AI4DiTraRe: Building the BFO-Compliant Chemotion Knowledge Graph

Ebrahim Norouzi[1,2,*], Nicole Jung[3], Anna M. Jacyszyn[1], Jörg Waitelonis[1] and Harald Sack[1,2]

[1] *FIZ Karlsruhe – Leibniz Institute for Information Infrastructure, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany*

[2] *Karlsruhe Institute of Technology, Institute of Applied Informatics and Formal Description Methods, Kaiserstr. 89, 76133 Karlsruhe*

[3] *Karlsruhe Institute of Technology, Institute of Biological and Chemical Systems, Hermann-von-Helmholtz-Platz 1, 76344 Eggenstein-Leopoldshafen, Germany*

Abstract

Chemistry is an example of a discipline where the advancements of technology have led to multi-level and often tangled and tricky processes ongoing in the lab. The repeatedly complex workflows are combined with information from chemical structures, which are essential to understand the scientific process. An important tool for many chemists is Chemotion, which consists of an electronic lab notebook and a repository. This paper introduces a semantic pipeline for constructing the BFO-compliant Chemotion Knowledge Graph, providing an integrated, ontology-driven representation of chemical research data. The Chemotion-KG has been developed to adhere to the FAIR (Findable, Accessible, Interoperable, Reusable) principles and to support AI-driven discovery and reasoning in chemistry. Experimental metadata were harvested from the Chemotion API in JSON-LD format. The JSON-LD, as an RDF serialization, was ingested into the triple store and subsequently transformed into a Basic Formal Ontology-aligned graph through SPARQL CONSTRUCT queries. The source code and datasets are publicly available via GitHub. The Chemotion Knowledge Graph is hosted by FIZ Karlsruhe Information Service Engineering. Outcomes presented in this work were achieved within the Leibniz Science Campus “Digital Transformation of Research” (DiTraRe) and are part of an ongoing interdisciplinary collaboration.

Keywords

digitalisation, chemistry, ontology, knowledge graph

1. Introduction

The generation of FAIR data relies on the ability to easily apply standards and to produce well-structured, well-annotated datasets. Electronic Lab Notebooks (ELNs) are essential tools in promoting the digitalization of scientific research, as they help incorporate standards and structure into the workflows of experimental scientists [1].

In the field of chemistry, the development of an ELN that supports the creation of findable, accessible, interoperable, and reusable (FAIR) data is particularly challenging. This complexity arises from the nature of chemical research, which involves intricate and highly diverse experimental workflows, the use of various measurement devices that generate large volumes of data, and a wide range of data formats. Additionally, the handling of chemical structures which must be drawn, interpreted, and processed is indispensable for documenting and understanding chemical experiments. These requirements among others make chemistry data management especially demanding compared to other disciplines.

In recent years, a team of software developers and scientists at the Karlsruhe Institute of Technology (KIT) has been working on a research data management (RDM) environment specifically designed for chemistry. This environment, called Chemotion, comprises an ELN [2, 3] tailored to the specific needs of chemists, along with a research data repository [4] that interoperates seamlessly with the ELN. The

5th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, Nov 2024, Nara, Japan

[*] Corresponding author.

Ebrahim.Norouzi@fiz-Karlsruhe.de (E. Norouzi); nicole.jung@kit.edu (N. Jung); Anna.Jacyszyn@fiz-Karlsruhe.de (A. M. Jacyszyn); joerg.waitelonis@fiz-karlsruhe.de (J. Waitelonis); Harald.Sack@fiz-Karlsruhe.de (H. Sack)
0000-0002-2691-6995 (E. Norouzi); 0000-0001-9513-2468 (N. Jung); 0000-0002-5649-536X (A. M. Jacyszyn); 0000-0001-7192-7143 (J. Waitelonis); 0000-0001-7069-9804 (H. Sack)

repository allows researchers to publish their recorded and analyzed data in accordance with FAIR principles.

The combination of the Chemotion ELN and the interoperable repository Chemotion provides scientists with a comprehensive digital infrastructure that addresses their domain-specific requirements. It enables them to manage, organize, analyze, and publish their data in a way that enhances transparency and ensures the reusability of their research outcomes [5, 6, 7].

Driven by the needs of scientists for improved documentation and data management, particularly in response to the unique challenges posed by chemistry data, the Chemotion systems were initially developed without a systematic approach for integrating ontologies and semantic meaning. This limitation is now being addressed incrementally, through the implementation of a semantically coherent framework capable of representing research lifecycle stages, agent roles, provenance information, and domain-specific chemical entities in a machine-interpretable manner.

To support and enhance the semantic enrichment of Chemotion, an ontology-driven modeling approach grounds the Chemotion Knowledge Graph (Chemotion-KG) in upper-level semantics provided by the Basic Formal Ontology (BFO) using the Chemotion repository as a source as it includes well curated data assigned to publication metadata. Ontology Design Patterns (ODPs) are adopted as a systematic method to encode these structures, serving as reusable solutions to recurring ontology engineering problems [8]. Their use ensures not only the semantic consistency of datasets, creators, studies, and chemical substances in Chemotion-KG but also the reuse of established modeling practices that facilitate alignment with broader scientific knowledge graphs and ontology standards.

The Chemotion-KG, a synergy of Chemotion and semantics, is being created within the Leibniz Science Campus “Digital Transformation of Research” (DiTraRe)[1] [9, 10]. DiTraRe studies effects of digitalisation of research in a multilevel, interdisciplinary way. In this multidimensional exchange, one of the use cases is Chemotion which represents novel methods of data acquisition. The DiTraRe use cases are being analysed from different aspects by teams called *dimensions*. “Exploration and Knowledge Organisation” (AI4DiTraRe) is the dimension responsible for applying AI and studying its effects [11]. Collaboration of the aforementioned dimension and the Chemotion use case has by now provided outcomes described in this paper.

In brief, this work presents the construction of the BFO-compliant Chemotion-KG for semantically integrating experimental chemistry data. Section 2 describes the workflow for data harvesting, semantic transformation, and ontology alignment. Section 3 presents the resulting knowledge graph structures and instantiated entities generated using ontology design patterns. Section 4 discusses the broader impact of the Chemotion-KG and future directions. Finally, section 5 provides a summary of the achieved contributions.

The source code and datasets are publicly available[2]. The Chemotion-KG is hosted at https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/.

2. Knowledge Graph Construction Approach

This section describes the end-to-end pipeline for constructing the BFO-compliant Chemotion-KG, covering metadata harvesting, semantic enrichment, ontology alignment, reasoning, and materialization. The workflow is depicted in Figure 1, illustrating the transformation from schema-based metadata to a semantically aligned knowledge graph.

The initial metadata were harvested from the Chemotion repository[3], which provides structured schema.org-based descriptions of chemical research data. The repository’s data covered (in)organic reactions, analytical measurements such as nuclear magnetic resonance (NMR) spectroscopy, mass spectrometry (MS), IR-, and Raman spectroscopy data. The metadata were expressed primarily with `schema:Dataset`, `schema:ChemicalSubstance`, and `schema:Study`, which provide only lightweight

[1] DiTraRe web page, https://www.ditrare.de/en

[2] https://github.com/ISE-FIZKarlsruhe/chemotion-kg

[3] https://www.chemotion-repository.net

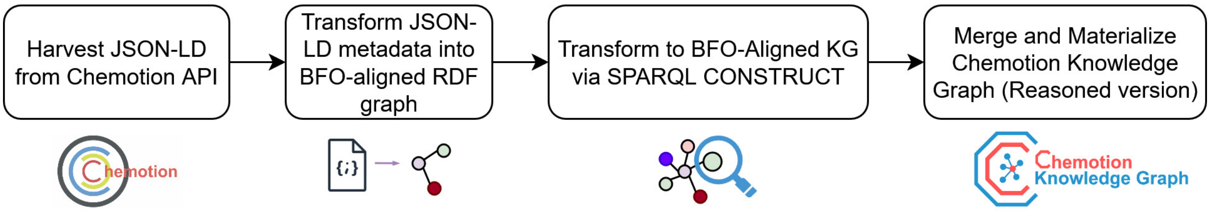
Harvest JSON-LD from Chemotion API
Transform JSON-LD metadata into BFO-aligned RDF graph
Transform to BFO-Aligned KG via SPARQL CONSTRUCT
Merge and Materialize Chemotion Knowledge Graph (Reasoned version)
Chemotion
{;}
Chemotion Knowledge Graph

Figure 1: Schematic workflow of the Chemotion Knowledge Graph construction.

descriptive semantics. Consequently, while adequate for basic metadata exchange, the schema.org representation did not capture the rich semantic structure required for ontology-driven reasoning, provenance modeling, and alignment with upper-level ontologies such as the BFO[4][12].

To address this, the harvested JSON-LD was converted into RDF graphs with canonical URIs and semantically enriched through a SPARQL CONSTRUCT transformation. The mapping aligned the schema-based metadata to the BFO, leveraging the NFDICore ontology[5][13] for research data structures, lifecycle modeling, and provenance, and reusing ChEBI[6][14] for chemical entities. The transformation queries were authored and documented using `shmarql`[7], a Linked Data publishing platform supporting SPARQL-based data pipelines.

The overall SPARQL CONSTRUCT query implementing the semantic transformation is openly available in our GitHub repository[8]. An example SPARQL UPDATE used to semantically enrich a dataset is shown below. It replaces a plain `schema:Dataset` description with NFDICore-aligned metadata, preserving original values while adding explicit BFO-compliant types. One of the essential steps in this transformation is the use of `BIND` to create new IRIs for elements such as descriptions, identifiers, titles, and URLs. Instead of representing these values only as plain literals, they are modeled as separate instances (e.g., description node, identifier node) with their own URIs. This approach follows BFO principles, where information content entities are treated as individual resources that can carry provenance and lifecycle information.

Creating explicit IRIs for these nodes allows the knowledge graph to keep stable references to metadata entities even when their literal values change. For instance, if a dataset's description or identifier is updated, the corresponding instance remains the same, making it possible to track changes over time and preserve provenance.

Listing 1: SPARQL CONSTRUCT for dataset transformation.

```
PREFIX schema: <http://schema.org/>
PREFIX nfdicore: <https://nfdi.fiz-karlsruhe.de/ontology/>
PREFIX obo: <http://purl.obolibrary.org/obo/>

CONSTRUCT {
  # Recast the dataset to BFO-aligned NFDICore class
  ?dataset a nfdicore:NFDI_0000009 ; # Dataset
          nfdicore:NFDI_0001027 ?creator ; # has creator (Person)
          nfdicore:NFDI_0000191 ?publisher ; # has publisher (Organization)
          obo:IAO_0000235 ?descriptionNode ; # description node
          nfdicore:NFDI_0001006 ?identifierNode ; # identifier node
          nfdicore:NFDI_0000142 ?license ; # license information
          nfdicore:NFDI_0000216 ?technique ; # associated measurement technique
          obo:IAO_0000235 ?nameNode ; # title node
```

[4] https://basic-formal-ontology.org/

[5] https://ise-fizkarlsruhe.github.io/nfdicore/

[6] https://www.ebi.ac.uk/chebi/

[7] https://github.com/epoz/shmarql

[8] https://github.com/ISE-FIZKarlsruhe/chemotion-kg/blob/main/processing/all-nfdicore.py

```
        obo:IAO_0000235 ?urlNode ; # landing page URL node
        nfdicore:NFDI_0001023 ?study ; # is output of a study
        obo:BFO_0000178 ?catalog . # is part of a registered catalog
}
WHERE {
  # Original Chemotion metadata described using schema.org
  ?dataset a schema:Dataset ;
        schema:creator ?creator ;
        schema:publisher ?publisher ;
        schema:description ?description ;
        schema:identifier ?identifier ;
        schema:license ?license ;
        schema:measurementTechnique ?technique ;
        schema:name ?name ;
        schema:url ?url ;
        schema:includedInDataCatalog ?catalog ;
        schema:isPartOf ?study .

  # Generate canonical URIs for literal-based nodes to ensure global identification
  BIND(IRI(CONCAT("https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/nodes/",
            ENCODE_FOR_URI(?description))) AS ?descriptionNode)
  BIND(IRI(CONCAT("https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/nodes/",
            ENCODE_FOR_URI(?identifier))) AS ?identifierNode)
  BIND(IRI(CONCAT("https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/nodes/",
            ENCODE_FOR_URI(?name))) AS ?nameNode)
  BIND(IRI(CONCAT("https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/nodes/",
            ENCODE_FOR_URI(?url))) AS ?urlNode)
}
```

To ensure that the resulting ontology design patterns and their interconnections are transparent and reusable, a standardized graphical representation was adopted. The Chemotion-KG modules and their links were visualized using Graffoo, a formal graphical notation for OWL ontologies [15].

3. Chemotion-KG Structure and Ontological Representation

The Chemotion-KG was constructed by transforming schema.org-based metadata into a BFO-aligned representation to ensure semantic interoperability with other scientific knowledge graphs. The choice of the Basic Formal Ontology (BFO) reflects its wide adoption as an upper ontology in the life sciences and research data management communities. Using BFO as a common top-level framework provides a consistent foundation for aligning domain-specific ontologies and supports integration with external resources that follow the same design principles. For modeling research data and its lifecycle, we adopted the NFDICore ontology, which is itself aligned with BFO. NFDICore was chosen because it already provides a well-defined vocabulary for metadata relevant to our setting—such as datasets, studies, creators, and affiliations—while also being developed and maintained within the German National Research Data Infrastructure (NFDI) context, ensuring institutional support and long-term sustainability. For representing chemical entities, we integrated the Chemical Entities of Biological Interest (ChEBI) ontology. ChEBI is a well-established community standard in chemistry, already aligned with BFO, and provides the necessary controlled vocabulary to represent molecules, substances, and their roles in experiments. Its wide acceptance in cheminformatics also facilitates linking the Chemotion-KG with other existing knowledge bases in chemistry.

Figure 2 illustrates the dataset representation, where schema.org `Dataset` instances were mapped to `nfdicore:NFDI_0000009` and enriched with identifiers, licensing information, catalog registration, and links to associated studies. The class nfdicore:NFDI_0000009 represents a dataset, modeled as an information content entity that denotes a structured collection of data, typically organized for a defined purpose such as analysis, research, or reference. In this context, a dataset constitutes structured information about a resource curated or provided by an individual researcher. The alignment introduces explicit semantics for provenance and lifecycle stages through NFDICore and BFO relations. Provenance is explicitly captured via relations such as nfdicore:NFDI_0001027 (*has creator*) and nfdicore:NFDI_0000191 (*published by*), while lifecycle stages are represented using BFO continuant–occurrent relations like obo:BFO_0000178 (*has continuant part*).

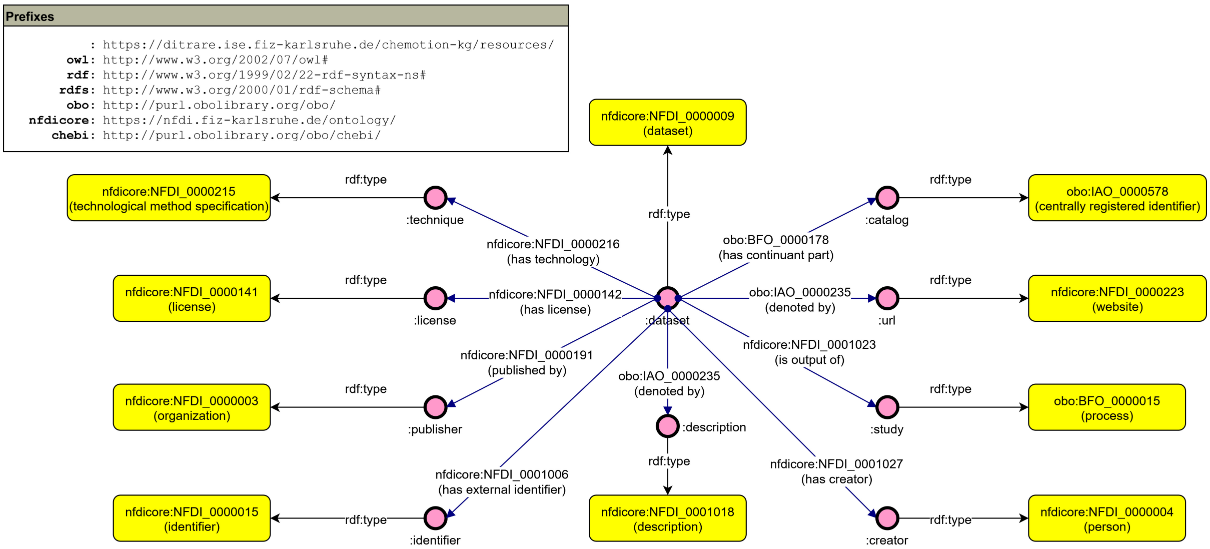
Prefixes
: https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/resources/
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
obo: http://purl.obolibrary.org/obo/
nfdicore: https://nfdi.fiz-karlsruhe.de/ontology/
chebi: http://purl.obolibrary.org/obo/chebi/
nfdicore:NFDI_0000009 (dataset)
rdf:type
:dataset
nfdicore:NFDI_0000215 (technological method specification)
rdf:type
:technique
nfdicore:NFDI_0000216 (has technology)
nfdicore:NFDI_0000141 (license)
rdf:type
:license
nfdicore:NFDI_0000142 (has license)
nfdicore:NFDI_0000003 (organization)
rdf:type
:publisher
nfdicore:NFDI_0000191 (published by)
nfdicore:NFDI_0000015 (identifier)
rdf:type
:identifier
nfdicore:NFDI_0001006 (has external identifier)
obo:IAO_0000235 (denoted by)
:description
rdf:type
nfdicore:NFDI_0001018 (description)
obo:BFO_0000178 (has continuant part)
:catalog
rdf:type
obo:IAO_0000578 (centrally registered identifier)
obo:IAO_0000235 (denoted by)
:url
rdf:type
nfdicore:NFDI_0000223 (website)
nfdicore:NFDI_0001023 (is output of)
:study
rdf:type
obo:BFO_0000015 (process)
nfdicore:NFDI_0001027 (has creator)
:creator
rdf:type
nfdicore:NFDI_0000004 (person)

Figure 2: Dataset representation in the Chemotion-KG.

The creator model (Figure 4) demonstrates the transformation of schema:Person metadata into a semantically rich structure. Individual names, ORCID identifiers, and organizational affiliations were represented as distinct resources, using nfdicore:NFDI_0000004 for persons and nfdicore:NFDI_0000003 for organizations. Roles and affiliations were explicitly modeled using the Process–Agent–Role ODP[9], aligning with the NFDICore and BFO modeling principles. This pattern establishes the relationship between *bfo:Process* (Occurrent), *nfdicore:Agent* (Independent Continuant), and *bfo:Role* (Specifically Dependent Continuant). Within the Chemotion-KG, *nfdicore:Agent* instances represent both organizations and persons involved in experimental chemistry workflows, including research institutions, and individual researchers. The pattern uses bfo:has_participant, bfo:realizes, and bfo:bearer_of to connect processes, agents, and their assigned roles, enabling explicit provenance tracking and role-based attribution for research data.

Studies were represented as obo:BFO_0000015 processes, connecting datasets, chemical substances, and associated publishing events (Figure 5). The publishing activities were explicitly modeled using nfdicore:NFDI_0000014, which is defined as a *process of making information available to the public either for sale or free access*[10]. Within the Chemotion-KG, this class was used to capture the temporal regions of publication via obo:BFO_0000199 (*occupies temporal region*) and to link standard profiles with each study through nfdicore:NFDI_0000207 (*has standard*).

Chemical entities (Figure 6) were aligned to ChEBI classes to ensure standardized representation of chemical knowledge. Generic chemical substances were modeled as obo:CHEBI_59999, capturing the abstract notion of a chemical entity within a study, while specific molecular structures were represented as obo:CHEBI_23367 molecular entities. To guarantee unambiguous identification and interoperability, the alignment included explicit mapping of InChI identifiers, InChIKeys, SMILES strings, and molecular formulas to structured nodes. Molecular weights were represented using the

[9] https://ise-fizkarlsruhe.github.io/mwo/docs/patterns/#pattern-1-process-agent-role
[10] https://nfdi.fiz-karlsruhe.de/ontology/NFDI_0000014

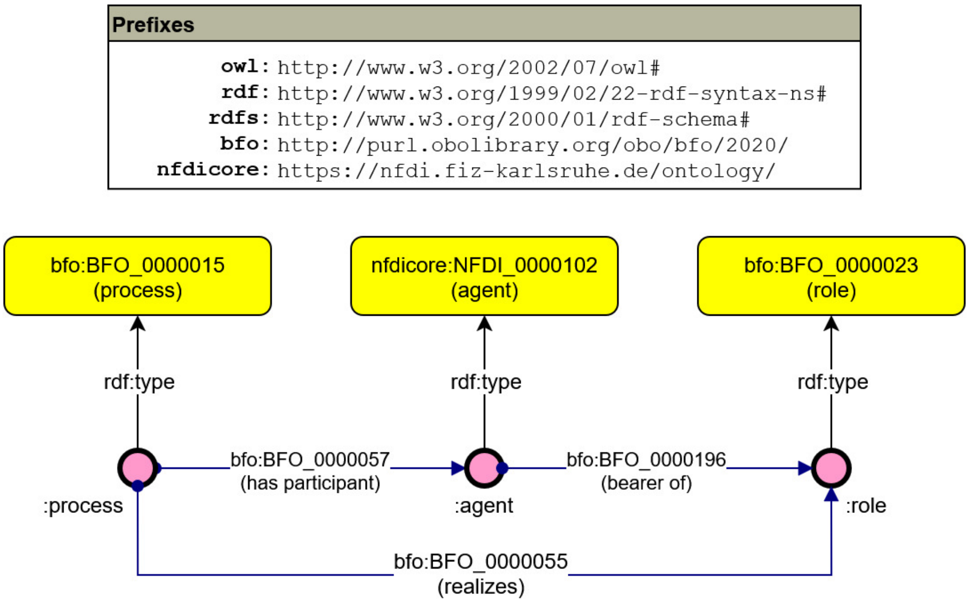
Prefixes
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
bfo: http://purl.obolibrary.org/obo/bfo/2020/
nfdicore: https://nfdi.fiz-karlsruhe.de/ontology/
bfo:BFO_0000015 (process)
nfdicore:NFDI_0000102 (agent)
bfo:BFO_0000023 (role)
rdf:type
rdf:type
rdf:type
:process
bfo:BFO_0000057 (has participant)
:agent
bfo:BFO_0000196 (bearer of)
:role
bfo:BFO_0000055 (realizes)

Figure 3: Process–Agent–Role Ontology Design Pattern.

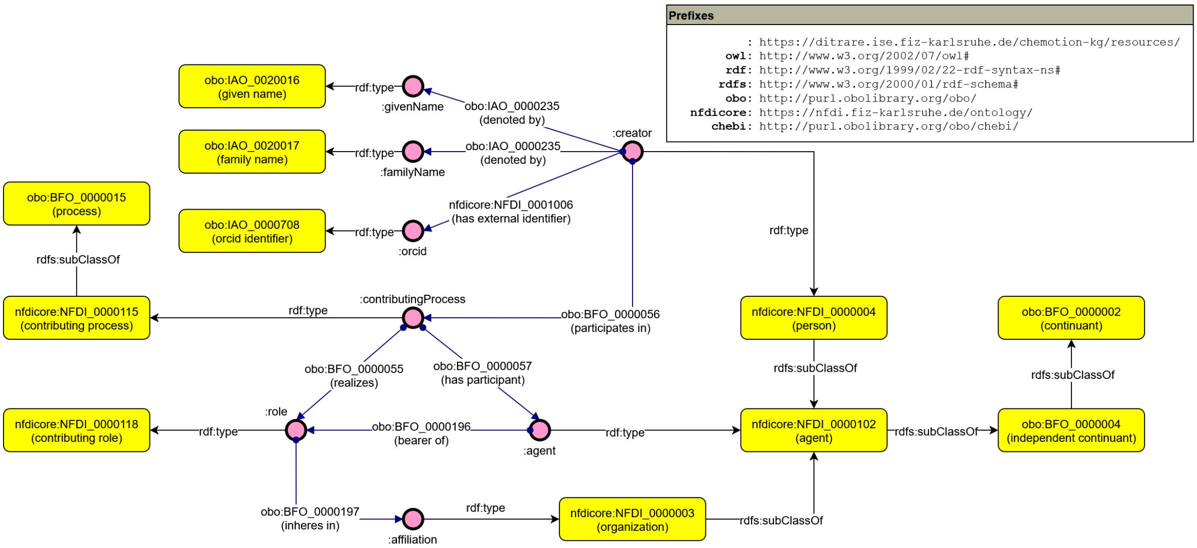
Prefixes
: https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/resources/
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
obo: http://purl.obolibrary.org/obo/
nfdicore: https://nfdi.fiz-karlsruhe.de/ontology/
chebi: http://purl.obolibrary.org/obo/chebi/
obo:IAO_0020016 (given name)
rdf:type
:givenName
obo:IAO_0000235 (denoted by)
obo:IAO_0020017 (family name)
rdf:type
:familyName
obo:IAO_0000235 (denoted by)
:creator
obo:BFO_0000015 (process)
nfdicore:NFDI_0001006 (has external identifier)
obo:IAO_0000708 (orcid identifier)
rdf:type
:orcid
rdfs:subClassOf
rdf:type
nfdicore:NFDI_0000115 (contributing process)
rdf:type
:contributingProcess
obo:BFO_0000056 (participates in)
nfdicore:NFDI_0000004 (person)
obo:BFO_0000002 (continuant)
obo:BFO_0000055 (realizes)
obo:BFO_0000057 (has participant)
rdfs:subClassOf
rdfs:subClassOf
:role
nfdicore:NFDI_0000118 (contributing role)
rdf:type
obo:BFO_0000196 (bearer of)
:agent
rdf:type
nfdicore:NFDI_0000102 (agent)
rdfs:subClassOf
obo:BFO_0000004 (independent continuant)
obo:BFO_0000197 (inheres in)
:affiliation
rdf:type
nfdicore:NFDI_0000003 (organization)
rdfs:subClassOf

Figure 4: Creator description aligned to NFDICore and BFO patterns, enabling explicit representation of roles and affiliations.

obo:BFO_0000019 *quality* pattern combined with obo:IAO_0000109 *measurement datum*, linking to obo:IAO_0000003 measurement units following the BFO measurement modeling principles. Structural images and external references were attached via nfdicore:NFDI_0000223 URL nodes to maintain provenance and facilitate linking to external chemical databases.

The resulting Chemotion-KG provides a semantically enriched integration of experimental chemistry data, employing formal ODPs for each core entity type, including datasets, creators, studies, and chemical substances. The transformation pipeline applies a SPARQL CONSTRUCT-based approach that preserves the original schema.org metadata while aligning it with BFO semantics and enriching it through NFDICore and ChEBI ontologies. This results in a BFO-compliant graph that supports semantic interoperability, logical reasoning, and linkage to external knowledge resources. The overall Chemotion-KG integrates all ODPs into a coherent and interconnected model, capturing datasets, creators, studies, and chemical substances within a BFO-compliant framework. As illustrated in Figure 7, each ODP is instantiated as a modular pattern and linked via well-defined object properties to form a semantically

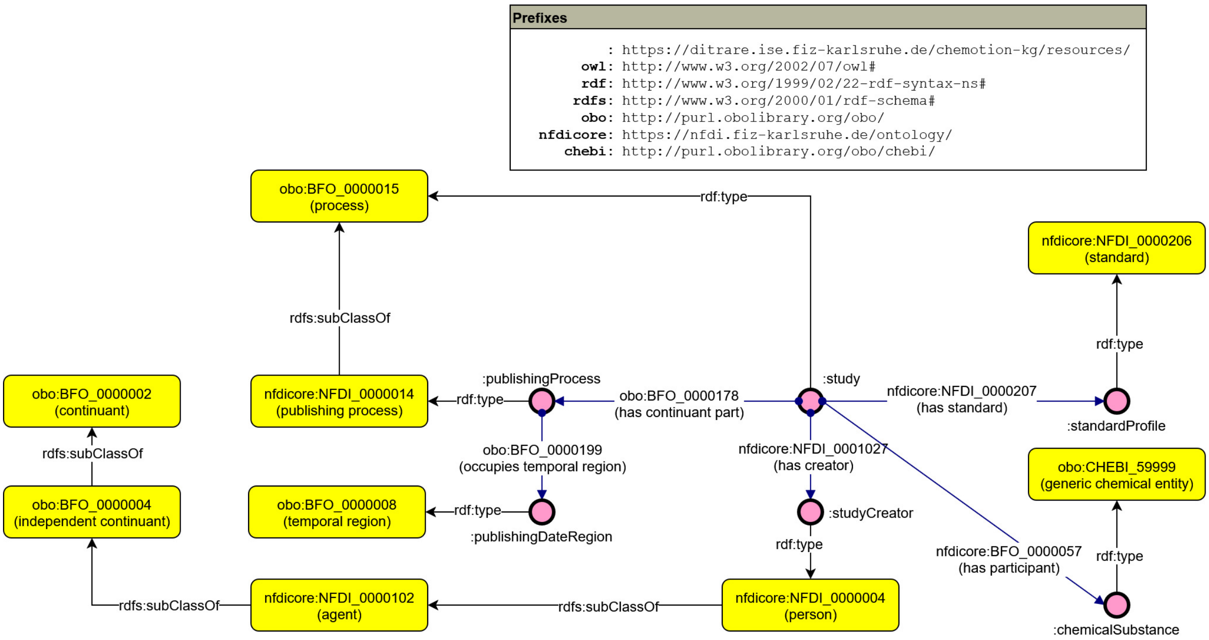
Prefixes
: https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/resources/
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
obo: http://purl.obolibrary.org/obo/
nfdicore: https://nfdi.fiz-karlsruhe.de/ontology/
chebi: http://purl.obolibrary.org/obo/chebi/
obo:BFO_0000015 (process)
rdf:type
nfdicore:NFDI_0000206 (standard)
rdfs:subClassOf
rdf:type
obo:BFO_0000002 (continuant)
nfdicore:NFDI_0000014 (publishing process)
:publishingProcess
rdf:type
obo:BFO_0000178 (has continuant part)
:study
nfdicore:NFDI_0000207 (has standard)
:standardProfile
rdfs:subClassOf
obo:BFO_0000199 (occupies temporal region)
nfdicore:NFDI_0001027 (has creator)
obo:CHEBI_59999 (generic chemical entity)
obo:BFO_0000004 (independent continuant)
obo:BFO_0000008 (temporal region)
rdf:type
:publishingDateRegion
:studyCreator
rdf:type
nfdicore:BFO_0000057 (has participant)
rdf:type
nfdicore:NFDI_0000102 (agent)
rdfs:subClassOf
rdfs:subClassOf
nfdicore:NFDI_0000004 (person)
:chemicalSubstance

Figure 5: Study representation with explicit modeling of publishing processes, temporal regions, and standard profiles.

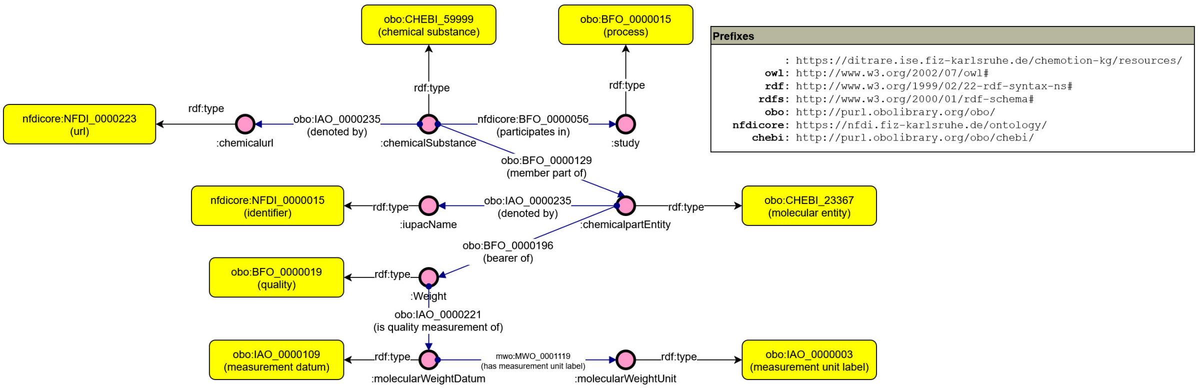
obo:CHEBI_59999 (chemical substance)
obo:BFO_0000015 (process)
Prefixes
: https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/resources/
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
obo: http://purl.obolibrary.org/obo/
nfdicore: https://nfdi.fiz-karlsruhe.de/ontology/
chebi: http://purl.obolibrary.org/obo/chebi/
rdf:type
rdf:type
nfdicore:NFDI_0000223 (url)
rdf:type
:chemicalurl
obo:IAO_0000235 (denoted by)
:chemicalSubstance
nfdicore:BFO_0000056 (participates in)
:study
obo:BFO_0000129 (member part of)
nfdicore:NFDI_0000015 (identifier)
rdf:type
:iupacName
obo:IAO_0000235 (denoted by)
:chemicalpartEntity
rdf:type
obo:CHEBI_23367 (molecular entity)
obo:BFO_0000196 (bearer of)
obo:BFO_0000019 (quality)
rdf:type
:Weight
obo:IAO_0000221 (is quality measurement of)
obo:IAO_0000109 (measurement datum)
rdf:type
:molecularWeightDatum
mwo:MWO_0001119 (has measurement unit label)
:molecularWeightUnit
rdf:type
obo:IAO_0000003 (measurement unit label)

Figure 6: Chemical substance representation aligned to ChEBI and NFDICore, with explicit molecular entity and measurement modeling.

consistent network. This approach ensures that dataset descriptions, experimental studies, agents with roles, and chemical entities are harmonized through reusable modeling practices. The figure highlights how these ODPs are composed together to create a unified, semantically enriched representation of the Chemotion experimental data.

The Chemotion-KG is continuously updated, with new data from the Chemotion repository ingested into the graph on a daily basis. Instances are managed using a named graph that organizes resources according to their submission date. Each resource IRI encodes the year and month of submission alongside the original Chemotion identifier, ensuring temporal provenance and stable referencing. For example, the dataset representing Raman spectroscopy data is available at [11], where the path segments 2014/05 capture the submission date and the suffix encodes the Chemotion-specific identifier. The corresponding source entry in the Chemotion repository[12], ensuring traceability between the knowledge graph and its originating records in the research data repository.

Access to the graph is provided through a public SPARQL endpoint at https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/sparql. This architecture combines automated metadata harvesting, ontology-based

[11] https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/resources/2014/05/10.14272/VRYFQVRFMNXTJS-UHFFFAOYSA-N/Raman

[12] https://www.chemotion-repository.net/inchikey/VRYFQVRFMNXTJS-UHFFFAOYSA-N/Raman

semantic enrichment, and fine-grained provenance modeling to create a high-quality, AI-ready knowledge graph for experimental chemistry. The constructed Chemotion-KG as of July 2025 comprises a total of 1,462,187 RDF triples, reflecting the semantic integration of experimental chemistry data and metadata. The graph contains 87,782 instantiated entities, including 20,701 datasets, 20,563 studies, and 3,746 molecular entities. The knowledge graph models 250 individual creators with explicit provenance, and integrates chemical information through 4,923 instances of obo:CHEBI_59999 for generic chemical substances.

The Chemotion-KG evolves continuously as new data are submitted to the Chemotion repository. Updates are handled through daily ingestion runs that add new instances and enrich existing ones, while preserving provenance by assigning each submission to a dedicated named graph organized by date. Entities are generally not removed; instead, updated resources are versioned through their temporal context in the named graph, ensuring reproducibility and historical traceability. IRIs are minted systematically by combining the Chemotion identifier with a submission timestamp, which mitigates the risk of collisions and guarantees long-term stability. To prevent excessive length, descriptive elements are restricted to standardized identifiers (e.g., InChIKey for molecules), while additional metadata are attached via well-defined properties rather than embedded directly in the IRI string. IRIs in the Chemotion-KG are minted by combining the original Chemotion identifier with temporal submission information (year and month). This strategy provides globally unique and traceable references that preserve provenance across updates and reduces the likelihood of collisions. At present, descriptive identifiers (e.g., InChIKeys for molecules) are embedded in the IRI structure to maximize human interpretability and support external linking. However, this approach can result in relatively long IRIs, particularly for chemical description nodes, which may pose challenges for readability and downstream applications. To address this, we plan to adopt a Universally Unique Identifier (UUID) in future releases.

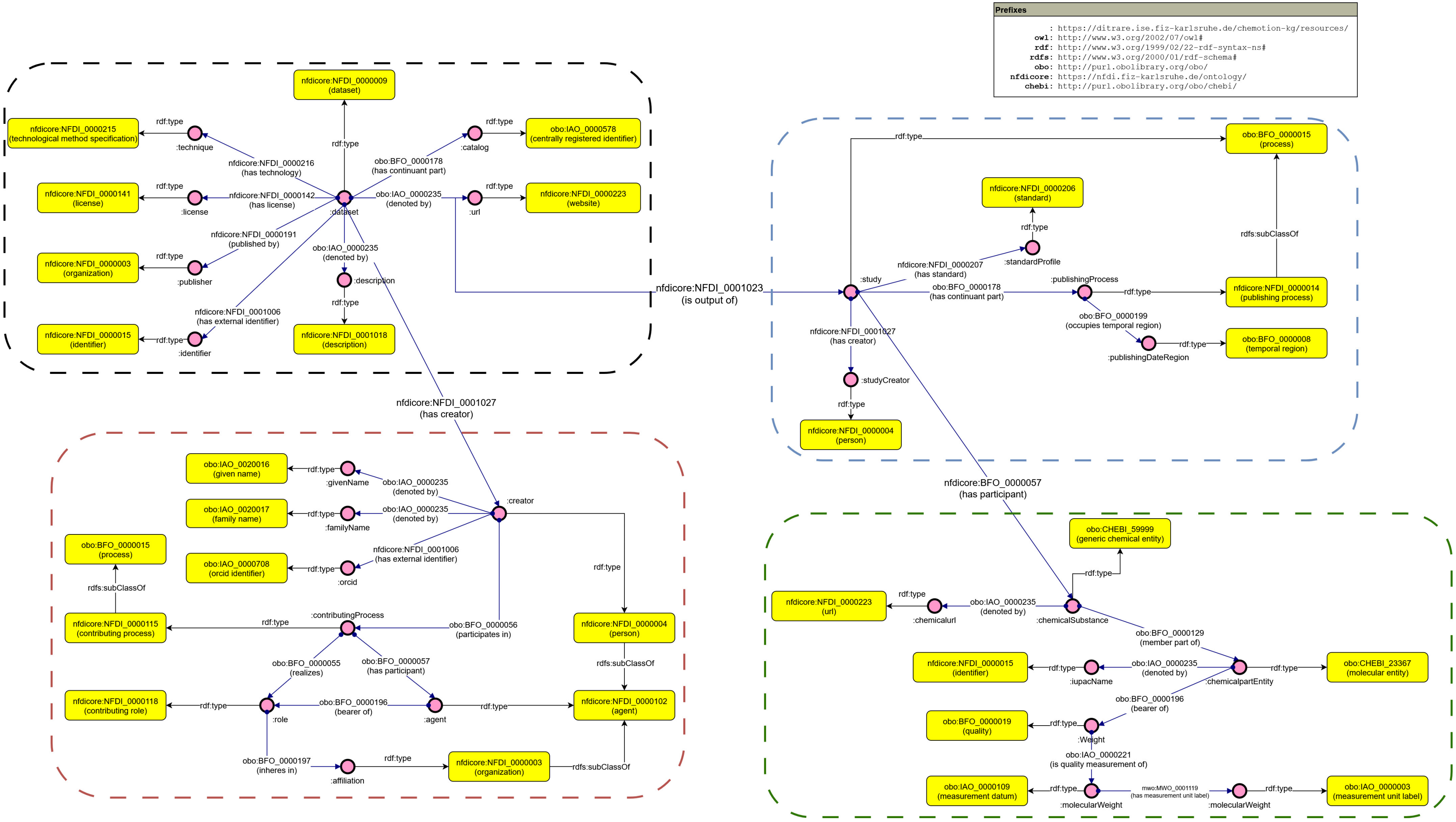
Prefixes
: https://ditrare.ise.fiz-karlsruhe.de/chemotion-kg/resources/
owl: http://www.w3.org/2002/07/owl#
rdf: http://www.w3.org/1999/02/22-rdf-syntax-ns#
rdfs: http://www.w3.org/2000/01/rdf-schema#
obo: http://purl.obolibrary.org/obo/
nfdicore: https://nfdi.fiz-karlsruhe.de/ontology/
chebi: http://purl.obolibrary.org/obo/chebi/
nfdicore:NFDI_0000009 (dataset)
nfdicore:NFDI_0000215 (technological method specification)
nfdicore:NFDI_0000141 (license)
nfdicore:NFDI_0000003 (organization)
nfdicore:NFDI_0000015 (identifier)
nfdicore:NFDI_0001018 (description)
obo:IAO_0000578 (centrally registered identifier)
nfdicore:NFDI_0000223 (website)
:technique
:license
:publisher
:identifier
:dataset
:description
:catalog
:url
rdf:type
nfdicore:NFDI_0000216 (has technology)
nfdicore:NFDI_0000142 (has license)
nfdicore:NFDI_0000191 (published by)
nfdicore:NFDI_0001006 (has external identifier)
obo:IAO_0000235 (denoted by)
obo:BFO_0000178 (has continuant part)
nfdicore:NFDI_0001023 (is output of)
nfdicore:NFDI_0001027 (has creator)
obo:BFO_0000015 (process)
nfdicore:NFDI_0000206 (standard)
nfdicore:NFDI_0000014 (publishing process)
obo:BFO_0000008 (temporal region)
nfdicore:NFDI_0000004 (person)
:study
:standardProfile
:publishingProcess
:publishingDateRegion
:studyCreator
rdfs:subClassOf
nfdicore:NFDI_0000207 (has standard)
obo:BFO_0000199 (occupies temporal region)
nfdicore:NFDI_0001027 (has creator)
nfdicore:BFO_0000057 (has participant)
obo:IAO_0020016 (given name)
obo:IAO_0020017 (family name)
obo:IAO_0000708 (orcid identifier)
obo:BFO_0000015 (process)
nfdicore:NFDI_0000115 (contributing process)
nfdicore:NFDI_0000118 (contributing role)
nfdicore:NFDI_0000004 (person)
nfdicore:NFDI_0000102 (agent)
nfdicore:NFDI_0000003 (organization)
:givenName
:familyName
:orcid
:creator
:contributingProcess
:role
:agent
:affiliation
obo:IAO_0000235 (denoted by)
nfdicore:NFDI_0001006 (has external identifier)
obo:BFO_0000056 (participates in)
obo:BFO_0000055 (realizes)
obo:BFO_0000057 (has participant)
obo:BFO_0000196 (bearer of)
obo:BFO_0000197 (inheres in)
obo:CHEBI_59999 (generic chemical entity)
nfdicore:NFDI_0000223 (url)
nfdicore:NFDI_0000015 (identifier)
obo:BFO_0000019 (quality)
obo:IAO_0000109 (measurement datum)
obo:CHEBI_23367 (molecular entity)
obo:IAO_0000003 (measurement unit label)
:chemicalurl
:chemicalSubstance
:iupacName
:chemicalpartEntity
:Weight
:molecularWeight
obo:BFO_0000129 (member part of)
obo:IAO_0000221 (is quality measurement of)
mwo:MWO_0001119 (has measurement unit label)

Figure 7: Integrated view of the Chemotion Knowledge Graph.

4. Impact and Future Work

The Chemotion-KG establishes a semantically enriched, BFO-compliant representation of experimental chemistry data derived from the Chemotion repository. By combining schema.org-based metadata with NFDICore and ChEBI ontologies through a SPARQL CONSTRUCT-driven pipeline, the approach ensures semantic interoperability, provenance preservation, and reasoning support. The daily ingestion workflow and ontology design patterns for datasets, creators, studies, and chemical substances provide a scalable foundation for AI-driven discovery and integration with external resources. The public SPARQL endpoint further facilitates community access, reuse, and advancement of FAIR chemical research data.

Future work will address two major directions: broadening the data scope and establishing systematic evaluation. On the data side, we will extend beyond metadata to incorporate additional information from the Chemotion repository model (e.g., raw instrument logs, measurement data, dataset details, and chemical substance information). We also plan to link the Chemotion-KG with external chemical knowledge graphs and databases, including PubChem RDF[13][16], ChemSpider[14][17], and the NFDI4Chem Knowledge Graph[15], thereby enhancing interoperability and enabling cross-resource reasoning. On the evaluation side, we will define competency questions that reflect typical use cases in chemistry, develop SHACL shapes to validate the RDF graphs generated in the Chemotion-KG pipeline, and conduct query performance benchmarks to assess scalability. In addition, we envision user studies with chemists to evaluate usability and scientific relevance. Finally, we will explore leveraging the Chemotion-KG for automated knowledge extraction and enrichment by integrating it with heterogeneous data sources and applying machine learning techniques for semantic alignment and knowledge discovery. In this context, we also plan to investigate the integration of large language models (LLMs) with structured scientific data to support advanced query answering.

The overarching goal of this work is to prepare the Chemotion system for integration with AI-based methods while embedding it into the broader ecosystem of semantically enriched research data management systems. In this context, knowledge graphs can be considered a component of symbolic AI, since they provide machine-readable semantics, enable logical entailments, and support query answering over structured knowledge. The Chemotion-KG contributes to this paradigm by supplying formally aligned, provenance-rich descriptions of experimental chemistry data that can serve as a foundation for both reasoning tasks and data-driven approaches. Example reasoning tasks include identifying equivalent substances across studies, detecting inconsistencies in dataset annotations, validating metadata through SHACL constraints, and inferring missing provenance relations. Beyond symbolic reasoning, we are exploring the use of AI methods such as natural language processing for automated curation of reaction descriptions and analysis modules within the Chemotion repository [18, 19]. Linking Chemotion-KG to external chemical resources (e.g., PubChem, ChemSpider) will also enable cross-resource enrichment and the training of machine learning models for property prediction and reaction condition optimization. These efforts aim to bridge symbolic and statistical AI, ultimately supporting AI-assisted chemistry and laying the groundwork for self-driving laboratories [20].

Merging chemistry with AI has an immense potential for research. Studying this process and its effects is one of the foci of the Leibniz Science Campus DiTraRe. On a larger scale, DiTraRe brings an interdisciplinary perspective by investigating the influence of digitalisation of research. It sheds light on the potentials, challenges, and risks of the digital transformation through an interactive exchange between numerous research areas. Activities of the dimension "Exploration and Knowledge Organisation" in the Chemotion use case are strikingly similar to efforts with which researchers deal in other disciplines. More than the others, material science could serve here as a prime example, as it is already advancing the application of ontologies and knowledge graphs for experimental research [21, 22]. The corresponding German National Research Data Infrastructure, NFDIMatWerk, has already created a rich publication database concerning many topics which can be related to Chemotion-KG

[13] https://pubchem.ncbi.nlm.nih.gov/docs/rdf-federated-query
[14] https://www.chemspider.com/
[15] https://knowledgebase.nfdi4chem.de/knowledge_base/

[23]. DiTraRe plans to serve here as a platform connecting researchers from varying disciplines with an aim to exchange ideas, identify common problems, and create generalised solutions. This way, by raising above the level of individual research disciplines, we will study influence and effects of applied AI as an additional dimension to the digital transformation of research.

5. Summary

In this paper, a comprehensive pipeline for constructing the BFO-compliant Chemotion Knowledge Graph has been presented, integrating experimental chemistry data from the Chemotion Repository into a semantically enriched, ontology-driven representation. By leveraging ontology design patterns and aligning schema-based metadata with NFDICore and ChEBI ontologies, the approach establishes a reusable and interoperable framework for chemical research data. The generated knowledge graph comprises over 1.4 million triples and more than 87 thousand instances.

Acknowledgments

The Leibniz Science Campus "Digital Transformation of Research" (DiTraRe) is funded by the Leibniz Association (W74/2022). The authors would like to thank Pei-Chi Huang and Chia-Lin Lin (KIT IBCS) for their work on JSON-LD in Chemotion.

Declaration on Generative AI

During the preparation of this work, the author(s) used GPT-4 in order to: Grammar and spelling check. After using these tool(s)/service(s), the author(s) reviewed and edited the content as needed and take(s) full responsibility for the publication's content.

References

[1] C. Steinbeck, O. Koepler, F. Bach, et al., Nfdi4chem-towards a national research data infrastructure for chemistry in germany, Research ideas and outcomes 6 (2020) e55852.

[2] P. Tremouilhac, A. Nguyen, Y.-C. Huang, et al., Chemotion eln: an open source electronic lab notebook for chemists in academia, Journal of Cheminformatics 9 (2017) 54.

[3] S. Kotov, P. Tremouilhac, N. Jung, et al., Chemotion-eln part 2: adaption of an embedded ketcher editor to advanced research applications., Journal of Cheminformatics 10 (2018). URL: https://doi.org/10.1186/s13321-018-0292-9. doi:10.1186/s13321-018-0292-9.

[4] P. Tremouilhac, C.-L. Lin, P.-C. a. a. Huang, The repository chemotion: Infrastructure for sustainable research in chemistry, Angewandte Chemie International Edition 59 (2020) 22771–22778. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/anie.202007702. doi:https://doi.org/10.1002/anie.202007702. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/anie.202007702.

[5] P. Tremouilhac, P.-C. Huang, C.-L. Lin, Y.-C. Huang, A. Nguyen, N. Jung, F. Bach, S. Bräse, Chemotion repository, a curated repository for reaction information and analytical data, Chemistry methods 1 (2021) 8–11. doi:10.1002/cmtd.202000034.

[6] P. Huang, C. Lin, P. Tremouilhac, et al., Using the chemotion repository to deposit and access fair research data for chemistry experiments, Nature Protocols 20 (2025). URL: https://doi.org/10.1038/s41596-024-01074-z. doi:10.1038/s41596-024-01074-z.

[7] Y. Huang, P. Tremouilhac, A. Nguyen, et al., Chemspectra: a web-based spectra editor for analytical data, Journal of Cheminformatics 18 (2021). URL: https://doi.org/10.1186/s13321-020-00481-0. doi:10.1186/s13321-020-00481-0.

[8] A. Gangemi, V. Presutti, Ontology design patterns, in: Handbook on ontologies, Springer, 2009, pp. 221–243.

[9] M. Razum, F. Bach, S. Brünger-Weilandt, et al., Proposal for a Leibniz ScienceCampus – Digital Transformation of Research (DiTraRe), 2023. doi:10.5281/zenodo.11109406, project proposal.

[10] A. M. Jacyszyn, F. Bach, H. Sack, et al., Interim report of the leibniz science campus ”digital transformation of research” (ditrare), 2025. URL: https://doi.org/10.5281/zenodo.14941635. doi:10.5281/zenodo.14941635.

[11] A. M. Jacyszyn, H. Sack, D.-S. Group, et al., Ditrare: Ai on a spider's web. interweaving disciplines for digitalisation, in: 4th International Workshop on Scientific Knowledge: Representation, Discovery, and Assessment, volume Vol-3780, 2024. URL: https://ceur-ws.org/Vol-3780/paper5.pdf. doi:10.5281/zenodo.13862017.

[12] R. Arp, B. Smith, A. D. Spear, Building ontologies with Basic Formal Ontology, MIT Press, 2015.

[13] O. Bruns, T. Tietz, J. Waitelonis, et al., Nfdicore 2.0: A bfo-compliant ontology for multi-domain research infrastructures, arXiv preprint arXiv:2410.01821 (2024).

[14] P. Del Nostro, G. Goldbeck, D. Toti, Chameo: An ontology for the harmonisation of materials characterisation methodologies, Applied Ontology (2022) 1–21.

[15] R. Falco, A. Gangemi, S. Peroni, et al., Modelling owl ontologies with graffoo, in: European Semantic Web Conference, Springer, 2014, pp. 320–325.

[16] S. Kim, J. Chen, T. Cheng, et al., Pubchem 2025 update, Nucleic Acids Research 53 (2024) D1516–D1525. URL: https://doi.org/10.1093/nar/gkae1059. doi:10.1093/nar/gkae1059. arXiv:https://academic.oup.com/nar/article-pdf/53/D1/D1516/60743708/gkae1059.pdf.

[17] H. E. Pence, A. Williams, Chemspider: An online chemical information resource, Journal of Chemical Education 87 (2010) 1123–1124. URL: https://doi.org/10.1021/ed100697w. doi:10.1021/ed100697w. arXiv:https://doi.org/10.1021/ed100697w.

[18] D. Punjabi, Y. Huang, L. Holzhauer, et al., Infrared spectrum analysis of organic molecules with neural networks using standard reference data sets in combination with real-world data, Journal of Cheminformatics 17 (2025). URL: https://doi.org/10.1186/s13321-025-00960-2. doi:10.1186/s13321-025-00960-2.

[19] Y.-C. Huang, P. Tremouilhac, S. Kuhn, et al., (semi-) automatic review process for common compound characterization data in organic synthesis, ChemRxiv (2024). doi:10.26434/chemrxiv-2024-1r9tb, preprint.

[20] P. M. Maffettone, P. Friederich, S. G. Baird, et al., What is missing in autonomous discovery: open challenges for the community, Digital Discovery 2 (2023) 1644–1659. URL: http://dx.doi.org/10.1039/D3DD00143A. doi:10.1039/D3DD00143A.

[21] H. Beygi Nasrabadi, E. Norouzi, H. Sack, B. Skrotzki, Performance evaluation of upper-level ontologies in developing materials science ontologies and knowledge graphs, Advanced Engineering Materials n/a (????) 2401534. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/adem.202401534. doi:https://doi.org/10.1002/adem.202401534. arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/adem.202401534.

[22] E. Norouzi, J. Waitelonis, H. Sack, The landscape of ontologies in materials science and engineering: a survey and evaluation, in: Proceedings of the First International Workshop on Semantic Materials Science (SeMatS 2024): Harnessing the Power of Semantic Web Technologies in Materials Science, volume 3760 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2024, pp. 78–100. URL: https://ceur-ws.org/Vol-3760/paper3.pdf.

[23] Proceedings NFDI-MatWerk Conference 2023, Zenodo, 2024. URL: https://doi.org/10.5281/zenodo.11353339. doi:10.5281/zenodo.11353339.