FAIR Big Data in the Materials Design Domain

Patrick Lambrix, Rickard Armiento, Anna Delin, Huanyu Li

Definitions

To speed up the progress in the field of materials design, a number of challenges related to big data need to be addressed. This entry discusses these challenges and shows the semantic technologies that alleviate the problems related to Variety, Variability, Veracity and FAIRness.

Overview

Materials design and materials informatics is central for technological progress, not the least in the green engineering domain. Many traditional materials contain toxic or critical raw materials, whose use should be avoided or eliminated. Also, there is an urgent need to develop new environmentally friendly energy technology. Presently, relevant examples of materials design challenges include energy storage, solar cells, thermoelectrics, and magnetic transport (Ceder and Persson (2013); Jain et al (2013); Curtarolo et al (2013)).

The space of potentially useful materials yet to be discovered — the so-called *'chemical white space'* — is immense. The possible combinations of, say, up to six different elements, constitute many billions. The space is further extended by possibilities of different phases, low-dimensional systems, nanostructuring, and so forth, which adds several orders of magnitude. This space was traditionally explored by experimental techniques, *i.e.*, materials synthesis and subsequent experimental characterization. Parsing and searching the full space of possibilities this way is, however, hardly practical. Recent advances in condensed matter theory and materials modeling make it possible to generate reliable materials data by means of computer simulations based on quantum mechanics (Lejaeghere et al (2016)). High-throughput simulations

combined with machine learning can speed up progress significantly and also help to break out of local optima in composition space to reveal unexpected solutions and new chemistries (Gaultois et al (2016)). The progress brought by the combination of machine learning models and databases of materials data, is now so rapid that it can be discussed as a lead-up to a *singularity* for the field of materials design (Armiento (2020)).

This development has led to several global efforts to assemble and curate databases that combine experimentally known and computationally predicted materials properties, along with a desire to make them interoperable (e.g., OPTIMADE, `https://www.optimade.org/`). These efforts have collectively been referred to as the Materials Genome Initiative (`https://www.mgi.gov/`). A central idea is that materials design challenges can be addressed by searching these databases for entries with desired combinations of properties. Nevertheless, these data sources also open up for *materials informatics*, i.e., the use of big data methodology and data mining techniques to discover new physics from the data itself. A workflow for such a discovery process can be based on a typical data mining process, where key factors are identified, reduced and extracted from heterogeneous databases, similar materials are identified by modeling and relationship mining and properties are predicted through evaluation and understanding of the results from the data mining techniques (Agrawal and Alok (2016)). The use of the data in such a workflow requires addressing problems with data integration, provenance, and semantics, which remains an active field of research.

Even when a new material has been invented and synthesized in a lab, much work remains before it can be deployed. Production methods allowing manufacturing the material at large scale in a cost effective manner need to be developed, and integration of the material into the production must be realized. Furthermore, life-cycle aspects of the material need to be assessed. Today, this post-invention process takes typically about two decades (Mulholland and Paradiso (2016); Jain et al (2013)). Shortening this time is in itself an important strategic goal, which could be realized with the help of an integrated informatics approach (Jain et al (2013)).

To summarize, it is clear that materials data, experimental as well as simulated, has the potential to speed up progress significantly in many steps in the chain starting with materials discovery, all the way to marketable product. However, the data needs to be suitably organized and easily accessible, which in practice is highly nontrivial to achieve. It requires a multidisciplinary effort and the various conventions and norms in use need to be integrated. Materials data is highly heterogeneous and much of it is currently hidden behind corporate walls (Mulholland and Paradiso (2016)).

Big and FAIR Data Challenges

To implement the data-driven materials design workflow, we need to deal with several of the big data properties (e.g. Rajan (2015)).

Volume refers to the quantity of the generated and stored data. The size of the data determines the value and potential insight. Although the experimental materials science does not generate huge amounts of data, computer simulations with accuracy comparable to experiments can. Moreover, going from state-of-the-art static simulations at temperature $T = 0$ K towards realistic descriptions of materials properties at temperatures of operation in devices and tools will raise these amounts as well.

Variety refers to the type and nature of the data. The materials databases are heterogeneous in different ways. They store different kinds of data and in different formats. Some databases contain information about materials crystal structure, some about their thermochemistry, others about mechanical properties. Moreover, different properties may have the same names, while the same information may be represented differently in different databases.

Velocity refers to the speed at which the data is generated and processed to meet the demands and challenges that lie in the path of growth and development. In computational materials science new data is generated continuously, by a large number of groups all other the world. In principle, one can store summary results and data streams from a specific run as long as one needs (days, weeks, years) and analyze it afterwards. However, to store all the data indefinitely may be a challenge. Some data needs to be removed as the storage capacity is limited.

Variability deals with the consistency of the data. Inconsistency of the data set can hamper processes to handle and manage it. This can occur for single databases as well as data that was integrated from different sources.

Veracity deals with the quality of the data. This can vary greatly, affecting accurate analysis. The data generated within materials science may contain errors, and it is often noisy. The quality of the data is different in different databases. It may be challenging to have provenance information from which one can derive the data quality. Not all the computed data is confirmed by lab experiments. Some data is generated by machine learning and data mining algorithms.

Furthermore, to be able to enable machines to automatically find and use the data, and individuals to easily reuse the data, we need to make our data *FAIR* (Findable, Accessible, Interoperable, and Reusable, Wilkinson et al (2016)). Findable refers to the fact that data and metadata should be easy to find, accessible to the fact that it should be clear how to access the data, interoperable to the fact that the data needs to be integrated with other data and be usable by applications and workflows, and reusable to the fact that data and metadata are well described such that the data can be replicated or combined in different settings (`https://www.go-fair.org/fair-principles/`). Also in the materials science domain an awareness regarding the importance of such principles for data storage and management is developing and research in this area is starting (Draxl and Scheffler (2018)).

Sources of data and Semantic Technologies

Although the majority of materials data that has been produced by measurement

or through predictive computation have not yet become organized in general easy-to-use databases, several sizable databases and repositories do exist. However, as they are heterogeneous in nature, semantic technologies are important for the selection and integration of the data to be used in the materials design workflow. This is particularly important to deal with Variety, Variability and Veracity. These technologies have also been proposed to play a significant role in making data FAIR.

In materials science the interest in using semantic technologies is growing rapidly and the development of ontologies and standards is pursued by more and more groups. Ontologies aim to define the basic terms and relations of a domain of interest, as well as the rules for combining these terms and relations. They standardize terminology in a domain and are a basis for semantically enriching data, integration of data from different databases (Variety), reasoning over the data (Variability and Veracity) and making data FAIR.

Furthermore, standards for exporting data from databases and between tools are being developed. These standards provide a way to exchange data between databases and tools, even if the internal representations of the data in the databases and tools are different. They are a prerequisite for efficient materials data infrastructures that allow for the discovery of new materials (Austin (2016)). In several cases the standards formalize the description of materials knowledge (and thereby create ontological knowledge).

In the remainder of this section a brief overview of databases, ontologies and standards in the field is given.

Databases and repositories

The Inorganic Crystal Structure Database (ICSD, `https://icsd.fiz-karlsruhe.de/`) is a frequently utilized database for completely identified inorganic crystal structures, with nearly 200k entries (Belsky et al (2002); Bergerhoff et al (1983)). The data contained in ICSD serve as an important starting point in many electronic structure calculations. Several other crystallographic information resources are also available (Glasser (2016)). A popular open access resource is the Crystallography Open Database (COD, `http://www.crystallography.net/cod/`) with nearly 400k entries (Grazulis et al (2012)). Closely related to COD is the Predicted Crystallography Open Database (PCOD, `http://www.crystallography.net/pcod/`) with over 1 million predicted crystal structures.

At the International Centre for Diffraction Data (ICDD, `http://www.icdd.com/`) a number of databases for phase identification are hosted. These databases have been in use by experimentalists for a long time.

Springer Materials (`http://materials.springer.com/`) contains among many other data sources the well-known Landolt Bornstein database, an extensive data collection from many areas of physical sciences and engineering. Similarly, The Japan National Institute of Material Science (NIMS) Materials Database MatNavi (`http://mits.nims.go.jp/index_en.html`) contains a wide collection of mostly experimental but

also some computational electronic structure data.

Thermodynamical data, necessary for computing phase diagrams with the CALPHAD method, exists in many different databases (Campbell et al (2014)). Open access databases with relevant data can be found through OpenCalphad (http://www.opencalphad.com/databases.html).

Databases of results from electron structure calculations have existed in some form for several decades. In 1978, Moruzzi, Janak, and Williams published a book with computed electronic properties such as, e.g., density of states, bulk modulus and cohesive energy of all metals (Moruzzi et al (2013)). Only in the last few years, however, has the idea of collecting computed results at a large scale in publicly available databases for general use become widespread. Prominent examples of databases or repositories that appeared early during the present trend are Electronic Structure Project (ESP) (http://materialsgenome.se), Aflow (Curtarolo et al (2012), http://aflowlib.org/), the Materials Project (Jain et al (2013), https://materialsproject.org/), the Open Quantum Materials Database (OQMD, http://oqmd.org/) (Saal et al (2013)), and the NOMAD repository (https://repository.nomad-coe.eu/).

There is now a growing demand for open science from funding agencies, regulatory bodies, the scientific community and the general public. Data management plans are becoming mandatory, and making research data, also raw data, available is now expected and becoming the norm in research. This has lead to an explosion of available materials science datasets and archived data of varying quality and usefulness. Many of the above mentioned repositories have made their frameworks available, see, e.g., Automated Interactive Infrastructure and Database for Computational Science (AiiDA, http://www.aiida.net/) (Pizzi et al (2016)), the Atomic Simulation Environment (ASE, https://wiki.fysik.dtu.dk/ase/) (Larsen et al (2017)), and the high-throughput toolkit (httk, http://www.httk.org) (Faber et al (2016)). Popular repositories also include Zenodo (https://zenodo.org/), a catch-all repository for EC funded research developed within the OpenAIRE project, and Materials cloud (https://www.materialscloud.org/), which is specifically built to enable seamless sharing and dissemination of resources in computational materials science.

Ontologies

A number of ontologies in materials science have been developed and we show some characteristics in Table 1.

EMMO (European Materials & Modelling Ontology (https://github.com/emmo-repo/EMMO) is a top level ontology with the purpose to develop a standard representational ontology framework based on knowledge of materials modeling and characterization. Most ontologies, however, are domain ontologies that focus on specific sub-domains of the materials field (Domain in Table 1) and have been developed with a specific use in

mind (Application Scenario in Table 1). MatOnto (Cheung et al (2008)), based on the upper ontology DOLCE, aims to represent structured knowledge, properties and processing steps relevant to materials for data exchange, reuse and integration. MatOWL (Zhang et al (2009)) is extracted from MatML schema data to enable ontology-based data access. The Materials Ontology in Ashino (2010) was designed for data exchange among thermal property databases, particularly focusing on representing knowledge relevant to material processing, measurement methods and manufacturing processes. The NanoParticle Ontology (Thomas et al (2011)), based on the upper ontology BFO, and the eNanoMapper ontology (Hastings et al (2015)) are two ontologies in the nanotechnology domain. The former represents properties of nanoparticles to design new nanoparticles, while the latter focuses on assessing risks caused by the use of nanomaterials in engineering. Extensions to these ontologies are computed in Li et al (2019). The MMOY ontology (Zhang et al (2016)) captures metal materials knowledge from Yago. The Materials Design Ontology (Li et al (2020), `https://w3id.org/mdo/`), inspired by OPTIMADE, aims to enable semantic and integrated querying over multiple heterogeneous materials databases such as Materials Project, OQMD, NOMAD and AFLOW.

From the knowledge representation perspective, the basic terms defined in materials ontologies involve materials, properties, performance, and processing in specific sub-domains. All presented ontologies use OWL as a representation language (Language in Table 1). The number of OWL classes ranges from a few to several thousands (Ontology Metrics in Table 1). Some ontologies have more classes than properties (e.g., MatOnto, Materials Ontology, NanoParticle Ontology, MMOY and EMMO), while some have much more properties (e.g., MDO). Several ontologies are developed in a modular fashion (Modularity in Table 1).

Standards

Early efforts for standards including ISO standards and MatML achieved limited adoption according to Austin (2016). The standard ISO 10303-45 includes an information model for materials properties. It provides schemas for material properties, chemical compositions and measure values (Swindells (2009)). ISO 10303-235 includes an information model for product design and verification. MatML (Kaufman and Begley (2003), `https://www.matml.org/`) is an XML-based markup language for materials property data which includes schemas for such things as materials properties, composition, heat, and production.

Some other standards that have received more attention are, e.g., ThermoML and CML. ThermoML (Frenkel et al (2006, 2011)) is an XML-based markup language for exchange of thermophysical and thermochemical property data. It covers over 120 properties regarding thermodynamic and transport property data for pure compounds, multicomponent mixtures, and chemical reactions. CML or Chemical Markup Language (Murray-Rust and Rzepa (2011); Murray-Rust et al

Table 1 Characteristics of some materials ontologies

Ontologies	Knowledge Representation Perspective			Materials Science Perspective	
	Ontology Metrics	Language	Modularity	Domain	Application Scenario
MatOnto Cheung et al (2008)	78 classes, 10 properties, 24 individuals	OWL	✓	Crystals	Materials discovery
MatOWL Zhang et al (2009)	(not available)	OWL		Materials	Semantic querying
Materials Ontology Ashino (2010)	606 classes, 31 properties, 488 individuals	OWL	✓	Thermal properties	Data exchange, search
ELSSI-EMD ontology CEN (2010)	35 classes, 37 properties, 33 individuals	OWL	✓	Materials testing	Standardization
NanoParticle Ontology Thomas et al (2011)	1904 classes, 81 properties	OWL		Nanotechnology	Data integration, search
eNanoMapper Hastings et al (2015)	12781 classes, 5 properties 464 individuals	OWL	✓	Nanotechnology	Data integration
MMOY Zhang et al (2016)	2325 classes, 9 properties, 1738 individuals	OWL		Metals	Knowledge extraction
MDO Li et al (2020)	37 classes, 64 properties	OWL	✓	Materials design	Semantic querying over multiple databases
EMMO	309 classes, 35 properties, 3 individuals	OWL	✓	Materials science	Upper ontology

(2011)) covers chemistry and especially molecules, reactions, solid-state, computation and spectroscopy. It is an extensible language that allows for the creation of sub-domains through the convention construct. Furthermore, the dictionaries construct allows for connecting CML elements to dictionaries (or ontologies). This was inspired by the approach of the Crystallographic Information Framework or CIF (Bernstein et al (2016), `http://www.iucr.org/resources/cif`).

The European Committee for Standardization (CEN) organized workshops on standards for materials engineering data (Austin (2016)) of which the results are documented in CEN (2010). The work focuses specifically on ambient temperature tensile testing and developed schemas as well as an ontology (the ELSSI-EMD ontology from above).

Another recent approach is connected to the European Centre of Excellence NOMAD (Ghiringhelli et al (2016)). The NOMAD repository's (`https://repository.nomad-coe.eu/`) metadata structure is formatted to be independent of the electronic-structure theory or molecular-simulation code that was used to generate the data and can thus be used as an exchange format.

Conclusion

The use of the materials data in a materials design workflow requires FAIR data and solutions for big data problems including Variety, Variability and Veracity. Semantic technologies are a key factor in tackling some of these problems. Efforts have started in creating materials databases, ontologies and standards. However, much work remains to be done. To make full use of these resources there is a need for integration of different kinds of resources which use different database models and application programming interfaces, and reasoning capabilities should be used, as in the bioinformatics field in the 1990s (Lambrix et al (2009)). Databases could use ontologies to define their schemas and enable ontology-based query-

ing. For existing databases mappings between ontologies and the existing schemas can be created. Integration of databases is enabled by the use of ontologies. However, when databases have used different ontologies, alignments between different ontologies are needed as well (Euzenat and Shvaiko (2007)). Furthermore, more effort should be put on connecting ontologies and standards (as started in the CML, CEN and NOMAD approaches), which may also lead to connections between different standards. Reasoning can be used in different ways. When developing resources reasoning can help in debugging and completing the resources leading to higher quality resources (Ivanova and Lambrix (2013)). Reasoning can also be used during querying of databases as well as in the process of connecting different resources.

References

Agrawal A, Alok C (2016) Perspective: materials informatics and big data: realization of the Fourth paradigm of science in materials science. APL Materials 4:053,208:1–10, DOI 10.1063/1.4946894

Armiento R (2020) Database-driven high-throughput calculations and machine learning models for materials design. In: Schutt KT, Chmiela S, von Lilienfeld OA, Tkatchenko A, Tsuda K, Muller KR (eds) Machine Learning Meets Quantum Physics, Springer International Publishing, Cham, pp 377–395, DOI 10.1007/978-3-030-40245-7˙17

Ashino T (2010) Materials Ontology: An Infrastructure for Exchanging Materials Information and Knowledge. Data Science Journal 9:54–61, DOI 10.2481/dsj.008-041

Austin T (2016) Towards a digital infrastructure for engineering materials data. Materials Discovery 3:1–12, DOI 10.1016/j.md.2015.12.003

Belsky A, Hellenbrandt M, Karen VL, Luksch P (2002) New developments in the Inorganic Crystal Structure Database (ICSD): accessibility in support of materials research and design. Acta Crystallographica Section B: Structural Science 58(3):364–369, DOI 10.1107/S0108768102006948

Bergerhoff G, Hundt R, Sievers R, Brown ID (1983) The inorganic crystal structure data base. Journal of Chemical Information and Computer Sciences 23(2):66–69, DOI 10.1021/ci00038a003

Bernstein HJ, Bollinger JC, Brown ID, Grazulis S, Hester JR, McMahon B, Spadaccini N, Westbrook JD, Westrip SP (2016) Specification of the crystallographic information file format, version 2.0. Journal of Applied Crystallography 49:277–284, DOI 10.1107/S1600576715021871

Campbell CE, Kattner UR, Liu ZK (2014) File and data repositories for Next Generation CALPHAD. Scripta Materialia 70(Supplement C):7–11, DOI 10.1016/j.scriptamat.2013.06.013

Ceder G, Persson KA (2013) How Supercomputers Will Yield a Golden Age of Materials Science. Scientific American 309

CEN (2010) A guide to the development and use of standards compliant data formats for engineering materials test data European Committee for standardization

Cheung K, Drennan J, Hunter J (2008) Towards an Ontology for Data-driven Discovery of New Materials. In: McGuinness D, Fox P, Brodaric B (eds) Semantic Scientific Knowledge Integration AAAI/SSS Workshop, pp 9–14

Curtarolo S, Setyawan W, Wang S, Xue J, Yang K, Taylor R, Nelson L, Hart G, Sanvito S, Buongiorno-Nardelli M, Mingo N, Levy O (2012) AFLOWLIB.ORG: A distributed materials properties repository from high-throughput ab initio calculations. Computational Materials Science 58(Supplement C):227–235, DOI 10.1016/j.commatsci.2012.02.002

Curtarolo S, Hart G, Buongiorno-Nardelli M, Mingo N, Sanvito S, Levy O (2013) The high-throughput highway to computational materials design. Nature Materials 12(3):191, DOI 10.1038/nmat3568

Draxl C, Scheffler M (2018) NOMAD: The FAIR concept for big data-driven materials

science. MRS Bulletin 43(9):676–682, DOI 10.1557/mrs.2018.208

Euzenat J, Shvaiko P (2007) Ontology Matching. Springer

Faber F, Lindmaa A, von Lilienfeld A, Armiento R (2016) Machine Learning Energies of 2 Million Elpasolite $(AB–C˝˙–2˝–D˝˙–6˝)$ Crystals. Physical Review Letters 117(13):135,502, DOI 10.1103/PhysRevLett.117.135502

Frenkel M, Chiroco RD, Diky V, Dong Q, Marsh KN, Dymond JH, Wakeham WA, Stein SE, Konigsberger E, Goodwin ARH (2006) XML-based IUPAC standard for experimental, predicted, and critically evaluated thermodynamic property data storage and capture (ThermoML) (IUPAC Recommendations 2006). Pure and Applied Chemistry 78:541–612, DOI 10.1351/pac200678030541

Frenkel M, Chirico RD, Diky V, Brown PL, Dymond JH, Goldberg RN, Goodwin ARH, Heerklotz H, Konigsberger E, Ladbury JE, Marsh KN, Remeta DP, Stein SE, Wakeham WA, Williams PA (2011) Extension of ThermoML: The IUPAC standard for thermodynamic data communications (IUPAC Recommendations 2011). Pure and Applied Chemistry 83:1937–1969, DOI 10.1351/PAC-REC-11-05-01

Gaultois MW, Oliynyk AO, Mar A, Sparks TD, Mulholland GJ, Meredig B (2016) Perspective: Web-based machine learning models for real-time screening of thermoelectric materials properties. APL Materials 4(5):053,213, DOI 10.1063/1.4952607

Ghiringhelli LM, Carbogno C, Levchenko S, Mohamed F, Huhs G, Lueders M, Oliveira M, Scheffler M (2016) Towards a Common Format for Computational Materials Science Data. PSI-K Scientific Highlights July

Glasser L (2016) Crystallographic Information Resources. Journal of Chemical Education 93(3):542–549, DOI 10.1021/acs.jchemed.5b00253

Grazulis S, Dazkevic A, Merkys A, Chateigner D, Lutterotti L, Quiros M, Serebryanaya NR, Moeck P, Downs RT, Le Bail A (2012) Crystallography Open Database (COD): an open-access collection of crystal structures and platform for world-wide collaboration. Nucleic Acids Research 40(Database issue):D420–D427, DOI 10.1093/nar/gkr900

Hastings J, Jeliazkova N, Owen G, Tsiliki G, Munteanu CR, Steinbeck C, Willighagen E (2015) enanomapper: harnessing ontologies to enable data integration for nanomaterial risk assessment. Journal of Biomedical Semantics 6(1):10, DOI 10.1186/s13326-015-0005-5

Ivanova V, Lambrix P (2013) A unified approach for debugging is-a structure and mappings in networked taxonomies. Journal of Biomedical Semantics 4:10:1–10:19, DOI 10.1186/2041-1480-4-10

Jain A, Ong SP, Hautier G, Chen W, Richards WD, Dacek S, Cholia S, Gunter D, Skinner D, Ceder G, Persson KA (2013) Commentary: The Materials Project: A materials genome approach to accelerating materials innovation. APL Materials 1(1):011,002, DOI 10.1063/1.4812323

Kaufman JG, Begley EF (2003) MatML: A Data Interchange Markup Language. Advanced Materials And Processes 161:35–36

Lambrix P, Stromback L, Tan H (2009) Information Integration in Bioinformatics with Ontologies and Standards. In: Bry F, Maluszynski J (eds) Semantic Techniques for the Web, Springer, Berlin, Heidelberg, pp 343–376, DOI 10.1007/978-3-642-04581-3˙8

Larsen AH, Mortensen JJ, Blomqvist J, Castelli IE, Christensen R, Dulak M, Friis J, Groves MN, Hammer B, Hargus C, Hermes ED, Jennings PC, Jensen PB, Kermode J, Kitchin JR, Kolsbjerg EL, Kubal J, Kaasbjerg K, Lysgaard S, Maronsson JB, Maxson T, Olsen T, Pastewka L, Peterson A, Rostgaard C, Schitz J, Schutt O, Strange M, Thygesen KS, Vegge T, Vilhelmsen L, Walter M, Zeng Z, Jacobsen KW (2017) The atomic simulation environment - a Python library for working with atoms. Journal of Physics: Condensed Matter 29(27):273,002, DOI 10.1088/1361-648X/aa680e

Lejaeghere K, Bihlmayer G, Bjorkman T, Blaha P, Blugel S, Blum V, Caliste D, Castelli IE, Clark SJ, Corso AD, Gironcoli Sd, Deutsch T, Dewhurst JK, Marco ID, Draxl C, Dulak M, Eriksson O, Flores-Livas JA, Garrity KF, Genovese L, Giannozzi P, Giantomassi M, Goedecker S, Gonze X, Grnas O, Gross EKU, Gulans A, Gygi F, Hamann DR, Hasnip PJ, Holzwarth NaW, Iusan D, Jochym DB, Jollet F, Jones D, Kresse G, Koepernik K, Kucukbenli E,

Kvashnin YO, Locht ILM, Lubeck S, Marsman M, Marzari N, Nitzsche U, Nordstrom L, Ozaki T, Paulatto L, Pickard CJ, Poelmans W, Probert MIJ, Refson K, Richter M, Rignanese GM, Saha S, Scheffler M, Schlipf M, Schwarz K, Sharma S, Tavazza F, Thunstrom P, Tkatchenko A, Torrent M, Vanderbilt D, van Setten MJ, Speybroeck VV, Wills JM, Yates JR, Zhang GX, Cottenier S (2016) Reproducibility in density functional theory calculations of solids. Science 351(6280):aad3000, DOI 10.1126/science.aad3000

Li H, Armiento R, Lambrix P (2019) A method for extending ontologies with application to the materials science domain. Data Science Journal 18(1), DOI 10.5334/dsj-2019-050

Li H, Armiento R, Lambrix P (2020) An ontology for the materials design domain. In: Pan JZ, Tamma VAM, d'Amato C, Janowicz K, Fu B, Polleres A, Seneviratne O, Kagal L (eds) The Semantic Web - ISWC 2020 - 19th International Semantic Web Conference, Athens, Greece, November 2-6, 2020, Proceedings, Part II, Springer, Lecture Notes in Computer Science, vol 12507, pp 212–227, DOI 10.1007/978-3-030-62466-8\˙14

Moruzzi VL, Janak JF, Williams ARAR (2013) Calculated electronic properties of metals. Pergamon Press, New York

Mulholland GJ, Paradiso SP (2016) Perspective: Materials informatics across the product lifecycle: Selection, manufacturing, and certification. APL Materials 4(5):053,207, DOI 10.1063/1.4945422

Murray-Rust P, Rzepa HS (2011) CML: Evolution and design. Journal of Cheminformatics 3:44:1–44:15, DOI 10.1186/1758-2946-3-44

Murray-Rust P, Townsend JA, Adams SE, Phadungsukanan W, Thomas J (2011) The semantics of Chemical Markup Language (CML): dictionaries and conventions. Journal of Cheminformatics 3:43, DOI 10.1186/1758-2946-3-43

Pizzi G, Cepellotti A, Sabatini R, Marzari N, Kozinsky B (2016) AiiDA: automated interactive infrastructure and database for computational science. Computational Materials Science 111(Supplement C):218–230, DOI 10.1016/j.commatsci.2015.09.013

Rajan K (2015) Materials Informatics: The Materials "Gene" and Big Data. Annual Review of Materials Research 45:153–169, DOI 10.1146/annurev-matsci-070214-021132

Saal JE, Kirklin S, Aykol M, Meredig B, Wolverton C (2013) Materials Design and Discovery with High-Throughput Density Functional Theory: The Open Quantum Materials Database (OQMD). JOM 65(11):1501–1509, DOI 10.1007/s11837-013-0755-4

Swindells N (2009) The representation and exchange of material and other engineering properties. Data Science Journal 8:190–200, DOI 10.2481/dsj.008-007

Thomas DG, Pappu RV, Baker NA (2011) Nanoparticle ontology for cancer nanotechnology research. Journal of Biomedical Informatics 44(1):59–74, DOI 10.1016/j.jbi.2010.03.001

Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, Blomberg N, Boiten JW, da Silva Santos LB, Bourne PE, Bouwman J, Brookes AJ, Clark T, Crosas M, Dillo I, Dumon O, Edmunds S, Evelo CT, Finkers R, Gonzalez-Beltran A, Gray AJ, Groth P, Goble C, Grethe JS, Heringa J, 't Hoen PA, Hooft R, Kuhn T, Kok R, Kok J, Lusher SJ, Martone ME, Mons A, Packer AL, Persson B, Rocca-Serra P, Roos M, van Schaik R, Sansone SA, Schultes E, Sengstag T, Slater T, Strawn G, Swertz MA, Thompson M, van der Lei J, van Mulligen E, Velterop J, Waagmeester A, Wittenburg P, Wolstencroft K, Zhao J, Mons B (2016) The FAIR guiding principles for scientific data management and stewardship. Scientific data 3:160,018:1–9, DOI 10.1038/sdata.2016.18

Zhang X, Hu C, Li H (2009) Semantic query on materials data based on mapping matml to an owl ontology. Data Science Journal 8:1–17, DOI 10.2481/dsj.8.1

Zhang X, Pan D, Zhao C, Li K (2016) MMOY: Towards deriving a metallic materials ontology from Yago. Advanced Engineering Informatics 30:687–702, DOI 10.1016/j.aei.2016.09.002