Modelling Scientific Activities: Proposal for a global schema for integrating metadata about scientific observation.

Authors: Martin Doerr, Chryssoula Bekiari, Athina Kritsotaki, Gerald Hiebel, Maria Theodoridou
Conference: CIDOC - International Documentation Committee of ICOM: Content metadata - Administrative metadata - Technical metadata - Legal metadata (long paper) -Conference 6th-11th Sept. 2014 in Dresden/Germany

Abstract

The wide deployment and use of the CIDOC CRM for information exchange and integration between heterogeneous sources of cultural heritage information in recent years have highlighted that cultural information extends to other disciplines as well, such as biology, geology and others. The cultural discourse includes information from all sorts of sciences and product of sciences, such as digital productions, biological samples, specimen of physical objects (materials, fluids etc.). Scientific activities themselves are part of the human culture. In this paper we present a model about concepts of scientific observation and how this model is related to ISO21121. This model has been being developed bottom up from specific metadata examples from biodiversity, geology, archaeology, cultural heritage conservation and clinical studies. It has so far been validated in several national and international projects by implementing it in slightly different versions together with application-specific extensions and by mapping to and from related standards. The present version has been produced by FORTH and collaborators and describes a consolidated version from this experience, with the aim to present it for review and further adoption to the widest possible community. The model presented here describes, together with the CIDOC CRM, a discipline neutral level of genericity, which can be used to implement effective management functions and powerful queries for related data. It aims at providing super classes and super properties for any discipline-specific extensions, such that any entity referred to by a compatible extension can be reached with a more general query based on this model. We propose to open the discussions in CIDOC about conceptual modelling of products of human activities in general. We suggest to CIDOC to approve that modelling scientific activities is a valid scope for CIDOC and could be a working item for the CRM-Special Interest Group.

Introduction

Scientific observation is the most fundamental practice of any scientific method or process (Wenning 2009). The core skills of scientist are to make observation. Observation consists of receiving knowledge of the outside world through our senses, or by recording information using scientific tools and instruments, such as eyes, ears, telescopes, microscopes, photographic sensors, questionnaires, and a

myriad of other ingenious inventions designed to make the invisible visible, the evanescent permanent, the volatile concrete and quantifiable. In all scientific disciplines, the observers have devised ways to tackle the unknown and thereby redefining what is under investigation by the way in which it is investigated. Observation discovers the world anew, such as archaeology discovers the past (Daston 2011).

The ICS-FORTH, collaborators and members of CRM-SIG involved in several research infrastructure oriented projects in various disciplines i.e. Geology, Biology and archaeological excavations which were aiming at data integration for publishing linked open data about scientific observations, seeking to find an ontology and studying existing standards realized that these standards have common concepts and relationships with the CRM either explicitly or implicitly-hidden under other concepts or services. Also considering that the event centric approach used in CIDOC CRM and the CRM itself could be used as an adhesive substance for representing metadata of scientific observations have initiated discussions with domain experts about mappings concepts and relationships of the most common standards used for scientific observation such as INSPIRE, OBOE and Darwin Core without loss of meaning to CIDOC CRM in order to develop one consistent and generic model with a few disciplinary specializations. Building this model, special focus was given to support the required by domain experts functionality of integrated data for scientific queries. At the same time the theoretical framework of scientific observation and methods in various disciplines were investigated when it is confirmed by the current practice.

Related Work and Practice

INSPIRE , an earth science oriented standard promoted by the European Commission for interoperability of location-related data in Europe, employs a Generic conceptual schema which is not event-oriented (especially the schema that uses the observation-measurement specification)(INSPIRE). There is no class or element connecting directly actors, objects, place and time involved in one event – INSPIRE assigns temporal or spatial properties by using parts of foundation schemas such as ISO 19108:2006 Temporal Schema for time or ISO 19107:2003 Spatial Schema for place which are implemented in XML by ISO 19136 GML. There is no clear cut distinction between the notion of an action and its result, and an Observation is also considered a Feature. For different sub-disciplines, INSPIRE contains three different, unrelated and mutually incompatible models of observation.

The Science Environment for Ecological Knowledge (SEEK) is a knowledge environment that is being developed to address many of the current challenges associated with data accessibility and integration in the biodiversity and ecological sciences. The SEEK and SPiRE projects have developed a collection of ontologies for describing ecological organisms, systems, and observations.

SEEK Extensible Observation Ontology (OBOE) is a formal ontology for capturing ecological observational and measurement data and provides basic concepts and relationships for describing observational datasets, including field, experimental, simulation and monitoring data. It is compatible and supplements the Ecological Metadata Language (EML)(Madin et al, 2007b). OBOE can be used to characterize

the context of an observation (e.g., space and time), and clarify inter-observational relationships such as dependency hierarchies (e.g., nested experimental observations) and meaningful dimensions within the data (e.g., axes for cross-classified categorical summarization). It also enables the robust description of measurement units (e.g., grams of carbon per litter of seawater), and can facilitate automatic unit conversions (e.g., pounds to kilograms). The ontology can easily be extended with specialized domain vocabularies. Observation is used as a unifying concept for capturing the basic semantics of ecological data. Observations are distinguished at the level of the observed entity (e.g., location, time, thing, concept), and characteristics of an entity (e.g., height, name, color) are measured (named or classified) as data. Basic concepts are Observation, Measurement, (Ecological) Entity, Characteristic, and Measurement Standard (e.g., physical units) and six properties labelled *hasContext*, *ofEntity*, *hasMeasurement*, *hasValue*, *hasPrecision*, *usesStandar*d, and *ofCharacteristi*c. (Madin et al, 2007a)

The Darwin Core standard itself is a general-use metadata schema that defines fields which can be used to facilitate the sharing of information about biological diversity. The fields are organized into nine categories (often referred to as "classes, six of which cover broad aspects (event, location, geological context, occurrence, taxon, and identification) of the biodiversity domain. The remaining categories cover relationships to other resources, measurements, and generic information about records. Especially for the record level, Darwin Core recommends the use of a number of terms from Dublin Core (type, modified, language, rights, rights Holder, access Rights, bibliographic Citation, references). Darwin Core was designed to be minimal and flat, i.e. without nested elements. (Darwin Core, 2013)

The above competitive models focus on the act of attentive watching, perceiving, or noticing and the data measured, collected, perceived or noticed, especially during an experiment. They have been designed to facilitate the semantic annotation process of data sets. By this process of semantic annotation mappings are created between data in a data set and an ontology or metadata standard. In the case of OBOE, data values in columns of data sets are semantically annotated with ontology concepts or metadata, then by traversing relationships in the ontology (e.g. 'isA', 'part-of', and 'has characteristic'), a query about a concept could find data from different data sets since the values of the datasets are instances to particular concepts of the ontology.(Madin, 2007b). The OBOE can be used to suggest appropriate data summarizations, when a particular summarization is "sensible". It provides a logical structure, constraints and guidance for testing the usefulness of various statistical operations and modelling procedures, and automates data aggregation and summary (Madin, 2007a).

One of the major drawbacks of Darwin Core in the Semantic Web context is the lack of a well-defined ontology - a formal definition of relationships between the kinds of entities ("core schema") of the biodiversity domain including its scientific processes (Stucky et al 2013). A Darwin core data record leaves the interpretation of the relationships between the whole record and one of its fields to the intuition of the human reader. Depending on the use case, the same field can have completely different meaning and role. In other words it cannot be used to merge data from different records and to draw logical conclusions (e.g., consistency, equivalence) without human intervention or an overly complex interpretation framework

biologists do not dispose of. Also complex, causally related events (or composite events) cannot be described. Darwin Core can quite well serve as a data entry questionnaire.. A recent attempt “Darwin –sw” to turn Darwin Core into ontology for exposing data in RDF appears to us a direct, naïve interpretation of Darwin Core fields and retains the same ambiguities of the original.(Darwin-SW), (Webb 2011), (Webb 2013)

The above models actually model observation isolated from actions that are preceding or following an observation event. This isolation limits the kinds of inferences that can be drawn by reasoning tools or by humans on these data representation, even though, their maintainers advertise that they support the transformation of values, data integration and data discovery (Madin 2007). In particular, these models leave out information that would allow for later assessing, the quality and precision of the results or for re-evaluating existing measuring data due to new evidence which would not require redoing the measurement itself, if suitable raw data were provided. Even though they are using the above standards to publish data in repositories (e.g. to share data with other researchers), they typically lack the required information to facilitate effective long-term preservation and interpretation of data.

Observing in different applications how scientists use and reason about their data sets we conclude that:

(a) Theories are formalized sets of concepts that organize observations and predict and explain phenomena and demand a solid empirical base of evidence (Sagan 1997)

(b) Raw data provided by the data sets per se are of little use, and no scientific journal will publish long lists of data but the deductions and conclusions based on scientific observations (Wenning ,2009)

(c) Scientific observation forms the basis for understanding the phenomena being studied and it is a process by which we systematically advance our understanding of the world.

(d) The different fields of science do not arrive at conclusions in the same way, however common to all sciences is the workflow of forming of a hypothesis to perform and explain observations that are made, the gathering of data, and based upon this data the drawing of conclusions that confirm or deny the original hypothesis (Wenning, 2009), (Explorable, 2014).

(e) The difference between the types of sciences is in what is considered data, and how data is gathered and processed (Sagan, 1997).

From these empirical observations we decide bottom up to look for generic patterns that are followed and that allow for effective integrated management of research data, adequate to the scientific process regarding and regardless discipline. In contrast to other models, we do not stay at the surface of handling digital artefacts, but we try to understand and model the steps and reasons of the knowledge production processes in different sciences using real examples and the relative role and significance of information elements in the reasoning processes. Therefore our proposed model provides explicit references to hypotheses used for specific results in order to support the necessary monitoring to understand scientific results manually.

Therefore the highest level of description of scientific data is the one of historical records and things that have happened in the past, regardless if measurement data and models are used to simulate or predict future. This has been widely confirmed by the on-going convergence of Digital Provenance models on an event-centric approach.(Theodoridou 2010)

The above considerations form the following requirements:

- The vast amount of scientific data cannot be understood without knowledge about the meaning of the data and the ways and circumstances of their creation.
- Scientific data and metadata can be considered as historical records. Therefore relevant observation data cannot be found and understood without metadata about their context.
- Scientific observation and machine-supported processing is initiated on behalf of and controlled by human activity.
- Things, data, people, times and places in their contexts are causally related by events.
- Data Evaluation is based on observation records and hypotheses
- Data Simulation may be based on initial observation records or data evaluation.

The model

The proposed formal ontology is intended to be used as a global schema for integrating metadata about scientific observation satisfying the above requirements. Therefore it includes concepts and properties describing measurements and processed data in descriptive and empirical sciences such as biodiversity, geology, geography, archaeology, cultural heritage conservation and others in research IT environments and research data libraries. It shares the same primary purpose with CIDOC CRM, which is the facilitation of management, integration, mediation, interchange and access to research data by description of semantic relationships, in particular causal ones. It is not primarily a model to process the data themselves in order to produce new research results, even though its representations offer themselves to be used for some kind of processing. It provides the appropriate concepts and relationships for monitoring of the events and activities. It allow us to separate the sample taking event from measurement.

It uses and extends the CIDOC CRM (ISO21127) as a general ontology of human activity, things and events happening in spacetime. In order to integrate some of the new concepts, we have to introduce two new superclasses to existing CRM concepts; the *observable entity* and the *material substantial*. This model uses the same encoding-neutral formalism of knowledge representation (“data model” in the sense of computer science) as the CIDOC CRM, and can be implemented in RDFS, OWL, on RDBMS and in other forms of encoding. Since the model reuses, wherever appropriate, parts of CIDOC Conceptual Reference Model, we consider as part of this model all constructs used from ISO21127, together with their definitions following the version 5.1.2 maintained by CIDOC.

The CRMsci model comprises concepts and relationships for describing metadata about:

- The human observer (robots are not human!)
- The object of observation (a "thing", "something", a process or a state?)
- The observation hypothesis (choice of parameters)
- The identity of the object, if any
- The environment, time and location
- The condition of the thing
- The instrumentation and method used
- The identity, authenticity and transmission of the produced records
- Processes of Human argumentation about strengthening or weakening hypotheses about material facts (Doerr 2011).

Special focus is given to scientific events. Scientific events include classes and relationships about scientific activities. In CRMsci the scientific activities (Figure: 1) are classified in three major groups: in types of attribute assignment (Observation, Inference Making), types of handling matter (Matter Removal, Modification, Production) and types of changes of states of matter, natural or not (Alteration, Physical Genesis)

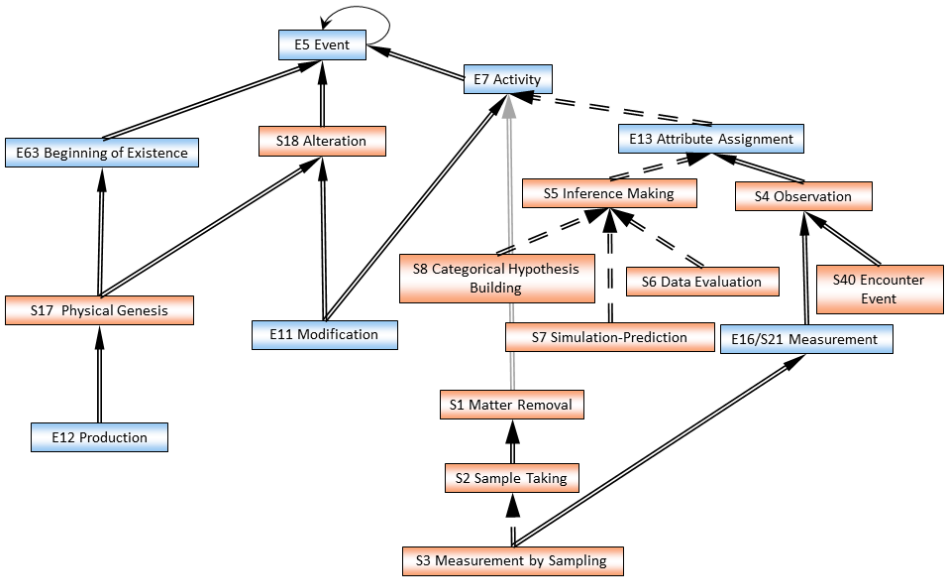

Figure: 1 Classification of Scientific Activities

The classes as "*S5 Inference Making*", "*S4 Observation*", *"S8 Categorical Hypothesis Building", "S6 Data Evaluation", "S40 Encounter Event","*E16 Measurement" are all subclasses of "E13 Attribute Assignment" of CIDOC CRM which comprises the actions of making assertions about properties of an object or any relation between two items or concepts.

The Classes "S2 Sample Taking, "*S3 Measurement by Sampling*", "*E80* Part Removal" are subclasses of "*S1 Matter Removal*" comprises the activities of a component or a piece removal from a physical object of an archaeological or geological layer, or taking a tissue sample from a body or a sample of fluid from a

body of water. By documenting the condition (P44 has condition), the constituent (P45 consists of) material the composition (P46 is composed of) and the place of removal (O15 occupied), we may argue about if an instance of S11 Amount of matter has been removed of a particular Material Substantial.

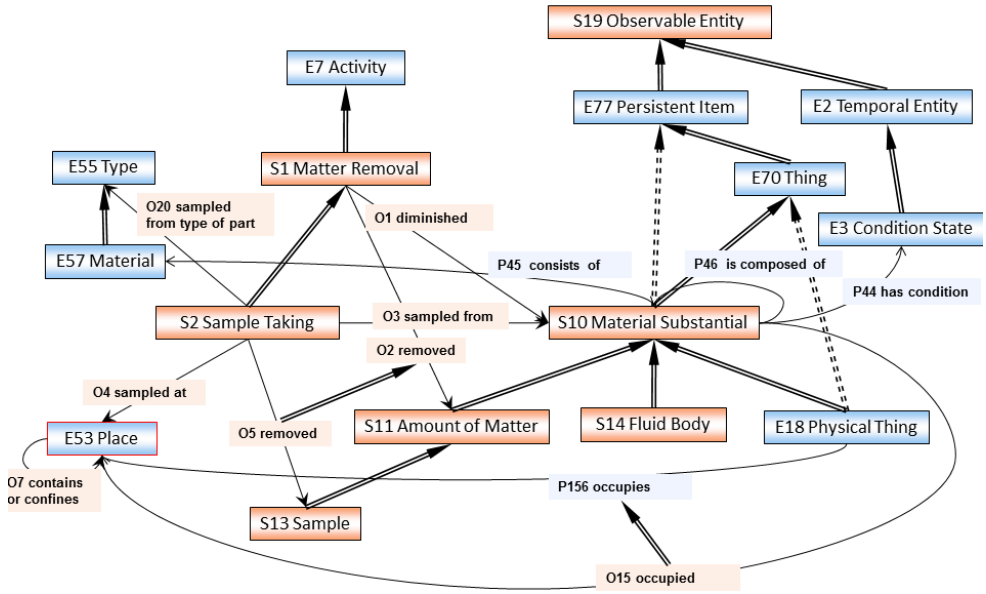

Figure: 2. Classes and Properties involved in an argument about removing an amount of matter.

The classes "E5 Event", "S8 Alteration" "S17 Physical Genesis" "S16 State" support propositions about causality in the sense of necessary conditions. An event can activate (O13 triggers) other event/s; in that sense it is interpreted as a cause, the triggering factor of a situation in tension (a system); a reaction between events. An Event may initialize (O14 initializes) the persistence of a particular value range of the properties of a particular thing or things over a time-span (*S16 States*).

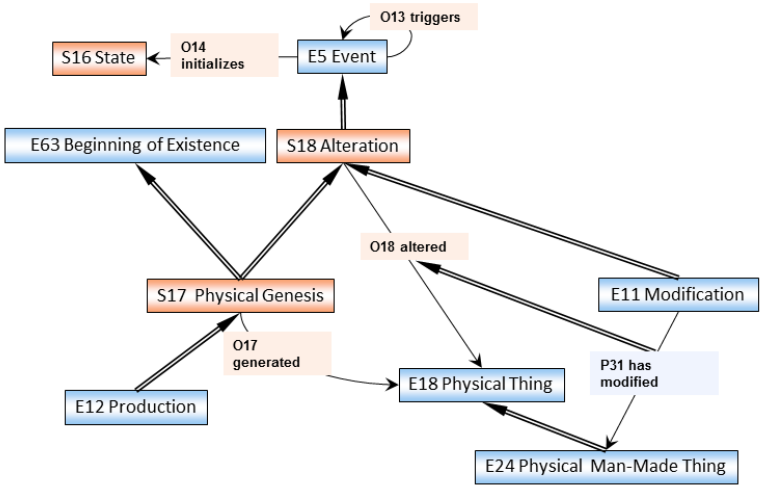

Figure: 3. Classes and properties involved in propositions about necessary conditions.

The "*S4 Observation*" class supports the documentation of human activities of gaining scientific knowledge about particular states of physical reality by empirical evidence, experiments and by measurements. We define observation in the sense of

natural sciences, as a kind of human activity: *at some Place* and *within some Time-Span*, *certain Physical Things*, *features*, *phenomena*, *their behavior and their interactions are observed*, *either directly by human sensory impression, or enhanced with tools and measurement devices*.
Measurements, witnessing of events and encountering objects are special cases of observations. Observations result in a belief about certain propositions. The degree of confidence in the observed properties is regarded to be "true" per default, but could be described differently by adding a property P3 has note to an instance of S4 Observation, or by reification of the property O16 observed value. Primary data from measurement devices are regarded in this model to be results of observation and can be interpreted as propositions believed to be true within the (known) tolerances and degree of reliability of the device.
Considering that Observations represent the transition between reality and propositions, when these propositions are instances of a formal ontology, they can be subject to data evaluation by the use of the "*O9 observed property type (property type was observed by*)". This cross- categorical property provides a value or evidence for an observed property e.g. the "Concentration of nitrate" observed in the water from a particular borehole.

The "S40 Encounter Event" is an innovation in this model. In particular it generalizes over the Darwin Core notion of Occurrence and the archaeological concept of "finds". Whereas the definition of "find" is relative to a state of knowledge, encounter is objective. Whereas in biology the object may escape or be killed, in archaeology it is inanimate. The relevant, necessary property is the creation of a record of the encounter that serves as later evidence. We define it as: S40 Encounter Event comprises observation activities where an E39 Actor encounters an instance of E18 Physical Thing of a kind relevant for the mission of the observation or regarded as potentially relevant for some community (identity). This observation produces knowledge about the existence of the respective thing at a particular place in or on surrounding matter. This knowledge may be new to the group of people the actor belongs to. In that case we would talk about a discovery. The observer may recognize or assign an individual identity of the thing encountered or regard only the type as noteworthy in the associated documentation or report.
In archaeology there is a particular interest if an object is found "in situ", i.e. if its embedding in the surrounding matter supports the assumption that the object was not moved since the archaeologically relevant deposition event. The surrounding matter with the relative position of the object in it as well as the absolute position and time of the observation may be recorded in order to enable inferences about the history of the E18 Physical Thing.
In Biology, additional parameters may be recorded like the kind of ecosystem, if the biological individual survives the observation, what detection or catching devices have been used or if the encounter event supported the detection of a new biological kind ("taxon").
This class supports a generic reasoning on the existence and trajectory of physical things, living or dead, in spacetime regardless discipline.
Finally for drawing evaluations, calculations and interpretations based on mathematical formulations and propositions the classes "*S8 Categorical Hypothesis*

*Building", "S6 Data Evaluation", "S40 Encounter Event","*E16 Measurement" can be used.
The following Figure 4 presents an example of modelling with S40 Encounter Event with real data.

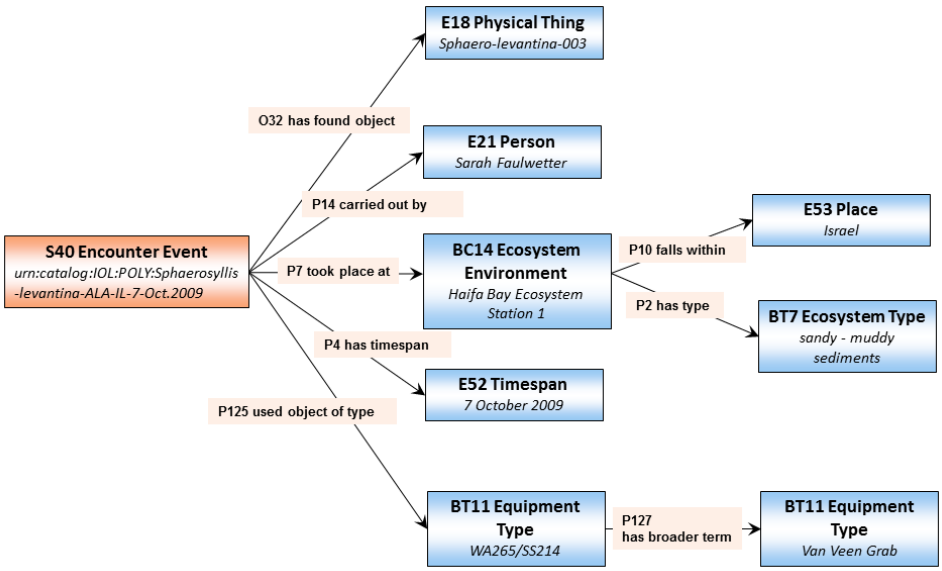

Figure: 4 An example of an S40 Encounter Event

Conclusions

The increasing deployment and use of the CIDOC CRM for information exchange and integration between heterogeneous sources of cultural heritage information in recent years have highlighted that cultural information extends to other disciplines as well, such as biology (Tzitzikas et al, 2013), geology and others. The cultural discourse includes information from all sorts of sciences and product of sciences, such as digital productions, biological samples, specimen of physical objects (materials, fluids etc.), and science itself is part of the human culture.
On the other side, scientific activities are part of the human culture and clearly subject of museums and other memory institutions.
Under this consideration, ICS-FORTH, collaborators and members of CRM SIG have been engaging in mapping the above scientific data standards without loss of meaning to one consistent and generic model, CRMsci. This model generalizes over disciplinary specializations and has been deployed and tested in several research infrastructure oriented projects in very different, specific disciplines. The model is now under test for all scientific investigation methods employed in archaeology. It is CRM compatible, and more detailed and powerful than any competitive scientific standard of this genericity. In particular, it enables unambiguous coherent and consistent information integration of scientific and cultural information.

Besides application-specific extensions, the CRMsci model is intended to be complemented by CRMgeo, a more detailed model and extension of the CIDOC CRM of generic spatiotemporal topology and geometric description, also currently available in a first stable version (Doerr 2013). CRMgeo links consistently CIDOC CRM

to the OGC standard of GeoSPARQL and OPENGIS, enabling to combine spatiotemporal relations derived from geometric computation with those derived from causal-semantic reasoning. CRMgeo should be used for spatiotemporal descriptions using explicit reference frames. CRMSci will further be extended by CRMarcheo, still under revision by CRM-SIG, a model of archaeological excavation (CRMarchaeo (2014). Still to be developed under CRMSci are models of the structures for describing quantities, such as IHS colors, volumes, velocities etc.

The CRMsci, together with the CIDOC CRM, can be used to implement more effective generic management functions and powerful queries for related scientific data than other standards. It allows for fully connecting scientific data with culturally relevant contexts. It aims at providing super classes and super properties for any discipline-specific extensions, such that any entity referred to by a compatible extension can be reached with a more general query based on this model.
This model aims at staying harmonized with the CIDOC CRM, i.e., its maintainers submit proposals for modifying the CIDOC CRM wherever adequate to guarantee the overall consistency, disciplinary adequacy and modularity of CRM-based ontology modules.
Finally, we propose to open the discussions in CIDOC about conceptual modelling of products of human activities in general and we suggest to CIDOC to approve that modelling scientific activities is a valid scope for CIDOC and could be a working item for the CRM-Special Interest Group.

References

CRMarchaeo (2014): the Excavation Model. An Extension of CIDOC-CRM to support archaeological excavations. Contributers: : Paul Cripps, Martin Doerr, Sorin Hermon, Gerald Hiebel, Athina Kritsotaki, Anja Masur, Keith May, Wolfgang Schmidle, Maria Theodoridou, Despoina Tsiafaki, http://www.ics.forth.gr/isl/index_main.php?l=e&c=711 (accessed by 15.08.2014)

Darwin Core (2013), Darwin Core, Available from http://rs.tdwg.org/dwc/index.htm, [Accessed, 30th July 2014]

Darwin-SW, Semantic web terms for biodiversity data, based on Darwin Core". Retrieved 29/08/2014, from https://code.google.com/p/darwin-sw/

Daston L. (2011) Histories of Scientific Observation, Max Planck Institute for the History of Science, feature stories, (March), Available form http://www.mpiwg-berlin.mpg.de/en/news/features/feature18/en/main.pt

Doerr M., Hiebel G. and Eide Ø., 2013. CRMgeo: Linking the CIDOC CRM to GeoSPARQL through a Spatiotemporal Refinement TECHNICAL REPORT: ICS-FORTH. http://www.ics.forth.gr/tech-reports/2013/2013.TR435_CRMgeo_CIDOC_CRM_GeoSPARQL.pdf

Doerr, M., Kritsotaki, A., & Boutsika, A. (2011). Factual argumentation - a core model for assertions making. Journal on Computing and Cultural Heritage (JOCCH) , 3(3), 34, New York, NY, USA : ACM .

Explorable(2014), Scientific Observation, available from https://explorable.com/scientific-observation

Hugh G. Gauch Jr(2003)., Scientific Method in Practice, Cornell University, Cambridge University Press, Cambridge, UK

INSPIRE, Infrastructure for Spatial Information in the European Community. Available from http://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:32008R1205&from=EN [Accessed, 30th July 2014]

Madin, J., Bowers, S., Schildhauer, M., Krivov, S., Pennington, D., and Villa, F. (2007a) "An ontology for describing and synthesizing ecological observation data. Ecological Informatics, 2, 279–96.

Madin J., Bowers S, Schildhauer M., Jones M. (2007b) , "Advancing ecological research with ontologies, Trends in Ecology & Evolution", Volume 23, Issue 3, March 2008, Pages 159–168, Elsevier Ltd 2007.

Sagan C., (1997) *The Demon-Haunted World: Science As a Candle in the Dark*. Ballantine Books, (March) ISBN 0-345-40946-9, 480

Stucky B., Deck J, Guralnick R, Conlin T (2013), BiSciCol, Triples, and Darwin Core, Available from http://biscicol.blogspot.gr/ [Accessed, 30th July 2014]

Theodoridou M., · Tzitzikas Y., Doerr M. · Marketakis Y, Melessanakis V.(2010), "Modeling and querying provenance by extending CIDOC CRM", Distributed and Parallel Databases, Volume 27, Number 2, 169-210, DOI: 10.1007/s10619-009-7059-2, Springer Netherlands

Tzitzikas Y. Alloca C., Bekiari C., Marketakis Y., Fafalios P., Doerr M., Minadakis N., Patkos T. Candela L. (2013) Integrating Heterogeneous and Distributed Information about Marine Species through a Top Level Ontology, Proceedings of the 7th Metadata and Semantic Research Conference, MTSR'13, Thessaloniki, Greece, (November)

Webb Campbell, Baskauf Steven, (2011) "Darwin-SW: Darwin Core data for the SemanticWeb", TDWG Annual Meeting; 2011-10-18 (accessed from "http://www.tdwg.org/fileadmin/2011conference/slides/Webb_DarwinSW.pdf")

Webb Campbell & Steven Baskauf (2013). (TDWG Annual Meeting 2013-11-01). "Using Darwin-SW to answer questions about Biodiversity Resources". Retrieved 29/08/2014 from "http://www.tdwg.org/fileadmin/2013conference/slides/Baskauf_Darwin-SW.pdf"

Wenning C. (2009) Scientific epistemology: How scientists know what they know. J. Phys. Tchr. Educ. Online, 5(2), (Autumn). Available from http://www.phy.ilstu.edu/pte/publications/scientific_epistemology.pdf [Accessed, 30th July 2014]