ISEIS

Journal of Environmental Informatics 23(2) 10-23 (2014)

Journal of Environmental Informatics

www.iseis.org/jei

An Ontology Driven Relational Geochemical Database for the Earth's Critical Zone: CZchemDB

X. Z. Niu [1,*] J. Z. Williams[1], D. Miller[1], K. Lehnert[2], B. Bills[1], and S. L. Brantley[1]

[1]Earth and Environmental Systems Institute, Penn State University, University Park, PA 16802, USA
[2]Lamont-Doherty Earth Observatory, Columbia University, Palisades, NY 10964, USA

Received 22 January 2013; revised 12 November 2013; accepted 8 Febuaray 2014; published online 20 June 2014

ABSTRACT. Multiple Critical Zone Observatories (CZO) have been established in recent years in the U.S.A. and elsewhere to conduct collaborative, multidisciplinary research on the earth's critical zone (CZ). As a result, a large amount of scientific data over space and time has been collected. However, heterogeneities in data documentation impede our ability for cross-site comparisons and for integrated analysis. To promote efficient data sharing, publishing, and integration, we developed a sample-based measurement ontology (SMO) to formalize data structures and unify variable terms in data documentations for the CZOs. "Sample" is the core of the SMO. Each sample is part of a "Medium" that represents an entity of the CZ such as soils. Characteristics of a sample are analyzed and the results are reported as values and errors. Based on the concepts of the SMO and geochemical data model of Lehnert et al. (2000), we created a relational database to accommodate CZ regolith geochemical data, namely CZchemDB, to bridge the gap between data collection, documentation and sharing among the CZOs. The CZchemDB has now been successfully implemented in the MS Access database management system for individual or small group uses. However, our ultimate goal is to integrate the CZchemDB with the online global geochemistry data portal, EarthChem, for broader data accessibility and reusability. Finally, we emphasize that the SMO is extensible to all media within the CZ and so CZchemDB can be used to store any sample-based chemical data measured on any medium such as minerals, water, gas, or biota in the CZ.

Keywords: critical zone, sample-based measurement ontology, geochemical database, CZchemDB, EarthChem

1. Introduction

The Critical Zone (CZ) is defined as the layer of Earth extending from the outer edge of the vegetation canopy down to lower limits of groundwater (National Research Council (U.S.). Committee on Basic Research Opportunities in the Earth Sciences, 2001; Brantley et al., 2007). To understand how physical and chemical processes interact within this zone, six Critical Zone Observatories (CZO) have been established in the U.S. and several such observatories are funded or planned in international settings. Operating at the watershed scale, CZOs serve as natural laboratories for measurements of chemical, physical, and biological characteristics of the earth surface.

The reservoirs within CZ that are currently investigated include, at the broadest level, the atmosphere, biota, water, and rock/regolith (Figure 1). Among these reservoirs, some are well mixed and relatively homogeneous in nature (e.g. atmosphere, plotted to the left on Figure 1), while others are poorly mixed and relatively heterogeneous (e.g. rock/regolith, plotted to the right). Reservoirs along the poorly- to well-mixed continuum represent intermediately mixed characteristics (e.g. groundwater, biota).The well-mixed reservoirs change rapidly over time but more slowly over space. Therefore, measurements of the well-mixed reservoirs are often made using sensors that collect data at rates that sometimes approach real time, producing large volumes of time-series data. In contrast, the poorly-mixed reservoirs vary greatly over space but change relatively slowly over time. The characteristics of these poorly mixed reservoirs are measured more intermittently and generally by sample collection and multivariate laboratory analysis, and therefore, tend to comprise spatially sparse datasets.

Given the nature of the different reservoirs within the CZ, data describing the reservoirs with relatively homogeneous characteristics on the left of Figure 1 depend strongly on time and less strongly on spatial position, whereas data for reservoirs with heterogeneous characteristics on the right depend strongly on spatial position and less on time. In addition, every sample that is collected from the reservoirs on the right side of Figure 1 is generally measured for a number of chemical or biological or physical parameters. Thus every box on the diagram that is labeled "chemistry" can be expanded as shown in Figure 1b and 1c. Similarly, other such expanded diagrams could be provided for types of (molecular) biologi-

[*] Corresponding author. Tel.: +1 (814) 863-9540 ; fax: +1 (814) 865-3191.
E-mail address: xzniu@psu.edu (X. Z. Niu).

ISSN: 1726-2135 print/1684-8799 online

doi: 10.3808/jei.201400266

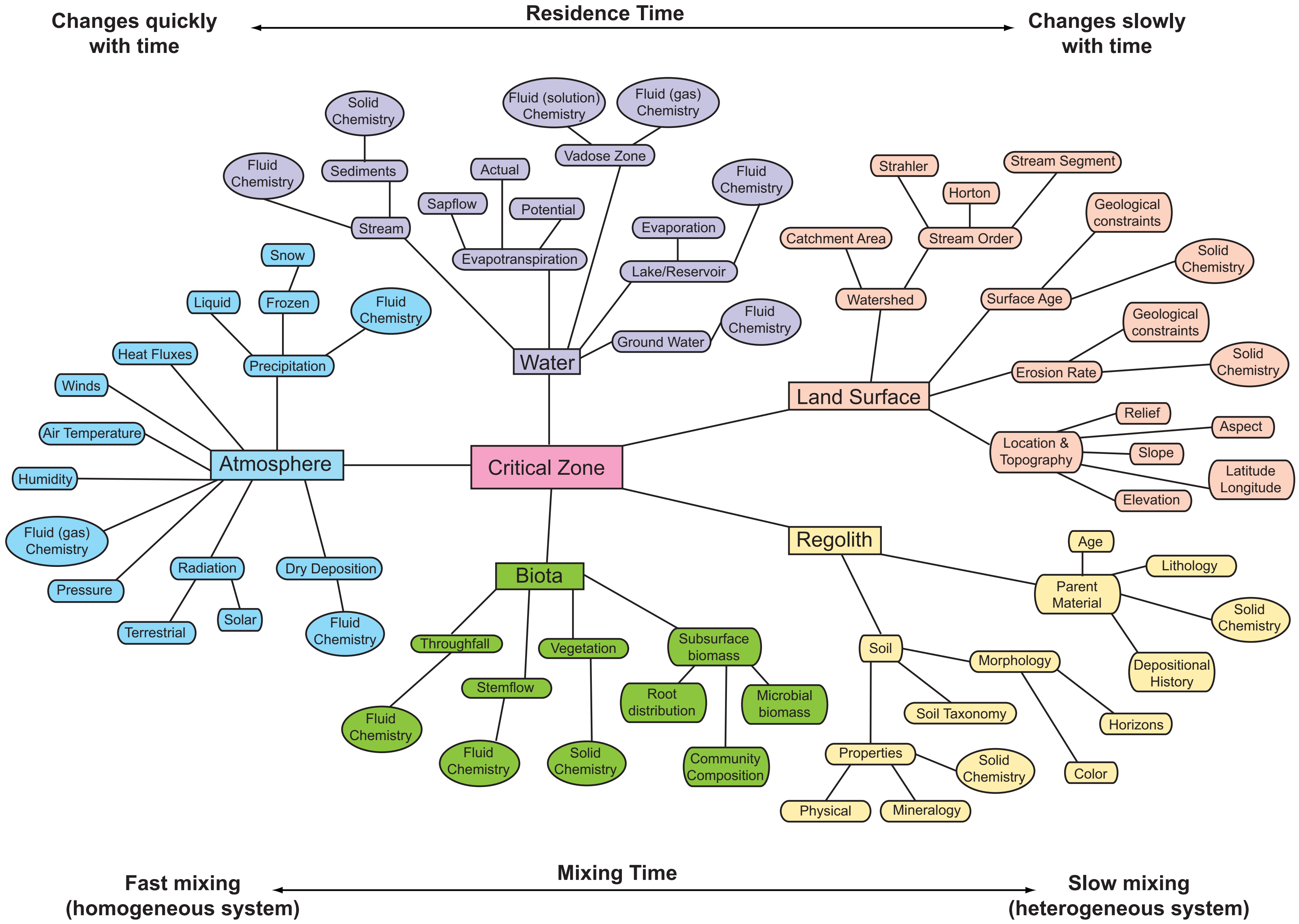

Figure 1. Diagram of major entities of the earth's critical zone (a).

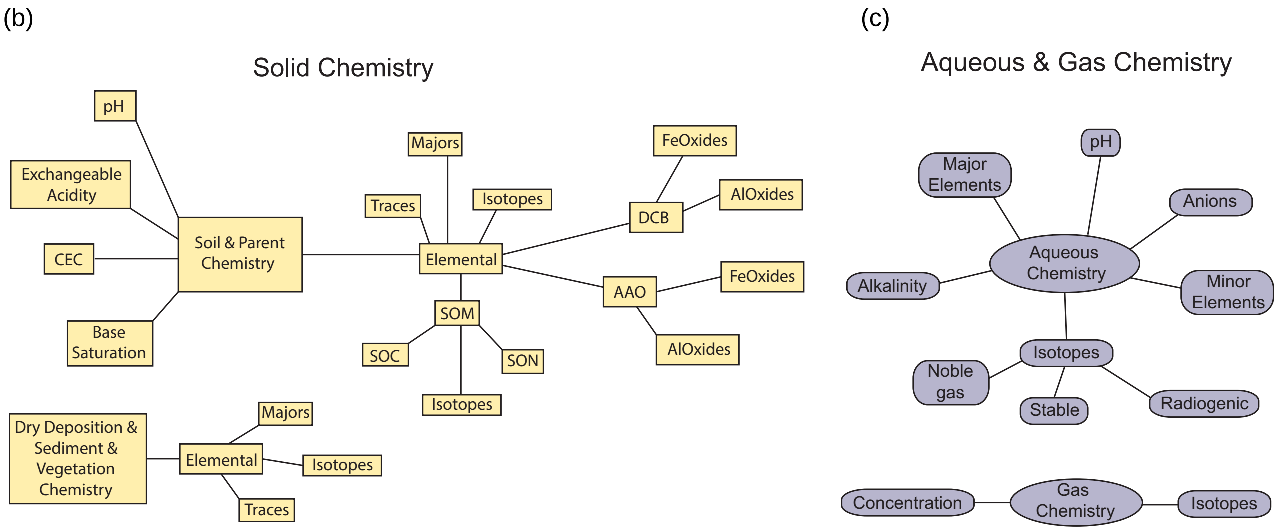

Figure 1 (cont'd). (b) Chemistry attributes measured from aqueous; (c) Measured from gaseous samples.

cal measurements. Furthermore, sample-based measurements are often performed on subsamples that are prepared by different treatments (e.g. different particle size, digestions, etc.) or analyzed using different techniques. Thus, the nature of sensor-derived data (toward the left of Figure 1) and sample-derived data (toward the right) necessitates different data structures and data models.

As the CZOs have grown, a large volume of data has been produced, and moreover, many such CZ data already exist in the published or online literature. The needs for efficiently managing and sharing data are increasingly recognized within the CZ science community for integrated CZ studies (Hofmockel et al., 2007; Niu et al., 2011; Zaslavsky et al., 2011). Currently, most of the CZO observation data are stored/managed in different file types (e.g. ASCII text file, spreadsheet file, or database file) with different format (e.g. cross-tab tables versus serial lists), largely depending on the data sources (sensor or sample/lab-derived) and the investigator's preference (e.g. using different software). Lack of consistent data structure and/or format greatly inhibits the scientists' ability to share and integrate data efficiently cross disciplines and from different investigators. In addition, insufficient metadata that describe analytical method, data quality and data sources and ambiguous terminology used in different files (or databases) also presents major challenges in effective data sharing, discovering, and data integration across scientific domains (Lehnert et al., 2000; Horsburgh et al., 2009). This becomes especially difficult when data are shared across international boundaries.

To solve some of the problems related to these issues, researchers have worked to develop strategies and/or databases for easy data managing, publishing, and sharing across multiple disciplines and temporal/spatial scales. For example, the hydrological community has established a service-oriented system, the Hydrologic Information System (CUAHSI HIS) (http://www.cuahsi.org/), to accommodate hydrological observations (mostly time-series data that are often measured by sensors) through CUAHSI (Consortium of Universities for the Advancement of Hydrologic Science, Inc.). In the CUAHSI HIS, an Observation Data Model (ODM) was created to provide a consistent format for the storage and retrieval of point environmental observations in a relational database (Horsburgh et al., 2008).

To facilitate the preservation, discovery, and visualization of global geochemical datasets, rock/sediment geochemical databases such as PetDB (http://www.petdb.org), GEOROC (http://georoc.mpch-mainz.gwdg.de), and SedDB (http://www.seddb.org) have been developed by the geochemistry community (Lehnert et al., 2007a, b; Sarbas and Nohl, 2008) using a common relational data model (Lehnert et al., 2000). These and other geochemical databases use EarthChemXML (http://www.earthchemportal.org/schema/earthchem_schema.xsd) as a common data exchange format to encode their data for inclusion in the EarthChem Portal (http://www.earthchem.org), which provides a central access point to geochemical data in distributed partner databases.

Another popular database related to CZ is SSURGO (Soil Survey Geographic database), a soil database that was established by the United States Department of Agriculture Natural Resources Conservation Service (NRCS) to house the sample-based soil survey data. SSURGO also employs a relational data model to link spatially referenced mapping units to associated attribute tables, including soil physical and chemical properties (http://soildatamart.nrcs.usda.gov/). With the database, users can store, retrieve and analyze soil data, as well as integrate the data with the mapping unit IDs in a geographic information system (GIS). In addition to SSURGO (the most detailed soil data-base at the finest scale county level), two similar databases are also available at the state level, STATSGO (State Soil Geographic database), and at national levels, NATSGO (National Soil Geographic database).

Many of these databases are well established for their respective purposes and widely used by different scientific communities. However, they are also somewhat inadequate to be directly used for CZO projects, especially for the regolith (soil) geochemical data. For example, CUAHSI HIS is sui- table for time-series observational data, but its current structure is not efficient for storing analytical values measured for samples collected from different media type at the same site (e.g. from surface or ground water, soils, biota-creating a data redundancy problem) and data generated from subsamples that represent splits from the same sample for different treat- ment or analysis. These issues are all related to the fact that HIS is largely built for sensor data whereas the CZ chemical data addressed in this research paper is predominantly sample-based.

Likewise, EarthChem is a sample-oriented database. However, some important soil characteristics (e.g. soil horizon, bulk density, soil texture and structure, etc.) that are often concurrently measured along with geochemical properties from the same samples collected from CZOs are not included in the EarthChem schema. Furthermore, many of the CZ samples derive from sampling strategies that incorporate augered cores, pits, or wells, and aspects of these strategies were not well described within current EarthChem databases. On the other hand, SSURGO, another sample-based soil database, is mainly designed to store soil survey data, mostly from one-time measurements or averages, collected at a relatively coarse spatial resolution. It makes the SSURGO databases not suitable for CZO soil data that are often collected at much higher density and frequency and are often split into multiple subsamples for different analysis. CZ scientists needed to develop a new data model to efficiently store not only the primary chemical analytical values but also measurements of other soil characteristics (physical, mineral, and biological properties) and associated metadata, which are all important for integrated CZ analysis.

Managing data is important; however, ultimately the goal is to allow discovery, sharing and integration of data across multiple disciplines within earth and environmental studies (Hofmockel et al., 2007; Madin et al., 2007; Horsburgh et al., 2009). To promote data sharing and discovering, the Critical Zone Exploration Network (CZEN) has been established to create a network that would allow researchers to access and integrate data in a way that allows isolation of environmental variables and comparison of environmental effects across gradients of climate, time, lithology, anthropogenic disturbance, biological activity, topography, or other variables (http://www.czen.org). Currently, various types of data files are uploaded to and are available to be downloaded from the CZEN website. However, many of the datasets are in their original disparate formats (e.g. text files, Excel spreadsheets, or databases), and therefore, are not convenient for easy data integration. An integrated CZO information system that will facilitate seamless access and analysis of hydrology, geochemistry and geomorphology data within and across six CZO sites is under development (Whitenack et al., 2011).

To meet the challenges of multiple institutional collaborative efforts within the CZO projects and to collect large volumes of data across multiple disciplines and multiple temporal and spatial scales, an initial infrastructure for CZO data storing and sharing has been proposed (Hofmockel et al., 2007; Zaslavsky et al., 2011). In the initial design, data are managed at different complexity levels so that data can be available both in human-readable form at individual CZO websites as well as via web services from the central CZO data repository. In order to ensure the published data can be unambiguously interpreted and automatically harvested into a centralized data system, uniform data modeling, data description and formatting practices are needed. This requires all data sources not only to have common data structure but also to have common semantics; that is, to validate the harvested dataset from different sources against a set of shared (controlled) vocabularies and parameter ontology before they are integrated into a centralized standard compliant data services.

In this paper, we discuss the development of a strategy that is targeted for managing geochemical data that describe various properties of critical zone regolith (i.e., weathered rock and soil and saprolite). First, we present a sample-based measurement ontology (SMO) that will serve as a formal data model for all kind of sample-based measurements from CZOs, including geochemical data of regolith, water, or vegetation. Next, we describe a specifically designed conceptual model and the schema of the Critical Zone regolith geochemical database (CZChemDB). The strategies for implementation of the CZchemDB with relational database management systems (e.g. MS Access) and tools for data entry and data application are also discussed. Finally, we discuss the integration of CZ-chemDB with the EarthChem portal which we are designing to increase the online accessibility and usability, along with the future integration of CZchemDB into the proposed CZO data-sharing infrastructure.

2. Critical Zone Sample-based Measurement Ontology

An ontology is a formal model to define terms or concepts and their relationships within a scientific domain such as critical zone science or its subdisciplines (Madin et al., 2008). Ontologies have been used in biology, ecology and other earth sciences to improve data interpretation, discovery and integration based on meaning (Madin et al., 2007, 2008). Study of the CZ will benefit from development of an ontology because such a formal model will unify the description of measurements and their relationships, providing a comprehensive pathway toward data discovery and integration.

Study of the highly heterogeneous CZ reservoirs, such as regolith, often involves large amount of sample-based measurements. Samples are defined here as raw materials collected from any particular media within the CZ, such as bulk soil samples or soil pore water. Comparing with the sensor-based data, sample-based measurements are more complicated because each sample can be split into multiple subsamples (and be reused) for different treatments and analysis. The “Sample” is the central part of the sample-based measurment ontology (SMO, Figure 2). Each sample is collected from a medium

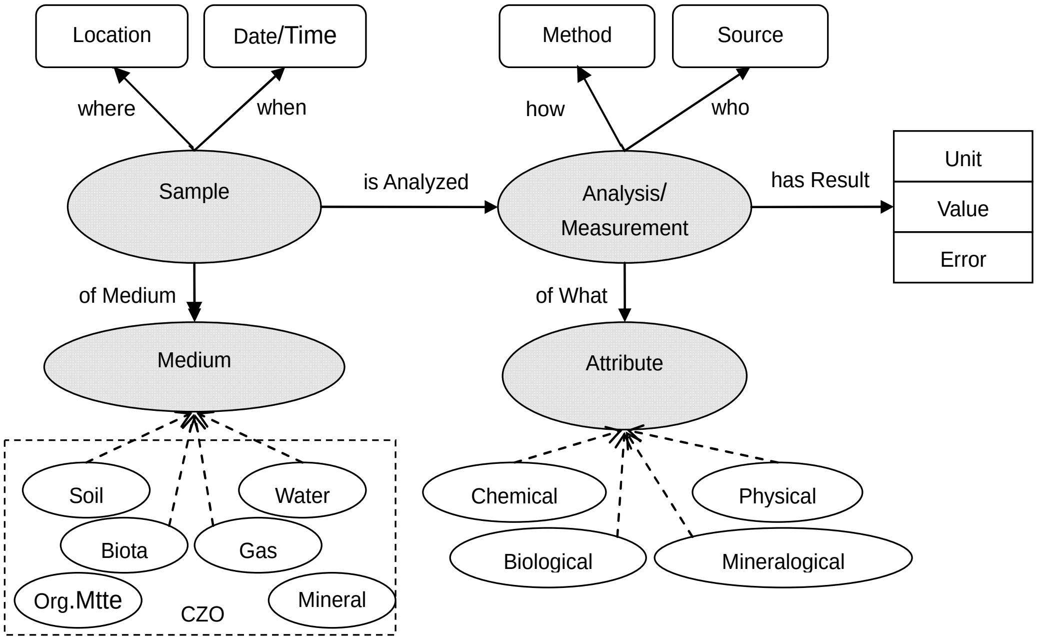

Figure 2. Sample-based measurement ontology (SMO).

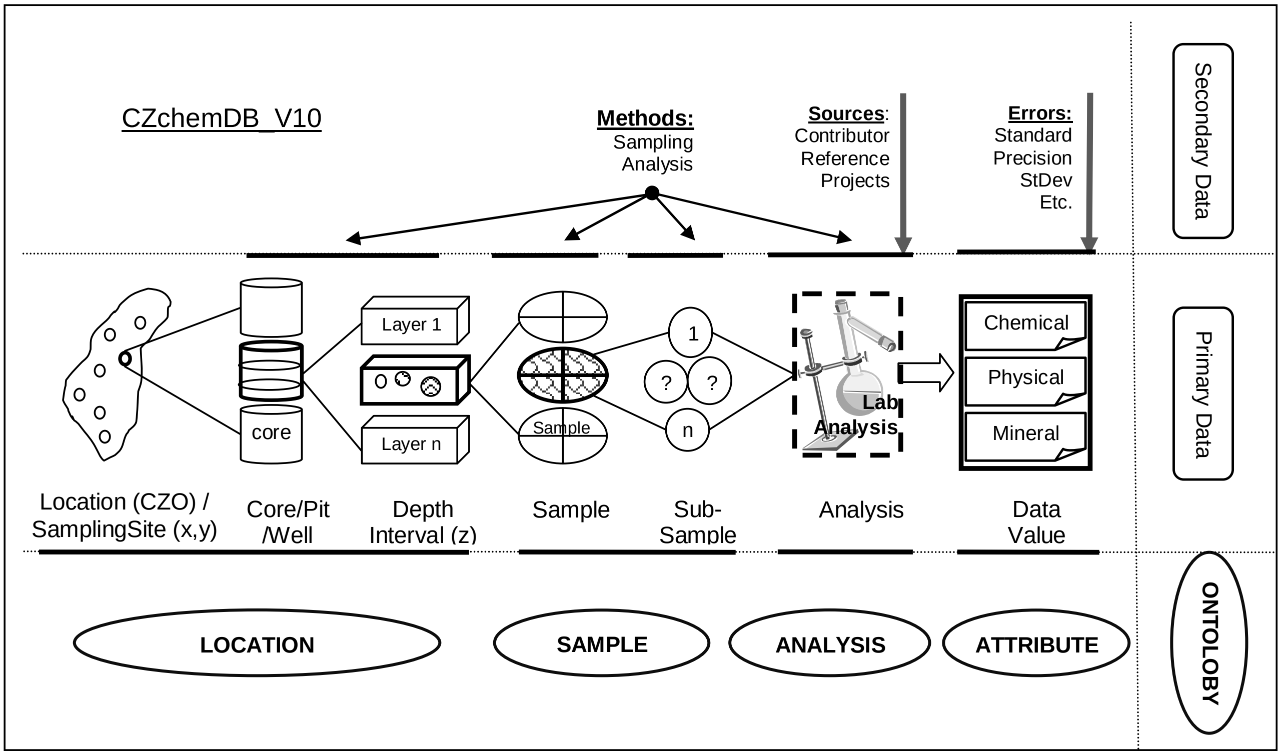

Figure 3. Conceptual model of the Critical Zone geochemistry database – CZChemDB.

within the CZ, which serves as an extension point to connect to different entities identified in the CZ ontology (see Figure 1). Meanwhile, each sample is associated with a sampling "Location" where it has unique environmental conditions (temperature, precipitation, etc.) and geographical features (landscape, vegetation, etc.). Samples (including subsamples) are then analyzed or measured for different "Attributes" (i.e. the chemical, physical, mineralogical or biological characteristics). The results of analysis (or measurements) are reported as data “Values”, “Units”, and “Errors”. In addition, categories “Date/Time”, “Methods”, and “Source” in the SMO are to describe when and how the sample was taken and who performed the sampling and analysis, which are all important metadata for future data integration and data analysis.

In the following, we will focus our discussion on the development of a Critical Zone geochemical DataBase (CZchemDB) that will be used to store geochemical data measured from CZO regolith/soil samples. However, as we discussed above, the same concepts can be applied to samples taken from any other medium or reservoirs identified in the CZ ontology diagram (see Figure 1). Unlike other studies that use ontology to facilitate integration of different existing data sources (Madin et al., 2008), we developed the SMO to provide a framework so that all kinds of sample-based measurement data collected from different CZOs or by different groups (or disciplines) can be managed with the same data structure and use the same terminology for easy future data integration and dis covery.

3. Conceptual Model for the Critical Zone Geochemical Database

A conceptual model shown in Figure 3 is essentially an expanded version of the SMO with more details describing features of the geochemical data measured from CZOs. For example, depending on the heterogeneity of the regolith/soil conditions across the CZO and the purpose of the study, samples can be taken from one sampling site to represent the entire region or from multiple sites along a transect to study the spatial variation. In each sampling site (i.e. a “fuzzy point” or a small area with a nominally homogeneous condition), one or multiple cores (pits or wells) can be drilled for sampling. Each core can then have one or multiple depth intervals (layers). Therefore, the category “Location” in the SMO can be expressed at various levels of detail in the conceptual model. With such a hierarchical data structure, data sets with different levels of detail can be easily integrated by aggregating data from higher levels of detail (e.g. depth-interval) to a lower level of detail (e.g. core/pit/well). Similar concepts can be applied to other categories, such as “Sample”, which can be split into different levels of subsamples for different physical and/ or chemical treatments and analysis.

For convenience, we group all variables in the conceptual model into two categories: primary data and secondary data. The primary data comprises major information directly related to samples, including where (i.e. Location, Site, Core/Pit/Well, Depth-interval), when (Date/Time), and how (Methods) the samples were collected and what (Attribute) was measured or analyzed. The secondary data include data sources (who, i.e. data contributor or authors of journal articles), and data qualities (i.e. error terms). In addition, a set of controlled vocabularies (such as variable names, units, and sampling method or analysis technique names) are also included as part of the metadata to ensure semantic consistency of the database. This conceptual model provides a framework in which the Critical Zone geochemical DataBase (CZchemDB) is developed following the rules of the SMO.

4. Structure of the CZChemDB Database

Based on the conceptual model in Figure 3, we developed a relational database for critical zone geochemical data CZChemDB. The schema of the CZChemDB is shown in Figure 4, which consists of a total of 21 interrelated tables. The relationships between tables are established through primary keys (a unique identifier of the data entry in a table) and foreign keys (an identifier that links to the corresponding primary key of another table) are represented by arrows in the schema. In correspondence with the conceptual model, data tables in the schema are grouped as primary data, secondary data and controlled vocabulary tables. In the following, we will define and describe the major functions and content of each data group. Detailed description of table contents and table attribute contents are listed in Table 1 and Table 2, respectively.

4.1. Primary Data Tables

The primary data tables include Location, SamplingSite, CorePitWell, DepthInterval, Sample, SubSample, Analysis, and DataValue. A Location is defined as a study area that has relatively homogeneous environmental conditions and is used for certain research purposes. One example of a location is the Shale Hills CZO which features a small, temperate, forested catchment in central Pennsylvania in which the regolith is primarily developed from homogeneous shale. The table Location contains general information about the study area, including annual mean temperature and annual precipitation.

A SamplingSite is a “fuzzy point” (or a small area) within a Location where samples are taken. A sampling site might be one position along a hillslope or a position randomly selected from a Location, depending on the purpose of the research. The table SamplingSite stores geographical information about this “point” (latitude, longitude, datum, elevation, aspect) and other geological (exposure age), biological (vegetation) and landscape features (elevation, slope, landuse etc.).

At each sampling site (SamplingSite), one or multiple Cores/Pits/Wells (CPW) may be implemented to make measurements or collect samples. The table Core/Pit/Well describes the collection or implementation methods, date, time, and physical dimension of each CPW that has been drilled (augered, or dug). In addition, each CPW is registered in the System for Earth Sample Registration (SESAR) as a parent sample (or parent object) with a unique International Geo Sample Number (IGSN) for unambiguous data-sample linking or tracking between different published articles or data systems (http://www.geosamples.org).

Each CPW has one or multiple depth intervals (i.e. sampling layers or horizons) that may have different physical features, such as color and texture, from the layers above and beneath. Table DepthInterval stores information regarding the depths from some datum (i.e. the land surface or the interface between the mineral and organic horizons) to the top and

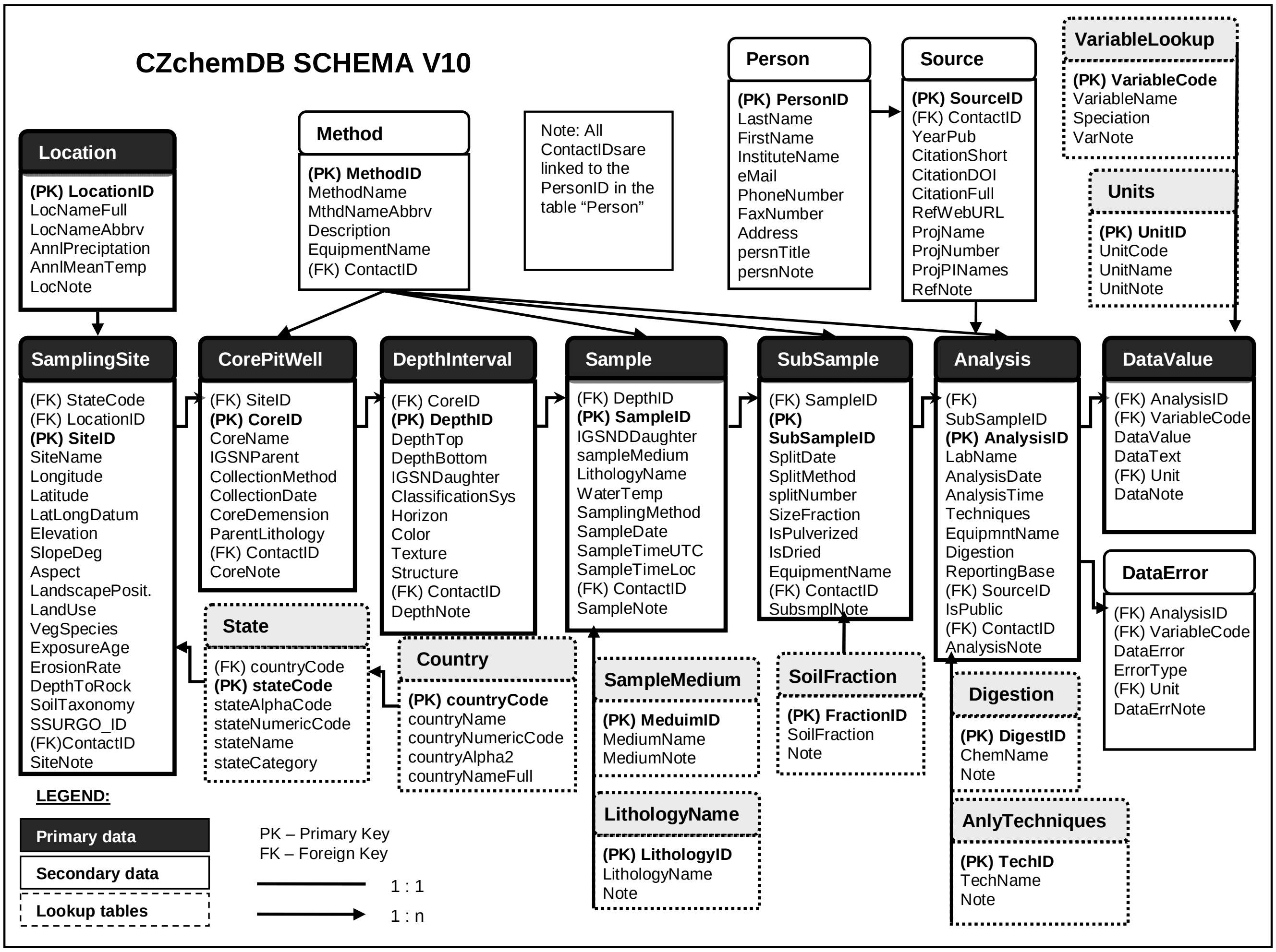

Figure 4. Schema of the CZChemDB database.

bottom of each layer, as well as soil physical characteristics such as horizon name, color, texture, and structure, if available. Each Depth-Interval is registered in the SESAR system as the "daughter object" of a CPW (the "parent object").

Sample is defined as the smallest "uniform" unit of raw material, such as bulk soil, rock, biota, gas, or liquid (water) collected from a DepthInterval for measurements of chemical, biological and mineral characteristics. Sample is the core concept of this database. Attributes in the Sample table include sampling methods, date/time, lithology name, and the sampling medium (soil, or water, or biota), which serves as a foreign key to link to the Sample-Medium table (i.e. an extension point to link to different media identified in the CZO ontology). Each sample is registered in the SESAR system as the daughter's daughter of a CPW.

Subsamples are normally splits of a Sample, which have been subjected to different physical or chemical splitting treatments before analysis. These treatments might include grinding /sieving, mineralogical separation, and others. Methods of splitting are included in this table or linked to a Method look-up table.

The Analysis table summarizes analytical methods applied to subsamples (or directly to samples) for chemical analysis. Given the myriad number of analytical strategies for both sample preparation and analysis, a decision was made to include only limited commonly used treatments applied prior to chemical analysis, such as digestion and ashing into this table. A foreign key, SourceID, is included in this table to link the analysis (data) to the source of a published article or directly refer to a person who is the data contributor.

The DataValue table stores the results of an analysis, i.e. analytical values and the associated error estimates for the analyzed items (element, oxides, isotope ratio, etc.). This is also the table that houses the information about what species have been measured: for example, the element of interest (e.g. Na or Sr), the isotope of interest (e.g. ^{56}Fe or ^{54}Fe), or the species of interest (e.g., nitrate versus nitrite).

4.2. Secondary Data

Secondary data tables mainly include tables describing data source (Source), sampling and analysis methods (Me-

Table 1. Description of Table Contents in the CZchemDB Database

Table Name	Description of Table Content
Location	List of locations. A location is any geographic area that has homogeneous environmental conditions, such as a watershed or a critical zone observatory (CZO)
SamplingSite	List of sampling sites. A sampling site is a "fuzzy point" where sampling cores or wells are drilled or pits are dug
CorePitWell	List of cores (or pits or wells) drilled from each sampling site
DepthInterval	List of intervals (also known as layers) that have homogeneous physical features, such as soil horizon, color, texture, and structure
Sample	List of samples taken from each intervals
SubSample	List of subsamples that are splits of a sample for different physical or chemical treatments/analysis
Analysis	List of methods used for lab analysis
DataValue	List of analytical values of measured element, oxide, isotope ratio, etc.
DataError	List of information about the precision/error of the analytical data
Source	List of bibliographical information and/or related project information
Person	List of persons that are referred to as authors, project PIs (principle investigator), or data contributors
Method	List of methods for core-collection, sampling, sample-splitting, analysis, etc.
Controlled vocabulary	
VCtl_Country	List of country names
VCtl_CollectionMethod	List of methods for core-drilling or pit-digging
VCtl_Digestion	List of chemical digestions
VCtl_LithologyName	List of lithology names
VCtl_ReportingBasis	List of reporting basis such as ashed or as-received
VCtl_SampleMedium	List of sample medium, such as soil, water
VCtl_SoilFraction	List of soil particle size
VCtl_State	List of state/province names
VCtl_TechniqueName	List of technique names for chemical analysis
VCtl_Units	List of data units
VCtl_VariableList	List of analyte and other physical and mineral attributes

Table 2. Description of Major Table Attributes in the CZchemDB Database

Field Name	Description	Type
Location		
LocationID	Unique identification number for a location	Integer
LocNameAbbrv	Abbreviation of a location name	Text
LocNameFull	Full location name	Text
AnnlPrecipitation	Annual precipitation of a location/watershed (mm)	Single
AnnlMeanTemp	Annual mean temperature of a location/watershed (^{o}C)	Single
LocNote	Comments	Memo
SamplingSite		
LocationID	A foreign key linked to the LocationID in the Location table	Integer
SiteID	Unique identification number of a sampling site (point)	Integer
SiteName	Name of a sampling site	Text
StateCode	A foreign key linked to the code of a State/Province in which a sampling site locates	Text
Longitude	Longitude of a sampling site in degree	Single
Latitude	Latitude of a sampling site in degree	Single
LatLongDatum	Spatial reference system of the latitude and longitude coordinates	Text
Elevation	Elevation of a sampling site – meters above sea level	Single
Slope	Slope of a sampling site in degree	Single
Aspect	Aspect of a sampling site	Text
LandscapePosition	Landscape position of a sampling site. e.g. ridgetop, hillslope, valley-floor, etc.	Text
LandUse	Landuse of a site. e.g. agricultural land, forest, urban, etc.	Text
VegSpecies	Vegetation genus and species	Text
ExposureAge	Exposure age, kys	Integer
ErosionRate	Erosion rate of a site, m/Myrs	Single

Table 2 (cont'd). Description of Major Table Attributes in the CZchemDB Database

Field Name	Description	Type
DepthToRock	Depth to bedrock, meters	Single
SoilTaxonomy	Soil taxonomy (U.S. NRCS soil classification system)	Text
SSURGO_ID	Identification number from the NRCS SSURGO soil database	Text
ContactID	A foreign key linked to the person (in table Person) who provides this site information	Integer
SiteNote	Comments	Memo
CorePitWell		
SiteID	A foreign key linked to the SiteID of the samplingSite table	Integer
CorePitWellID	Unique ID number of a core, pit, or well (CPW)	Integer
CorePitWellName	Name of a core, pit, or well	Text
IGSN_ParentCPW	International Geo Sample Number (IGSN) of a parent sample, i.e. core/pit/well (CPW)	Text
CollectionMethod	Description of method and equipment used to drill cores or dig pits	Text
CollectionDate	Date when cores (pits) were drilled, mm/dd/yyyy	Date/Time
CollectionTime	Local time when cores were collected, hh:mm	Date/Time
CorePitWellDimension	Core/Pit/Well dimension (or size)	Text
ContactID	A foreign key linked to the person (in table Person) who provides the information about cores/pits	Integer
CoreNote	Comments	Memo
DepthInterval		
CorePitWellID	A foreign key linked to CorePitWellID in the CorePitWell table	Integer
DepthIntervalID	Unique identification number of a depth interval (or a layer)	Integer
IGSN_DaughterCPW	IGSN of a depth interval which is a daughter of the parent sample (i.e. a CPW)	Text
DepthTop	Depth from the surface to the top of a depth-interval, cm	Single
DepthBottom	Depth from the surface to the bottom of a depth-interval, cm	Single
ClassificationSystem	A system for soil characteristic classifications (e.g. U.S NRCS system)	text
Horizon	Horizon name of a depth interval (e.g. Ap, EB, Bt1, C, etc.)	Text
Color	Color of a depth interval (e.g. 7y/r)	Text
Texture	Texture class of a depth interval	Text
Structure	Structure class of a depth interval	Text
ContactID	A foreign key linked to the person (in table Person) who provides the information about the depth-interval	Integer
DepthNote	Comments	Memo
Sample		
DepthIntervalID	A foreign key linked to the depthIntervalID in the DepthInterval table	Integer
SampleID	Unique ID number of a sample taken from a depth interval	Integer
IGSN_DDaughterCPW	IGSN number of a sample, which is the daughter's daughter of the parent sample (i.e. CPW)	Text
SamplingMethod	Description of the sampling method	Text
SamplingDate	Date when samples were taken, mm/dd/yyyy	Date/Time
SampleLocTime	Local time when samples were taken, hh:mm	Date/Time
Medium	Medium (material) name of a sample, e.g. soil, water, rock, etc.	Text
WaterTemp	Water temperature if a liquid sample	Single
LithologyName	Lithology name of a sample	Text
ContactID	A foreign key linked to the person (in table Person) who provides the information about the sample	Integer
SampleNote	Comments	Memo
SubSample		
SampleID	A foreign key linked to SampleID in the Sample table	Integer
SubSampleID	Unique ID number of subsamples, i.e. splits of a sample for different treatments or analysis	Integer
SplittingMethod	Description of method to split samples	Text
SplittingDate	Date when samples were split, mm/dd/yyyy	Date/Time
Fraction	Soil particle size (e.g. <2mm, bulk, etc.)	Text
IsPulverized	Sample is grinded (yes/no)	Text
IsDried	Drying status of samples (dried or as received)	Text
ContactID	A foreign key linked to the person (in table Person) who provides the information about the subSample	Integer
SubSampleNote	Comments	Memo

Table 2 (cont'd). Description of Major Table Attributes in the CZchemDB Database

Field Name	Description	Type
Anaylsis		
SubSampleID	A foreign key linked to the SubsampleID in SubSample table	Integer
AnalysisID	Unique ID number of an analysis	Integer
TechniqueName	Description of techniques used for analysis	Text
Digestion	Description of sample digestion methods	Text
ReportingBasis	Reporting basis (e.g. ashed or as-received)	Text
Equipment	Equipment used for analysis	Text
LabName	Name of laboratory where analysis is conducted	Text
AnalysisDate	Date of sample analysis, mm/dd/yyyy	Date/Time
SourceID	A foreign key linked to the SourceID in the Source table where data source information is given	Integer
IsPublic	Status of data: public or private	Yes/No
ContactID	A foreign key linked to the person (in table Person) who provides the information about the Analysis	Integer
AnalyNote	Comments	Memo
DataValue		
AnalysisID	A foreign key linked to the AnalysisID in the Analysis table	Integer
VariableCode	Variable (analyte) code	Text
DataValue	Data values	Single
UnitCode	Data units	Text
DataNote	Comments	Memo
DataError		
AnalysisID	A foreign key linked to the AnalysisID in the Analysis table	Integer
VariableCode	Variable Name (Code)	Text
DataError	Errors of a dataset	Single
ErrorType	Type of error terms, including absolute error and relative error, etc.	Single
UnitCode	Data units	Text
ErrorNote	Comments	Memo
Source		
SourceID	Unique ID number of a data source	Integer
ContactID	A foreign key linked to the person (in table Person) who is either the data contributor or corresponding author of an article or the PI of the related project	Integer
CitationShort	Short citation of a publication (journal article or book, etc.)	Text
CitationFull	Full citation of a publication	Text
CitationDOI	Digital Object Identifier of an article or a book	Text
RefWetURL	A website link of a reference	Text
ProjectName	Name of the Project related to the dataset	Text
ProjNameAbbrv	Abbreviation of a project name	Text
ProjectNumber	Project number	Text
ProjectSponsor	Project sponsor's name	Text
ProjectPIname	Project PI names	Text
SourceNote	Comments	Memo
Person		
PersonID	Unique ID number of a person	Integer
LastName	Last name	Text
FirstName	First name	Text
InstituteName	Institution name	Text
DepartmentName	Department name	Text
EMail	Person's eMail address	Text
PhoneNumber	Person's contact phone number	Text
FaxNumber	Fax number	Text
PersnAddress	Mailing address	Text
PersnTitle	Person's Title	Text
PersnNote	Notes	Memo

Table 2 (cont'd). Description of Major Table Attributes in the CZchemDB Database

Field Name	Description	Type
Method		
MethodID	Unique ID number of a method	Integer
MethodName	Name of a method	Text
MethdAbbrv	Abbreviation of a method name	Text
EquipmentName	Equipment names	Text
MethdDescription	Description of a method	Memo
ContactID	A foreign key linked to the person (in table Person) who provides the information about the method	Integer
MethdNote	Comments	Memo

thod), and data qualities (DataError). The Source table lists bibliographical information for publications reporting the relevant data. This information includes the title, journal, volume, page number, authors, and/or the citation DOI. In addition, data source can also include project-related information (principal investigator, sponsor, project number) or be a listing of people who provide the data (i.e. data contributor).

The Method table lists most commonly used methods and descriptions for core (pit, well) collection, sampling, sample splitting and preparation (physical or chemical treatments) and analytic techniques. The Method table is used to either directly link to a primary data table for detailed method description or serve as a lookup table for certain techniques that are directly incorporated into the respective primary data tables such as collection methods in CorePitWell table and analytical techniques in the Analysis table. Data quality information, provided in the DataError table, includes detection limits, standard deviations or absolute/relative errors.

4.3. Controlled Vocabulary (Lookup Tables)

To unify and limit terms used for variable names, units, and techniques, we have also created a set of tables to store controlled vocabulary for certain table attributes. These tables include Country, State/Province, Analyte and other variables (physical, chemical, and mineralogical characteristics), Unit, Sample Medium, Soil Fractions, Digestion, Analytical techniques, and Lithology names. The terms, words, and phrases used in the controlled vocabulary tables have been selected and compiled from existing databases such as CUAHSI HIS and EarthChem, and through communication among the critical zone geochemistry scientists. Each of the terms or phrases within the controlled vocabulary has a unique meaning to eliminate ambiguous interpretations of the variable names, units, and techniques. Additionally, using the controlled vocabulary can help to improve the accuracy and performance of data searching and data integration. One of the goals of this paper is to provide this controlled vocabulary in print format for the CZ community, which is available online in the Appendix (I ~ IV).

5. Implementation of the CZchemDB Database

The schema in Figure 4 represents the conceptual (or relational) structure of the CZchemDB database. It can be physically implemented in any relational database management system. In the following, we will discuss its current implementation in the Microsoft Access database management system and the future integration into an online geochemical database portal (i.e. EarthChem) and the CZO data center for broader web accessibility.

5.1. Implementation of the CZchemDB with Microsoft Access

Microsoft Access (MS Access) is a relational database management system that comes as part of the Microsoft Office package. During a pilot application phase, we successfully implemented the structure of the CZchemDB in MS Access®. To date (October 2013), there are more than 35,000 data values populated in the database. The database includes 276 soil cores (or soil pits) from 47 unique locations, including data from the Susquehanna Shale Hills CZO, Luquillo CZO, Jemez River Basin and Santa Catalina Mountains CZO, and Boulder Creek CZO. In addition to the CZO data, public datasets from published papers designating critical zone chemistry have been also compiled and are included. Currently, 23 scientists from 4 CZOs have contributed their data into the master database. The current MS Access version of the CZChemDB is available online at http://www.czo.psu.edu/data_agreement.html.

5.2. Incorporation of the CZchemDB into EarthChem and the CZO Data Center

The CZchemDB has been designed so that it can be seamlessly integrated into the web-based EarthChem portal to increase its accessibility and reusability. The EarthChem portal provides a central access point for geochemical data of distributed partner databases, which use EarthChemXML as the common language (format) to describe respective primary data and metadata. These partner databases currently include data for ocean floor sediment samples and rock samples worldwide, but none of these are designed for geochemical data measured from soil samples. Given the unique characteristics of the CZO soil geochemical data as discussed above, the EarthChem team is working on the extension of the EarthChemXML to accommodate the geochemical data from CZ-

ChemDB for integration into EarthChem system. As part of this integration process, for example, we have registered samples collected from the Susquehanna Shale Hills Critical Zone Observatory at the System for Earth Sample Registration (SESAR, www.geosamples.org) to obtain the unique International Geo Sample Number (IGSN, http://www.igsn.org) that will allow advanced data integration in the EarthChem data discovery system. A web-version of the CZchemDB database and integrated access of CZchemDB data at the EarthChem portal is expected to be available in 2015.

Meanwhile, the EarthChem team is working with the national CZO data team to build interoperability between EarthChem and the future CZO data sharing infrastructure. Until it is integrated into EarthChem, the MS Access version of the CZchemDB will be maintained and available through the national CZO website for download by individual users or for harvest in the future by the CZO centralized data system.

5.3. Data Management Tools and Applications

In an effort to facilitate the CZO data preparation, an MS Excel template file has been created with built-in self-explanation and self-validation fields so that users can easily enter their data through drop-down lists to comply with the controlled vocabulary generated from the SMO ontology. In the front-end of the MS Access version of the CZchemDB database file, an interface and tools were also developed to easy data exchange between MS Access and MS Excel programs. These include functions to automatically import/append new datasets from the Excel template files to the Access database and tools to export queried datasets from the Access database to Excel for future analysis (Appendix V).

To demonstrate the usefulness of this database for data searching and integration, we consider a typical query and the subsequent results. For example, if an investigator were interested in data for soil samples from reaching a certain depth (e.g. > 1 m), located in any CZOs, and wanted to get the chemistry available by depth interval for those specific soils, a simple search by querying "LocNameFull = like "*CZO*" from table Location and "DepthBottom > 100 cm" from table DepthInterval (with all related tables linked) would provide the data. Using our current available database, query results indicate 92 total cores with depths greater than 100 cm with 26 of those cores located at CZOs. A further query of those 26 CZO cores, to the variable level, reveals various subsets of 61 analytes were measured on those soils such that a total of 3,420 data values are available. When a specific analyte (e.g. P_2O_5) is queried for the same 26 cores, the search results display the information that P_2O_5 data are available for 12 of the 26 cores. Interestingly, the meta-data (methods) shows that analytical techniques which produced the phosphate data are also different, which provides more useful and sometime critical information for data (or cross-site) comparison. Among those twelve cores, seven were analyzed by the inductively coupled plasma atomic emission spectroscopy (ICP-AES) technique following a lithium metaborate (LiBO4) digestion, while the other five cores utilized X-ray fluorescence (XRF). If land use was of interest to the user, than a query of "landuse" from the table for SamplingSite produces all desired datasets measured from Forest Land, Rangeland, Urban/Industrial, or Agricultural Land. At this time, the database does not require all metadata fields, yet provides the opportunity for investigators to archive/preserve the metadata collected during field campaigns and laboratory investigations, ultimately increasing the search capacity.

In addition to easy data management, searching, and sharing, another important part of this study is to use the database to help investigators solve problems related to critical zone science. One example of such applications is the development of a τ (tau) calculator (Appendix V), which is part of the front-end of the CZchemDB Access database. With the tau-calculator, one can easily calculate τ values for any selected locations (for cross-site comparison) or for any element of interest to model weathering processes (Brimhall and Dietrich, 1987). In the tau-calculator, there are two options for parent layer (i.e. parent material) selection: the deepest layer from the same soil core or any reference layer of user's choice. When users choose to use a reference layer (from any soil cores) as the parent layer, the program automatically add such layer to the profile of soil of interest as a virtual deepest layer, and then calculate the tau values the same way as using the deepest layer as parent materials.

6. Discussion and Conclusions

The importance of data management, data sharing and discovery are increasingly recognized by scientific communities. Many funding agencies, such as the National Science Foundation, now require specific data sharing and management plans for grant applications. To meet the data management challenges, an initial data sharing infrastructure for critical zone observatory was proposed (Zaslavsky et al., 2011). As part of the datasharing infrastructure, EarthChem is designated to be the portal to accommodate critical zone soil/regolith geochemistry data.

The EarchChem portal provides a central access point for geochemical data of distributed partner databases. In this study, we developed a relational database for CZO soil chemistry data, namely CZchemDB, which will be integrated into the EarthChem system in the near future. The broader impacts and significance of the CZChemDB are multifold. First, the database complements current EarthChem data systems (PetDB, SetDB, and NavDat and GEOROC) with an integrated compilation of geochemical data of samples from the Critical Zone. Second, success in integration of CZChemDB with EarthChem will enable all rock and regolith geochemical data collected from CZOs and other sites to be accessed, discovered, and reused by a diverse community now and in the future. Finally, this effort will establish a model to bridge the connections between data acquisition, data management, data sharing, and data discovery that are all essential but weak in terms of linkages within most geosciences research projects.

Currently, the CZchemDB is implemented in a standalone MS Access relational database management system for individual and small group uses. The MS Access is selected because 1) it is part of the MS Office package and users do not need to invest extra money for the software; and 2) it is convenient for individuals or small research groups to start organize and populate their ongoing research data into the database before it becomes public. In addition, advanced users can take advantage of the mature techniques built in the MS Access to create tools/macros to facilitate data manipulation and exchange between MS Access and MS Excel, a program that is often used by (and is familiar to) current CZO scientists for data management and analysis. The Excel template file we developed in this study is one of examples of our efforts to create a user friendly data management environment.

Beside the different type and format of data files, heterogonous data structure and ambiguous terminology used in the database is another big challenge related to data sharing and integration. An ontology is often used as a generalized data model to help characterize the context, clarify the relationships, and unambiguously interpret the inherent "meaning" of datasets in a domain; therefore, the ontology is used to link or integrate different databases (Madin et al., 2008). Similarly, we developed a sample-based measurement ontology (SMO) in this study to describe concepts and their relationships of the sample-based measurements of critical zone characteristics. However, instead of developing a formal ontology database, we first used the SMO as guidance to structure a relational data model for critical zone soil geochemical data. This also sets an example for other sample-based observations, such as chemistry data measured from liquid (e.g. porewater) and biological samples (e.g. plant material). In the future, the SMO, in combination with the critical zone ontology, can be used to easily integrate different sample-based measurement databases among CZOs and among different disciplines.

In the long run, geochemists may desire to set standards for CZO geochemical data. We have not broached the subject here of data standards in terms of quality control or in terms of agreed-upon analysis protocols. Obviously, if we compare geochemical data across CZEN sites (http://www.czen.org), it would be best to agree upon criteria for sampling, preparation, and analysis. Such agreement is notoriously difficult to achieve. This lack of agreement can be understood in terms of Figure 1b and 1c: much of the geochemical analysis is focused upon the development of new analytical tools. Thus, measurement protocols are always changing to allow better precision and accuracy and to allow analysis of new analytes. This inherent characteristic of geochemical research is antithetical to the establishment of measurement standards. However, constantly evolving protocols make it hard to make comparisons of CZ behavior as a function of variables that range across environmental gradients.

In this regard, we follow the philosophy that the first step should not be establishment of such protocols: *that agreement will simply be too difficult in advance of CZ geochemists recognizing the utility of sharing data*. On the other hand, as demonstrated by the success of EarthChem and its publication record, as data sharing becomes the norm within the CZ community, it is possible that protocols for sampling collection, preparation, and analysis will be developed as the community itself shares data and begins to demand standards. This latter situation must occur before a database developer or team can require such protocols. As described in the literature (Brantley, 2007; Hofmockel et al., 2007), the geochemical community is not quite ready yet for such demands. In the future, some balance will likely be achieved between standards and growth of new techniques and protocols.

Finally, we conclude that CZchemDB is a relational database to accommodate soil/regolith geochemical data collected from CZOs. Comparing with its previous version (V6) published in Applied Geochemistry (Niu et al., 2011), the new version of CZchemDB (V10) is modified to be fully compliant with the principle of the sample-based measurement ontology (SMO) developed in this study. It is critical because this update would allow users to easily integrate such geochemical data with other sample-based measurements collected from different locations and/or from different media (e.g. samples from water, vegetation, etc.) for cross site comparison and integrated analysis. Other updates of CZchemDB V10 include additions of "CorePitWell" and "DepthInterval" tables to increase levels of detailed information of sampling locations, elimination of "Preparation" table (combined with "Analysis" table), and simplification of "Reference" related tables for users' data input convenience. When compared with plain text files or spreadsheet files that are often used by CZ scientists, the CZchemDB also has many advantages, including the capability of storing multiple values for the same data field (e.g. analytical values of different elements under the same analysis) and its powerful functionality to search, manipulate, and reorganize data for future analysis. More importantly, the integration of the CZchemDB with EarthChem and the CZO data-sharing infrastructure would greatly increase its accessibility and reusability to a broader science community. Finally, we emphasize that, although we only discussed the application of the CZchemDB to CZO soil geochemical data in this paper, any other sample-based dataset, including soil pore water data, vegetation chemistry, and other such biological data may also be usefully stored in the database. Thus CZChemDB might grow in the future to contain expanded data types.

Acknowledgments. This research was supported in part by the U.S. National Science Foundation under grants EAR-0533644; EAR-0725019; and EAR-0819857. The authors would also like to thank colleagues from other CZOs for their valuable comments and suggestions during the development of the CZchemDB.

References

Brantley, S.L. (2007). Why geochemists never agree: It's all about the data! *Elements.* 2.

Brantley, S.L., Goldhaber, M.B., and Ragnarsdottir, K.V. (2007). Crossing disciplines and scales to understand the critical zone. *Elements,* 3(5), 307-314. http://dx.doi.org/10.2113/gselements.3.5.307

Brimhall, G., and Dietrich, W.E. (1987). Constitutive mass balance relations between chemical composition, volume, density, porosity, and strain in metasomatic hydrochemical systems: Results on weathering and pedogenisis. *Geochim. Cosmochim. Acta,* 51, 567-587. http://dx.doi.org/10.1016/0016-7037(87)90070-6

Hofmockel, M., Richter, D., Miller, D., and Brantley, S.L. (2007). Building critical zone research cyberinfrastructure. *EOS Trans. AGU,* 88(50), 560. http://dx.doi.org/10.1029/2007EO500005

Horsburgh, J.S., Tarboton, D.G., Maidment, D.R., and Zaslavsky, I (2008). A relational model for environmental and water resources data. *Water Resour. Res.,* 44(5), W05406. http://dx.doi.org/10.1029/2007WR006392

Horsburgh, J.S., Tarboton, D.G., Piasecki, M., Maidment, D.R., Zaslavsky, I., Valentine, D., and Whitenack, T. (2009). An integrated system for publishing environmental observations data. *Environ. Model. Software.,* 24(8), 879-888. http://dx.doi.org/10.1016/j.envsoft.2009.01.002

Lehnert, K., Su, Y., Langmuir, C.H., Sarbas, B., and Nohl, U. (2000). A global geochemical database structure for rocks. *Geochem. Geophys. Geosyst.,* 1. http://dx.doi.org/10.1029/1999GC000026

Lehnert, K.A., Goldstein, S.L., Johansson, A., Murray, R.W., Pisais, N., Vinayagamoorthy, S., and Djapic, B. (2007a). SedDB - A new information system to facilitate use of marine sediment geochemistry in science and education. *MARGINS Newsl.,* 18, 9-11.

Lehnert, K.A., Langmuir, C.H. (2007b). The PetDB data collection: Impact on science. *Geol. Soc. Am. Programs Abstracts*, 39(6), 153.

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F. (2007). An ontology for describing and synthesizing ecological observation data. *Ecol. Inf.,* 2(3), 279-296. http://dx.doi.org/10.1016/j.ecoinf.2007.05.004

Madin, J.S., Bowers, S., Schildhauer, M.P., and Jones, M.B. (2008). Advancing ecological research with ontologies. *Trends Ecol. Evol.,* 23(3), 159-168. http://dx.doi.org/10.1016/j.tree.2007.11. 007

National Research Council (U.S.). Committee on Basic Research Opportunities in the Earth Sciences. (2001). Basic research opportunities in earth science. National Academy Press, Washington, D.C.

Niu, X., Lehnert, K.A., Williams, J., and Brantley, S.L. (2011). CZChemDB and EarthChem: Advancing management and access of critical zone geochemical data. *Appl. Geochem.,* 26, S108-S111. http://dx.doi.org/10.1016/j.apgeochem.2011.03.042

Sarbas, B., Nohl, U. (2008). The GEOROC database as part of a growing geoinformatics network. In: Brady, S.R., Sinha, A.K., Gundersen, L.C. (Eds.), Geoinformatics 2008 - Data to Knowledge, Proceedings: US Geological Survey Scientific Investigation Report 2008-5172, 42-43.

Whitenack, T., Zaslavsky, I., Williams, M.W., Lehnert, K.A., Tarboton, D.G., Schreuders, K., Aufdenkampe, A.K., and Mayorga, E. (2011). Prototype cross-domain cyberinfrastructure for the Critical Zone Observatories. Abstract IN11C-1305 presented at 2011 Fall Meeting, AGU, San Francisco, CA.

Zaslavsky, I., Whitenack, T., Williams, M., Tarboton, D.G., Schreuders, K., and Aufdenkampe, A. (2011). The initial design of data sharing infrastructure for the critical zone observatory. Proceedings of the Environmental Information Management Conference, Santa Barbara, CA, 145-150.