Data Descriptor

Not peer-reviewed version

OntoNanoMat: A Semantic Dataset and Ontology for Green-Synthesized Nanomaterials in Environmental Remediation

Carolina L. Recio-Colmenares , Roxana B. Recio-Colmenares , F. E. Castillo-Barrera , Cesar A. Garcia-Garcia
*

Posted Date: 29 January 2026

doi: 10.20944/preprints202601.2271.v1

Keywords: nanomaterials; green synthesis; environmental remediation; ontology; semantic web; FAIR data; knowledge graphs; data descriptor; sustainable nanotechnology; linked data

Data Descriptor

OntoNanoMat: A Semantic Dataset and Ontology for Green-Synthesized Nanomaterials in Environmental Remediation

Carolina L. Recio-Colmenares [1], Roxana B. Recio-Colmenares [1], F. E. Castillo-Barrera [2] and Cesar A. Garcia-Garcia [3,*]

[1] Departamento de Ciencias Básicas y Aplicadas, Centro Universitario de Tonalá, Universidad de Guadalajara, Guadalajara, Mexico
[2] Facultad de Ingeniería, Universidad Autónoma de San Luís Potosí, San Luís Potosí, Mexico
[3] Departamento de Ciencias de la Información y Desarrollos Tecnológicos, Centro Universitario de Tonalá, Universidad de Guadalajara, Guadalajara, Mexico
* Correspondence: cesar.ggarcia@academicos.udg.mx

Abstract

Background: Research on green-synthesized nanomaterials (GSNs) for environmental remediation is growing rapidly, yet data remains fragmented in non-interoperable formats. **Methods:** We present OntoNanoMat, a comprehensive semantic resource consisting of a modular OWL 2 DL ontology and a curated dataset of case studies. The data was structured into five thematic modules: Identification, Synthesis, Mechanism, Performance, and Provenance. **Results:** The dataset is provided in three interoperable formats: CSV for tabular analysis, JSON for web applications, and Turtle (RDF) for Semantic Web integration. Technical validation was performed using SHACL shapes and SPARQL query libraries to ensure logical consistency and data integrity. **Conclusions:** OntoNanoMat provides a FAIR-compliant (Findable, Accessible, Interoperable, and Reusable) foundation for future machine learning applications and knowledge graph integration in sustainable nanotechnology.

Dataset: https://doi.org/10.5281/zenodo.18201276

Dataset License: CC-BY 4.0

Keywords: nanomaterials; green synthesis; environmental remediation; ontology; semantic web; FAIR data; knowledge graphs; data descriptor; sustainable nanotechnology; linked data

1. Summary

The rapid expansion of green nanotechnology has led to a vast but fragmented body of literature regarding the use of green-synthesized nanomaterials (GSNs) for environmental remediation [1]. While these materials offer sustainable alternatives to traditional chemical synthesis, the lack of standardized data structures makes it difficult to perform cross-study comparisons, assess "greenness" objectively, or reuse data for large-scale meta-analyses [2]. To address these challenges, we developed OntoNanoMat, a semantic resource designed to externalize and formalize knowledge in the GSN domain.

The OntoNanoMat dataset was collected through a systematic review of recent scientific literature focusing on green synthesis routes (e.g., biogenic reagents, plant extracts) and their application in removing contaminants like organic dyes and heavy metals from water [3]. The dataset was structured using a modular OWL 2 DL ontology, ensuring that every data point is linked to its chemical precursors, synthesis conditions, and performance indicators (such as removal efficiency and recyclability) [4].

doi:10.20944/preprints202601.2271.v1

This dataset is a core component of a broader research effort at the University of Guadalajara to digitalize material knowledge and promote FAIR (Findable, Accessible, Interoperable, and Reusable) principles in nanotechnology. The creation of this resource was motivated by the need to provide a "ground truth" for semantic integration frameworks, such as the one described in our corresponding research article [5], where we demonstrate how this ontology-based approach enables the interoperable assessment of material sustainability and performance.

By releasing this dataset in multiple formats (CSV, JSON, and Turtle), we aim to provide a ready-to-use resource for the scientific community. Potential benefits include the facilitation of automated data discovery, the training of machine learning models for predicting nanomaterial efficiency, and the integration of GSN data into global Knowledge Graphs for sustainable chemistry.

2. Data Description

The OntoNanoMat dataset is a curated collection of case studies focusing on the environmental application of green-synthesized nanomaterials. The resource is structured to provide high interoperability between tabular processing, web development, and semantic reasoning. All data files are available in the Zenodo repository https://doi.org/10.5281/zenodo.18201276.

2.1. Dataset Files and Formats

The dataset is distributed in three main formats to support different use cases:

- **dataset_case_studies.csv**: A UTF-8 encoded tabular file containing the primary data for statistical analysis.
- **dataset_case_studies.json**: A machine-readable JSON array, ideal for integration into web platforms or NoSQL databases.
- **dataset_case_studies.ttl**: An RDF serialization in Turtle format. This file links the data instances to the classes and properties defined in the green_nanomaterials_ontology.ttl file.

2.2. Tabular Data Structure

The CSV file consists of 35 columns (attributes) per entry. Table 1 describes the main fields and their interpretation.

Table 1. Description of the attributes included in the dataset_case_studies.csv file.

Attribute Group	Column Name	Data Type	Description
Identification	case_id	String	Unique identifier for each case study (e.g., CS1, CS2).
	nanomaterial_name	String	Common name of the synthesized material.
	nanomaterial_type	String	Categorization (e.g., Magnetic nanocomposite, Photocatalyst).
Synthesis	synthesis_route	String	Description of the green synthesis procedure.
	solvent_greenness	String	Qualitative assessment of the solvent (e.g., Low-toxicity).
	renewable_precursor	Boolean	True if biogenic or renewable reagents were used.
Process	mechanism	String	Remediation process (Adsorption or Photocatalysis).
	contaminant_name	String	Name of the target pollutant (e.g., Methylene blue).
	pH	Float	Operational acidity/alkalinity during the process.

Performance	removal_efficiency_percent	Float	Maximum removal percentage achieved.
	qmax_mg_per_g	Float	Maximum adsorption capacity (for adsorption cases).
	cycles	Integer	Number of successful recyclability tests reported.
Provenance	provenance_publication_doi	String	DOI link to the original source of the data.

2.3. Semantic Mapping and Interpretation

The data is designed to be interpreted through the lens of the OntoNanoMat Ontology. In the Turtle (.ttl) version, each record is transformed into an individual of the class gsn:Nanomaterial.

- **Logic Links:** The synthesis descriptors are mapped to the gsn:SustainabilityProfile class, while performance metrics are linked to gsn:PerformanceIndicator.
- **Units:** All numerical values follow standard units: Temperature in Kelvin (K), concentration in g/L, and adsorption capacity in mg/g, as defined by the ontology's datatype properties.

2.4. Validation Resource

In addition to the dataset, the repository includes green_nanomaterials_queries.rq, a library of SPARQL queries. These queries serve as an "executable documentation" that demonstrates how to retrieve and filter data based on multi-dimensional criteria (e.g., finding materials with high efficiency that also use renewable solvents).

3. Methods

The development of the OntoNanoMat resource followed a three-stage methodology: data acquisition through systematic curation, semantic modeling (ontology design), and data transformation (serialization).

3.1. Data Acquisition and Curation

The case studies included in the dataset were retrieved through a systematic search of peer-reviewed literature published between 2018 and 2025. Search queries were conducted in databases such as Scopus, Web of Science, and Google Scholar using combinations of keywords including "green synthesis", "nanomaterials", "environmental remediation", and "sustainable nanotechnology".

Data extraction was performed manually to ensure the high fidelity of technical parameters. For each case study, we recorded:

1. **Synthesis parameters:** Solvent type, precursors, and energy indicators.
2. **Experimental conditions:** pH, temperature, and dosage.
3. **Performance metrics:** Removal efficiency and adsorption capacity (q_{max}).

Numerical values were normalized to standard units (e.g., converting all temperatures to Kelvin and concentrations to mg/L) to facilitate interoperability and comparison.

3.2. Ontology Development

The OntoNanoMat Ontology was developed using the OWL 2 DL (Web Ontology Language) standard. The modeling process followed an iterative approach using Protégé 5.6.x.

- **Modularity:** The ontology was organized into five core modules (Material, Synthesis, Process, Performance, and Provenance) to allow for independent updates.
- **Reusability:** Where possible, classes and properties were aligned with existing vocabularies such as PROV-O for provenance and CHEO or ENM for chemical entities.

doi:10.20944/preprints202601.2271.v1

- **Axiomatization:** Logical restrictions (SubClassOf and EquivalentTo) were implemented to enable automatic classification of "green-synthesized" materials based on their sustainability profiles.

3.3. Data Transformation and RDFization

To generate the multi-format dataset, we followed these steps:

1. **Tabular Structuring:** The curated data was first organized into a master CSV file.
2. **Semantic Mapping:** Using a custom Python-based mapping script, each CSV row was transformed into an RDF individual (instance).
3. **Serialization:** The data was exported into **JSON** for web accessibility and **Turtle (.ttl)** for semantic reasoning. The Turtle version explicitly uses the gsn: namespace defined in the ontology to ensure that the instances are logically bound to their semantic definitions.

3.4. Technical Validation Setup

Validation was not limited to syntax checking. We developed a set of SHACL (Shapes Constraint Language) files to enforce data integrity constraints (e.g., ensuring that any material labeled as "adsorbent" must have an associated q_{max} value). Finally, a library of SPARQL queries was created to verify that the graph could answer complex competency questions regarding material performance and greenness.

4. Technical Validation

The technical quality and integrity of the OntoNanoMat resource were evaluated through a multi-layered validation pipeline.

4.1. Syntactic and Structural Validation

All RDF serializations in Turtle (.ttl) format were validated using the Apache Jena RIOT tool to ensure compliance with W3C standards. This step confirmed that the data is free of syntax errors and ready for ingestion by any standard triplestore. The JSON and CSV files were also checked for schema consistency and UTF-8 encoding integrity.

4.2. Logical Consistency and Reasoning

The green_nanomaterials_ontology.ttl was subjected to automated reasoning using the HermiT 1.4.3 reasoner within the Protégé environment. No logical inconsistencies or unsatisfiable classes were detected. The hierarchy correctly infers individuals into their respective subclasses (e.g., a material with a "biogenic precursor" property is correctly classified as a gsn:GreenSynthesizedMaterial).

4.3. Semantic Validation (SHACL)

To ensure the dataset follows the required structural constraints, we applied Shapes Constraint Language (SHACL). The validation shapes (provided in the repository) verify that:

- Each Nanomaterial entry is linked to at least one RemediationMechanism.
- Quantitative indicators (like removal_efficiency_percent) are restricted to numerical ranges (0–100).
- Mandatory provenance metadata (DOI and year) is present for every record.

4.4. Competency Question Testing

A library of eight SPARQL queries was used to validate the functional utility of the data. These queries successfully retrieved complex cross-referenced information, such as identifying nanomaterials that achieve >90% efficiency while maintaining a "low-toxicity" solvent profile. This confirms that the resource can answer the domain-specific questions for which it was designed.

5. Usage Notes (or User Notes)

The OntoNanoMat resource is designed for researchers in nanotechnology, environmental science, and data engineering.

5.1. Accessing and Exploring the Data

The dataset and ontology can be accessed via the GitHub repository (for version control and issue tracking) or the Zenodo archive (for the stable, citable version).

- **For Nanotechnologists:** The dataset_case_studies.csv file can be opened in any spreadsheet software (Excel, Google Sheets) or R/Python environments for quick benchmarking.
- **For Knowledge Engineers:** The .ttl file should be loaded into Protégé or a triplestore like Apache Jena Fuseki or GraphDB. Users can then execute the provided SPARQL queries to filter materials by specific green chemistry or performance criteria.

5.2. Integration and Extensibility

The modular nature of the ontology allows it to be easily extended. Researchers can add new remediation mechanisms (e.g., membrane filtration) or additional nanomaterial characterization parameters by defining them as subclasses of the existing core classes.

5.3. Software Requirements

No specialized software is required to view the primary data (CSV). However, to fully leverage the semantic features:

1. **Protégé (v5.5 or higher)** is recommended for ontology visualization.
2. **Python (rdflib library)** is suggested for those wishing to programmatically integrate this dataset into machine learning pipelines or larger Knowledge Graphs.

Author Contributions: Conceptualization, C.L.R.-C. and C.A.G.-G.; methodology, C.L.R.-C.; software, C.L.R.-C. and C.A.G.-G.; validation, C.L.R.-C., R.B.R.-C., F.E.C.-B. and C.A.G.-G.; formal analysis, C.L.R.-C.; investigation, C.L.R.-C. and R.B.R.-C.; resources, C.A.G.-G.; data curation, C.L.R.-C.; writing—original draft preparation, C.L.R.-C.; writing—review and editing, R.B.R.-C., F.E.C.-B. and C.A.G.-G.; visualization, C.L.R.-C.; supervision, C.A.G.-G.; project administration, C.A.G.-G.; funding acquisition, C.A.G.-G. All authors have read and agreed to the published version of the manuscript.

Funding: This research was supported in part by the Secretaría de Ciencia, Humanidades, Tecnología e Innovación (SECIHTI) through the SNII scholarship No. 244413, as well as by the Universidad de Guadalajara (UdeG) through institutional funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data and ontology presented in this study are openly available in Zenodo at https://doi.org/10.5281/zenodo.18201276.

Acknowledgments: During the preparation of this manuscript, generative artificial intelligence (Gemini 3.0) was used exclusively to improve the clarity, style, and readability of the text. All content generated with the assistance of this tool was carefully reviewed, validated, and approved by the human authors. The authors take full responsibility for the scientific content, interpretations, and conclusions presented in this work.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

Abbreviations

The following abbreviations are used in this manuscript:

CSV	Comma-Separated Values
DL	Description Logic
DOI	Digital Object Identifier
FAIR	Findable, Accessible, Interoperable, and Reusable
GSN	Green-Synthesized Nanomaterial
IRI	Internationalized Resource Identifier
JSON	JavaScript Object Notation
OWL	Web Ontology Language
RDF	Resource Description Framework
SHACL	Shapes Constraint Language
SPARQL	SPARQL Protocol and RDF Query Language
TTL	Terse RDF Triple Language (Turtle)
W3C	World Wide Web Consortium

References

1. Recio-Colmenares, C.L.; Recio-Colmenares, R.B.; Castillo-Barrera, F.E.; Garcia-Garcia, C.A. An Ontology-Based Framework for Semantic Integration and Interoperable Assessment of Green-Synthesized Nanomaterials for Environmental Remediation. Appl. Sci. 2026, submitted.
2. Arshadi, M.; Faraji, A.R.; Mehravar, M. Green synthesis of magnetic nanoparticles and their application in environmental remediation. J. Clean. Prod. 2023, 410, 137254. DOI:10.1016/j.jclepro.2023.137254.
3. Schweizer, C.; Thomas, A.; Janka-Ramm, M. Digitalizing Material Knowledge: A Practical Framework for Ontology-Driven Knowledge Graphs in Process Chains. Appl. Sci. 2024, 14, 11683. DOI:10.3390/app142411683.
4. Labra-Gayo, J.E.; Iglesias-Préstamo, Á.; Martín-Fernández, D.; Arnaud, M.A. rudof: A Rust Library for handling RDF data models and Shapes. CEUR Workshop Proc. 2024, 3828, paper 32.
5. Wilkinson, M.D.; Dumontier, M.; Aalbersberg, I.J.; Appleton, G.; Axton, M.; Baak, A.; et al. The FAIR Guiding Principles for scientific data management and stewardship. Sci. Data 2016, 3, 160018. DOI:10.1038/sdata.2016.18.
6. Recio-Colmenares, C.L.; Recio-Colmenares, R.B.; Castillo-Barrera, F.E.; Garcia-Garcia, C.A. OntoNanoMat: A Semantic Dataset and Ontology for Green-Synthesized Nanomaterials. Zenodo 2026. DOI:10.5281/zenodo.18201276.
7. Berners-Lee, T.; Hendler, J.; Lassila, O. The Semantic Web. Sci. Am. 2001, 284, 34–43.
8. Titocci, J.; Pulieri, M.; Rosati, I.; Karam, N. Enhancing Trait Thesauri Interoperability Using a Manual and Automated Alignment Approach. Appl. Sci. 2025, 15, 12484. DOI:10.3390/app152312484.
9. Noy, N.F.; McGuinness, D.L. Ontology Development 101: A Guide to Creating Your First Ontology; Stanford Knowledge Systems Laboratory Technical Report KSL-01-05; Stanford University: Stanford, CA, USA, 2001.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.