The Extended Metadata Registry (XMDR) Project and Water Resource Modeling

John McCarthy and Bruce Bargmeyer
Lawrence Berkeley National Laboratory

Computational Methods for Water Resources
Panel on Ecoinformatics for Water Resource Modeling

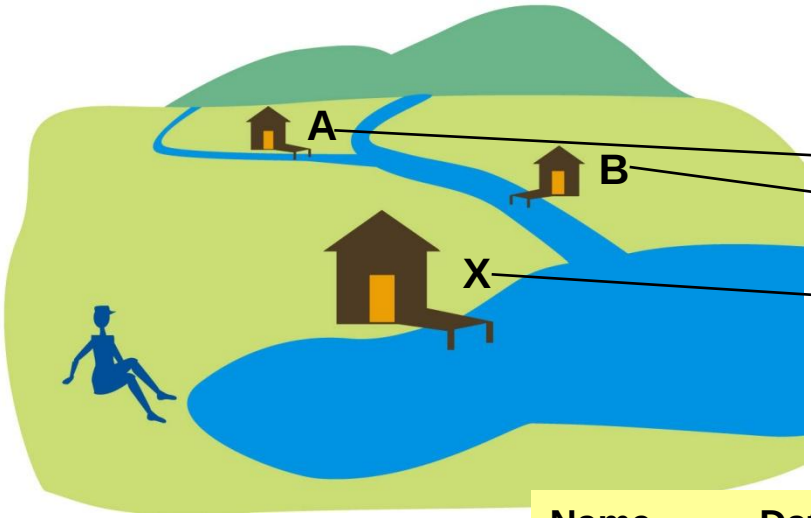
8 July, 2008

Westin San Francisco Market Street Hotel
San Francisco, California

XMDR Presentation Outline

- What is (are?) metadata & why is it important?
- Metadata for water research & policy
  - current projects, technology & issues
- What are metadata registries?
  - why are metadata registry standards necessary?
  - what are current registry strengths & weaknesses?
- XMDR Project
  - motivation & goals – adding more formal semantics
  - XMDR Prototype System – architecture & test contents
- Cancer Data Standards Repository example
- Possible uses of metadata registries for water?

Data, Metadata and Ancillary Data



Data

ID	Date	Temp	Hg
A	06-09-13	4.4	4
B	06-09-13	9.3	2
X	06-09-13	6.7	78

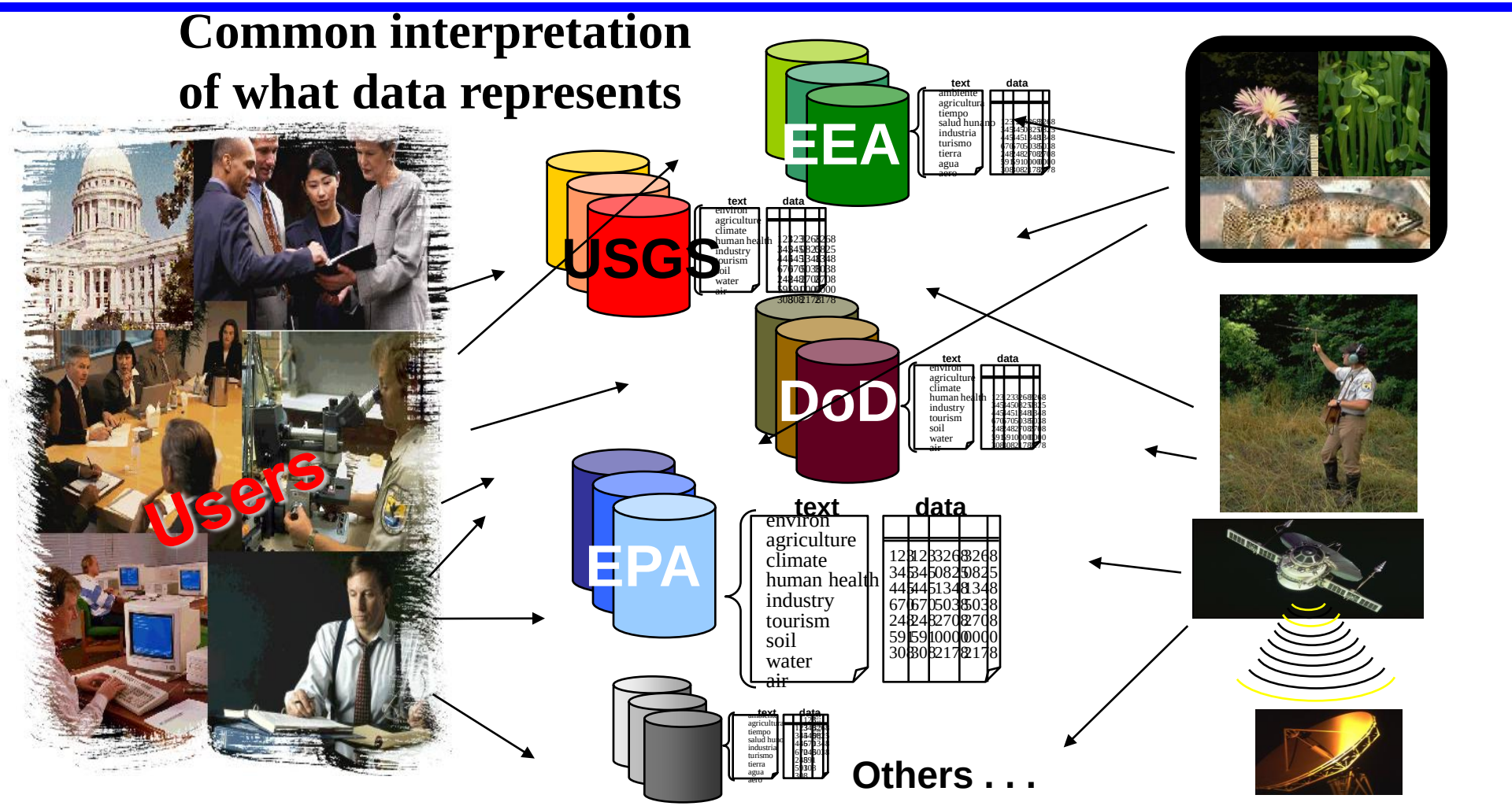
Metadata

Name	Datatype	Definition	Units
ID	text	Monitoring Station Identifier	not applicable
Date	date	Date	yy-mm-dd
Temp	number	Temperature (to 0.1 degree C)	degrees Celsius
Hg	number	Mercury contamination	micrograms per liter

Ancillary Data

Site Code	Site Name	Latitude	Longitude	Elevation	State	County
A	BEAR RIVER BLW SMITHS FORK, NR COKEVILLE, WY	42.1266021	-110.973243	1872	WY	Lincoln
B	BEAR RIVER AT SODA SPRINGS, ID	42.6138103	-111.583556	1756.1	ID	Caribou

Metadata can help achieve common understanding of meaning between Data Creators and Data Users



What is the current state of metadata technology for water data?

- **EPA STORET (STOrage and RETrieval) epa.gov/storet/**
  - **Characteristics Group Summaries & detail by Source Organization**
- **USGS NWIS National Water Information System nwis.waterdata.usgs.gov**
- **European Environmental Agency**
  - **Water Data Centre & Water Information System for Europe (WISE)**
- **CUAHSI - Consortium of Universities for the Advancement of Hydrologic Science, Inc. cuahsi.org (NSF)**
- **CUAHSI Hydrologic Information System (CUAHSI-HIS) his.cuahsi.org**
- **Observations Data Model (ODM) water.usu.edu/cuahsi/ODM & "Controlled Vocabularies"**

12 CUAHSI Controlled Vocabularies for Observation Data Model (ODM)

- **CensorCodeCV:** Used to populate the CensorCode field of the DataValues table
- **DataTypeCV:** Used to populate the DataType field of the Variables table
- **GeneralCategoryCV:** Used to populate the GeneralCategory field in the Variables table
- **QualityControlLevels:** Defines QualityControl LevelID used in the DataValues table
- **SampleMediumCV:** Used to populate the SampleMedium field in the Variables table
- **SampleTypeCV:** Used to populate the SampleType field in the Samples table
- **SpatialReferences:** Defines the coordinate systems used in the Sites table
- **TopicCategoryCV:** Used to populate the TopicCategory field in the ISOMetadata table
- **Units:** Defines the units used in the Variables and Offset types tables
- **ValueTypeCV:** Used to populate the ValueType field in the Variables table
- **VariableNameCV:** Used to populate the VariableName field in the Variables table
- **VerticalDatumCV:** Used to populate the VerticalDatum field im the Sites table

221 ODM Variables are listed in the Variable Name Controlled Vocabulary

ODM — A COMMUNITY DATA MODEL FOR HYDROLOGIC OBSERVATIONS

CUAHSI universities allied for water research

UtahState UNIVERSITY

Home | ODM Downloads | Comments | Controlled Vocabularies

VariableNameCV Values Back to CV Page VariableNameCV

Used to populate the VariableName field in the Variables table

New	Term	Definition
Edit	19-Hexanoyloxyfucoxanthin	The phytoplankton pigment 19-Hexanoyloxyfucoxanthin
Edit	9 cis-Neoxanthin	The phytoplankton pigment 9 cis-Neoxanthin
Edit	Acid neutralizing capacity as CaCO3	Acid neutralizing capacity as calcium carbonate
Edit	Agency code	Code for the agency which analyzed the sample
Edit	Alkalinity, bicarbonate as CaCO3	Bicarbonate Alkalinity as calcium carbonate
Edit	Alkalinity, carbonate as CaCO3	Carbonate Alkalinity as calcium carbonate
Edit	Alkalinity, hydroxide as CaCO3	Hydroxide Alkalinity as calcium carbonate
Edit	Alkalinity, total as CaCO3	Total Alkalinity as calcium carbonate
Edit	Alloxanthin	The phytoplankton pigment Alloxanthin

http://water.usu.edu/cuahsi/ODM/downloads/

ODM CUASHI Controlled Vocabularies & Definitions Evolve via Change Form

VariableNameCV change request

Edit values below

Term: 9 cis-Neoxanthin

Delete this entry

Definition: The phytoplankton pigment 9 cis-Neoxanthin

Reason for request:

We may need to contact you to discuss this request

Your name:

Email:

Submit Changes | Cancel

What are metadata registries? How are they used? By whom?

- What? Standardized databases for metadata
  - data elements from diverse databases & systems
  - reusable value sets (e.g., country codes, cause of death)
  - provenance information (source, derivation, etc.)
  - metadata, parameters & equations for simulations
- How? Data Administration & Integration *(design + run time)*
  - manage definitions, data relationships, etc. over time
  - combine data & concepts from diverse sources
  - discover hidden relationships between data
  - support data entry forms, navigation & federated queries
- Who? Large organizations (e.g., EPA, NCI, DOD)
  - for interoperation & data discovery, which also require
  - *standards for sharing across disciplines & organizations*

*ISO/IEC 11179 Metadata Registry Standard facilitates inter-operability

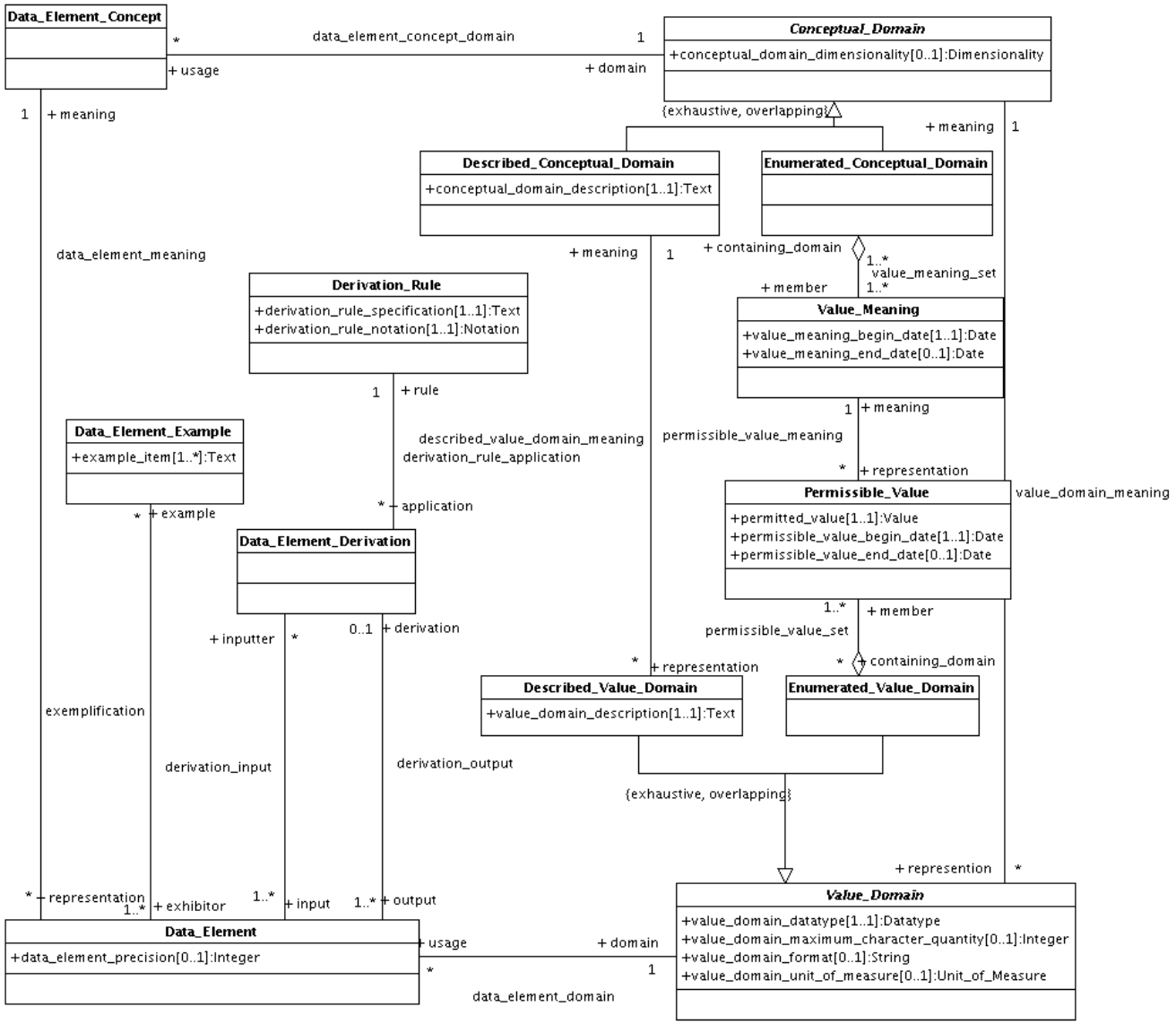
- Part 1: Framework (32 pages)
- Part 2: Classification (14 pages)
- Part 3: Registry metamodel and basic attributes (108 pages)
- Part 4: Formulation of data definitions (16 pages)
- Part 5: Naming and Identification Principles for Data Elements (25 pages)
- Part 6: Registration (72 pages)

Work on part 3 began ~1988;ed 1 became an ISO standard 199

Work on new part 3, ed. 3 began in 2003 – target is 2009

Publicly Available from:
http://isotc.iso.ch/livelink/livelink/fetch/2000/2489/
lttf_Home/PubliclyAvailableStandards.htm??Redirect=1

11179 ed. 3 metamodel has more fine-grained metadata entities & attributes



Example *Data Element* Metadata from EPA Environmental Data Registry

Responsible Agency Type Code

Definition: The code that represents the name of the type of federal agency primarily responsible for a facility

XMLTag: ResponsibleAgencyTypeCode

Type: enumerated (1)

Identifier:1-89872:1

Status: Standard/Final

Origin: US EPA Data Standard EX000020.2 item 8.2

Data Steward: Larry Fitzwater Environmental Information Office

Permissible Value and Value Meaning

Value	Value Meaning	Begin Date	End Date
C	Civilian		
M	Military		

Metadata registry success & growth has led to new semantic challenges

- **Natural language descriptions are too limited**
  - imprecise and fuzzy, even for human users
  - computer software cannot process unambiguously
  - does not help identify what is known and not known
  - require too much intervention by expensive humans
- **Weak integration of concepts with metadata**
  - e.g., "nutrients" with particular measured variables
- **Relationships are not well-specified**
  - e.g., "is-a", "part-of", "broader-term", etc.
- **Limited scalability**
  - for multiple terminologies & many databases
- **Limited relationships with other standards**
  - e.g., terminologies, ontologies, OMG, etc.
  - no formal axioms to specify relationships, etc.

XMDR project & 3rd edition of 11179 address current MDR limitations

- **Use concepts to unify different types of metadata**
  - evolution requires increasing granularity & details
  - combine strengths of data dictionaries/registries and ontologies
- **Add more rigorous & formal specification for**
  - concepts and concept systems (including ontologies, thesauri, taxonomies)
  - relationships between metamodel components (concepts, elements, values)
  - formal axioms for conceptual & structural relationships
- **Register & manage complex semantic metadata (i.e., concepts)**

in more formal, systematic ways (e.g., description logic)
to facilitate machine processing of semantics in order to

  - link together data elements & terms across multiple systems
  - discover relationships among data elements, terms & concepts
  - create and manage names, definitions, terms, etc.
  - support software inference, aggregation, and agent services

What is the XMDR Project? (eXtended Metadate Registries)

- **Set of *collaborative initiatives* with shared goals & funding**
  - **EPA, NCI, DOD, LBNL, USGS, Ecoterm, UNEP, … (major 11179 users)**
    - **XMDR project at LBNL began in 2003**
    - **principals have been meeting in Berkeley since 2004**
  - **ISO-IEC JTC1/SC32/WG2 & ANSI/INCITS L8 working on 11179 ed. 3**
    - **Joint Technical Committee 1, Subcommittee 32, Working Group 2**
    - **metadata registry standards work began in 1980's re data dictionaries & codesets**
    - **Coordinate/harmonize related metadata standards (e.g., OMG, ODM, CWM, etc.)**
- **Open source *reference implementation & testbed system***
  - **test implementations of proposed extensions to 11179 metamodel**
    - **add more formal semantic metadata on concepts & relationships to data**
  - **assemble semantic metadata from diverse sources & structures**
    - **terminologies, ontologies, etc. for environment, geography, health, …**
  - **explore emerging semantic technologies (e.g., RDF, OWL, CL, …)**
  - **demonstrate new capabilities**
    - **e.g., ontology lifecycle management & harmonization**

XMDR Prototype implementation helps test & demonstrate new parts of 11179

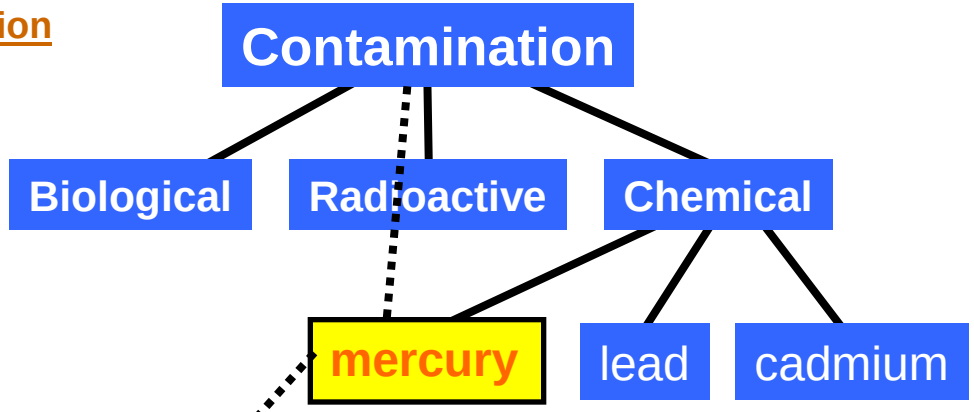
- Demonstrate *feasibility & utility* of proposed revisions to ISO/IEC 11179
- Provide open-source reference implementation with XMDR capabilities
  - Determine the necessary features to leverage semantic interoperability between 'concept' systems and 'data elements'
  - e.g., for ontology lifecycle management & harmonization
- Explore benefits of representing XMDR content using emerging semantic technologies (e.g., RDF, OWL, CL, …)
  - integrate open source tools to create, maintain, deploy XMDR standards
  - test capabilities and performance of candidate tools
- Assemble semantic metadata with different structures from diverse sources to test various semantic technologies
  - terminologies, thesauri, ontologies, …
  - from health, environment, geography, … water?
- Help identify ways to resolve registration & harmonization issues for different metadata standards, including ODM & MMF

XMDR Protype facilitates basic types of inference using concepts plus metadata

Inference Search Query:

"find water bodies downstream from Fletcher Creek where chemical contamination was over 10 micrograms per liter between December 2001 and March 2003"

Concept System (multi-lingual):



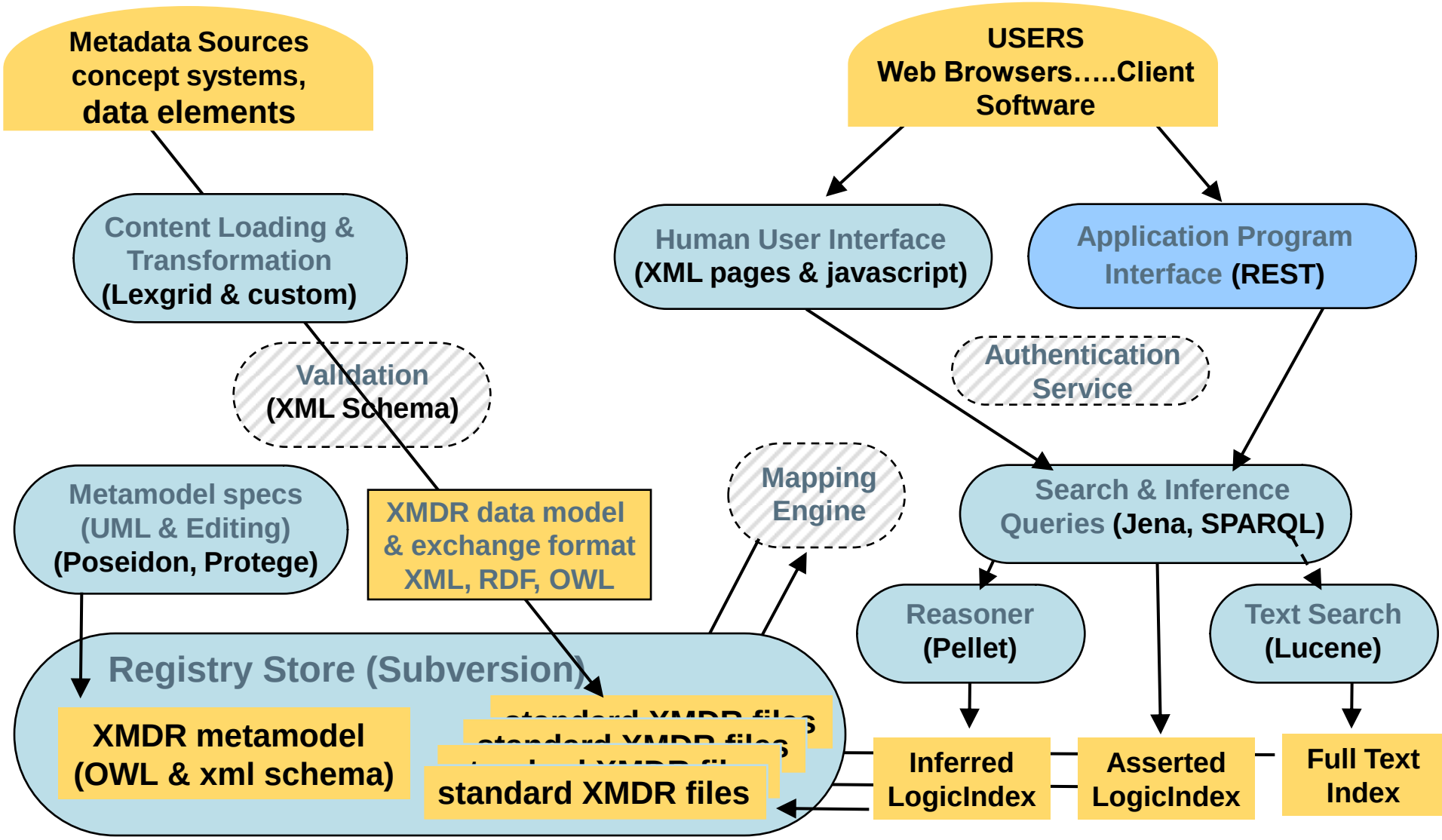
Data:

ID	Date	Temp	Hg
A	06-09-13	4.4	4
B	06-09-13	9.3	2
X	06-09-13	6.7	78

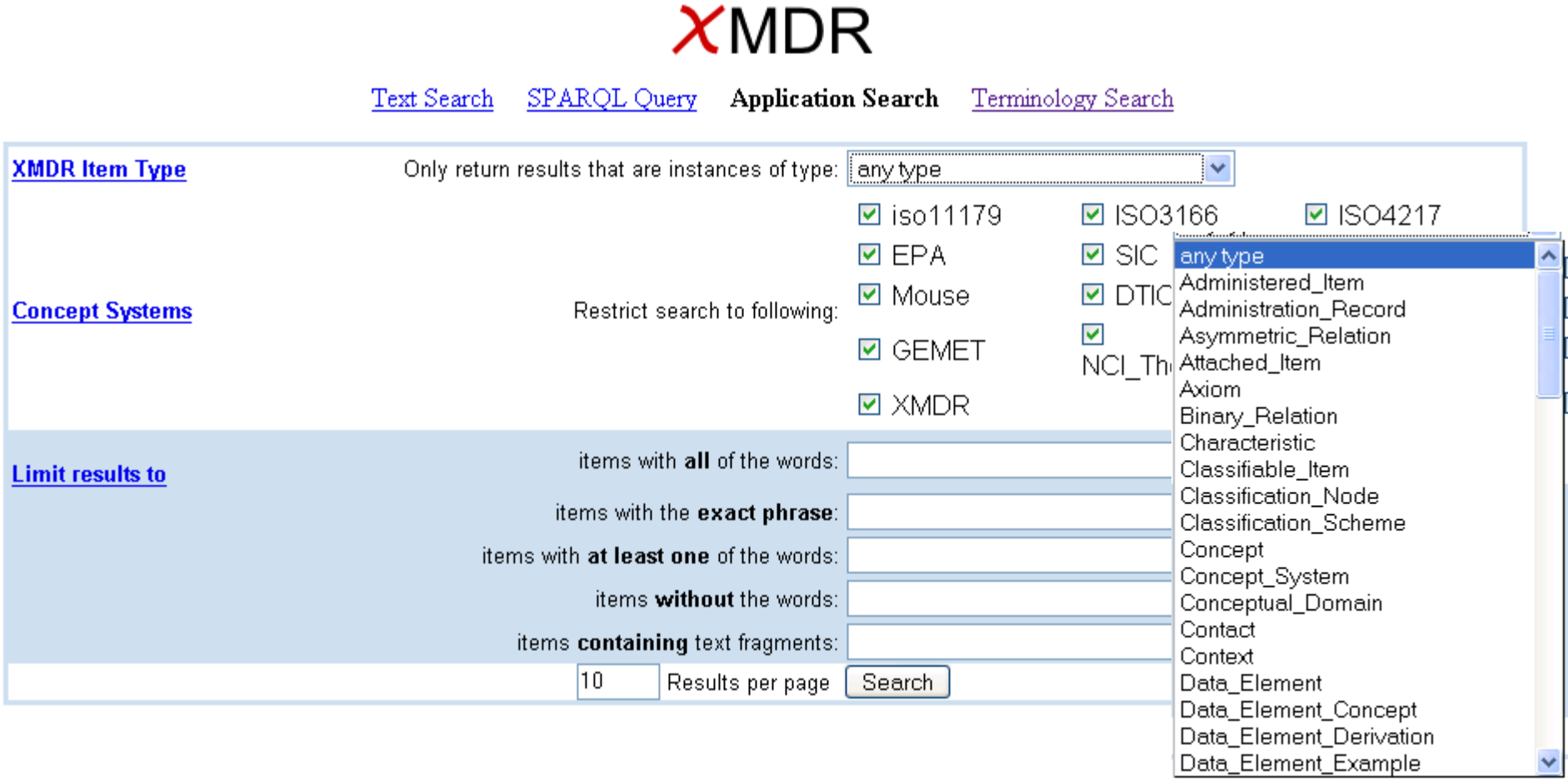
Metadata:

Name	Datatype	Definition	Units
ID	text	Monitoring Station Identifier	not applicable
Date	date	Date	yy-mm-dd
Temp	number	Temperature (to 0.1 degree C)	degrees Celsius
Hg	number	Mercury contamination	micrograms per liter

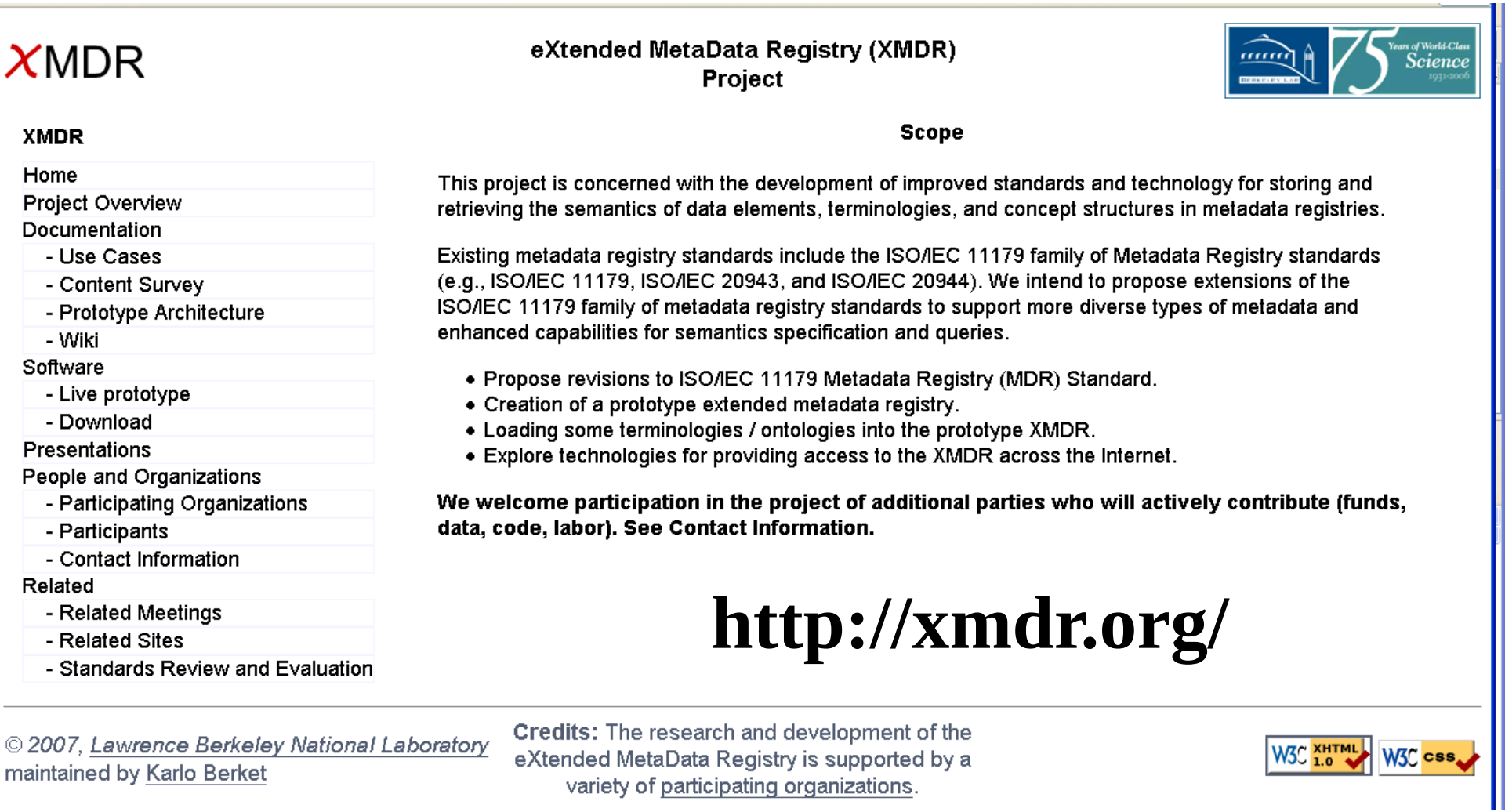
XMDR Prototype Rest-Style Modular Architecture uses open source software



XMDR Prototype supports several different types of search interfaces



XMDR Web Site has documentation downloadable software & content



XMDR

eXtended MetaData Registry (XMDR) Project

XMDR

- Home
- Project Overview
- Documentation
  - Use Cases
  - Content Survey
  - Prototype Architecture
  - Wiki
- Software
  - Live prototype
  - Download
- Presentations
- People and Organizations
  - Participating Organizations
  - Participants
  - Contact Information
- Related
  - Related Meetings
  - Related Sites
  - Standards Review and Evaluation

Scope

This project is concerned with the development of improved standards and technology for storing and retrieving the semantics of data elements, terminologies, and concept structures in metadata registries.

Existing metadata registry standards include the ISO/IEC 11179 family of Metadata Registry standards (e.g., ISO/IEC 11179, ISO/IEC 20943, and ISO/IEC 20944). We intend to propose extensions of the ISO/IEC 11179 family of metadata registry standards to support more diverse types of metadata and enhanced capabilities for semantics specification and queries.

- Propose revisions to ISO/IEC 11179 Metadata Registry (MDR) Standard.
- Creation of a prototype extended metadata registry.
- Loading some terminologies / ontologies into the prototype XMDR.
- Explore technologies for providing access to the XMDR across the Internet.

We welcome participation in the project of additional parties who will actively contribute (funds, data, code, labor). See Contact Information.

http://xmdr.org/

Credits: The research and development of the eXtended MetaData Registry is supported by a variety of participating organizations.

Many aspects of the XMDR project have been inspired by NCI's caDSR

- **Cancer Data Standards Registry began in 1997**
  - **began with breast cancer treatment trials data**
- **Based on first edition of ISO/IEC 11179**
  - **conceptual model (3) AND administrative procedures (6)**
- **First project to combine concepts & metadata**
  - **NCI Metathesaurus currently includes 67,102 concepts**
  - **Enterprise Vocabulary Service also includes 7 other concept systems (GO, SNOMED, LOINC, MedDRA,…)**
    - **12,000 concepts tied to data elements**
  - **Data element registry > 30,00 variables (data elements)**
    - **4,000 of these are Standard, 148 Preferred Standards**
- **Thousands of participating projects/PI's**
- **Open source MDR tools by the end of 2008!**

Why should water data people learn more about 11179 & XMDR?

- **Leverage existing technology for metadata**
  - **possible use of open-source tools from XMDR & NCI**
- **Inter-operability with other metadata registries**
  - **European Environmental Agency & WISE**
  - **National Cancer Institute, EPA, DoD, etc.**
  - **other NSF-sponsored scientific metadata registries?**
- **CUAHSI HIS Project already moving in this direction**
  - **Data Observation Model**
  - **Controlled Vocabularies**
- **Opportunities to collaborate & help set standards**
  - **XMDR Project**
  - **11179 Metadata Registry Standard, ed. 3**

Thanks & Acknowledgements

- **Bruce Bargmeyer, principal investigator** BEBargmeyer@LBL.gov
- **Kevin Keck, initial & current designer & implementor**
- **Fred Gey, concept mapping, etc.**
- **Anirban Sen, coding, implementation, documentation, …**
- ***Frank Olken, initial concepts & metamodel extensions***
- ***Karlo Berkett, implementation, user interface, data loading***
- **Harold Solbrig, Lexgrid, model development, etc!**
- **L8 and SC 32/WG 2 Standards Committees**
- **Major XMDR Project Sponsors and Collaborators**
  - **National Science Foundation (Grant #0637122)**
  - **U.S. Environmental Protection Agency**
  - **Department of Defense**
  - **National Cancer Institute**
  - **U.S. Geological Survey**
  - **And others!** ***for more info, see xmdr.org***

Other Topics?
Extra Slides below here

- **This is the end of the presentation**
- **Slides following this one can be**
  - **folded back into the mainline presentation,**
  - **Held in reserve if questions arise they can help**
  - **Dropped altogether**

Introduction to the XMDR Project: selected overview documents

- www.xmdr.org/
- hpcrd.lbl.gov/SDM/XMDR/overview.html **(link from xmdr.org)**
- hpcrd.lbl.gov/SDM/XMDR/presentations/XMDR_Elevator_Summary_rough_draft.ppt **(overview)**
- **xmdr.lbl.gov/xmdr/ (prototype system)**
- hpcrd.lbl.gov/SDM/XMDR/arch/index.html **(architecture)**
- erdos.lbl.gov/mediawiki/index.php/Main_Page **(project wiki)**
- hpcrd.lbl.gov/SDM/XMDR/presentations/ **(esp recent ones)**
- hpcrd.lbl.gov/SDM/XMDR/presentations/XMDR-Prototype-Status-Oct-2005.ppt **(status report)**

How does XMDR Prototype differ from current registry technology?

- **Evolutionary aspects**
  - **Finer-grained, more formal metadata**
    - **e.g., distinct attributes for measurement units**
    - **rather than just part of textual description**
  - **Machine inference complements text searching**
- **Revolutionary aspects**
  - **Use of formal ontologies, logic, and inference**
    - **to specify 11179 metamodel**
    - **to store, search, retrieve and display metadata**
  - **Logic engines & machine reasoning**

Example concept system content currently loaded in XMDR Prototype

via Lexgrid (from Mayo Clinic & Harold Solbrig)

- GEMET 2001.0 Multilingual Environmental Thesaurus
- National Biological Information Infrastructure biodiversity
- NCI Thesaurus_06.02d health concepts system
- ISO4217_1981 currency codes
- ISO3166_V-10 country codes (only 2 letter codes)
- Mouse_1.32 anatomy
- Defense Technology Information Center 1.0 Thesaurus
- Portions of EPA controlled vocabulary
- NAICS and SIC industrial classification codes

via special purpose scripts

- Omega ontology

Additional candidate metadata content to test 11179 metamodel expressivity

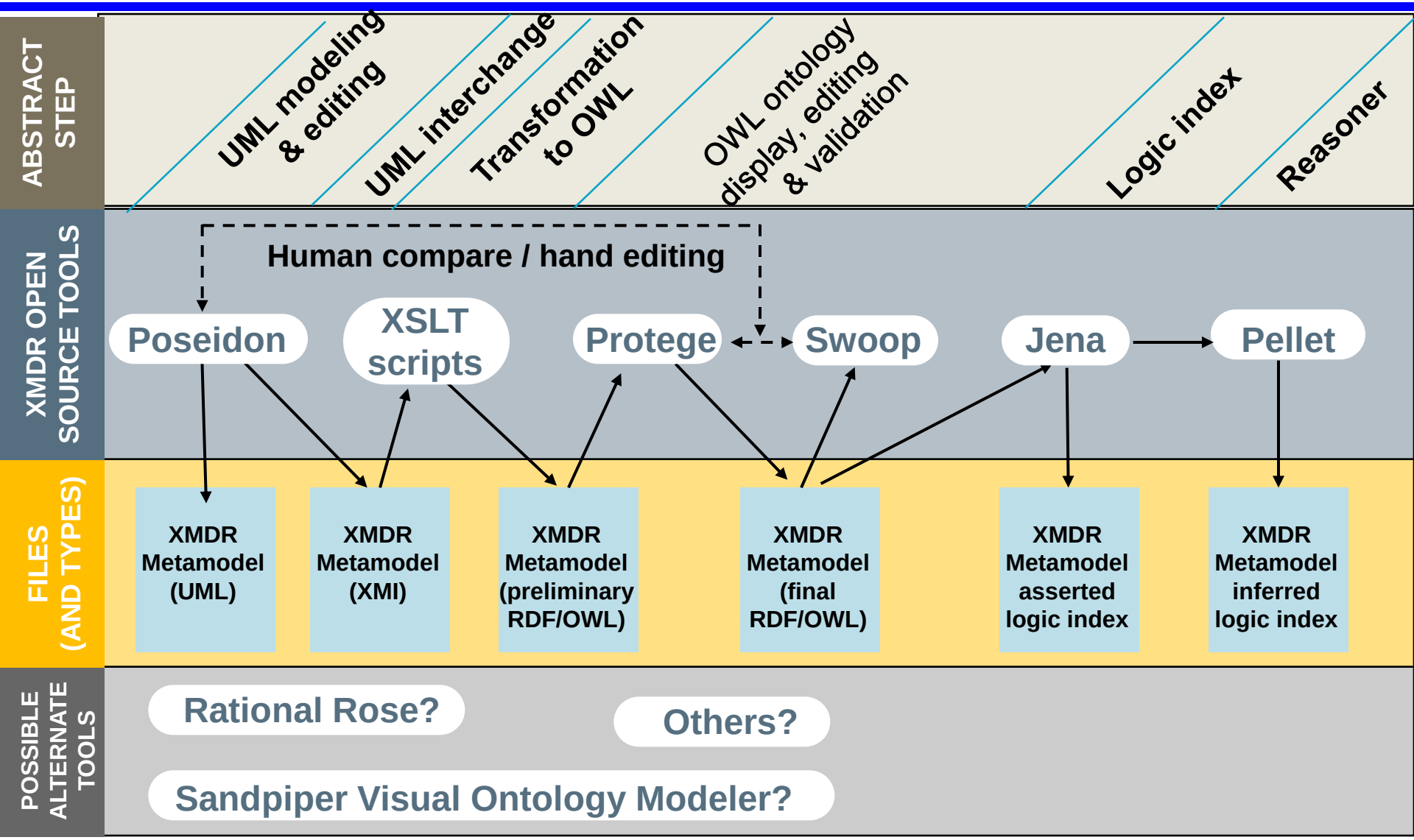
Current 11179 Data Element Registries

- caDSR (full NCI Cancer Data Standards Registry)
- EDR (EPA Environmental Data Registry)

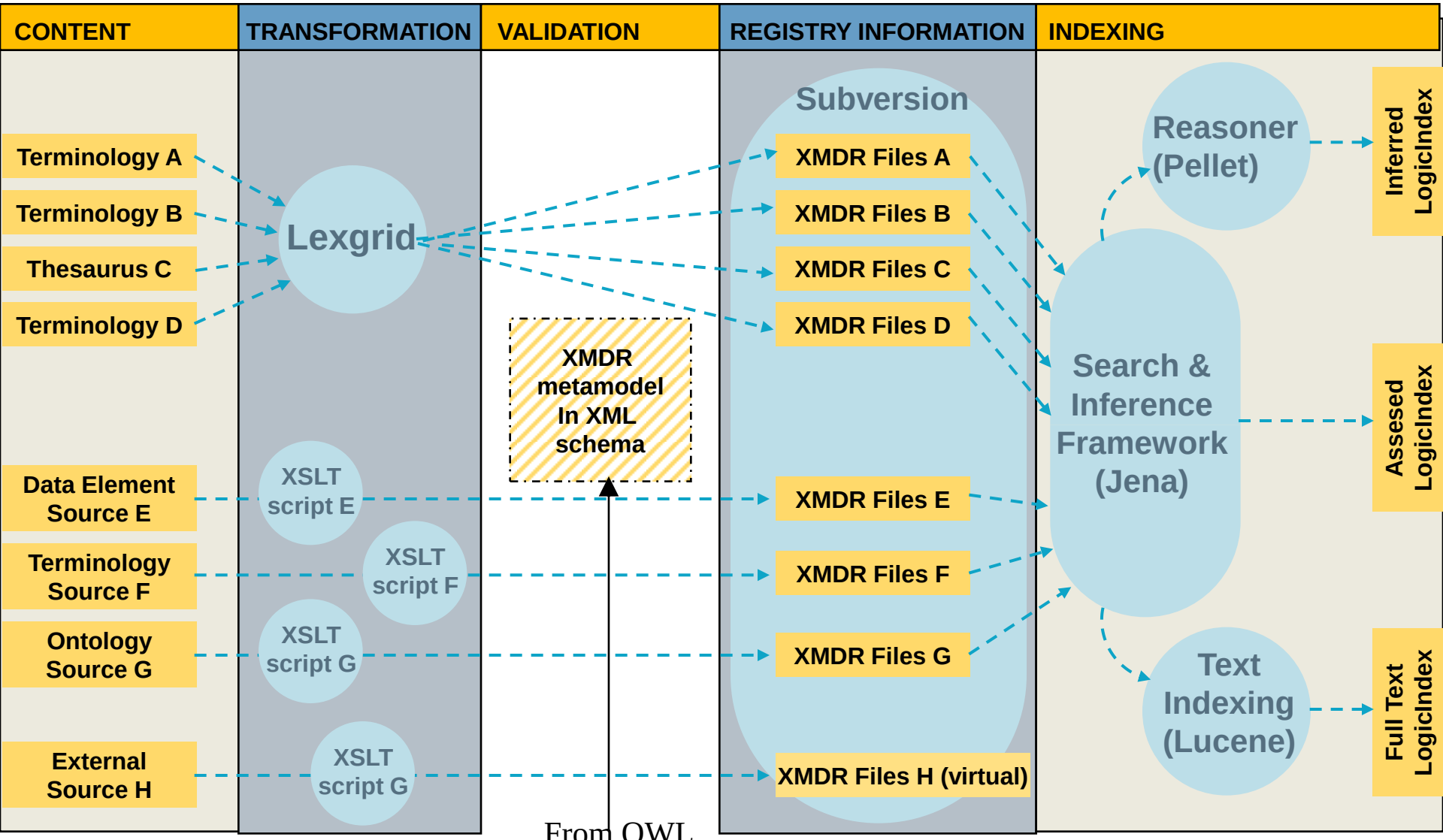
Candidate Additions to Concept Systems and Ontologies

- NASA SWEET (Semantic Web Earth & Environmental Terminologies)
- IETF RFC 3066 Language Codes
- USGS Geographic Names Information System
- Getty Thesaurus of Geographic Names
- I.T.I.S. - Integrated Taxonomic Information System
- Foundational Model of Anatomy
- EPA Chemical Substance Registry
- GO (Gene Ontology), ….Agrovoc, …and possibly others
- OMV Ontology Metadata Vocabulary (European NeON consortium & Stanford NCBO)

XMI from 11179-ed3 UML is transformed to RDF/OWL for XML metamodel specification



Content loading, with XMDR metamodel used for inferred indexing and validation



*Concept Use & Integration for the Cancer Data Standards Repository