CHEMCONNECT: AN ONTOLOGY-BASED REPOSITORY OF EXPERIMENTAL DEVICES AND OBSERVATIONS

Edward S. Blurock

Blurock Consulting AB, Lund, Sweden

ABSTRACT

CHEMCONNECT is an ontology cloud-based repository of experimental, theoretical and computational data for the experimental sciences domain. Currently, the emphasis is on the chemical combustion community, but in future work (in collaboration with domain experts) the domain will be expanded. CHEMCONNECT goes beyond traditional meta-data annotated scientific result repositories in that the data is parsed and analysed with respect to an extensive chemical and combustion knowledge base. The parsed data is then inter-linked allowing for efficient searching and comparison. The goal is to link all data associated with experiments, including the device description, the intermediate data (both computed and measured), the associated interpretations, procedures and methodologies used to produce the data and the final published results and references. Having published data linked to its dependent measurements and constants, devices, subsystems, sensors and even people and laboratories provides an effective accountability and more confidence in the data. Data entry and availability can range from private user, to user defined consortia to general public. These concepts are implemented at http://www.connectedsmartdata.info.

KEYWORDS

Case Study, Ontology, Repository, Database, Experimental Devices, Experimental Results

1. INTRODUCTION

With the advent of the explosion of data produced within a project and with the increasing pressure to have an efficient means to communicate and disseminate this data, the role of a well defined data management plan within a proposal is expanding and even becoming mandatory. An essential part of a data management plan is a sustainable and manageable repository of data. CHEMCONNECT and its fundamental principles with an extensive knowledge base through the use of ontologies represents the next generation of data repositories. Data files are not just stored. The data within those files are interpreted and put into the context of the extensive domain knowledge. CHEMCONNECT provides the framework for extensive documentation of the entire scientific process from thorough descriptions of the measurement devices (including subsystems, components and sensors), documentation of the methodologies, from calibration to measurement to manipulation to final results, and research protocols linking together all aspects of the scientific process.

The goal of CHEMCONNECT is the be the tool at the center of the research ecosystem[1] and to promote the FAIR concept, that data is Findable, Accessible, Interoperable and Reusable [2]–[4]. The further goal of CHEMCONNECT is to provide the framework of experimental

Dhinaharan Nagamalai et al. (Eds) : SAI, NCO, SOFT, ICAITA, CDKP, CMC, SIGNAL - 2019
pp. 115-130, 2019. © CS & IT-CSCP 2019 DOI: 10.5121/csit.2019.90709

documentation, with protocols, methodologies and device descriptions, for Transparency and Openness Promotion (TOP[5]) trend in scientific publishing.

A further goal of CHEMCONNECT is to provide a natural framework for data entry through its cloud-based interface. The framework is meant to adapt to the researcher needs. CHEMCONNECT should accept data, including units and formats, as the researcher naturally presents them. The extensive knowledge base of CHEMCONNECT plays an essential role in achieving this.

What separates CHEMCONNECT from a traditional repository is the following:

KNOWLEDGE BASE: CHEMCONNECT distinguishes itself from multidisciplinary experimental data repositories with simple meta data[6]–[9]by having an extensive knowledge base of experimental devices, protocols and data, created in collaboration with experimentalists and modelers in the field. This knowledge base is represented in an extensive interconnected network of concepts and data. These concepts allow for convenient uploading of data, its subsequent interpretation and efficient search of chemical information. The knowledge is represented by an ontology[10], [11].

ADAPTS TO USER: The knowledge base gives context and meaning to the data that is uploaded to the database. It also allows the experimentalist and modeler to upload data in their format convenient to them. The goal of CHEMCONNECT is not to restrict input to any particular format. The knowledgebase puts all data, regardless of the original input format, on a common foundation.

INTERCONNECTION OF DATA: The interlinking of data and concepts[12] within CHEMCONNECT facilitates efficient and thorough search for data within the repository. The interface promotes linking data to the devices (even sensors and subsystems of the device), protocols (methodologies and procedures used to generate the final results), researchers (the institution, the lab and even who performed the experiment) and external sources and references (websites, publications and entries in other databases).

2. CONCEPTUAL STRUCTURE

At a conceptual level, CHEMCONNECT consists of the interaction of four levels of data, knowledge and visualization:

KNOWLEDGE BASE: A representation (through an ontology) of the knowledge related to experimental devices and data. This knowledge base gives information on how data should be read and how to present the information.

PERSISTENT STORAGE: The original source data, in the input formats provided directly from the researchers is stored in the persistent storage of the repository. These files are supplemented, even before interpretation, with supplementary data such as descriptions, purpose, general concepts and links to references, devices, organizations, external links, etc.

DATABASE: The database consists of two types of information provided by users:

SPECIFICATIONS: From templates from the knowledge base, specific user defined specifications are defined. These provide details about how the source data should be read in and interpreted, including where in the original data file the stored data is to be found, which parameters are used, their units and error specification.

DATA: Using the user defined specifications and the knowledge-base, user data is interpreted and entered into the database in a form providing a broader context to the data for efficient visualization, search, comparison and validation.

CLOUD-BASED USER INTERFACE: Regardless of physical location or device, the data can be inputted, visualized, compared and searched. The knowledge-base guides the interface for effective input, visualization and comparison of data.

2.1. KNOWLEDGE BASE ONTOLOGY

The knowledge-based ontology provides the generic knowledge and structural patterns on which specific instances catalog specifications and entities can be set up and entered in the database. For example, the Knowledge Base steers and sets up the user interface so the generic values can be replaced by specific values. The information inentities, such as devices, organizations or people, are filled in, stored in the repository and represent information about the specific devices, organizations and people. The data of specifications steer how source data is interpreted by given specific information about, for example, units or correspondences to source data.

The knowledge of the knowledge-base is captured using an ontology representation[13]from the semantic web[14].

2.2. PERSISTENT STORAGE

This is the repository of data files, images, spreadsheets, documents uploaded by the user. These files represent the traditional repository aspect of the CHEMCONNECT database. What separates CHEMCONNECT from traditional repositories is that with the help of the knowledge base, these files are parsed, interpreted and the information within these files are given a context within a larger set of concepts.

2.3. DATABASE

The database has the individual pieces of data parsed from the original data. The knowledge base provides templates and conceptual context to the data.

The database has two basic types of data. The first is the interpretation of the parsed information of the data inputted by the user. This could be observational data (matrices of results, observations, etc.), information about users, devices, organizations, etc. The second is the represents individualized templates, defined by the user, that makes the parsing of observational data files possible. These database objects help optimize and automate the input process.

2.4. USER INTERFACE

The cloud-based user interface provides the means for the user to interact with the repository and database, irrespective of device, operating system and user location. The knowledge base steers the user interface using the knowledge of data types and data context.\

3. ONTOLOGY STRUCTURE

Ontologies, or vocabularies[13], are used to classify the terms that can be used in a particular application, characterize possible relationships, and define possible constraints on using those terms. In CHEMCONNECT, the ontologies are used to represent the experimental knowledge and concepts needed to describe the elements of the database, drive the user interface and give context to all the data within the database.

3.1. CATALOG STRUCTURES

The primary database objects, describing the major elements of the data within the repository, are represented as catalog (a term associated with the DCAT ontology[15]) items within the data base. The primary objects of the CHEMCONNECT database are:

CATALOG HIERARCHY: The (directory) structure of the repository of original data and virtually with regard to all the derived data structures in the database. Each data item is put into a directory structure giving it further context. It can be likened to a file structure.

CONTACT DATA

PERSON: These are researchers connected or contributing to the repository

ORGANIZATION: Organization associated with data in the repository

DEVICE AND COMPONENT DATA

INSTRUMENT: The instrument producing the data

SUBSYSTEM: Instruments and devices are viewed as collections of subsystems

COMPONENTS: sensors, valves, baths, pistons, rotors, nozzles, etc.

SPECIFICATION/INTERPRETATION KNOWLEDGE BASE

PROTOCOL: Set of observations associated with a methodology protocol

CORRESPONDENCES: Correspondence between an observation and the source data parameters.

OBSERVATION SPECIFICATION: A specification of an observation consisting of several parameters include the type of parameter and its units

BLOCK DEFINITION: Isolation specification of a block of data within a source.

OBSERVATIONS

SINGLE DATABASE OBSERVATION: The result of interpretation of a source into a set of data parameters connected to the knowledge base.

ISOLATED MATRIX OBJECT: A matrix of data (associated with a data specification)

FULL MATRIX OBJECT: A matrix of data (typically from the original source)

There are basically three levels of data structures:

CATALOG STRUCTURES: These are the top-level data structures representing the main objects of the database.

RECORD STRUCTURES: The catalog structures are made up of a set of record structures (from **DCAT:RECORD** from the DCAT ontology). Each record structure contains several pieces of 'primitive' information.

PRIMITIVES: Primitive structures are single pieces of information within a record. They are basically a single (string) word.

3.2 COMMON INFORMATION FOR CATALOG OBJECTS

Catalog entities are the top-level data structures holding the repository data and specifications. In keeping with the philosophy of CHEMCONNECT to provide as many links and additional data to repository data as possible, some common information fields are available to supplement the information of each catalog entity.

Each catalog entity supplements their information with:

DATA LINKS: Data is never an independent entity and can be, for example, created by, defined by, related to, derived from or used by other entities in the repository. A link is a keyword specifying what type of link is to be made.This link type is chosen from the knowledge base of concepts.

REFERENCES: These are publication references, for example, with corresponding DOI number referring to the catalog entity. This could be the source of the data, where the device description is found, etc.

WEB LINKS: Another form of referencing could be from sites on the internet with further information as to source, description, etc. of the catalog entity.

DESCRIPTION (TITLE AND TEXTUAL SUMMARY): There are also fields for textual description, both as a title or a longer abstract and even keywords. This information also contains the accessibility of the entity and when it was created (which may be different than when the data entry was created).

DATA CONCEPT AND PURPOSE KEYS: These specific concept keywords, extracted from the knowledge-base hierarchy of concept keys and purpose keys, give added domain context to the data. These are standardized keywords which give added information about the nature of the data and also gives context of the data in relation to other data that lie in the same place in the concept tree.

OWNER AND VISIBILITY: The owner of the data is the researcher that originally entered the data. It is the owner who determines the visibility and editability (including removal) of the data by other users. The visibility can be just the owner, a consortium of users or even public.

PLACE IN CATALOG HIERARCHY: The place in the catalog hierarchy has three entries:

1) The base position in the hierarchy. This corresponds to a directory structure of data objects.
2) The second field indicates what type of data the object represents. This data key is determined/selected from the knowledge base.
3) The simple name of the data object (that is unique for the position within the hierarchy).

4. PARAMETERS

Parameters and their values are the essence of data in a data repository. One of the main purposes of a parameter is the condensation of the complex reality into a single value, for example an attribute describing a device or an experimental observation of a complex process. Within collaboration between researchers a parameter's utility is in the comparisons of the 'same' parameter in other similar situations elsewhere. Within a data source (file) the 'meaning' of a parameter is only implicitly implied by its context, i.e. in which data set it is found and with more information implied by the name, such as a column name. However, concepts rely on human interpretation, especially when they are to be related to other 'similar' or even the same parameters in other data sets.

The goal of the knowledge base of CHEMCONNECT is to formalize the concepts (through the use of ontologies) and take one step closer is automating the ambiguity and comparability of parameters coming different sources.

To illustrate, let us look at a parameter giving a temperature value. What temperature that is intended can usually be interpreted (not necessarily automatically) from the name of the parameter used, such as ‘experimental temperature’, ‘water bath temperature’, ‘measured temperature’, etc. The context, i.e. in which data source it is found, also provides information as to the intention and meaning of the parameter. Though the name label implies meaning, there are no universal standards, even within a domain community there is rarely a consensus. For example, just the label ‘temperature’ could be given as simply as ‘T’ or ‘Temp’ or even spelled out completely with or without capital letters. The label could also be complicated with the units within the label. Which brings up another source of ambiguity, namely the units used for the value of the parameter. The difficulty lies not in the ambiguity of the value, but in the comparison with other similar reported results which may not be in the same units. Though SI units should be used as the units of choice, sometimes for historical or convenience reasons within a community, they are not used. And even if an SI unit is used, there is still of choice of, for example as the case with energy, joules, millijoules, kilojoules, etc.

What the knowledge-base of CHEMCONNECT (through the ontology representation) does is to give a (standardized) parameter context and relationship to other information. The context of the parameter is given by the parameter’s placement of a hierarchy of concepts. For example, the pressure measurements in a rapid compression machine (RCM) experiment (RCM Compression Pressure) can be found in the hierarchy under RCM Pressure Measurement (all pressure measurements of a rapid compression machine) which in turn is under Pressure Parameter (all pressure measurements). In addition, the purpose of the RCM Compression Pressure is labeled as a fundamental experimental measurement (Fundamental Experimental Measurement) and the general concept associated with the parameter is that it is a experimental measurement of the rapid compression machine (RCM Experimental Measurement). The unit expected for the parameter is given in the parameter specification as a unit class, in this case being pressure (Pressure Or Stress Unit). It is only when the actual parameter is given that the specific units are specified, for bar or atmospheres. The knowledge base helps in conversion (based on the QUDT ontology[16]) by having conversions between the different units of the class.

A parameter is used and defined on several levels:

PARAMETER SPECIFICATION: Within the knowledge base, a parameter type with a specific label is defined with the unit type (unit class), uncertainty value type, a purpose, a concept and whether it is an input (dimension) or an output (measure).

ATTRIBUTE: Within a device definition as a description or in an observation as a parameter, the parameter concept is specified.

VALUE SPECIFICATION: In the definition of the observation specification further specification of the parameter is made through the specification of the specific unit of the parameter value and the correspondence to the source parameter.

VALUE: The actual value corresponding to the specification.

4.1. PARAMETER SPECIFICATION

The set of parameters used in the knowledge-base ontology is found within a tree of parameter concepts. At the top level, there are two fundamental kinds of parameters:
FIXED PARAMETER: This is where the label is used and is the parameter name.

DYNAMIC PARAMETER: This is where the label is specified dynamically in its use. A typical example is a chemical species parameter with the labels being the chemical species names.

The parameter specification is:

LABEL: This is the label used to identify the type of parameter, for example, within other specifications such as observation and device parameter specifications.

UNIT CLASS: This is class of units of the parameter (from the annotated QUDT ontology).

TYPICAL VALUE: A typical (default) value and unit for the parameter is given.

PURPOSE: The purpose of the parameter

CONCEPT: The general concept represented by the parameter

TYPE: Fixed or dynamic parameter

4.2. ATTRIBUTE

In the specification of a catalog object attribute descriptions, such as in a device or protocol specification, attribute parameter specifications are given. In the instantiation of the catalog object, for example, defining a characterization of a device such as a heat flux burner (**HeatFluxBurner**), the specific attribute unit and value of the description is given. For example, in the case of the heat flux burner, the parameter of the burner plate diameter (**BurnerPlateDiameter**) was deemed by the domain community as an important attribute to distinguish among similar devices. A default unit of centimeter (a typical unit of this parameter) and a default value of 3 (a typical value of this parameter) is given within the specification, the user interface allows changing these values.

4.3. VALUE SPECIFICATION

In an observation specification a set of parameter specifications are given representing the standard values for that particular observation. Within the observation, they are classified as input parameters (cube:dimension from data cube ontology[17]) or output parameters (cube:measure from data cube ontology[17]). For example, in the definition of the standard reporting from a rapid compression machine, the input parameters represent the experimental conditions (pressure and temperature) and the output parameters represent the measured values including ignition delay times.

4.4. VALUE

As an attribute or as a parameter in an observation, the specifications set forth in the catalog specifications are used to interpret the values given. In a catalog object instantiation, such as a device description or the values within an observation, with the correct corresponding units, are given through the interface. In an observation, the set of parameter values are given through the data source.

A set of observations corresponding to a protocol entails the specification of each parameter (as **Value Specification**) of the data to be found, particularly the units and the correspondence within the source file, within a data source file (as Value). Within the observation specification within the protocol specification within the knowledge base is the parameter specification (parameter concept).

In the standard reporting of ignition delay times from rapid compression machines, for example of ignition delay times from the files given at the University of Connecticut[18], two blocks of data are specified in the Rapid Compression Machine ignition delay time reporting protocol (**RapidCompressionMachineReportingProtocol**), the fuel composition, defined in and several parameters. In the value specification (**Value Specification**) of the ignition delay time (**IgnitionDelayTime**), the time units are milliseconds and the corresponds to the column of the spreadsheet labeled 'Ignition Delay (**msec**)'. In the parameter concept of the ignition delay time (**ExperimentalIgnitionDelayTime**), units are specified as time units (**qudt:TimeUnit**from the QUDT ontology) with the specific units being milliseconds (**MilliSecond**). In addition, the purpose is given as a fundamental experimental measurement (**FundamentalExperimentalMeasurement**).

4.5. Use of Quantities, Units, Dimensions and Data Types Ontologies (QUDT)

An important aspect of scientific observations are the units of the individual pieces of data. More often than not, the units used are implicit. The community usually report using a 'standard' unit common for the domain. Problem arise with the words 'usually use' and if the data should be used in another context or community. Fundamental knowledge about the units and conversions between units is supplied by the QUDT, Quantities, Units, Dimensions and Data Types[18], ontologies. Within CHEMCONNECT, the QUDT have been expanded where needed adding more domain specific units.

In CHEMCONNECT, a parameter specification lists the type of data, such as **qudt:TimeUnit** in the properties of the specification as a **qudt:SystemUnit**. Since it is common within a domain to have a 'typical' unit, under the annotations of the parameter specification, a specification unit in the class, for example, **qudt:MilliSecond**, can be listed as a **skos:example**property which, in turn, can be annotated with a **skos:example** value, such as 10.

Within the interface, when a specific unit type is selected, initially the example specific unit is viewed. However, the list of possible other specific units is made available through a pull-down menu.

5. Ontology Description of Concepts

A fundamental purpose of the ontology is to serve as a knowledge base about domain information. Within CHEMCONNECT, under the very general entity **skos:Concept[19]**, from the simple knowledge organization ontology, domain concepts are defined. There are many different types of concepts used within CHEMCONNECT. The domain concepts give context to 'keywords' used to describe domain concepts, supply template information about domain objects (such as filling in the catalog data elements outlined previously), and to drive the user interface for clear and efficient presentation and management.

In repositories a common method to categorize objects is to allow the user to enter keywords describing the object. Those objects with the same keyword would deem to be similar. Leaving the choice of keywords free has two disadvantages. The first is though semantically two keywords would be the same, if they are not exactly the same (for a keyword string match), the two similar objects would not be matched. The second disadvantage is that the interpretation of the keywords come purely from point of view of the researchers creating them and the researchers reading them. The keywords have no greater context.

Within CHEMCONNECT, keywords are arranged in a hierarchy of Concepts (under the **skos:Concept** ontological object). Most of the domain information is stored as a concept. Within the hierarchy of concepts there are several types:

CONCEPT KEYWORDS: These are concepts, under **ChemConnectConceptProperties**, within a hierarchy of concepts, with the hierarchy giving them context and meaning. The only extra information that can be associated with this type of concept are a label, for a more readable title and a comment offering a short explanation.

PURPOSE KEYWORDS: These are concepts, under the **ChemConnectPurpose** concept, that are a concept keyword relating to a describing purpose.

CLASSIFICATION CONCEPTS: These are concepts, under the **ChemConnectClassifications** concept, representing classifications. The sub-classes of these classification concepts are the specific choices of classifications (each being a concept keyword). Classifications are used, for example, in the interface to produce a pull-down list of choices.

LINK CONCEPTS: These are concepts associated with linking two objects, such as catalog objects. This concept can have an extra property limiting the structures it links to. This information, for example, is used by the interface to produce a list of choices from the database.

PARAMETER CONCEPT: This set of concepts specifies a standardized parameter name, but also gives the template of the parameter's properties. This is the essential information making up the parameter specification.

OBSERVABLE CONCEPT: This set of concepts give the name of standard observation data, but also specifies the observable data configuration, i.e. which parameters are needed in the observable. This is the essential information making up an observation as a set of parameters.

DATA STRUCTURE CONCEPT: This set of concepts define templates for catalog objects. The supplementary information in the concept fills the catalog objects with domain specific information. These can define, for example, the subsystem make-up of a device.

5.1. PROPERTIES WITHIN CONCEPTS

Some concepts, for example, catalog sub-objects or templates, are not just standardized keywords, but also have properties associated with them. CHEMCONNECT defines these properties in a few standardized ways to give additional information about the concept and are key to their display within the CHEMCONNECT interface.

Properties are added to concepts through ontology class restrictions. For example:

```
<rdfs:subClassOf>
<owl:Restriction>
<owl:onPropertyrdf:resource="http://purl.org/linked-data/cube#concept"/>
<owl:qualifiedCardinality
rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1</owl:qualifiedCardinality>
<owl:onClass rdf:resource="http://www.esblurock.info/dataset#HeatFluxBurnerBurnerDimensions"/>
</owl:Restriction>
</rdfs:subClassOf>
```

owl:onProperty is used to define the restriction, in this case cube:concept. In this case, the concept is linked to the concept **dataset:HeatFluxBurnerBurnerDimensions**. The property (restriction) can be specified in two ways:

Single: The property is only used once.

Multiple: The property can be repeated several times.
In the example given above, the property is singular. This is given by a singular cardinality:

```
<owl:qualifiedCardinality
rdf:datatype="http://www.w3.org/2001/XMLSchema#nonNegativeInteger">1</owl:qualifiedCardinality>
```

The meaning of this is that there is a single property of **cube:concept** of type **dataset:HeatFluxBurnerDimensions**.

The following example specifies a multiple property:

```
<rdfs:subClassOf>
<owl:Restriction>
<owl:onPropertyrdf:resource="http://www.w3.org/ns/dcat#record"/>
<owl:someValuesFrom rdf:resource="http://www.esblurock.info/dataset#DataSetReference"/>
</owl:Restriction>
</rdfs:subClassOf>
```

The property (**owl:onProperty**) of a **dcat:record** of **dataset:DataSetReference** can be multiply repeated using the specification:

```
<owl:someValuesFrom rdf:resource="http://www.esblurock.info/dataset#DataSetReference"/>
```

The meaning of this property specification is that there can be multiple references (this is found in the definition information of a standard Catalog object).

5.2. Templates

The information structure of CHEMCONNECT representing the interaction between the domain knowledge and the actual use of the domain knowledge to build database objects can be thought as having three levels:

The Catalog Object: This ontology object gives the basic structure of the object, i.e. which knowledge (ontology objects) should be given. This definition is not domain specific. These are the database entities and JAVA classes in the CHEMCONNECT implementation.

Domain Templates: The domain templates specify how some of the records in the catalog object should be filled. For example, attributes of devices, subsystems and sensors that describe a particular device or the set of parameters that make up an observation would be specified. This domain knowledge comes from case studies. These templates are defined in the ontology.

Database: The templates specify, for example, the attributes or parameters, but when the templates are used to, for example, characterize a device, then a database object is made and specific values for the attribute are filled in. The database object represents instantiations of the templated object.

Templates within the ontology give a pattern or a generic description of domain objects. The templates themselves are ontology objects, but they contain information to further characterize domain objects. Templates answer the question, for example, how should one characterize a typical device. In the example below, the template for a heat flux burner will be shown.

One can think of templates as the needed information to characterize an object and the database object itself is a particular example. For example, in characterizing a burner, the template would give as one of the describing attributes the diameter of the burner. It is implying that all burners have a diameter (the template), but particular burners could have particular diameter (the database value).

5.2.1. TEMPLATE EXAMPLE: HEAT FLUX BURNER DEVICE (SUBSYSTEM)

The heat flux burner is a device which is characterized by the **SubSystemDescription**catalog object which is used to characterize generic devices, subsystems, sensors and device components. The **SubSystemDescription** catalog object, other than the standard catalog object fields, has additional specialized fields. To create a database object of this type, these specialized fields should be filled in:

OBSERVATION SPECIFICATION: The set of observations coming from the subsystem.

PARAMETER VALUE: A set of attributes that describe the subsystem and differentiate this same subsystem from other similar subsystems. What attributes are listed is determined by domain analysis.

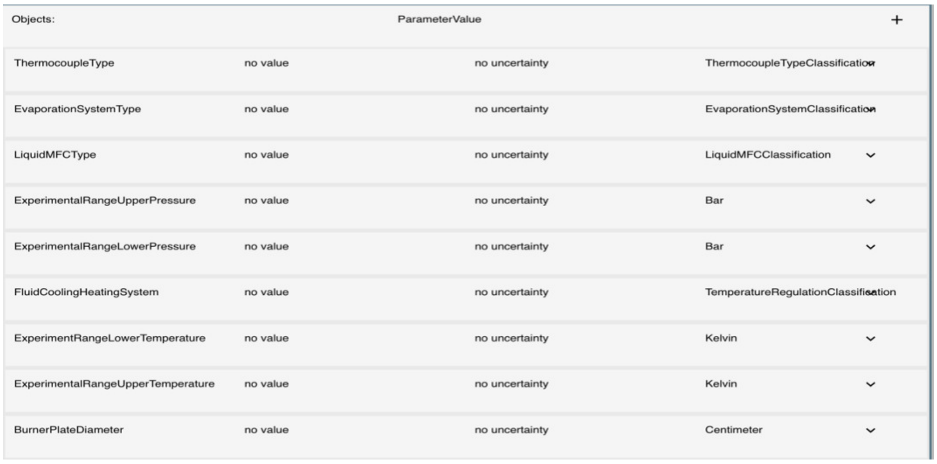
SYSTEM DESCRIPTION: The set of subsystems making up this subsystem. A device is considered to be a hierarchy of subsystems.

This is the generic subsystem description. The generic catalog object, described in the ontology, says that all subsystems have these quantities. These are general enough to specify a wide range of devices, subsystems, sensors and device components. It is the job of templates to give a more detailed specification, for example, which set of **ParameterValue**entities should be specified to describe the particular subsystem.

The task of the template is to, through domain knowledge, give patterns (templates) on how to fill in these elements. The domain knowledge is captured in another ontology object which specifies these parts.

The device, the **Heat Flux Burner**, has a set of subsystems, observations and parameters. In the template specification of **HeatFluxBurner** (an ontology concept), a set of 'attributes' (**cube:attribute** from the data cube ontology) are specified. Which parameters were deemed needed or appropriate were gleaned from the literature and derived from researchers working in the domain. The catalog object specification said that a subsystem is described by a set of **ParameterValues**. The template ontology object **HeatFluxBurner** specifies that these parameters (**cube:attribute**) should be used to characterize a typical heat flux burner. CHEMCONNECT uses this template information to set up the interface so these values can be supplied for a particular heat flux burner description that will be entered in the database.

This template is used to set up the user interface to input the actual values to characterize the particular Heat Flux Burner. Each attribute above corresponds to a user interface row for that particular value where the user can enter the property value. After the values are entered by the user, then the object is saved into the database. All database objects describing a heat flux burner have specific values of all these attributes. Having the same set of attributes for an object facilitates their comparison. The purpose of the template is to say that all objects of a given type should have these (common) properties.



Objects:	ParameterValue		
ThermocoupleType	no value	no uncertainty	ThermocoupleTypeClassification
EvaporationSystemType	no value	no uncertainty	EvaporationSystemClassification
LiquidMFCType	no value	no uncertainty	LiquidMFCClassification
ExperimentalRangeUpperPressure	no value	no uncertainty	Bar
ExperimentalRangeLowerPressure	no value	no uncertainty	Bar
FluidCoolingHeatingSystem	no value	no uncertainty	TemperatureRegulationClassification
ExperimentRangeLowerTemperature	no value	no uncertainty	Kelvin
ExperimentalRangeUpperTemperature	no value	no uncertainty	Kelvin
BurnerPlateDiameter	no value	no uncertainty	Centimeter

Figure 1. The CHEMCONNECT user interface derived from template concepts of the attributes of a device.

Another template description is a parameter. The essential elements of the parameter template are:

CONCEPT: The chosen concept, out of the hierarchy of concepts, associated with the particular parameter

PURPOSE: The chosen purpose, out of the purpose concepts, associated with the parameter.

ELEMENT KIND: This specifies whether the parameter is a fixed parameter or a dynamic parameter.

UNITSYSTEM: The unit class of the parameter, in the example is unit system LengthUnit.

The parameter template is used in two ways.

SPECIFICATION: The specification, for example, the specifications of parameters that make up an observation, where the values are left open. A specification is, for example, defines a column of a matrix of values. All of the elements of the column have the properties of the specification.

ATTRIBUTE: An attribute, for example as above, specifying the value (with units) of the object.

5.2.2. EXAMPLE:FROM CATALOG ENTITY TO REPOSITORY VALUE

Parameters are central to the storage of data within CHEMCONNECT and how they managed is another typical example of how CHEMCONNECT defines andusesthe ontology on different conceptual levels. The following example shows how starting with the catalog object, **ObservationCorrespondenceSpecification** and a concept template defining a specific set of observations for a domain, is used create a correspondence between a matrix of input values from the user and the standardized domain parameters defined in CHEMCONNECT.

When reading in repository data, the user is not required to use the standard names (including unit names) for the data values (typically a matrix of values). To provide meaning and expand the context of the repository data, a correspondence needs to be established between the CHEMCONNECT parameter concepts in the ontology and those found in the user's input (file). This is accomplished by the catalog object **ObservationCorrespondenceSpecification**, where a one-to-one connection is made between specifications of the record object, through **ObservationSpecification,** and that set of parameters in the user input. If the user input is a matrix, the **ObservationSpecification** can be viewed as the specification of the parameters for each column. The **ObservationSpecification** entity is defined as having multiple input (**DimensionParameterSpecification**) and multiple output (**MeasureParameterSpecification**) parameter specifications. Both **DimensionParameterSpecification** and**MeasureParameterSpecification** are derived from **ParameterSpecification**, where the individual specifications, such as units, labels, etc., of the parameter are defined.

The record object **ParameterSpecification** and the catalog and record objects mentioned previously define a data object with fields, which has a one-to-one correspondence with a data object in the database.However, it does not hold any specific domain information about a specific parameter.The domain information comes from the template concept information.

Associated with the **ParameterSpecification** record object is a hierarchy (under the ontology concept **ChemConnectParameters**) of templates representing domain information about specific parameters. The name of the parameter specification concept reflects what the parameter describes in the domain. For example, **ExperimentalTemperature**describes the temperature conditions of the experiment. Further information about the parameter is found in the properties, **cube:concept**, giving another concept keyword describing the parameter, and **hasPurpose**, a more specific keyword specifying the purpose of the parameter. These two properties bind the parameter to keyword concepts (standardized ontology concepts within a hierarchy of concept and purpose keywords). For example, **ExperimentalTemperature**has the concept **TemperatureOfExperiment** and a purpose of **ExperimentalCondition.**

A third property of the template concept lists the units of the parameter (**qudt:unitSystem**). This property points to the unit ontology entity within the QUDT hierarchy (which has CHEMCONNECT annotations as required by the domain knowledge). The properties and instances of the QUDT unit definition provide all the necessary examples of specific units and conversions between specific units. The **qudt:unitSystem** of **ExperimentalTemperature** is **qudt:TemperatureUnit**. Which has instances of temperature units such as **qudt:Kelvin** and **qudt:Celcius**.

In the template within the ontology, only the unit class is specified. If a specific parameter is to be described, then the specific unit should be chosen. This is done by creating an instance derived from the template specifying the unit. This instance is then stored in the database. This is the next level of parameter specification. This level would be used, for example, to describe a column in a matrix of data (see next example).

A typical data set, or observation (coming from **qb:Observation** in the data cube ontology) involves several parameters. For example, a pressure versus time graph involves two parameters, time and pressure. To describe experimental conditions in a chemical experiment, typically three or more parameters are needed, temperature, pressure and each of the specifies concentrations. A single data set can be thought of as a matrix of data, where each column is a parameter.

To describe a matrix of data, a **ParameterSpecification** of each column is needed. In CHEMCONNECT an entire matrix specification is described using the **ObservationSpecification** ontology record object in CHEMCONNECT. As described previously, this involves two sets of input and output specifications (subclasses of **ParameterSpecification**). To describe complete data sets, templates describing which parameters are involved are listed as **qb:dimension** or **qb:measure** properties for input and output properties, respectively. For example, the observation **PressureTrace**has **TimeInEvent** as a **qb:dimension** property and **ExperimentalPressure** as a **qb:measure** property. The **ObservationSpecification** also has the **qb:concept** and **hasPurpose** properties.

As before, the template concept for **ObservationSpecification** describes which parameters are to be found in the type of observation. But, to describe a specific matrix, all the **ParameterSpecification** entities must be filled in with the specific unit to be used and stored in the database. Once again, in CHEMCONNECT the **ObservationSpecification** is the data-type, the corresponding template provides the generic domain information and the database object describes a specific matrix of observations.

The database object for **ObservationSpecification** has the column information for a matrix. However, the **ObservationSpecification** uses the ontology standardized domain names for the parameters. The columns of the input matrix would not necessarily have these names. The names of the specification names have to be assigned to the corresponding columns in the input matrix. This is the job of the **ObservationCorrespondenceSpecification** data type which is set up in the user interface. The column headings are extracted, the user assigns each to a specification parameter name and this information is stored in an **ObservationCorrespondenceSpecification** database object.

The database object provides the key, so to speak, for interpreting a specific matrix. Typically a laboratory (or domain consortium) would have a standardized way of presenting data. The **ObservationCorrespondenceSpecification**database object would only have to be defined once for all the data produced by the laboratory.

6. USE OF STANDARD ONTOLOGIES

The ontology and terminology used by CHEMCONNECT was not created from scratch but is based on proven ontologies from the W3C community. The ontologies which have had a particular influence are:

- **Basic Concepts:**
  - Dublin Core Terms for metadata[20]: This is the source of the basic terminology and the basis of the other ontologies.
  - Simple Knowledge Organization System[19]: This is the basic ontology for the concept terms and their inter-relations.
- **Contacts: These ontologies serve as the basis for the contact information from**
  - Friend of a friend [21]
  - Organization[10]
- **Data**
  - Data Catalog Vocabulary (DCAT)[21]: This is the basic for the category and record entities. This serves as the basis for the basic data types and database objects used within CHEMCONNECT.
  - Data Cube Vocabulary (QB)[17]:
- **Domain Knowledge**
  - Semantic Sensor Network (SSN)[22]
  - Sensor Organization Sampling Actuator (SOSA)[23]
  - Quantities, Units, Dimensions and Data Types Ontologies (QUDT)[18]

7. Conclusions

This paper has outlined the guiding principles and the use of ontologies in the ongoing work to create a knowledge-based repository of experimental and modelling data, CHEMCONNECT (implementing at http://www.connectedsmartdata.info). The current domain is concentrating on experimental results from combustion research. But the structure, both in terms of the knowledge base and the structure of the database, is general enough to encompass many scientific experimental disciplines. The design goal of CHEMCONNECT is to provide a smart interlinked repository to promote collaboration and FAIR practices[2] among researchers in the experimental scientific community.

CHEMCONNECT uses ontologies, based on numerous established ontologies, to capture knowledge of domain concepts, experimental devices and observational data and personal and organizational contact data. This knowledge is instrumental in driving the user interface, data presentation and repository database. The knowledge base also provides the foundation for more efficient search.

Acknowledgements

The author would like to acknowledge the cooperation of domain experts within the combustion community that has been made possible through the SMARTCATS COST Action CM1303, Chemistry of Smart Energies Carriers and Technologies. In particular, the members of the working task group within the action for *Standard definition for data collection and mining toward a virtual chemistry of Smart Energy Carriers*, of which the author is the group leader. I would like to particularly acknowledge Nils Hansen of Sandia National Labs and Alexander Konnov of Lund University for their help in formulating initial prototypes of for combustion devices.

References

[1] S. C. Pandey and P. N. Pattnaik, “University Research Ecosystem: A Conceptual Understanding,” Rev. Econ. Bus. Stud., vol. 8, no. 1, pp. 169–181, Jun. 2015.

[2] S. Hagstrom, “The FAIR Data Principles,” FORCE11, 03-Sep-2014. [Online]. Available: https://www.force11.org/group/fairgroup/fairprinciples. [Accessed: 11-Jan-2019].

[3] M. D. Wilkinson et al., “The FAIR Guiding Principles for scientific data management and stewardship,” Scientific Data, 15-Mar-2016. [Online]. Available: https://www.nature.com/articles/sdata201618. [Accessed: 19-Feb-2018].

[4] “FAIR Principles - GO FAIR.” [Online]. Available: https://www.go-fair.org/fair-principles/. [Accessed: 15-Feb-2019].

[5] B. A. Noseket al., “Promoting an open research culture,” Science, vol. 348, no. 6242, pp. 1422–1425, Jun. 2015.

[6] “figshare - My data.” [Online]. Available: https://figshare.com/account/home. [Accessed: 15-Feb-2019].

[7] “Zenodo - Research. Shared.” [Online]. Available: https://zenodo.org/. [Accessed: 15-Feb-2019].

[8] “Dryad Digital Repository - Dryad.” [Online]. Available: http://datadryad.org/. [Accessed: 15-Feb-2019].

[9] “How-to Guides & Checklists | Digital Curation Centre.” [Online]. Available: http://www.dcc.ac.uk/resources/how-guides. [Accessed: 15-Feb-2019].

[10] “The Organization Ontology.” [Online]. Available: https://www.w3.org/TR/vocab-org/. [Accessed: 14-Feb-2019].

[11] “Ontology Use for Semantic e-Science | www.semantic-web-journal.net.” [Online]. Available: http://www.semantic-web-journal.net/content/ontology-use-semantic-e-science. [Accessed: 15-Feb-2019].

[12] "RDF - Semantic Web Standards." [Online]. Available: https://www.w3.org/RDF/. [Accessed: 15-Feb-2019].

[13] "Ontologies - W3C." [Online]. Available: https://www.w3.org/standards/semanticweb/ontology. [Accessed: 15-Feb-2019].

[14] "World Wide Web Consortium (W3C)." [Online]. Available: https://www.w3.org/. [Accessed: 15-Feb-2019].

[15] "Data Catalog Vocabulary (DCAT)." [Online]. Available: https://www.w3.org/TR/vocab-dcat/. [Accessed: 14-Feb-2019].

[16] "Quantities, Units, Dimensions and Data Types Ontologies," Quantities,Units, Dimensions and Data Types Ontologies. [Online]. Available: http://qudt.org/. [Accessed: 14-Feb-2019].

[17] "The RDF Data Cube Vocabulary." [Online]. Available: https://www.w3.org/TR/vocab-data-cube/. [Accessed: 14-Feb-2019].

[18] "RCM Database | Combustion Diagnostics Laboratory." [Online]. Available: https://combdiaglab.engr.uconn.edu/database/rcm-database/. [Accessed: 15-Feb-2019].

[19] "SKOS Simple Knowledge Organization System Namespace Document 30 July 2008 'Last Call' Edition." [Online]. Available: https://www.w3.org/TR/2008/WD-skos-reference-20080829/skos.html. [Accessed: 14-Feb-2019].

[20] "Dublin Core Metadata Initiative," Dublin Core Metadata Initiative. [Online]. Available: http://dublincore.org/. [Accessed: 14-Feb-2019].

[21] "FOAF Vocabulary Specification." [Online]. Available: http://xmlns.com/foaf/spec/. [Accessed: 14-Feb-2019].

[22] "Semantic Sensor Network Ontology." [Online]. Available: https://www.w3.org/TR/vocab-ssn/. [Accessed: 14-Feb-2019].

[23] "SOSA Ontology - Spatial Data on the Web Working Group." [Online]. Available: https://www.w3.org/2015/spatial/wiki/SOSA_Ontology. [Accessed: 14-Feb-2019].

AUTHOR

Edward S. Blurock has over 35 years of experience, both in chemical data and modelling which is the core of this projectand in applied artificial intelligence and machine learning techniques and software (*Research Institute for Symbolic Computation,* Austria and *Combustion Physics*, Sweden). Currently, he is recognized to be at the forefront of data management within the community by being the '*Standard definition for data collection and mining toward a virtual chemistry of Smart Energy Carriers*'Working Group leader in COST Action network CM1404, '*Chemistry of Smart Energy Carriers and Technologies'*, and leader of the CM1401 task to catalog data use within the community. He is co-editor and co-author of several chapters in the book *Cleaner Combustion: Developing Detailed Chemical Kinetic Models.*