Towards an Ontology for Data-driven Discovery of New Materials

Kwok Cheung[1], John Drennan[1], Jane Hunter[2]

[1]AIBN, The University of Queensland
St Lucia, Queensland, Australia
[2]ITEE, The University of Queensland
St. Lucia, Queensland, Australia
(kwokc, jane)@itee.uq.edu.au, j.drennan@uq.edu.au

Abstract

Materials scientists and nano-technologists are struggling with the challenge of managing the large volumes of multivariate, multidimensional and mixed-media data sets being generated from the experimental, characterisation, testing and post-processing steps associated with their search for new materials. In addition, they need to access large publicly available databases containing: crystallographic structure data; thermodynamic data; phase stability data and ionic conduction data. Materials scientists are demanding data integration tools to enable them to search across these disparate databases and to correlate their experimental data with the public databases, in order to identify new fertile areas for searching. Systematic data integration and analysis tools are required to generate targeted experimental programs that reduce duplication of costly compound preparation, testing and characterisation. This paper presents MatOnto – an extensible ontology, based on the DOLCE upper ontology, that aims to represent structured knowledge about materials, their structure and properties and the processing steps involved in their composition and engineering. The primary aim of MatOnto is to provide a common, extensible model for the exchange, re-use and integration of materials science data and experimentation.

Introduction and Objectives

The advent of the high-throughput, combinatorial and robotic laboratory instruments, atomic resolution microscopes and high speed modelling and simulation software tools are triggering an explosive growth in the magnitude and complexity of materials data. Materials science data ranges from complex compound preparation and processing workflows, to spectrographic analyses, 2D nano-scale microsocopy images, textual publications, numerical data, animations to 3D crystallographic structures and complex phase diagrams.

Materials informatics is emerging as a new discipline addressing the issues of data management, curation, integration and analysis that are challenging materials scientists. Materials informatics is defined as *the high speed robust acquisition, management, analysis and dissemination of diverse materials data.* Materials data access, acquisition, interoperability and curation were recently identified as critical cyberinfrastructure imperatives for the materials science community [1, 22].

Critical requirements include: persistent unique identifiers materials science resources; metadata standards for describing samples, processes, properties; common semantic models/ontologies to enable mapping between database schemas, information integration and semantic; laboratory information and provenance capture systems that capture the processes both in the laboratory as well as in the post-processing of the data. Semantic Web technologies are essential to addressing these issues, and the development of Materials Ontology (MatOnto) is a significant step towards an integrated solution.

Ontologies provide rich machine-processable semantic descriptions; formal definitions of domains by defining classes, properties and relationships between them in Web Ontology Language (OWL); and a basis to enable reasoning or deduction of new information. Ontologies enable semantic interoperability between resources, services, databases, and devices via inter-related knowledge structures.

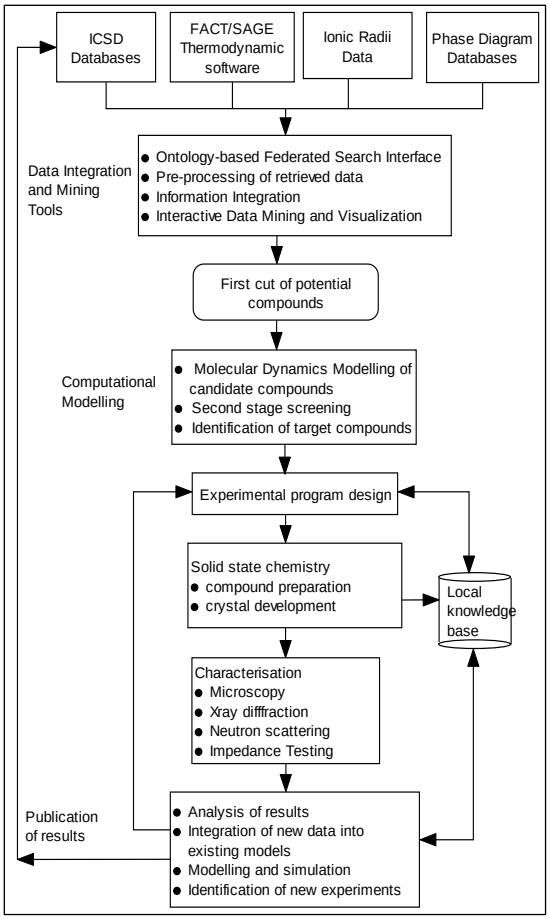
The remainder of this paper describes the MatOnto ontology, which aims to:

- Provide an extensible framework that encapsulates the top level structured knowledge of materials science;
- Enable integration of and mapping between disparate databases within the materials science domain;
- Enable the modelling and capture of precise provenance data in both the digital and physical domains. This is essential to enable verification, validation, comparison and re-use of experimental results;
- Enable the inferencing and extraction of new knowledge via SWRL rules and a reasoning engine.

Example Scenario

At the University of Qld, we are working with fuel cell scientists who are searching for novel oxygen ion conducting materials that can operate more efficiently at lower temperatures for longer durations. The electrolyte compound must have oxygen conductivities $> 10^{-1}$ Scm^{-1} and mechanical and chemical stability at

elevated temperatures (500°C). Based on past experience and intuitive knowledge, fuel cell scientists want rapid answers to queries such as: "Give me compounds that contain tungsten-oxygen-X (where X is a different cation), with bond lengths between Y and Z nm, with large anomalies and anisotropy in the positional parameters of oxygen, with bond angles between J° and K° and which are stable below 500 °C". Figure 1 below illustrates the scientific workflow and methodology for this this research project.



To answer such queries, scientists currently have to manually search, retrieve, process and correlate data from a number of related but disparate databases including: the Inorganic Crystal Database (ICSD); the thermodynamic FactSage database; the Ionic Radii Database; and NIST Phase Equilibria Diagrams database. One of the greatest hurdles to this process is that the search interfaces, metadata terms, data structures, formats and metrics are inconsistent across these databases. (For example, *temperature factors* can be represented in three different formats (Isotropic Temperature Factor (ITF), temperature factor (β) and the mean square amplitude of vibration (U)). Sophisticated database integration and mining tools are required so fuel cell scientists can more easily retrieve answers to such queries. Fast and intelligent tools are required to seamlessly interrogate, retrieve, integrate and present data so users can iteratively hone in on areas of interest. XML schemas (e.g., the Materials Markup Language MatML) can specify the data structures and formats within each database. However, machine-processable ontologies are required to provide the semantic mappings between related terms and to dynamically pre-process, integrate and correlate data from the disparate databases.

Related Work

There have been a number of independent efforts in ontology development within the materials science community. These prior efforts have focussed on: 1) developing ontologies for data integration; 2) extracting knowledge from text; 3) a cross-disciplinary classification scheme for nanoscale research; and 4) demonstrating ontology development based on object-oriented database schemas.

Ashino et al [2] presents an ontology for the material selection process, that aims to integrate material properties such as the creep property, with materials databases via a standardized XML schema. Second, The PLINIUS ontology [3] was designed specifically for the domain of ceramic materials and was aimed at semi-automatic knowledge extraction from texts. Tanaka [4] presents a meta-level ontology as a multi-disciplinary classification scheme within Nanoscience and Nanotechnology. Finally, Ono et al. [5] argues that a set of object-oriented classes in a domain provides an ideal framework for ontology development due to the high similarity between the specifications of an ontology and an object-oriented system.

Ashino's and PLINIUS's ontologies focus on the details of specific sub-disciplines within materials science while Tanaka provides a coarse classification of sub-disciplines within nanoscale research. All of them fail to mention the importance of extensibility of their ontologies. In contrast, MatOnto has been developed upon DOLCE [6] in order to address this issue. As a result of this design, other domain ontologies can be harmonized with MatOnto via the upper ontology.

Ashino's and Tanaka's ontologies are primarily extracted from materials data standards or prestigious data sources. While Ono's proposed phase diagram ontology includes reinvented terms for new concepts. The PLINIUS ontology develops complex concepts based on atomic concepts and construction rules. The PLINIUS approach may work within a limited scope of a problem domain, but it is doubtful that such an approach will satisfactorily capture every term within the materials science domain accurately. Consequently, we chose to develop MatOnto by merging Ashino's and Tanaka's approaches with DOLCE – and also calling on existing standards such as MatML.

MatOnto Development

MatOnto's design principles are to provide an ontology that: 1) is based on an upper ontology and so can easily be extended to accommodate and harmonize with both existing and evolving ontologies; 2) leverages existing peer-reviewed ontologies or vocabularies developed through community consensus; 3) enables integration of those high priority databases identified by our fuel cell collaborators. Below we describe the six steps in the process of developing the MatOnto ontology.

Firstly, we decided to use DOLCE [6] the upper ontology developed by the Laboratory for Applied

Ontology (LOA), as the upper basis for MatOnto. DOLCE stems from the *Entity* root class. *Entity* has three subclasses: E*ndurant, Perdurant* and *Abstract*, from which we define MatOnto subclasses. Figure 1 demonstrates MatOnto's top-level view. The green classes are from DOLCE, the yellow classes are from existing ontologies, the purple classes are our extensions, and the white classes are those we have developed specifically for MatOnto.

Secondly we leveraged a number of existing peer-reviewed ontologies and a classification system: Ontolingua's Standard Units and Dimensions [7] (classes with prefix *ontolingua)*; the Joint Academic Classification of Subjects (JACS) (classes with prefix *jacs)* [8]; W3C's Time Ontology in OWL [9] (classes with prefix w3c); and AIFB's Semantic Web for Research Communities (SWRC) ontology [10] (classes with prefix *swrc)*.

Thirdly, we extended EXPO [11] an ontology for describing scientific experiments with the ABC Metadata Ontology [12] in order to enrich EXPO with the concepts of *events* and *processes*. EXPO is primarily a taxonomy of scientific experiments, while the ABC Ontology models events in both the physical domain and a digital object's lifecycle. Figure 2 demonstrates the high-level view of the merger of EXPO and the ABC ontology. The purple nodes are EXPO's classes while the blue arrows are the ABC's object properties. In addition, there are also classes (yellow) from the reused ontologies and the customized MatOnto classes (white) required to complement the modelling of scientific experiments.

Fourthly, we developed the top-level ontology for materials science according to [13] and [14], beginning with the *matonto:Material* class, which is linked to *jacs:Materials Science* of the Joint Academic Classification of Subjects. We have identified five core properties associated with *matonto:Material* shown in Figure 3:
1) *matonto*:*Property* – the materials property;
2) *matonto:Family* – the materials classification;
3) *matonto*:*Process* – the materials manufacturing and measurement processes;
4) *matonto*:*Structure* – the material structure; and
5) *matonto*:*Measurement Data* – the data resulting from the materials measurement/characterisation process. We have drilled down to certain levels and structured the involved concepts in a logical way.

Fifthly, we developed a sub-disciplinary ontology describing the concepts associated with crystalline structures according to Crystallographic Information Framework [15] and subsumed it to under class *matonto:Crystalline*. Figure 4 demonstrates the top-level view of the Crystalline Structure Ontology.

Finally we developed a Scientific Data ontology to describe the different types of data. Figure 5 demonstrates a high-level view of the MatOnto ScientificData ontology.

Linking of all of these sub-ontologies via common classes generates the complete MatOnto ontology.

Discussion

Achievements

MatOnto satisfies the objectives outlined in the *Introduction and Objectives* section. It enables easy harmonization and incorporation of existing related sub-disciplinary and relevant ontologies through the DOLCE upper ontology and the top level materials science classes shown in Figure 1 and 3. The Crystalline Structure Ontology shown in Figure 4 enables the integration of and mapping between, the Inorganic Crystal Structure Database (ICSD) [16] and Ionic Radii databases [17]. The extended EXPO Ontology shown in Figure 2 enables the capture of precise provenance data and the inferencing of new knowledge (e.g., relationships between nodes that are not explicitly related). This aspect is used to automatically infer coarse-grained views of the scientific methodology from fine grained provenance trails, for publication or elearning purposes.

Ontological Assessment

MatOnto's quality has been assessed based on Gruber's five criteria [18]: clarity, coherence, extendibility, minimal encoding bias, and minimal ontological commitment - with satisfactory results. First, MatOnto is clear because its vocabulary is sourced from peer-reviewed ontologies and existing standardized taxonomies. Secondly, MatOnto does not have incoherency issues because no concepts are derived via inferencing. Thirdly, MatOnto is extensible because DOLCE together with JACS provides a platform for harmonizing disciplinary ontologies. The high-level materials science ontology provides a platform for integrating sub-disciplinary ontologies within materials science, e.g. the Crystalline Structure Ontology. MatOnto has no encoding bias because it is free of implementation details. MatOnto has low ontological commitment because we have reused existing peer-reviewed ontologies and extended them based on standardized vocabularies.

Evaluation

We have represented MatOnto in the Web Ontology Language (OWL) and we are in the process of evaluating it through its application within three software tools that we have developed and that are being user tested by our materials science collaborators: 1) a federated ontology-based search interface that enables materials scientists to search, retrieve and integrate data from the ICSD, Ionic Radii and Phase Diagram databases; 2) a scientific workflow system [19] that collects scientific results with provenance data during a fuel-cell manufacturing process, 3) SCOPE [20] a Scientific Compound Object Publishing and Editing tool that generates OAI-ORE compliant compound objects. This tool enables the visualization and exploration of provenance trails by expanding or collapsing links between nodes in the scientific workflow. A set of SWRL rules are being developed, specifically for materials science, that can be executed using the Pellet reasoning engine, to infer new implicit relationships and knowledge from explicit data. Inferencing is applicable to a number of aspects of materials science, including:

- Inferring relationships between processing parameters and structure
- Inferring relationships between structure and properties or behaviour
- Inferring structural features from automatic image analysis of microscopy images
- Inferring coarse grained views of provenance from fine-grained provenance trails.

We plan to further explore the application of semantic inferencing to knowledge extraction from materials science data, in the near future.

Conclusions

In this paper, we have described MatOnto - an ontological framework that encapsulates the knowledge structure of materials science and that can be easily extended to harmonize with and incorporate related ontologies. MatOnto enables materials scientists to search, retrieve and integrate data from heterogeneous and disparate data sources, based on a common set of ontological terms. It also enables the capture of processing steps and provenenace information both within the laboratory as well as within the computing environment. This enables the repeatability, exchange, comparison and re-use of experimental results. The MatOnto ontology also provides the potential for inferencing and extraction of new knowledge using SWRL rules defined by domain experts (e.g., fuel cell scientists) and a reasoning engine (e.g., Pellet). MatOnto provides an essential and fundamental component of the cyberinfrastructure requirements of the materials science community.

References

1. Hunt, W., *Materials informatics: Growing from the Bio World.* JOM Journal of the Minerals, Metals and Materials Society, 2006. **58**(7): p. 88-88.
2. Ashino, T. and M. Fujita, *Definition of a web ontology for design-oriented material selection.* Data Science Journal, 2006. **5**: p. 52-63.
3. van der Vet, P.E., P.-H. Speel, and N.J.I. Mars. *The PLINIUS Ontology of Ceramic Materials*. in *the Eleventh European Conference on Artificial Intelligence (ECAI'94) Workshop on Comparison of Implemented Ontologies*. 1994. Amsterdam: The Netherlands.
4. Tanaka, M. *Toward a Proposed Ontology for Nanoscience*. in *CAIS/ACSI 2005: Data, Information, and Knowledge in a Networked World*. 2005. The University of Western Ontario, London, Ontario.
5. Ono, N., et al. *Ontology for phase diagram databases*. in *Intelligent Processing and Manufacturing of Materials, 1999. IPMM '99. Proceedings of the Second International Conference on*. 1999.
6. Aldo, G., et al., *Sweetening Ontologies with DOLCE*, in *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management. Ontologies and the Semantic Web*. 2002, Springer-Verlag.
7. Gruber, T.R. and G.R. Olsen. *An Ontology for Engineering Mathematics*. in *Fourth International Conference on Principles of Knowledge Representation and Reasoning*. 1994. Gustav Stresemann Institut, Bonn, Germany: Morgan Kaufmann.
8. *The Joint Academic Classification of Subjects*. 2000 [cited 1 November 2007]; Available from: http://www.hesa.ac.uk/dox/jacs/JACS_complete.pdf.
9. *Time Ontology in OWL - W3C Editor's Draft*. 2005 20 September 2005 [cited 1 November 2007]; Available from: http://www.isi.edu/~pan/SWBP/time-ontology-note/time-ontology-note.html.
10. Sure, Y., et al., *The Ontology – Semantic Web for Research Communities*, in *Progress in Artificial Intelligence*. 2005. p. 218-231.
11. Soldatova, L.N. and R.D. King, *An ontology of scientific experiments.* Journal of The Royal Society Interface, 2006. **3**(11): p. 795-803.
12. Lagoze, C. and J. Hunter, *The ABC Ontology and Model.* Journal of Digital Information, 2001. **2**(2).
13. Ashby, M., H. Shercliff, and D. Cebon, *Materials : engineering, science, processing and design*. First ed. 2007, Amsterdam, Boston: Butterworth-Heinemann.
14. *Springer handbook of materials measurement methods*, ed. H. Czichos, T. Saito, and L. Smith. 2006: Springer.
15. IUCr. *Crystallographic Information Framework Version 1.1 Working Specification. International Union of Crystallography*. 2002 [cited 1 November 2007]; Available from: http://www.iucr.org/iucr-top/cif/spec/version1.1/index.html.
16. Belsky, A., et al., *New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design.* Acta Crystallographica Section B, 2002. **58**(3 Part 1): p. 364-369.
17. Shannon, R., *Revised effective ionic radii and systematic studies of interatomic distances in halides and chalcogenides.* Acta Crystallographica Section A, 1976. **32**(5): p. 751-767.
18. Gruber, T.R., *Toward Principles for the Design of Ontologies Used for Knowledge Sharing.* International Journal Human-Computer Studies, 1995. **43**(5-6): p. 907-928.
19. Hunter, J. and K. Cheung. *Generating eScience Workflows from Statistical Analysis of Prior Data*. in *APAC'05*. 2005. Royal Pines Resort, Gold Coast.
20. Cheung, K., et al. *SCOPE - A Scientific Compound Object Publishing and Editing System*. in *3rd International Digital Curation Conference*. 2007. Washington DC, USA.
21. Goble, C. *Position Statement: Musings on Provenance, Workflow and (Semantic Web) Annotations for Bioinformatics*. in *Workshop on Data Derivation and Provenance*. 2002.
22. From Cyberinfrastructure to Cyberdiscovery in Materials Science. Report from a workshop held in Arlington, Virginia August, 2006 http://www.mcc.uiuc.edu/nsf/ciw_2006/NSFDMRCyberreportFinal061128.pdf

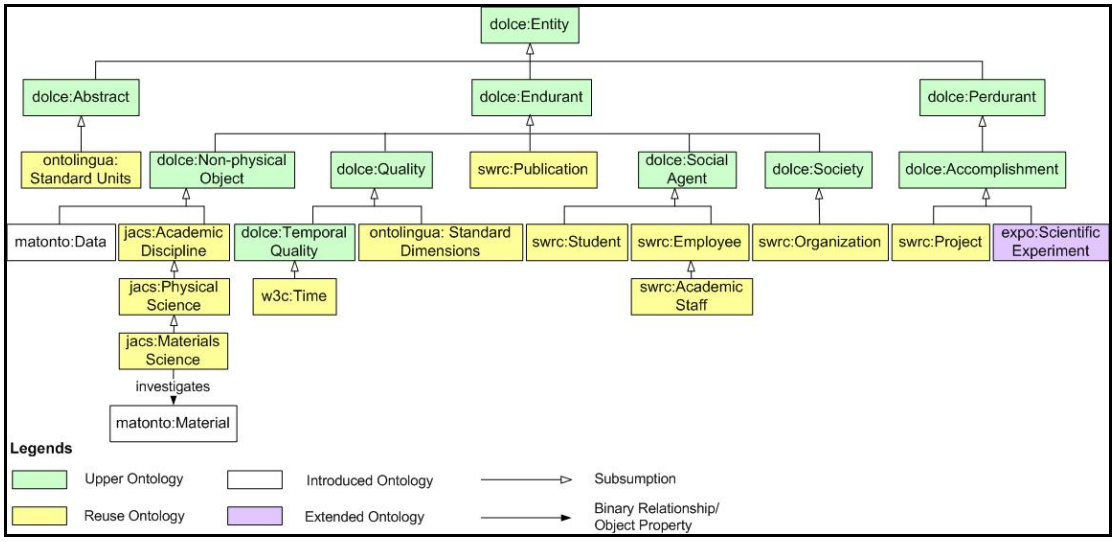


Figure 1 MatOnto's top level view

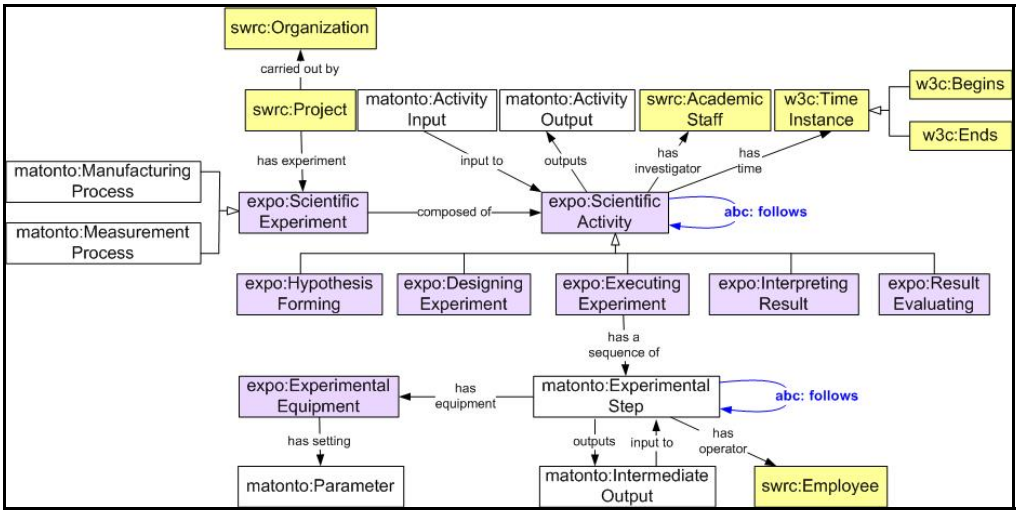


Figure 2 EXPO extended with the ABC Ontology

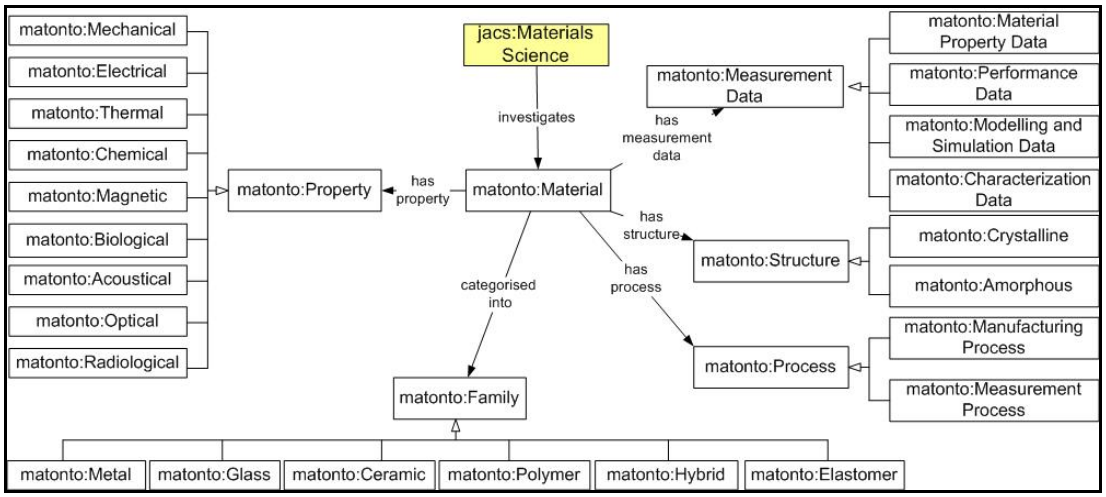


Figure 3 The Top-level Knowledge of Materials Science

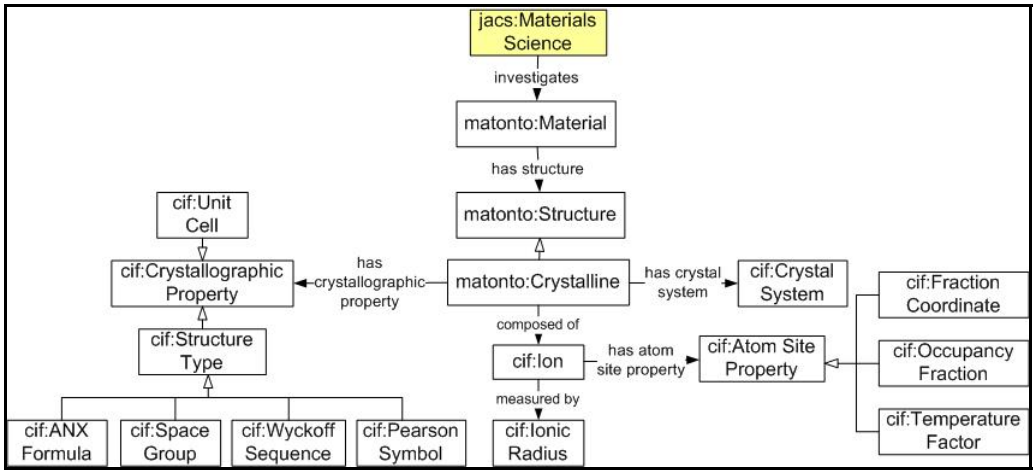


Figure 4 Crystalline Structure Ontology

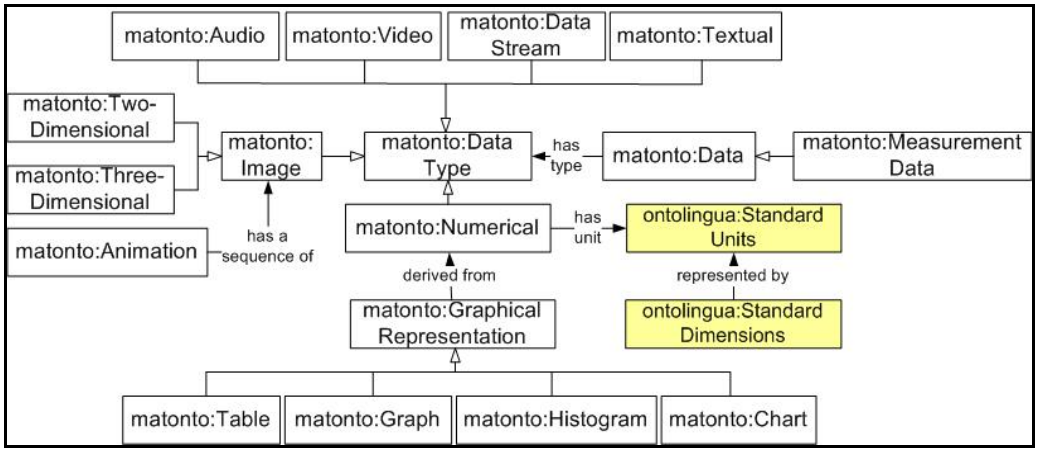


Figure 5 A high-level view of the scientific data ontology