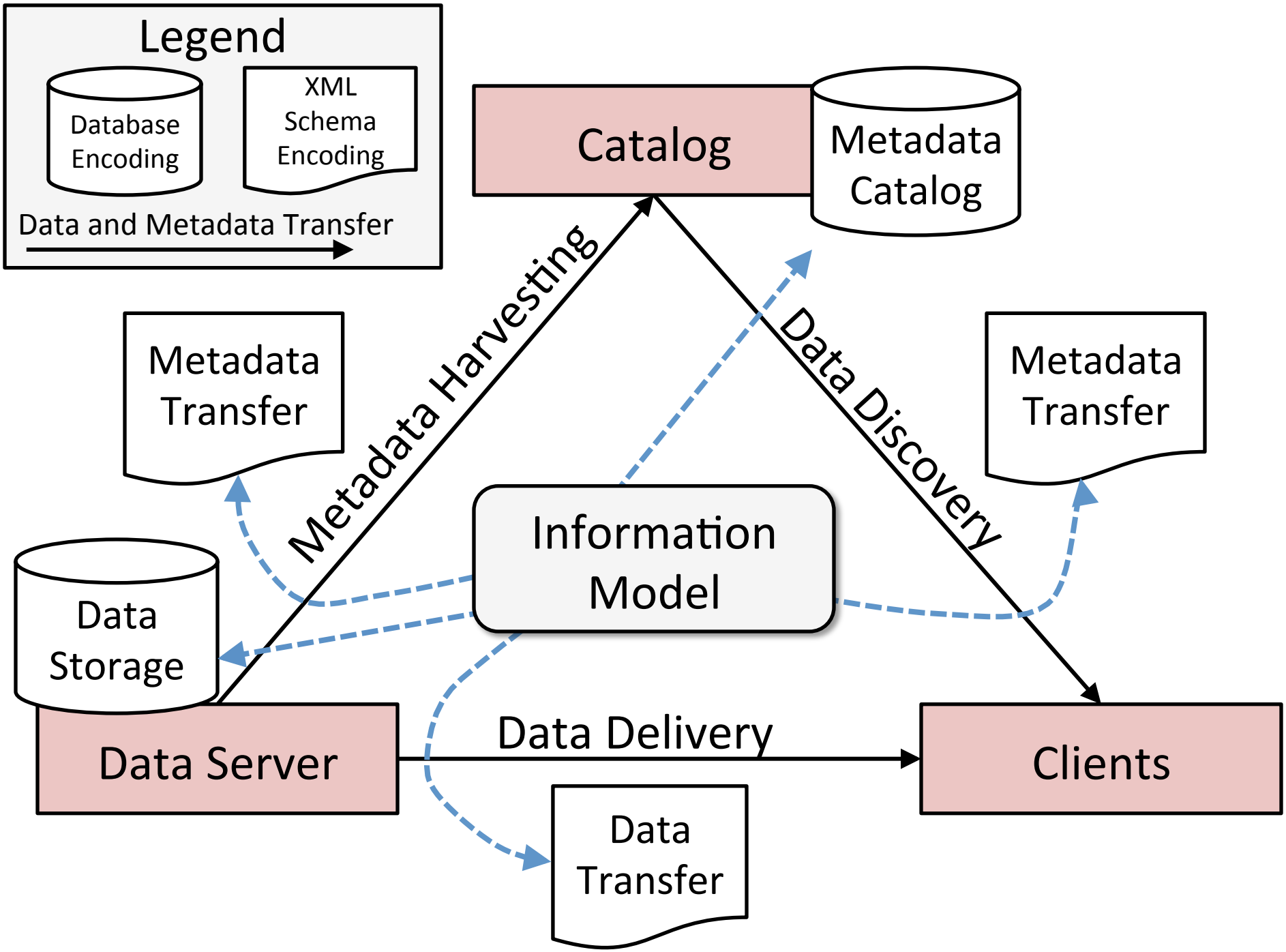
Jeffery S. Horsburgh

Anthony Aufdenkampe, Kerstin Lehnert, Emilio Mayorga,
David Tarboton, Ilya Zaslavsky

Support:
EAR 1224638

The ODM2 Team

- Utah State University
  - **Jeff Horsburgh, David Tarboton**
- Columbia U., Lamont-Doherty Earth Observatory and IEDA
  - **Kerstin Lehnert**
- Stroud Water Research Center
  - **Anthony Aufdenkampe**
- University of Washington, Applied Physics Laboratory
  - **Emilio Mayorga**
- San Diego Supercomputer Center
  - **Ilya Zaslavsky, David Valentine**

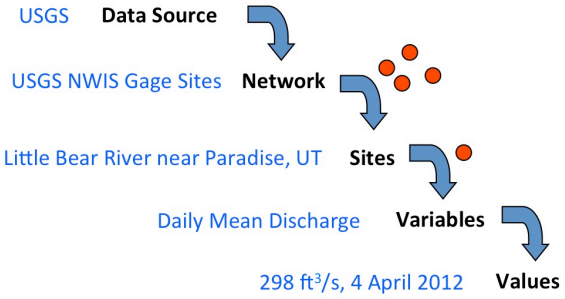

Legend
Database Encoding
XML Schema Encoding
Data and Metadata Transfer
Catalog
Metadata Catalog
Metadata Harvesting
Data Discovery
Metadata Transfer
Metadata Transfer
Information Model
Data Storage
Data Server
Data Delivery
Clients
Data Transfer

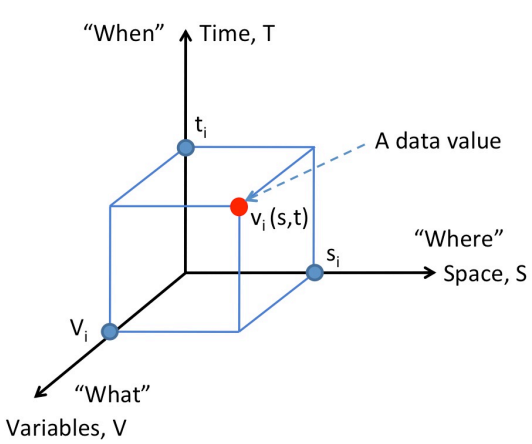
What is an "Information Model?"

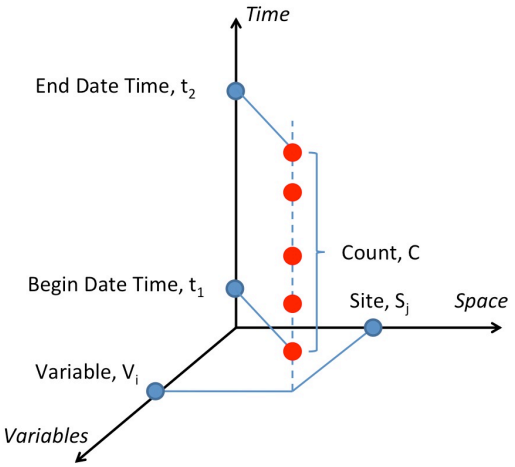
- Representation of concepts, relationships, constraints, rules, and operations that specify the semantics of data for a domain of discourse
  - Defines the domain's entity types and their attributes, relationships, and allowed operations on the entities
  - In a relational database implementation, entities become tables and their attributes become table columns
- Sharable, stable, and organized structure of information requirements for a domain context
  - Without constraining how that description is mapped to an actual implementation in software
  - **There may be many physical implementations** – e.g., relational database, XML schema, etc.

- Critically important to the effectiveness and interoperability of the cyberinfrastructure

Example Information Model

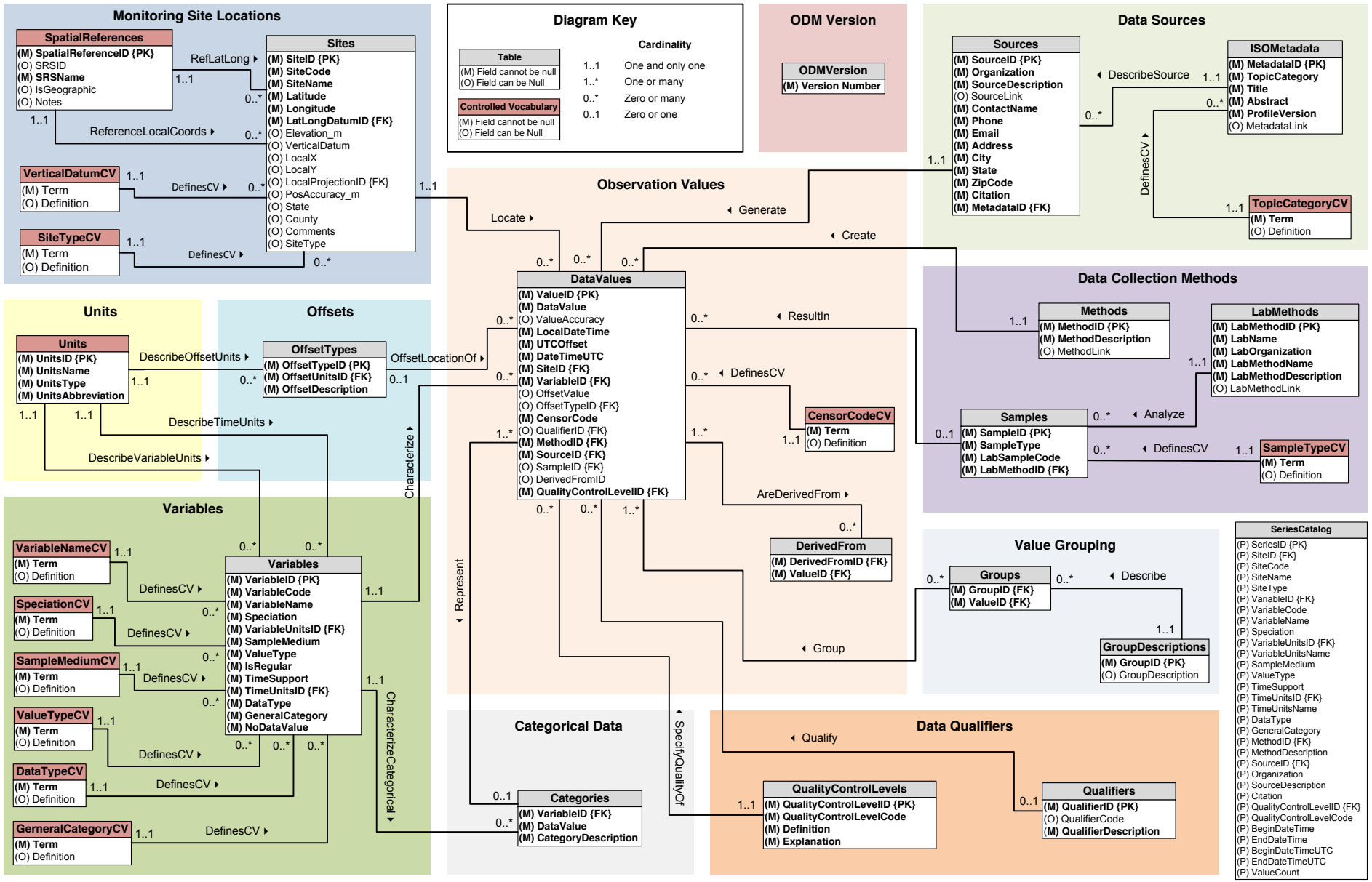






- An organization operates a network of monitoring sites
- At each monitoring site a number of variables are measured
- For each variable there is a time series of data values
- Each data series is made up of individual, time-indexed values that are each characterized by location (where the observation was collected), time (when the observation was collected), and variable (what the observation represents)

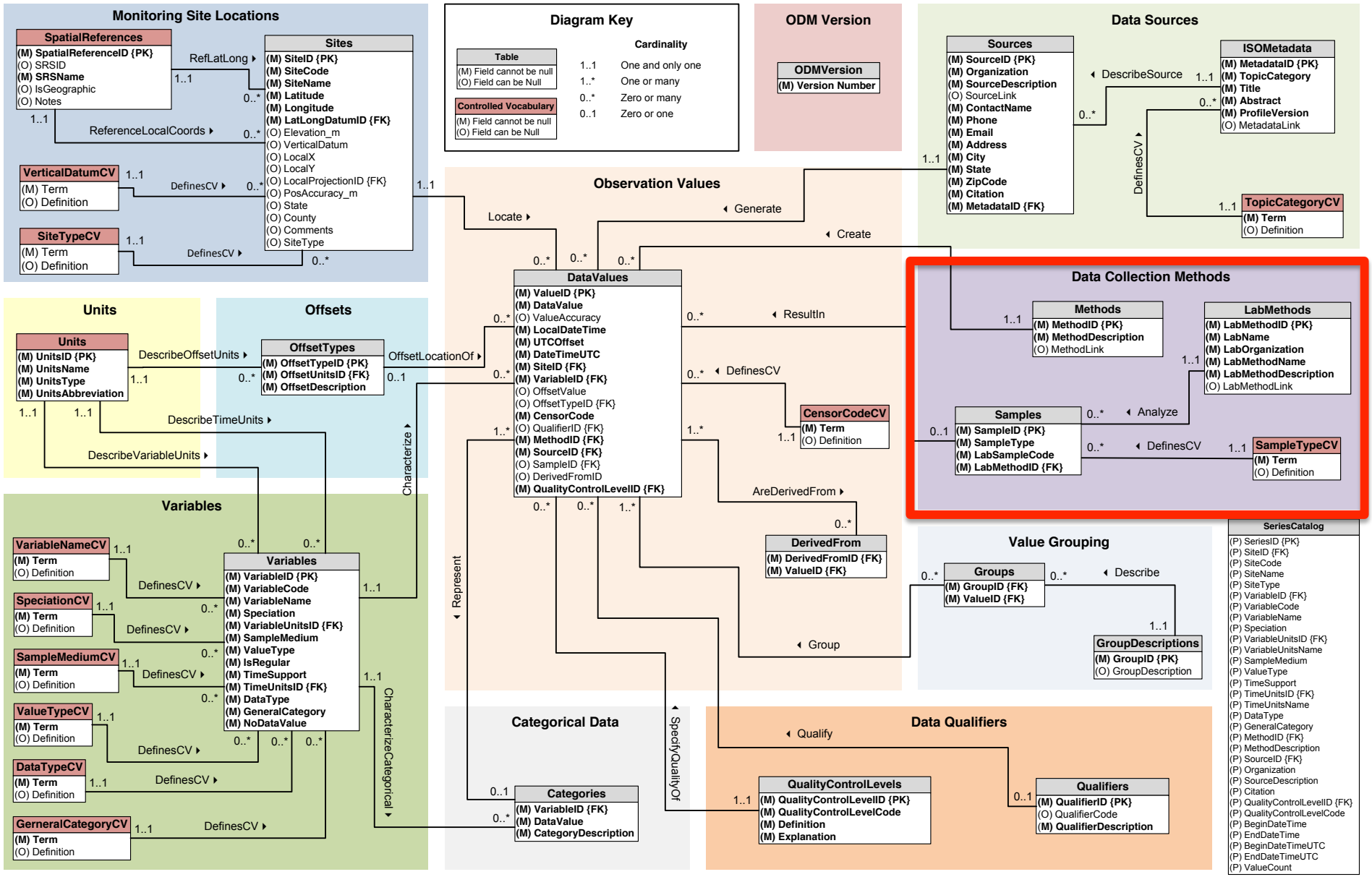
Example Physical Implementation: ODM 1.1.1



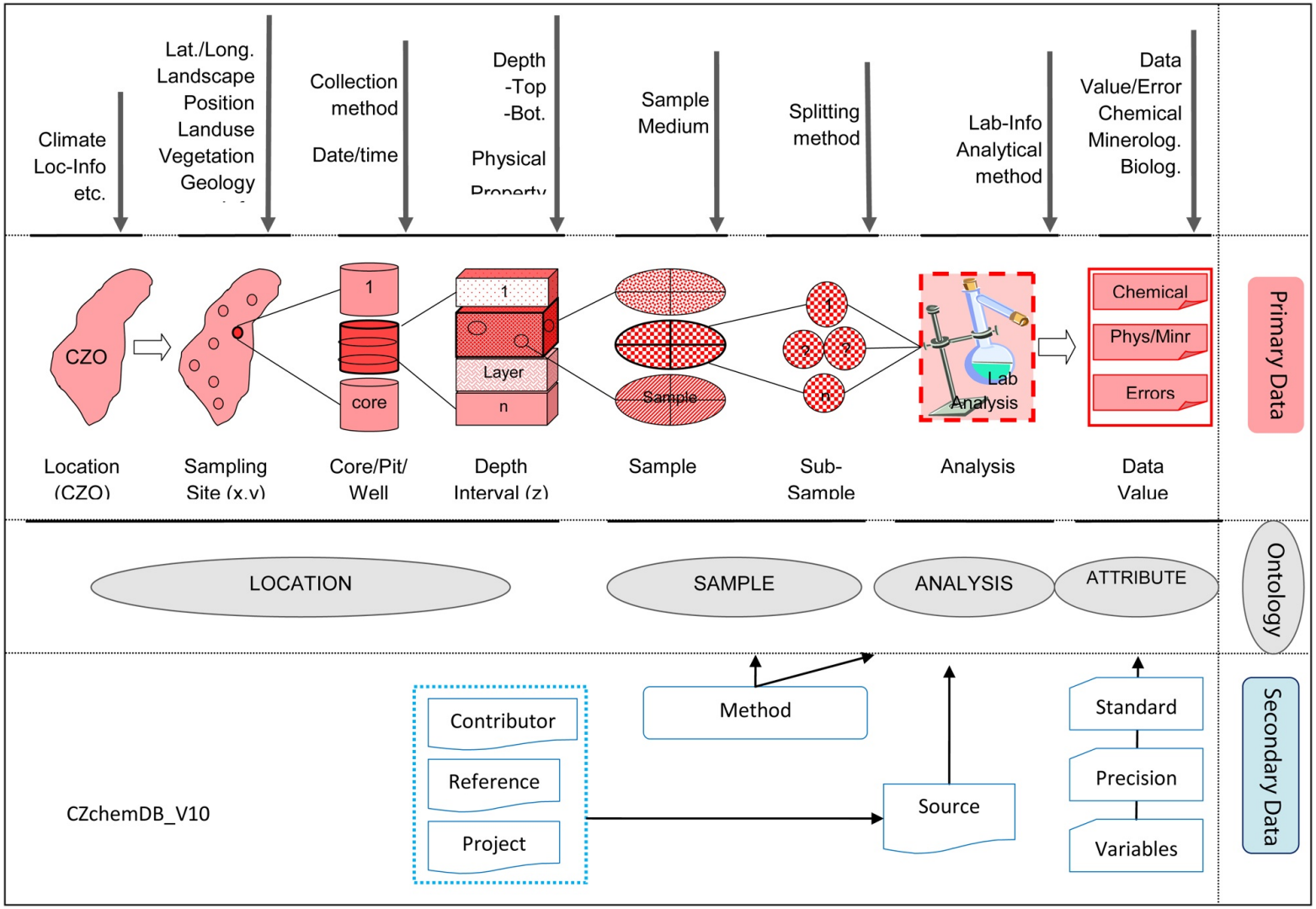
Some Limitations of ODM 1.1.1

- ODM supports only point-based observations
- ODM doesn't support sample-based data well
- Not all of the structure of ODM is required for each type of data
- Versioning and provenance of data can be difficult in ODM
- The central DataValues table doesn't support every functional use case (e.g., metadata catalog)

ODM 1.1.1: Sample Based Data

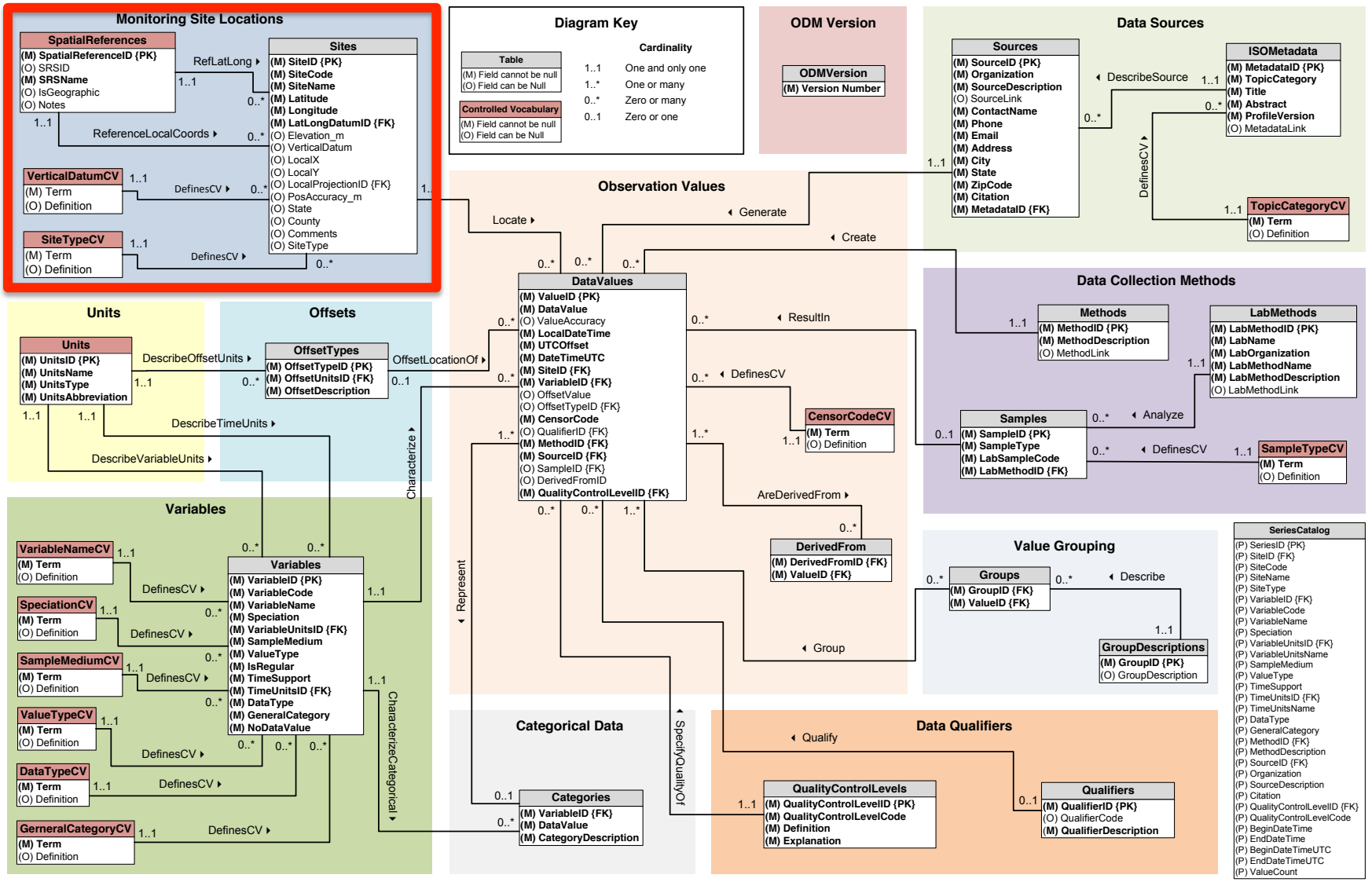


Another Example Information Model

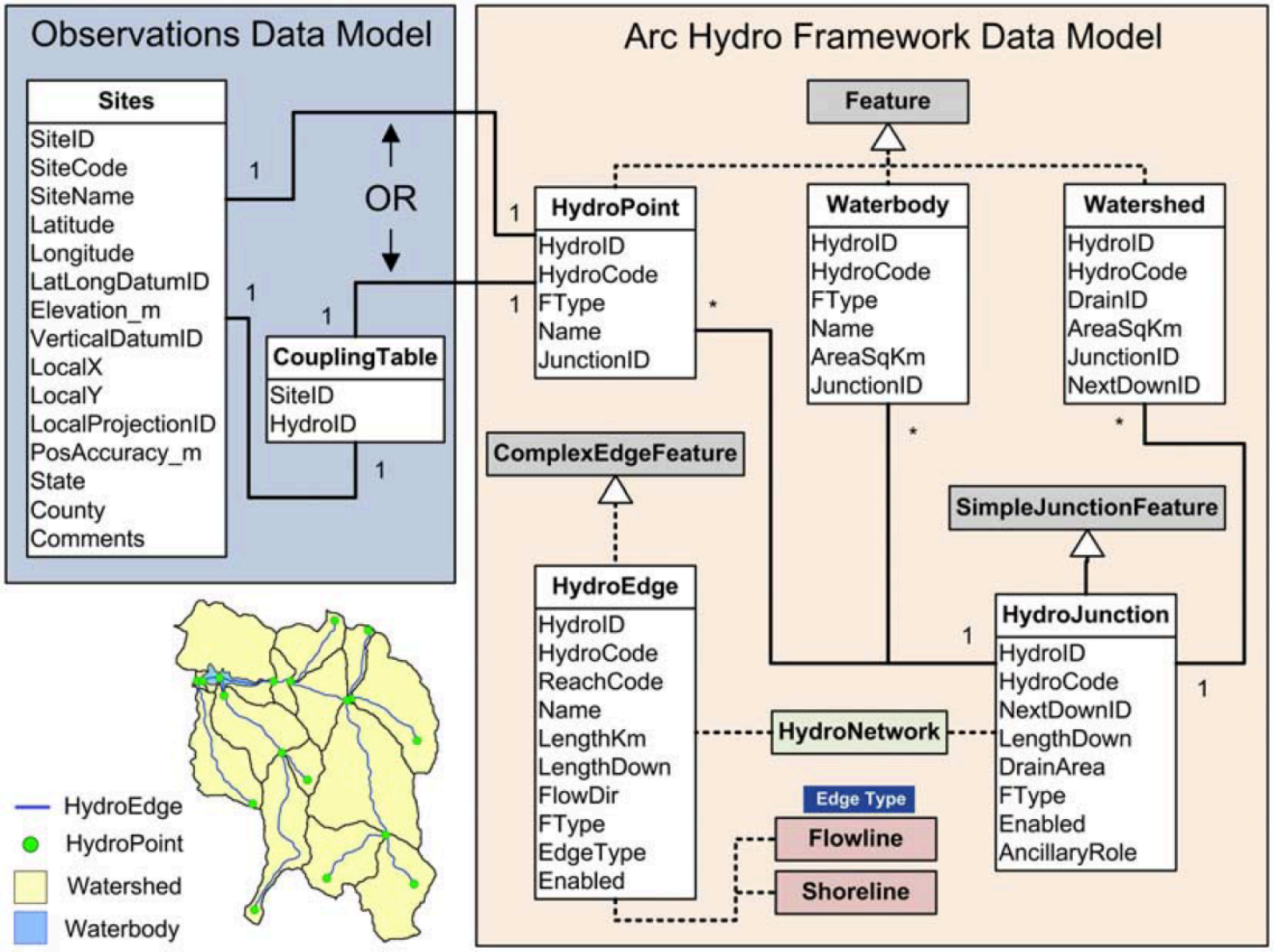


Xianzeng Niu, Jennifer Z. Williams, Doug Miller, Kerstin Lehnert, Brian Bills, Susan L. Brantley, (2013), An Ontology Driven Relational Geochemical Database for the Earth's Critical Zone: CZchemDB, Submitted to *Journal of Environmental Informatics*.

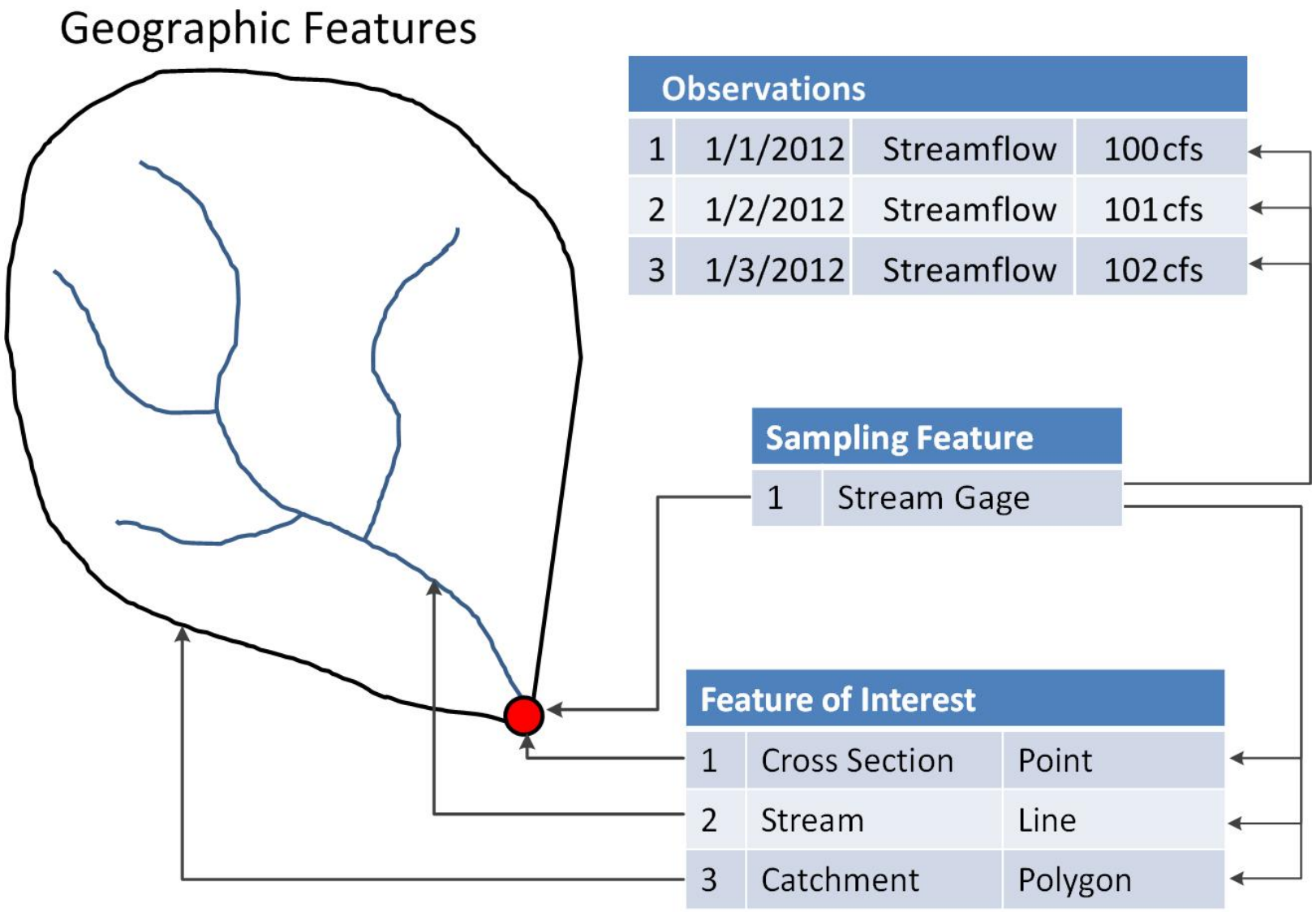
ODM 1.1.1: Linking Observations to the Geo-Environment



ODM 1.1.1: Linking Observations to the Geo-Environment



Linking Observations to the Geo-Environment



Overarching Goals for ODM2 Project

- Create an ***information model*** that is ***integrative*** and ***extensible***
  - Accommodating a wide range of observational data
  - Aimed at achieving interoperability across multiple disciplines and systems that support publication of earth observations
- Allow a diverse range of geoscience observations to be consistently shared, discovered, accessed, and interpreted

Project Objectives

1. Development of a community information model for spatially discrete, feature based earth observations
2. Engagement of geoscience communities in the design of the information model
3. Physical implementation of the information model in schemas for data storage, archival, transfer, and for cataloging metadata
4. Deployment of data publication prototypes using data from CZOData, CUAHSI HIS, EarthChem, and IOOS

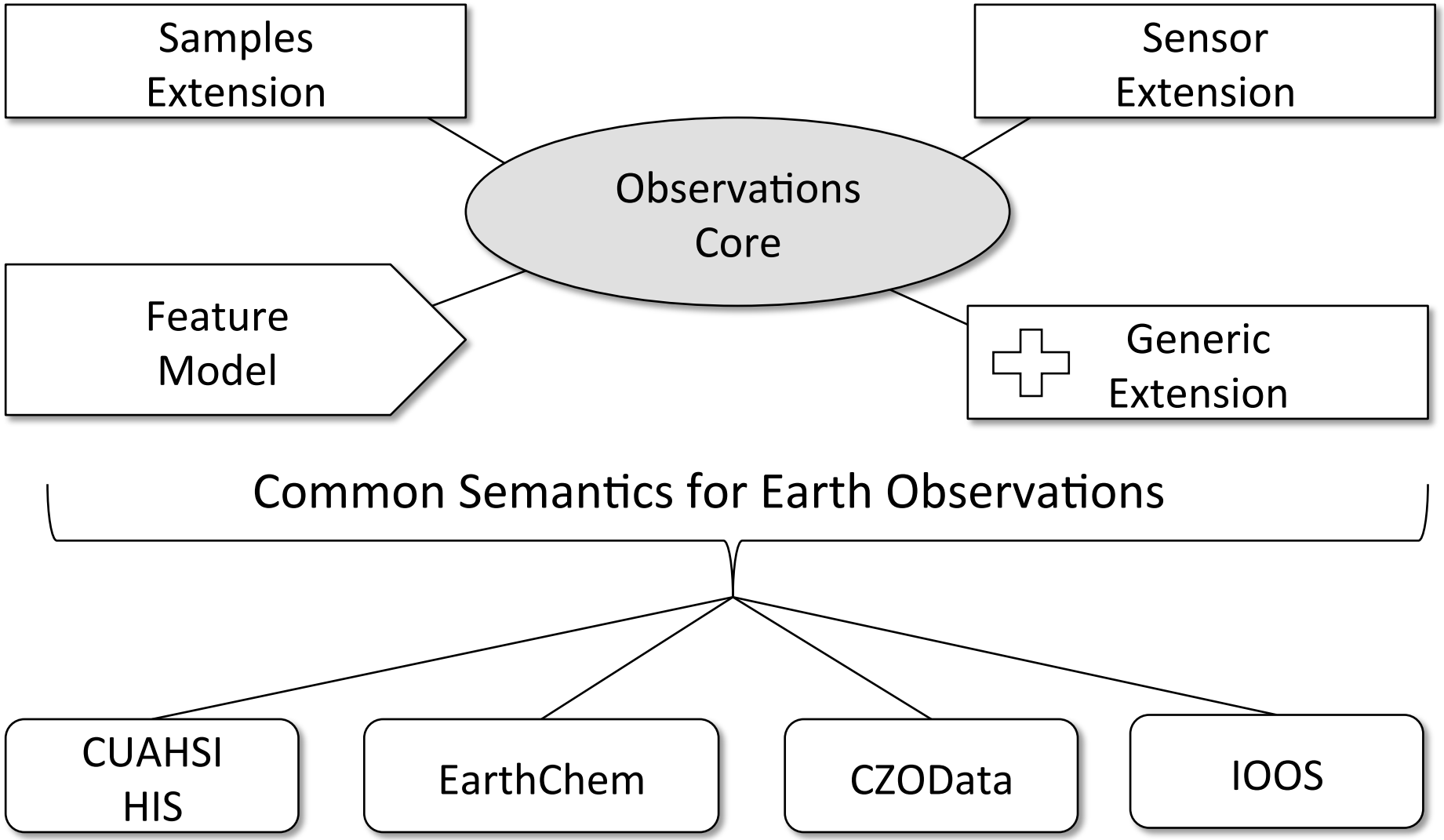
ODM2 Project Status

- August 1, 2012 – July 31, 2014
- Defined data use cases
  - Horsburgh – Hydrology/Water Quality Little Bear River (CUAHSI HIS)
  - Aufdenkampe - Christina River CZO
  - Lehnert – EarthChem ocean vent data
  - Mayorga – IOOS marine/coastal observations
- Reviewing existing information models and preliminary design work
- First community design workshop – March 2013

Implications of ODM2

- Better support for sample based data in CUAHSI HIS and related tools
- Support for and integration of sensor and sample-based data in the Critical Zone Observatory Integrated Data Management System (CZOData)
- Potential adoption of ODM within EarthChem
- Providing context and guidance for future WaterML 2.0 development (water quality samples)

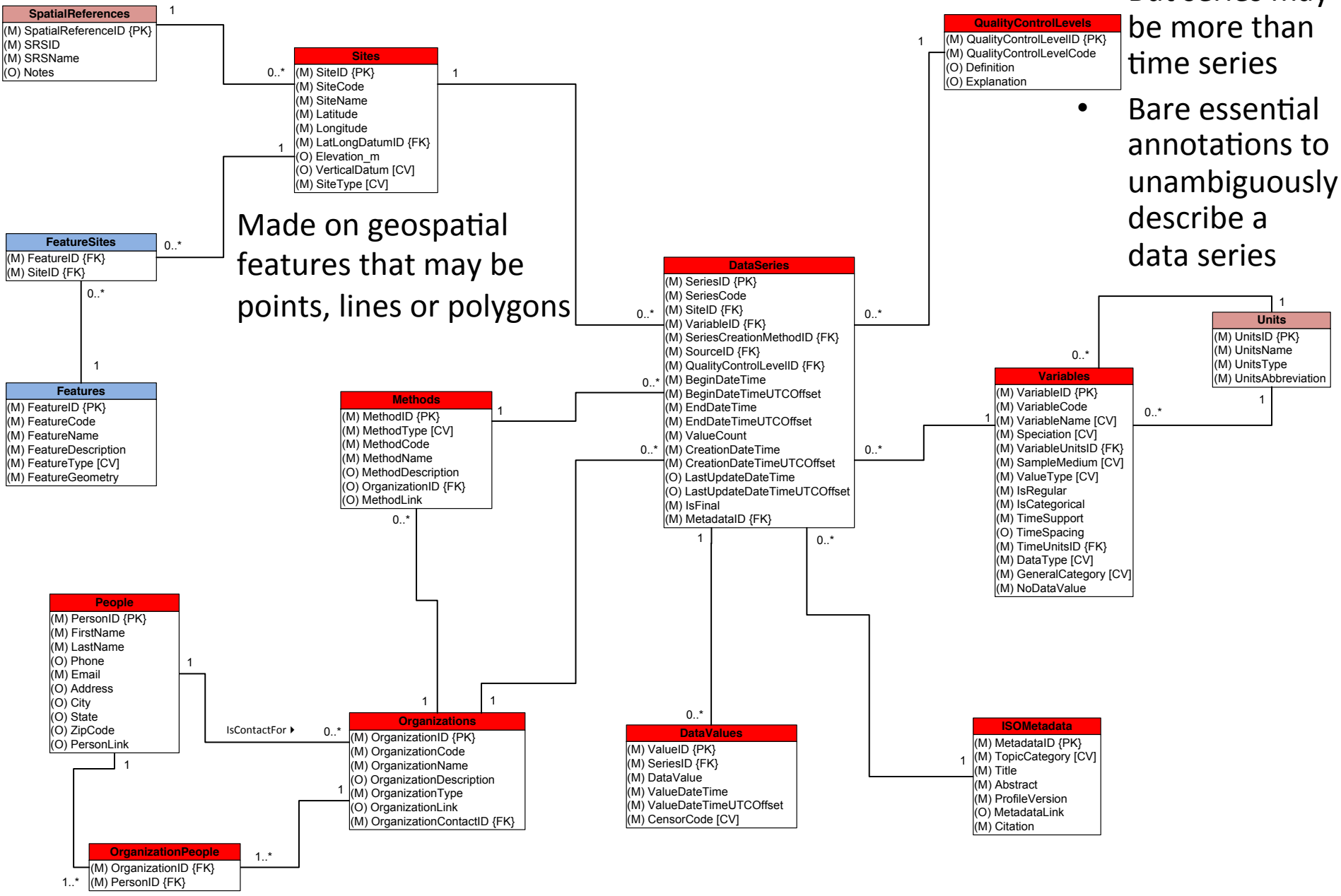
Approach: ODM2 Information Model



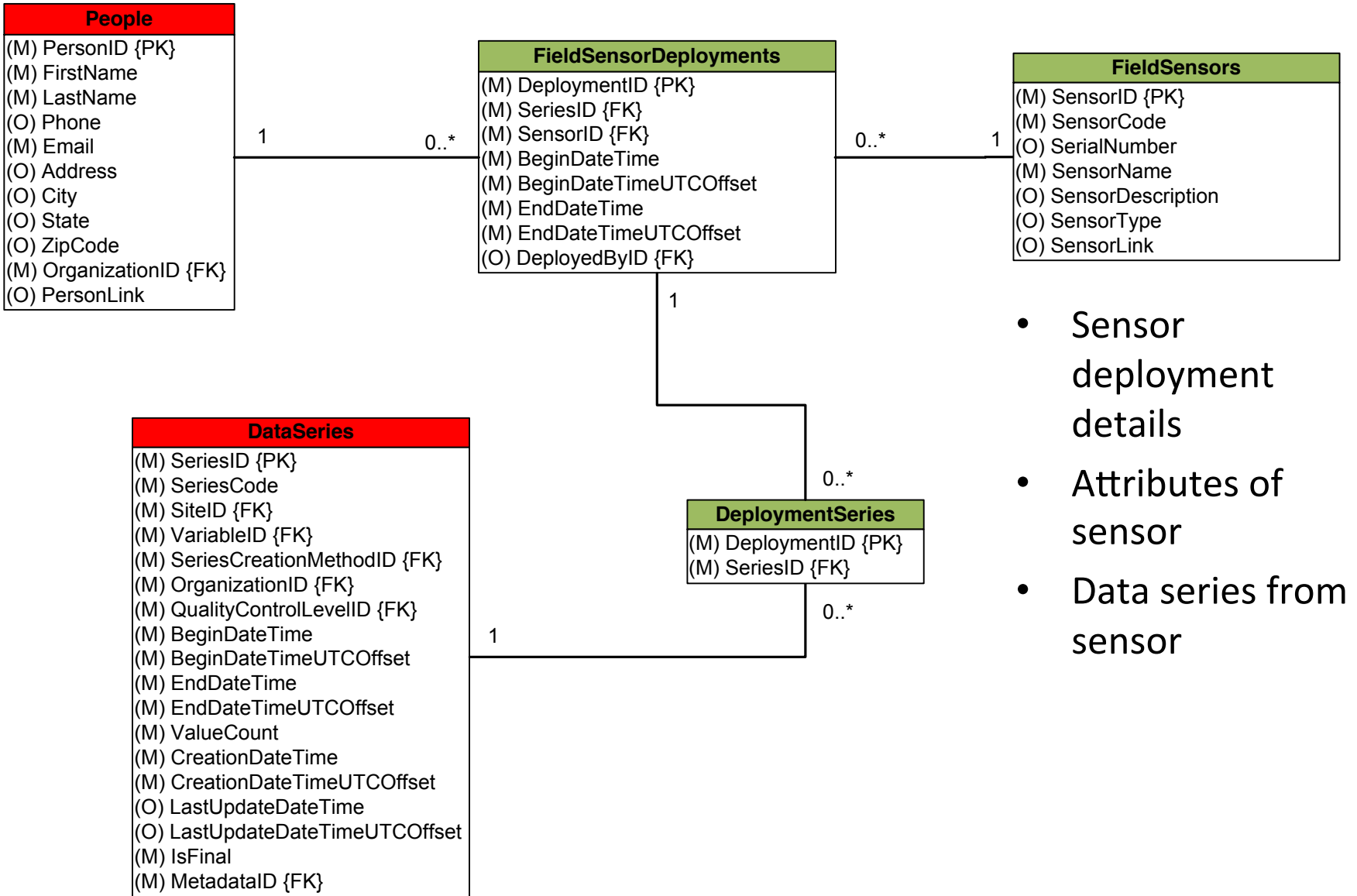
Domain Cyberinfrastructures

ODM2 - Core

- Series centric
- But series may be more than time series
- Bare essential annotations to unambiguously describe a data series

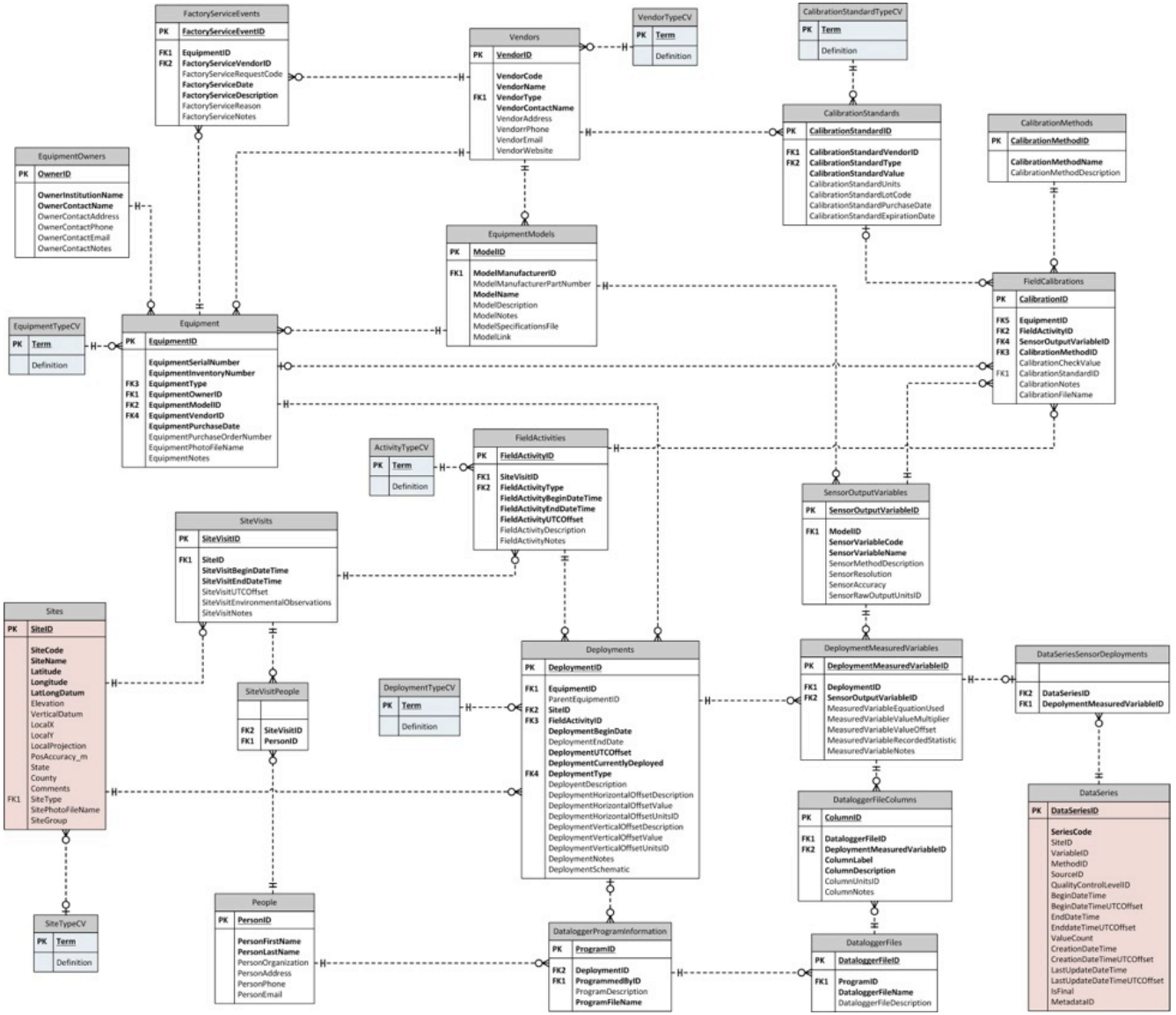


ODM2 Sensor Extension to support field sensor deployments and *in situ* observations

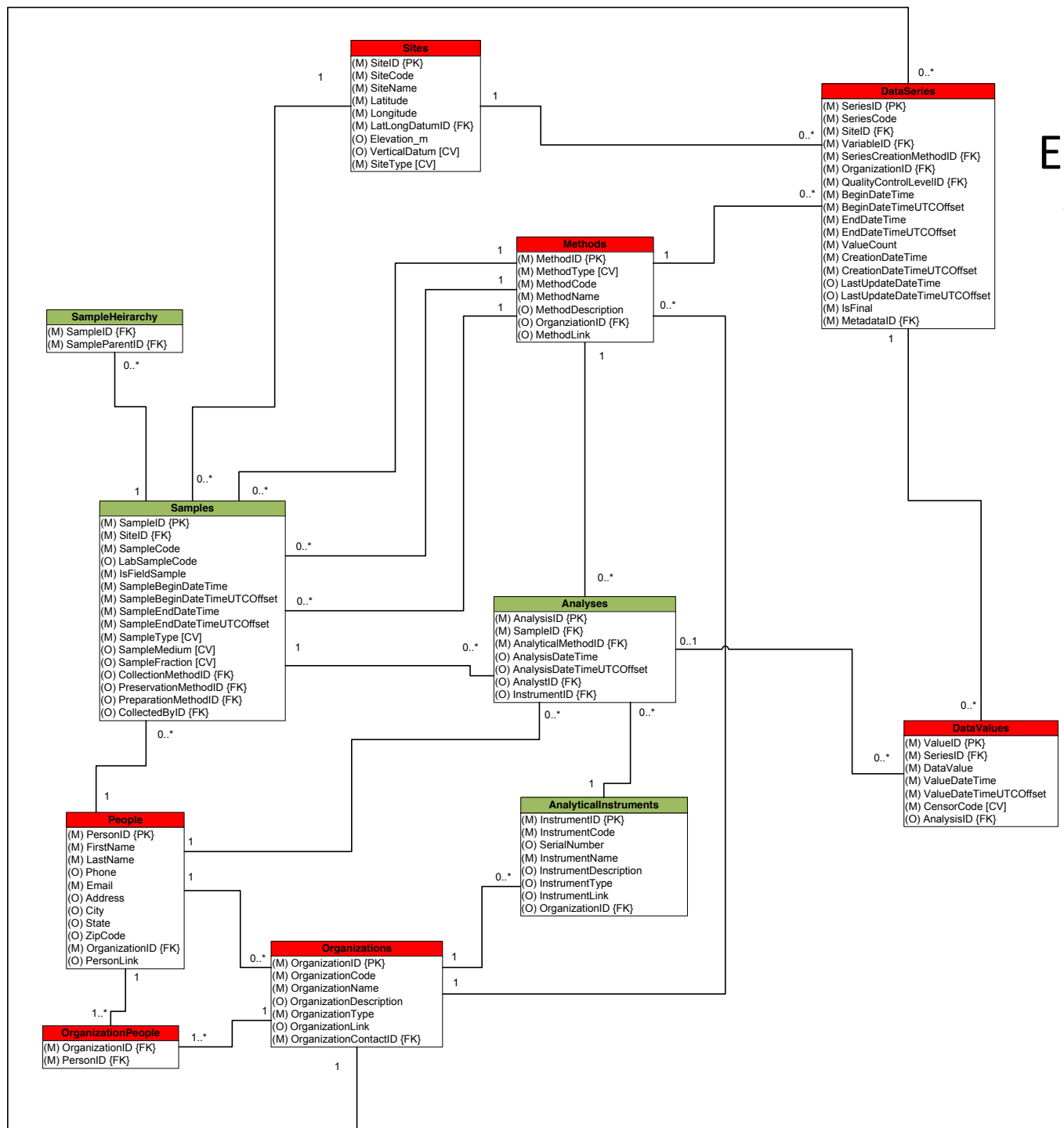


- Sensor deployment details
- Attributes of sensor
- Data series from sensor

ODM2 Sensors Extension



- Track physical infrastructure:
  - sensors
  - data loggers
  - batteries
- Track events:
  - deployments
  - calibrations
  - site visits
  - factory service

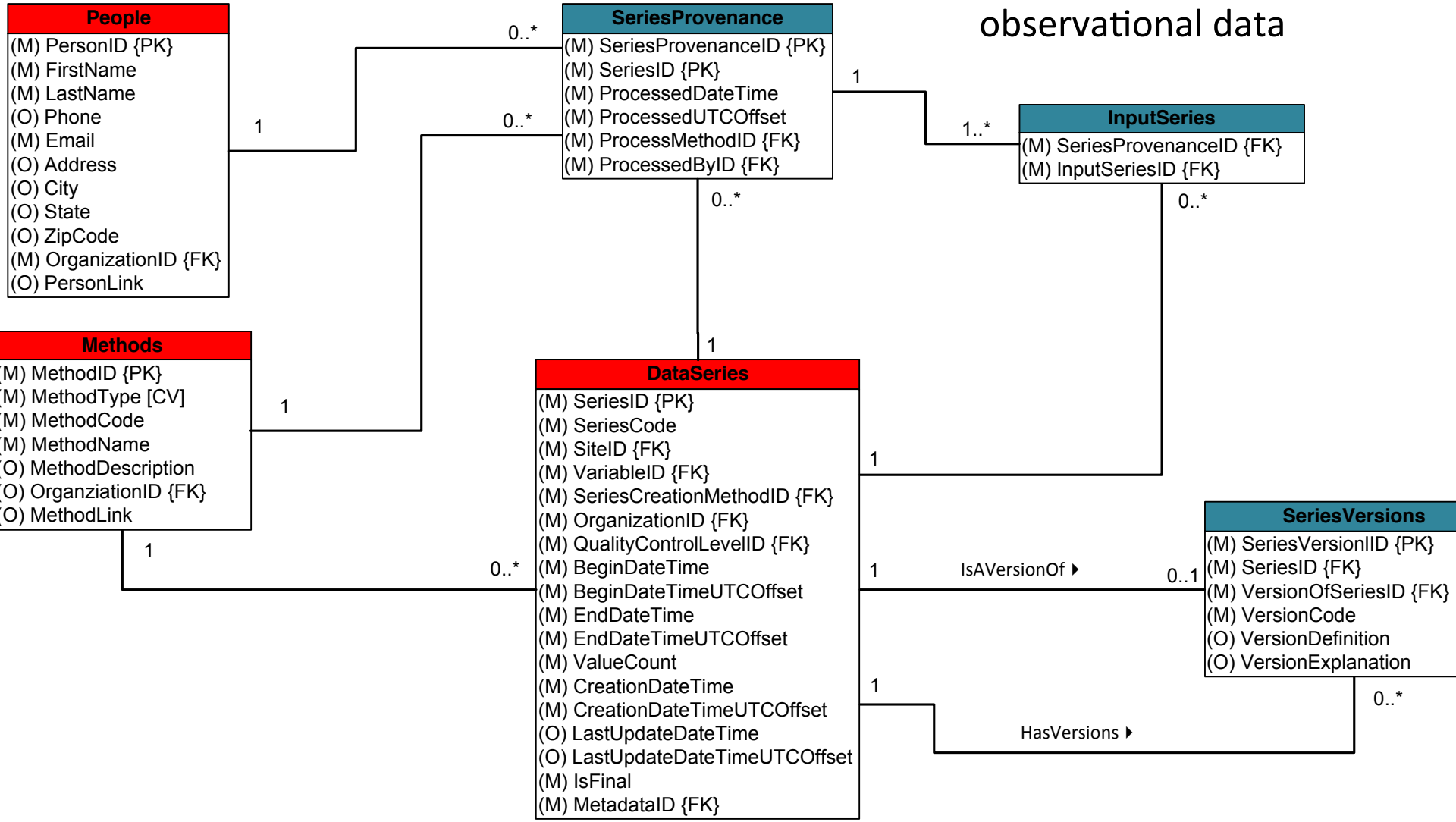


ODM2 Samples Extension to handle field samples and *ex situ* analyses

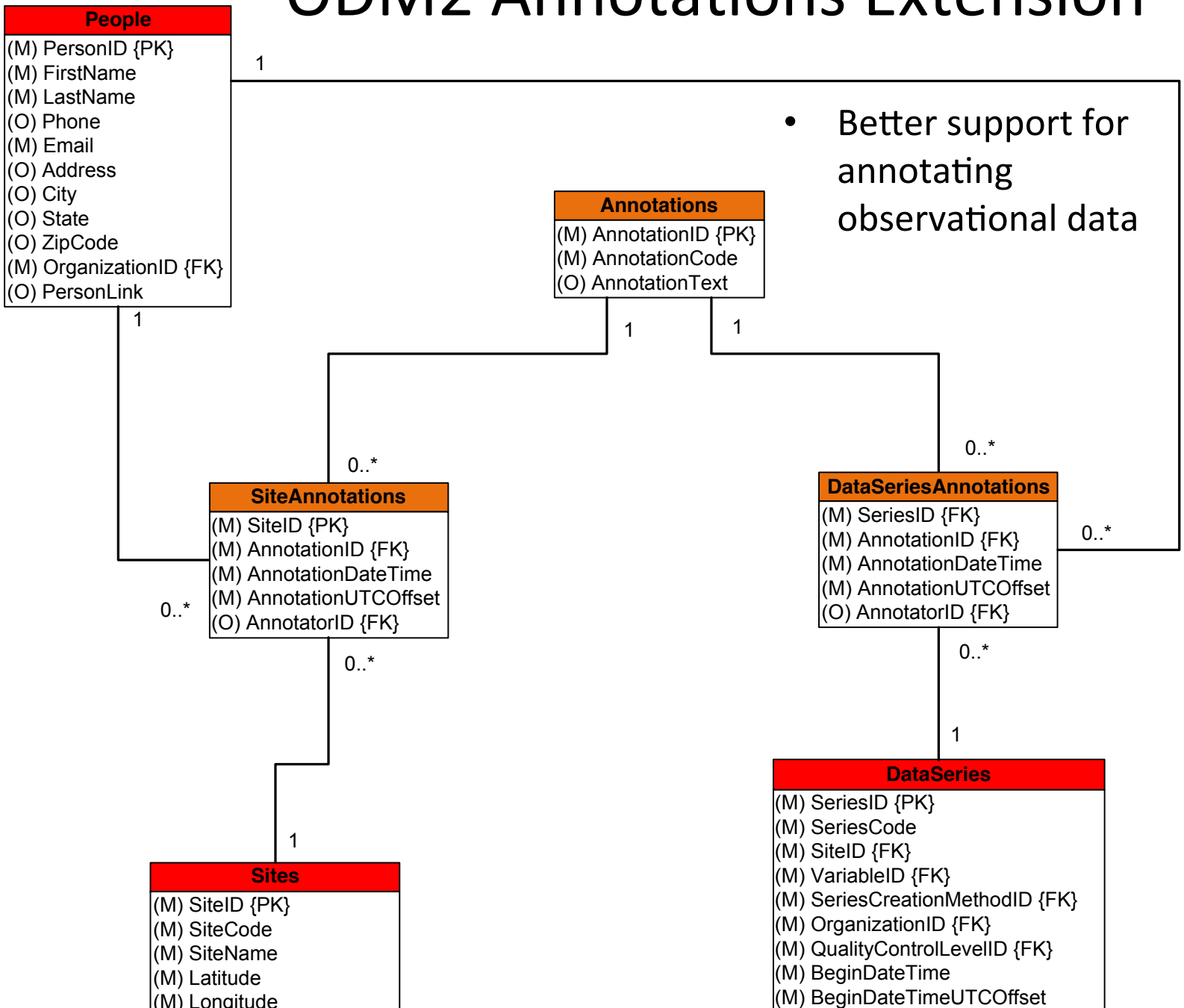
- Attributes of individual samples
- Relationships between samples and parents
- Attributes of a sample analysis
- Attributes of analytical instruments used in sample analysis

ODM2 Data Series Provenance Extension

- Better support for storing provenance of observational data



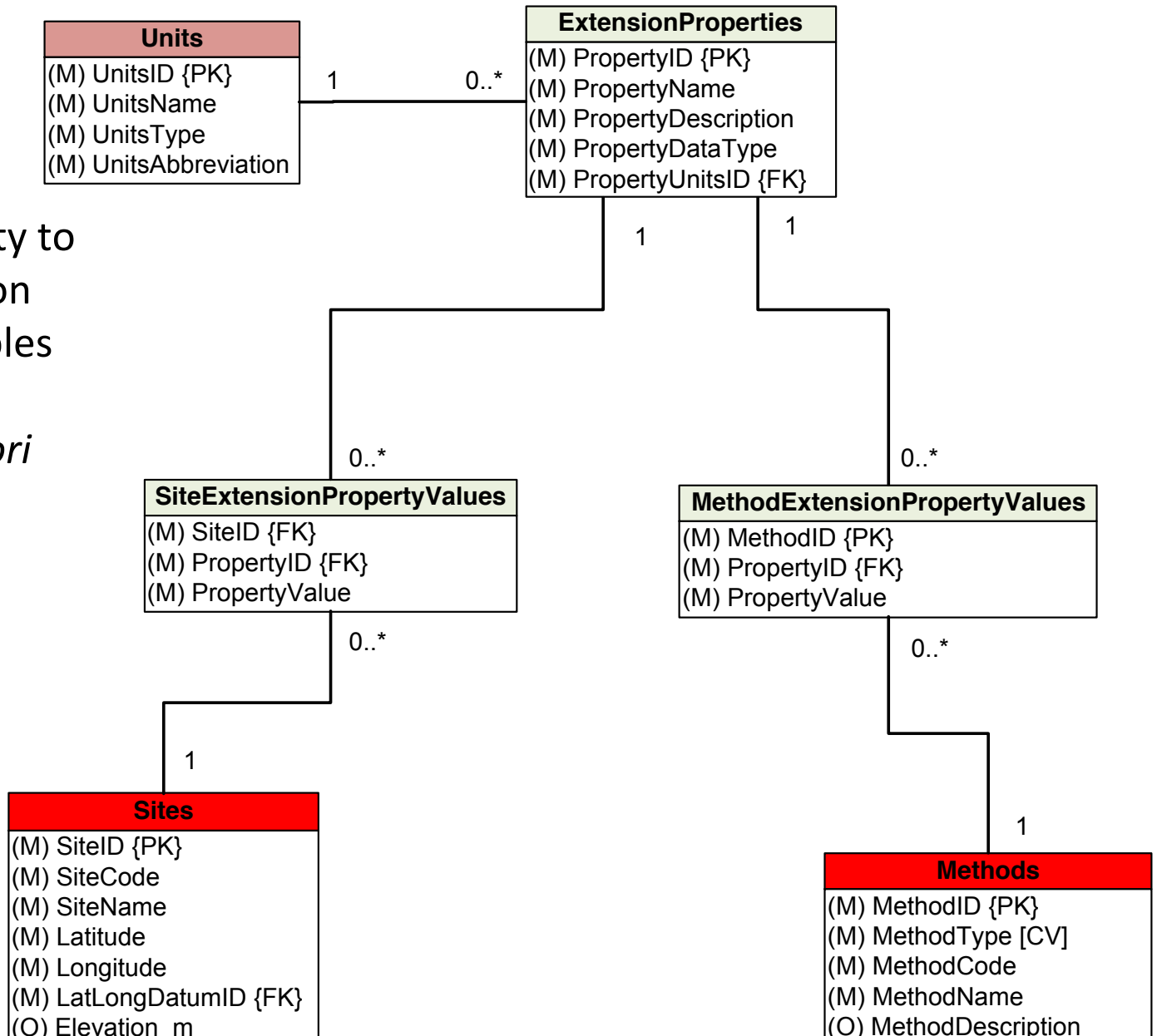
ODM2 Annotations Extension



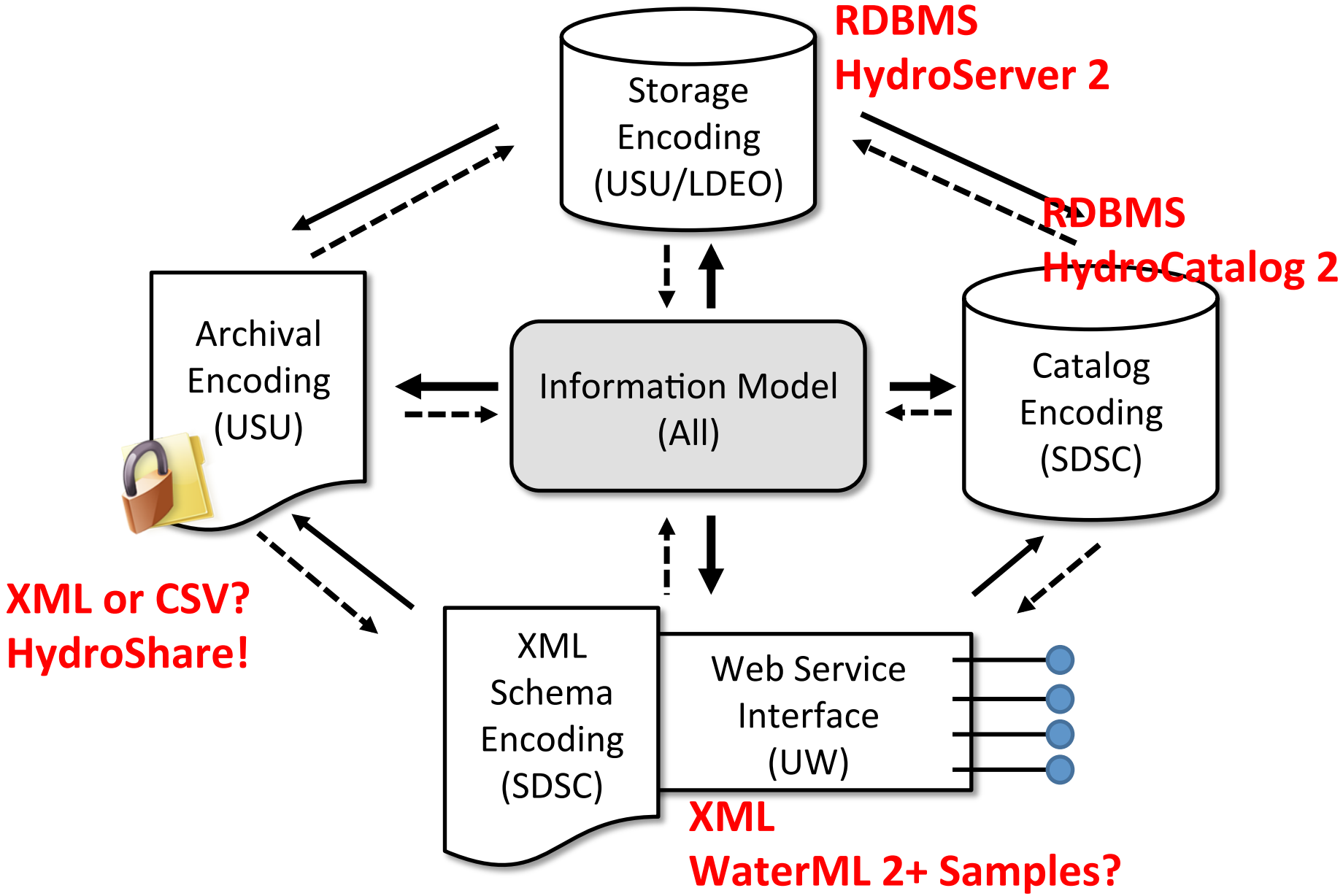
- Better support for annotating observational data

ODM2 General Extensibility

- Provides capability to record information (add fields) in tables that was not anticipated *a priori*



ODM2 Functional Use Cases



ODM2 Archival Encoding Requirements

- Must be self contained (no external pointers)
- File format must not become obsolete
- Non-proprietary
- Widely used
- Files opened and shared without special software or hardware
- Time series + other observational data (e.g., samples)

MyDB as an Archival Format?

- ✓ Must be self contained (no external pointers)
- ✓ Format must not become obsolete
- ✓ Non-proprietary
- ✓ Widely used
- ✓ Opened and shared without special software or hardware
- ✓ Time series + other observational data (e.g., samples)

- MyDB and ODM 1.1.1 share the same information model (**almost**)
- But, would need MyDB2 for ODM2
- CSV may not easily capture ODM2 extensibility mechanisms

WaterML 1.1 as an Archive Format?

- ✓ Must be self contained (no external pointers)
- ✓ Format must not become obsolete
- ✓ Non-proprietary
- ✓ Widely used
- ✓ Opened and shared without special software or hardware
- ✓ Time series + other observational data (e.g., samples)

- WaterML 1.1 and ODM 1.1.1 share the same information model (**almost**)
- But, would need WaterML for ODM2

WaterML 2 as an Archive Format?

- Must be self contained (no external pointers)
- ✓ Format must not become obsolete
- ✓ Non-proprietary
- ✓ Widely used
- ✓ Opened and shared without special software or hardware
- Time series + other observational data (e.g., samples)

- WaterML 2 is an exchange format, not an archival format
  - These are different functional use cases
- Future work: specification of an ex-situ profile
  - WaterML 2 currently focuses only on in-situ time series

ODM2 Archival Format

- XML likely a more appropriate encoding because of planned extensibility
  - Annotations
  - Provenance
  - General attribute extensibility
  - ODM2 extensions – ***not all data stored in ODM2 will be time series***

Proposal

- Initially accept MyDB OR WaterML 1.1 as the archival format for time series data
  - Accept both as input, but internally convert to one or the other so all data are stored consistently inside HydroShare
- When the ODM2 archival format becomes available – use it as the internal archival format for HydroShare
- Use the ODM2 archival specification to extend HydroShare archival support for sample-based data