Providing adverse outcome pathways from the AOP-Wiki in semantic web format to increase usability and accessibility of the content.

Marvin Martens[1], Chris T. Evelo[1,2], and Egon L. Willighagen[1]

[1]Department of Bioinformatics - BiGCaT, NUTRIM, Maastricht University, The Netherlands
[2]Maastricht Centre for Systems Biology (MaCSBio), Maastricht University, The Netherlands

January 5, 2021

Abstract

The AOP-Wiki is the main environment for the development and storage of Adverse Outcome Pathways. These Adverse Outcome Pathways describe mechanistic information about toxicodynamic processes and can be used to develop effective risk assessment strategies. However, it is challenging to automatically and systematically parse, filter, and use its contents. We explored solutions to better structure the AOP-Wiki content and to link it with chemical and biological resources. Together this allows more detailed exploration which can be automated.

We converted the complete AOP-Wiki content into Resource Description Framework. We used over twenty ontologies for the semantic annotation of property-object relations, including the ChemInformatics Ontology, Dublin Core, and the Adverse Outcome Pathway Ontology. The latter was used over 8,000 times. Furthermore, over 3,500 link-outs were added to twelve chemical databases and over 6,500 link-outs to four gene and protein databases.

SPARQL queries can be used against the Resource Description Framework to answer biological and toxicological questions, such as listing measurement methods for all Key Events leading to an Adverse Outcome of interest. The full power that the use of this new resource provides becomes apparent when combining the content with external databases using federated queries. For example, we can link genes related to Key Events with molecular pathway on WikiPathways in which they occur and find all Adverse Outcome Pathways caused by stressors that are part of a particular chemical group. Overall, the AOP-Wiki Resource Description Framework allows new ways to explore the rapidly growing Adverse Outcome Pathway knowledge and makes the integration of this database in automated workflows possible.

Keywords: Adverse Outcome Pathways, Resource Description Framework, Risk Assessment, Linked Data, SPARQL

1 Introduction

(max 750 words)

Since its establishment in 2010, the Adverse Outcome Pathway (AOP) concept has become a prominent tool for the risk assessment community [1, 2]. AOPs are a chain of biological processes, called Key Events (KEs), starting from a molecular perturbation with a stressor towards an Adverse Outcome (AO), connected by Key Event Relationships (KERs). AOPs exist to capture all mechanistic toxicological knowledge from literature and data, to direct future studies to fill gaps of existing knowledge, and to drive Integrated Approaches to Testing and Assessment (IATA) development [1, 3]. This was demonstrated with the AOP-based IATA for skin sensitization, resulting in various IATA with combinations of *in vitro* and *in silico* assays outperforming animal tests [4].

The majority of the AOPs are developed and stored in the AOP-Wiki (https://aopwiki.org/), which is part of the AOP Knowledge Base, released in 2014 as a result of the AOP development program initiated by the Organisation for Economic and Collaborative Development (OECD) [5]. This wiki is designed to facilitate collaborative development of qualitative AOP descriptions, and thereby promote their incorporation into risk assessments and stimulate effective reuse of mechanistic toxicological knowledge [6, 7].

The resulting AOPs describe much of the biological space surrounding toxicological processes, most of the information on genes, chemicals, biological pathways, and phenotypes, among other things, are already captured in specialised databases or ontologies outside of AOP-Wiki [8]. However, the AOP-Wiki has limited possibilities for linking of external information and data, mostly consisting of free-text descriptions and links to the US CompTox Chemistry Dashboard [9] and to NCBI for taxonomic applicability [10]. An initiative to make the reporting more consistent was the introduction of Key Event Components [11] for the annotation of Biological Processes (BPs), Biological Objects (BOs) and Biological Actions (BAs) for KEs, and annotations of cell types and organs in which KEs can occur.

Since the AOP-Wiki is the central repository for AOPs and therefore a key player in the shift towards animal-free testing strategies, it is essential that its contents can be queried and utilized effectively to answer biological questions and to reuse existing knowledge. However, accessing the data computationally or linking with other resources is hardly possible when only downloadable eXtensible Markup Language (XML) data dumps are provided that consist mostly of free text. Because of these aspects, parsing and querying the continuously growing amount of information in the AOP-Wiki is a complex, time-consuming task. This is a problem because it prevents the integration of AOP knowledge with other data and resources.

This could be resolved by applying Linked Open Data (LOD) solutions, such as structuring the data in a Resource Description Framework (RDF) model [12], introducing persistent identifiers and semantic annotations, and implementing Application Programming Interfaces (APIs) for accessing the data. RDF represents knowledge as semantic triples, in which a subject, predicate and object together define a statement and assist in the meaningful representation of knowledge in a machine-readable manner.

These concepts are generally in line with the FAIR principles[13] for data and knowledge management, developed to enhance the Findability, Accessibility, Interoperability, and Reusability of data and allow computational support of data usage. For example, such as the solutions applied by the Swiss Institute of Bioinformatics with the development of neXtProt Linked Data by implementing RDF annotations for easier exploration and retrieval of data through web services [14, 15].

Also, the use of ontologies and vocabularies for semantic annotations allows for the integration of data between resources, such as the direct linking of chemical or protein databases with WikiPathways [16, 17].

In this paper we want to show how using RDF makes the AOP-Wiki content more usable for automated exploration in combination with other existing semantic web based information sources. We describe our implementation of LOD solutions for the AOP-Wiki to introduce new, effective ways of accessing and using the data. These solutions will enhance the usefulness of the

AOP-Wiki to risk assessors, developers, and modelers, and facilitate answering complex research questions, also across databases or as part of automated workflows. We hypothesise that with the implementation of RDF, with the use of standard ontologies for semantic modelling of information captured in AOPs, the data can be better exploited [18]. Furthermore, the domain-specific AOPOntology (AOPO), in combination with other relevant ontologies, can be used to link various pieces of mechanistic toxicological information and thereby facilitate knowledge-based hazard identification using AOPs [19]. The use of persistent, unique and resolvable identifiers allows the interoperability with other related data sources. When combined with computational tools that can access experimental data these approaches can make AOP information a core element for predictive modelling [20].

2 Methods

2.1 AOP-Wiki XML

The AOP-Wiki XML quarterly download file of October 1st, 2020 (downloaded from `https://aopwiki.org/downloads`) was retrieved and used for this paper.

2.2 Code

The code for the XML-to-RDF conversion was written as a Jupyter notebook using Python version 3.7.3 in JupyterLab version 0.35.5, and is stored in GitHub (`https://github.com/marvinm2/AOPWikiRDF`) [21]. It downloads and parses the AOP-Wiki XML file with the ElementTree XML API Python library, and stores all its content in a Python nested dictionary data model, one for each of the main components which form the basis of the existing AOP-Wiki. The AOPs themselves, KEs, KERs, stressors, chemicals, taxonomy, cell-terms, organ-terms, and the KE Components, which comprise of BPs, BOs, and BAs.

2.3 Persistent identifiers

Prior to the development of the AOP-Wiki RDF, we registered the identifiers for the AOP, KE, KER, and Stressor in the Minimum Information Required In the Annotation of Models (MIRIAM) Registry [22] to allow Identifiers.org to resolve Internationalized Resource Identifiers (IRIs). In order to make all identifiers in the AOP-Wiki resolvable and linking to their corresponding database webpages, these IRIs, along with a variety of chemical and gene database identifier types, were implemented in the AOP-Wiki RDF.

2.4 Semantic annotation

Terms from common biomedical terminologies and standard metadata vocabularies were used as predicates. These terms were retrieved from BioPortal [23] or in the corresponding Web Ontology Language (OWL) [24] files stored in GitHub (Table S1). Furthermore, the IRIs for existing annotations for KE Components, cell-terms and organ-terms were added to improve their semantic meaning (Table S2).

2.5 Gene / protein identifier mapping

In order to increase the number of annotations and add more types of gene and protein identifiers for improved linking of data and repositories, the XML-to-RDF conversion includes two methods of mapping to gene and protein identifiers.

The first of which is based on BO annotations with PRotein ontology (PR) terms [25], which were mapped to identifiers from NCBI Gene, HUGO Gene Nomenclature Committee (HGNC) and UniProt with the PR mapping file, promapping.txt, downloaded from `https://proconsortium.org/download/current/` on May 10th, 2020.

The second method involved textual gene identifier mapping for KEs and KERs, for which we extracted approved symbols, names, and synonyms for all human genes from the HGNC (downloaded from `https://www.genenames.org/` [26] in July 2019). For KEs, textual gene identifier mapping was done on KE descriptions, and the MIE- and AO-specific sections. For the KERs, the mapping was performed for their descriptions, and the sections of biological plausibility and empirical support.

2.6 Chemical identifier mapping

On top of the chemicals already present in the AOP-Wiki and the genes and proteins IDs added to the RDF with the Jupyter notebook for RDF-to-XML conversion, we extended the coverage of external molecular databases using BridgeDb, an identifier mapping service for chemicals, genes, proteins, and interactions [27]. Therefore, the BridgeDb Docker image (version 2.3.3) was deployed, which contains the Metabolite BridgeDb ID Mapping Database (version 20190207) and gene identifier mapping file for human genes (version Ensembl 91). The "requests" Python library (version 2.22.0) was used for calling BridgeDb's "xref" function to perform identifier mapping for chemicals and HGNC IDs that resulted from the textual mapping. For chemicals, the CAS IDs from the AOP-Wiki XML were used as input to retrieve identifiers from ChEBI [28], ChemSpider [29], Wikidata [30, 31], ChEMBL [32], PubChem [33], Drugbank [34], KEGG [35], Lipid Maps [36], and HMDB [37]. For genes, the HGNC IDs [26] were used to request matching identifiers for NCBI Gene [38], UniProt [39] and Ensembl [40].

2.7 File creation

All AOP-Wiki content, persistent identifiers, ontology annotations, and additional information for chemicals, genes and proteins, were stored into three RDF files using Turtle (ttl) syntax. While the central AOP-Wiki RDF file contains all existing AOP-Wiki components plus added chemical identifiers and identifiers mapped from PR terms in BOs, the second file contains all the text-mapped gene IDs and matching identifiers added by BridgeDb. These files are accompanied by a metadata file which describes the datasets, code, and provenance, using standard vocabularies for semantic annotations of metadata, such as Dublin Core [41, 42], Data Catalog Vocabulary [43], Friend of a Friend [44], and Vocabulary of Interlinked Datasets [45]. The RDF files were validated with the IDLab Turtle validator [46], an open-source RDF validator for Turtle and Ntriples syntax and XSD datatype errors.

2.8 Validation

The AOP-Wiki RDF was tested locally by loading the data in a SPARQL endpoint using an openlink/virtuoso-opensource-7 Docker image and exploring the data with SPARQL queries using a Jupyter notebook. These SPARQL queries retrieve numbers for types of subjects, numbers of times that ontologies are used, and number of link-outs to the various databases [21].?

3 Results

The main result of this project is an RDF schema and scripts that lead to the production of RDF content for all AOP-Wiki content with additional semantic annotations, persistent identifiers and extended identifiers for genes and chemicals, and consists of 115,247 unique triples consisting of 14,727 unique subjects, 156 unique predicates, and 51,131 unique objects (Figure 1). The semantic annotation was done using eight standard metadata vocabularies and seventeen domain-specific ontologies and vocabularies. We here detail these results.

The metadata vocabulary we used most is Dublin Core, of which terms are present in 47,993 triples in the AOP-Wiki RDF. Its original set of terms was used to relate various subjects to their identifier, title, description, source, and creator, and the extended set of terms was used to

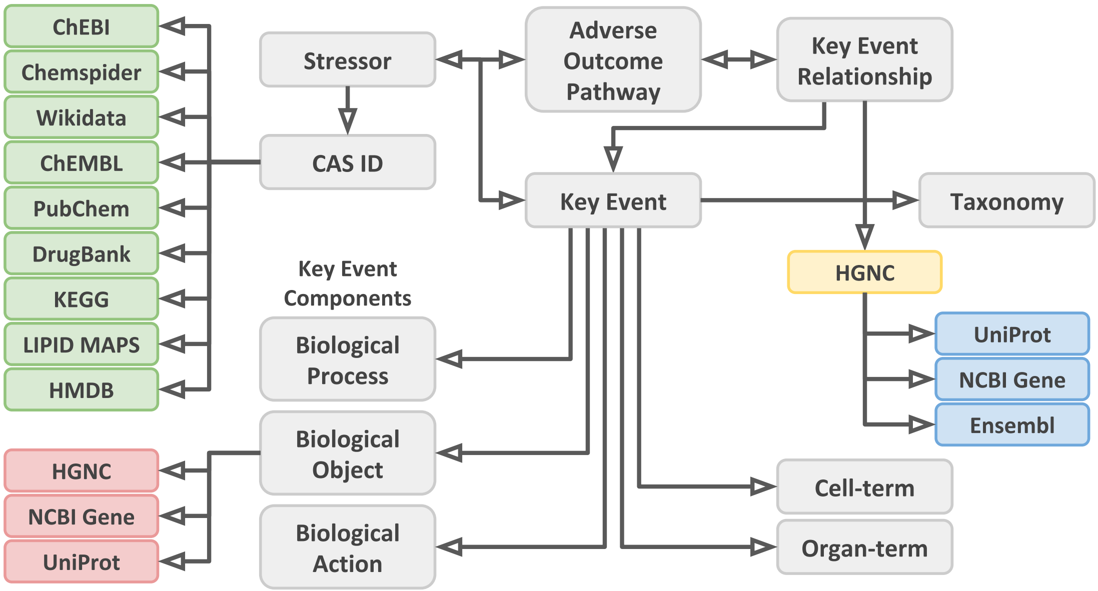


Figure 1: **General overview of the AOP-Wiki RDF scheme.** Arrows show the directional relationships described in the RDF. Grey boxes are the basic elements of the AOP-Wiki. Green boxes indicate added chemical IDs using BridgeDb. Red boxes indicate added gene/protein IDs using Protein Ontology mapping. The yellow box indicates the text-mapped gene IDs and the blue boxes indicate the added gene/protein IDs mapped from the text-mapped gene IDs using BridgeDb.

describe the alternative name, abstract, creation and modification date, and relational information to other subjects with 'dcterms:isPartOf'. Other standard vocabularies we extensively used are the RDF vocabulary to describe the type of subjects with the 'rdf:type' term which we used 21,057 times, and the RDF Schema vocabulary to describe the label of subjects with the 'rdfs:label' was used 3,261 times. Furthermore, the Friend Of A Friend vocabulary is used to define the webpage URLs of AOPs, KEs, KERs and Stressors with 'foaf:page' a total of 3,263 times, and the term 'skos:exactMatch' from the Simple Knowledge Organization System (SKOS) was used 8,539 times to map chemical and gene/protein identifiers to other database identifiers. The NCI Thesaurus is used a total of 2,858 times for predicates and 499 times for objects, using seven distinct terms in AOP, KE, KER and Stressor triples.

Developed for the AOP domain of research and facilitate consistent reporting, the AOPO has been used for the semantic annotations for AOP-specific elements. Terms of the AOPO were used to provide relational information for AOPs, KEs, KERs, and life-stage applicability, stressors, chemicals, and cell- and organ-terms. In total, the AOPO is used 8,615 times for predicate annotations, and 2,886 times as object annotations.

The AOP subjects are connected to KE triples with the AOPO terms of 'has_key_event', 'has_molecular_initiating_event' and 'has_adverse_outcome', and to KER triples with 'has_key _event_relationship'. Furthermore, AOPs are unique to contain an abstract, creator, and access rights, and they contain information on the AOP overall assessment including the overall assessment description, KE essentiality, applicability, weight of evidence, quantitative considerations and the potential applications of the AOP. Furthermore, they include annotations for stressors, and sex and life stage applicability (Figure 2).

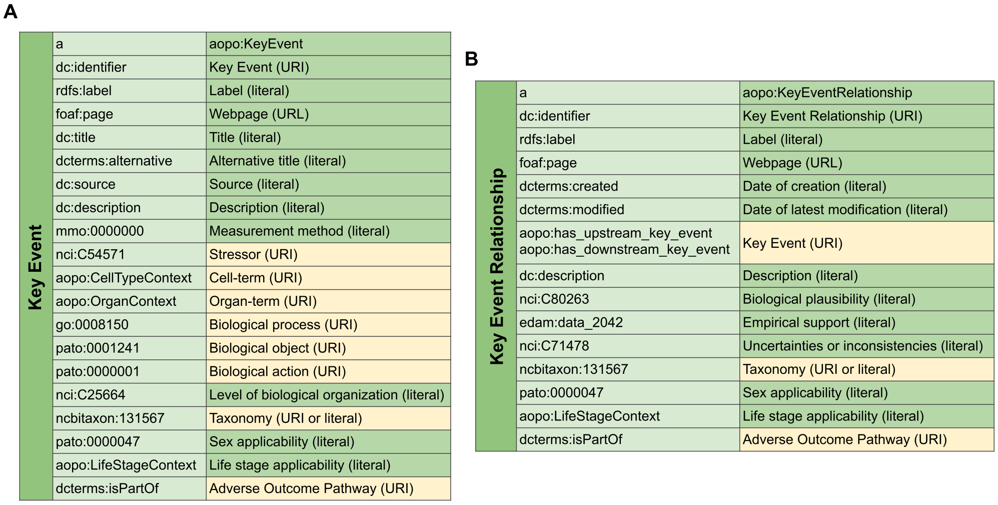
Adverse Outcome Pathway	a	aopo:AdverseOutcomePathway
	dc:identifier	Adverse Outcome Pathway (URI)
	rdfs:label	Label (literal)
	foaf:page	Webpage (URL)
	dc:title	Title (literal)
	dcterms:alternative	Alternative title (literal)
	dc:source	Source (literal)
	dcterms:created	Date of creation (literal)
	dcterms:modified	Date of latest modification (literal)
	dc:creator	Author (literal)
	dcterms:abstract	Abstract (literal)
	dc:description	Description (literal)
	nci:C54571	Stressor (URI)
	aopo:has_key_event aopo:has_molecular_initiating_event aopo:has_adverse_outcome	Key Event (URI)
	aopo:has_key_event_relationship	Key Event Relationship (URI)
	dc:accessRights	AOP status (literal)
	pato:PATO_0000047	Sex applicability (literal)
	aopo:LifeStageContext	Life stage applicability (literal)
	aopo:AopContext	Applicability (literal)
	edam:operation_3799	Quantitative considerations (literal)
	aopo:has_evidence	Weight of Evidence (literal)
	nci:C25725	Potential applications (literal)
	nci:C25217	Overall assessment (literal)
	nci:C48192	Key Event essentiality (literal)

Figure 2: **Adverse Outcome Pathways and their properties in RDF.** The middle column indicates predicates, and the right column describes the objects. Yellow boxes show the object IRIs that connect to other subjects in the RDF.

KEs have unique properties describing measurement methods, level of biological organization, and structured information on cell-terms, organ-terms, BPs, BOs and BAs. The measurement methods are coupled to KEs with the predicate 'mmo:0000000', which stands for measurement method, and level of biological organization is linked with the 'nci:C25664', which stands for its level. The ontological annotations of KEs are connected through IRIs of other subjects in the RDF. (Figure 3A).

KER-specific properties are the biological plausibility, empirical support, uncertainties, which we linked with the predicates 'nci:C80263', 'edam:data_2042' and 'nci:71478'. These stand for the rationale, evidence, and uncertainty, respectively. Also, the RDF connects the upstream and downstream KEs of KERs with the terms 'aopo:has_upstream_key_event' and 'aopo:has_downstream _key_event' (Figure 3B).

The RDF contains general stressor information such as descriptions and identifiers, and 64% of them are linked to chemicals with the predicate 'aopo:has_chemical_entity' (Figure 4A). These chemical subjects are annotated with CAS identifiers, but often also have properties for the



A

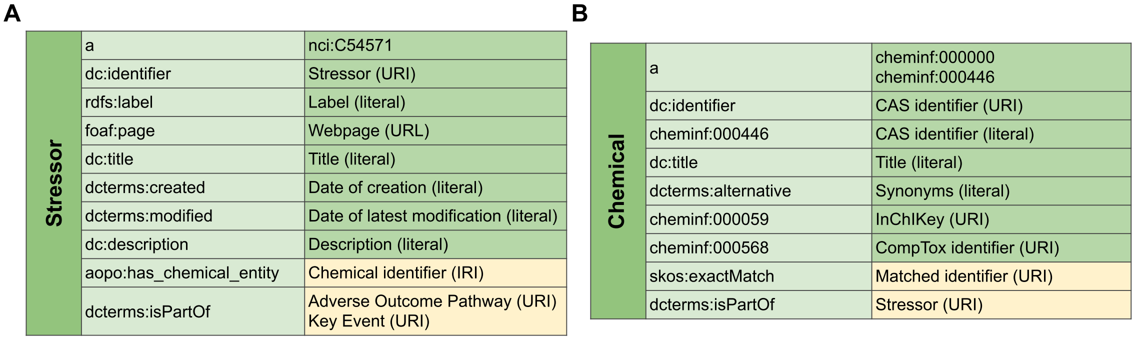
Key Event		
	a	aopo:KeyEvent
	dc:identifier	Key Event (URI)
	rdfs:label	Label (literal)
	foaf:page	Webpage (URL)
	dc:title	Title (literal)
	dcterms:alternative	Alternative title (literal)
	dc:source	Source (literal)
	dc:description	Description (literal)
	mmo:0000000	Measurement method (literal)
	nci:C54571	Stressor (URI)
	aopo:CellTypeContext	Cell-term (URI)
	aopo:OrganContext	Organ-term (URI)
	go:0008150	Biological process (URI)
	pato:0001241	Biological object (URI)
	pato:0000001	Biological action (URI)
	nci:C25664	Level of biological organization (literal)
	ncbitaxon:131567	Taxonomy (URI or literal)
	pato:0000047	Sex applicability (literal)
	aopo:LifeStageContext	Life stage applicability (literal)
	dcterms:isPartOf	Adverse Outcome Pathway (URI)

B

Key Event Relationship		
	a	aopo:KeyEventRelationship
	dc:identifier	Key Event Relationship (URI)
	rdfs:label	Label (literal)
	foaf:page	Webpage (URL)
	dcterms:created	Date of creation (literal)
	dcterms:modified	Date of latest modification (literal)
	aopo:has_upstream_key_event aopo:has_downstream_key_event	Key Event (URI)
	dc:description	Description (literal)
	nci:C80263	Biological plausibility (literal)
	edam:data_2042	Empirical support (literal)
	nci:C71478	Uncertainties or inconsistencies (literal)
	ncbitaxon:131567	Taxonomy (URI or literal)
	pato:0000047	Sex applicability (literal)
	aopo:LifeStageContext	Life stage applicability (literal)
	dcterms:isPartOf	Adverse Outcome Pathway (URI)

Figure 3: **Key Events and their properties in RDF.** The middle column indicates predicates, and the right column describes the objects. Yellow boxes show the object IRIs that connect to other subjects in the RDF.

InChIKeys and CompTox identifiers, all of which we annotated with the Cheminformatics Ontology (Figure 4B) [47]. Furthermore, the chemicals have predicate 'skos:exactMatch' to link to all mapped chemical subjects present in the RDF, providing link-outs to eight additional external databases (Figure 5C). We annotated these with 'rdf:type' and terms from the Cheminformatics Ontology. In total, there are 3,737 link-outs to twelve different chemical databases, allowing users to explore the AOP-Wiki by using their preferred type of chemical identifiers.



A

Stressor		
	a	nci:C54571
	dc:identifier	Stressor (URI)
	rdfs:label	Label (literal)
	foaf:page	Webpage (URL)
	dc:title	Title (literal)
	dcterms:created	Date of creation (literal)
	dcterms:modified	Date of latest modification (literal)
	dc:description	Description (literal)
	aopo:has_chemical_entity	Chemical identifier (IRI)
	dcterms:isPartOf	Adverse Outcome Pathway (URI) Key Event (URI)

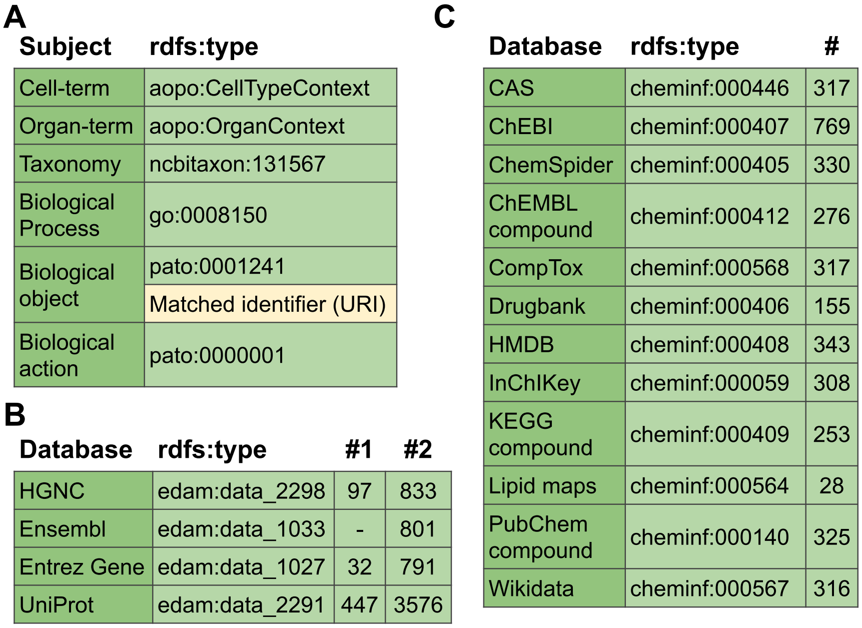
B

Chemical		
	a	cheminf:000000 cheminf:000446
	dc:identifier	CAS identifier (URI)
	cheminf:000446	CAS identifier (literal)
	dc:title	Title (literal)
	dcterms:alternative	Synonyms (literal)
	cheminf:000059	InChIKey (URI)
	cheminf:000568	CompTox identifier (URI)
	skos:exactMatch	Matched identifier (URI)
	dcterms:isPartOf	Stressor (URI)

Figure 4: **Stressors and chemicals and their properties in RDF.** The middle column indicates predicates, and the right column describes the objects. Yellow boxes show the object IRIs that connect to other subjects in the RDF.

Since the taxonomies, cell terms, organ terms, and the KE components all already have ontological annotations in the AOP-Wiki, they have the same properties that describe their type (5A), identifier, title and source. These titles are based on the user-provided entries in the AOP-Wiki. Unique for the biological objects annotated with the Protein Ontology is the inclusion of the 'skos:exactMatch' predicate linking to 576 matching identifiers from UniProt, HGNC and NCBI Gene (Figure 5B).

Extending the links of KEs and KERs with genes and proteins, RDF triples of 833 text-mapped gene identifiers on KEs and KERs are stored in a separate file. These make triples of KE and KER subjects to link to the mapped HGNC identifiers with the predicate 'edam:data_1025', which stands for Gene identifier. These HGNC identifiers are subjects themselves, and have the 'skos:exactMatch' predicate to link to matching identifiers from UniProt, NCBI Gene, and Ensembl, providing a total of 6,001 link-outs using this method (Figure 5B).



A

Subject	rdfs:type
Cell-term	aopo:CellTypeContext
Organ-term	aopo:OrganContext
Taxonomy	ncbitaxon:131567
Biological Process	go:0008150
Biological object	pato:0001241
	Matched identifier (URI)
Biological action	pato:0000001

B

Database	rdfs:type	#1	#2
HGNC	edam:data_2298	97	833
Ensembl	edam:data_1033	-	801
Entrez Gene	edam:data_1027	32	791
UniProt	edam:data_2291	447	3576

C

Database	rdfs:type	#
CAS	cheminf:000446	317
ChEBI	cheminf:000407	769
ChemSpider	cheminf:000405	330
ChEMBL compound	cheminf:000412	276
CompTox	cheminf:000568	317
Drugbank	cheminf:000406	155
HMDB	cheminf:000408	343
InChIKey	cheminf:000059	308
KEGG compound	cheminf:000409	253
Lipid maps	cheminf:000564	28
PubChem compound	cheminf:000140	325
Wikidata	cheminf:000567	316

Figure 5: **Ontology annotations and molecular identifiers.** A. Cell-terms, organ-terms, taxonomies, Key Event Components and type annotation in the RDF. The yellow box indicates the matching identifiers to other subjects in the RDF. B. Gene and protein databases, their type annotation, and the number of identifiers present in the RDF. Values in #1 are based on Protein Ontology mappings, and values in #2 are based on textual mapping with HGNC symbols. C. Chemical databases, their type annotation and the number of identifiers present in the RDF.

4 Discussion

The work described in this paper has led to the creation of AOP-Wiki RDF based on the existing AOP-Wiki XML, combined with a variety of ontologies and enriched with persistent identifiers. Besides, the data is extended with additional identifiers for chemicals, proteins and genes. The RDF has been validated with the IDLab Turtle validator [46] and was tested on a SPARQL endpoint using a Jupyter notebook. These developments made AOP-Wiki content ready for use in risk assessment workflows, through coding environments, or in federated SPARQL queries.

Because the AOPO is developed for consistent reporting in the domain of AOPs and allowing the integration of data and tools [19], it was an obvious choice to implement the AOPO for semantic annotations in the AOP-Wiki RDF. The ontology includes a variety of AOP-specific definitions for properties and classes which were directly applicable to AOP-Wiki content in the RDF. However, these do not fully cover all types of entities and relationships that exist in the AOP-Wiki. For example, while it has terms to describe the connections between AOPs, KEs and KERs, there is no annotation for the link with stressors. Similarly, terms are lacking for sex and taxonomic applicability, KE components, KER-specific information, and AOP assessment sections such as KE essentiality and quantitative considerations, among others.

For the terms missing in the AOPO, we selected definitions from a wide range of other ontologies and vocabularies for the semantification of predicates and subjects, including NCI Thesaurus, NCBI of Organismal Classification, Gene Ontology, and Measurement Method Ontology, among others. Whereas the majority of the AOP-Wiki contents are generic and can be described with well-established metadata ontologies, some properties of AOPs, KEs and KERs could not be found, leading to the selection of more general terms, lacking detail that would be preferred. With the conversion to RDF, most of the necessary terms have been uncovered and documented, and therefore are the logical base set of terms to be added to the AOPO.

Because the realm of AOPs includes many types of data, knowledge, repositories and services, the development and implementation of a central, community-wide vocabulary would facilitate their integration. Since the AOPO has been developed to fill that purpose, it could be extended to include descriptions of classes and properties for all ontology terms used in the AOP-Wiki RDF. Having a central, field-wide ontology for AOPs helps maintaining a high quality vocabulary through continuous development and involvement of the community. Such an ontology would facilitate the annotation and integration of data, resources and tools, as is done with the eNanoMapper ontology in the nanotoxicology community [48].

With the increased importance of consistent use of identifiers to integrate knowledge and data, our implementation of persistent identifiers for AOPs, KEs, KERs, Stressors, chemicals, proteins and genes will benefit the integration of the AOP-Wiki with other resources, data, and tools [49]. These persistent identifiers stored in the MIRIAM registry are stable, unique, resolvable, documented, and directly link to the corresponding entries in the databases [22, 50]. Furthermore, our efforts have introduced additional content to the AOP-Wiki RDF through ID mappings and text-mapping for chemicals and genes, providing more ways of extrapolating the data and linking with other resources and data.

While we have added molecular identifiers to increase the number of link-outs and improve the usefulness of the database, our addition of genes through textual ID mapping does introduce errors to the AOP-Wiki RDF. The automated process on free-text content assumes good practice in writing gene symbols and names according to the HGNC guidelines [51]. Although HGNC strives for stable gene symbols and makes justified changes for problematic ones [52], some gene symbols still overlap with free-text abbreviations in the AOP-Wiki and are therefore falsely recognised.

Opportunities exist to improve the AOP-Wiki machine-readability by having more structured text and annotations for molecular entities, pathways, organs, species and other biological concepts that are relevant for AOPs and not yet covered by the KE Components. Text-mining tools, such as ProMiner[53] and PolySearch2[54], could be implemented for extracting biological concepts and understanding associations to add more structured information in the AOP-Wiki and facilitate the integration with other databases and tools. Once such concepts are recognised and extracted, the RDF could be extended and increase the interoperability of the AOP-Wiki with external databases, such as pathway databases.

The AOP-Wiki RDF allows for new and efficient ways of accessing the data, and using it to answer questions. By loading the RDF in a SPARQL endpoint, SPARQL queries can be used to access the data and extract all necessary information. It allows complex queries across the complete AOP-Wiki database, optional filtering for any variable, and requesting an outputs suitable for answering the research question. Furthermore, these SPARQL queries can be executed from most coding environments as part of larger workflows or data pipelines. It also facilitates direct linkage of databases through federated SPARQL queries, which returns information across databases with a single query. Any database with a SPARQL endpoint can be used for such questions across databases, such as WikiPathways [17, 16], Wikidata [30, 31], neXtProt [15], UniProt [39], DisGeNET, Rhea [55], Pathway Commons [56], among others (see `https://github.com/marvinm2/AOPWikiRDF` for examples).

Another way of using the RDF to extract AOP-Wiki content is through a web service such as the git repository linked data API constructor (grlc) [57], which can build a Web API on top of a SPARQL endpoint with predefined SPARQL queries. While more straight-forward than SPARQL queries, the API is limited to the predefined SPARQL queries and variables implemented in these.

An advantage of creating RDF for the AOP-Wiki is the ability to link and expand AOP-Wiki content with information from other databases. For example, the AOP-DB combines knowledge from the AOP-Wiki with annotations of genes, chemicals, diseases, tissues, pathways, ontologies, and ToxCast data [58, 59]. Future work should focus on integrating such efforts by developing

RDF and thereby allow full integration of their data and tools [60] with the AOP-Wiki RDF and other databases.

The implementation of compact identifiers and development of a formal, machine-readable RDF schema makes the AOP-Wiki more findable and relatable to other components of the database, and by allowing SPARQL queries and API to explore the data, the AOP-Wiki database was made more accessible. Furthermore, the addition of link-outs to various chemical, gene and protein databases, as well as the data storage in an RDF format and implementing LOD standards, has made the data more interoperable with other databases and tools. Taken together with the addition of metadata and semantic information represented by ontology annotations, the content of the AOP-Wiki has been made more reusable. Therefore, the development of the AOP-Wiki RDF addresses all major FAIR principles [13].

Overall, the AOP-Wiki RDF allows for new ways of exploring the data, using it in automated workflows, from coding environments, or directly through a SPARQL endpoint. With the implementation also comes the possibility to execute federated queries to combine data of multiple resources and answer more elaborate questions.

5 Data links

All data and code used in this manuscript are publicly available. The main conversion code, statistics code and example SPARQL queries are available on `https://github.com/marvinm2/AOPWikiRDF`. The AOP-Wiki XML can be downloaded on `https://aopwiki.org/downloads`. The HGNC mapping file can be downloaded via `https://www.genenames.org/` and the Protein Ontology mapping file can be downloaded with `https://proconsortium.org/download/current/`.

6 Conflict of Interest

The authors declare that there is no conflict of interest.

7 Funding

This project has received funding from the European Union's Horizon 2020 (EU 2020) research and innovation program under grant agreement no. 681002 (EU-ToxRisk) and EINFRA-22-2016 program under grant agreement no. 731075 (OpenRiskNet).

References

[1] Gerald T. Ankley et al. "Adverse outcome pathways: A conceptual framework to support ecotoxicology research and risk assessment". In: *Environmental Toxicology and Chemistry* 29.3 (Mar. 2010), pp. 730–741. DOI: 10.1002/etc.34.

[2] Daniel Krewski et al. "Toxicity Testing in the 21st Century: Implications for Human Health Risk Assessment". In: *Risk Analysis* 29.4 (Apr. 2009), pp. 474–479. DOI: 10.1111/j.1539-6924.2008.01150.x.

[3] Marcel Leist et al. "Adverse outcome pathways: opportunities, limitations and open questions". In: *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3477–3505. DOI: 10.1007/s00204-017-2045-3.

[4] Catherine Willett. "The Use of Adverse Outcome Pathways (AOPs) to Support Chemical Safety Decisions Within the Context of Integrated Approaches to Testing and Assessment (IATA)". In: *Alternatives to Animal Testing* (2019), pp. 83–90. DOI: 10.1007/978-981-13-2447-5_11.

[5] Mathieu Vinken et al. "Adverse outcome pathways: a concise introduction for toxicologists". In: *Archives of Toxicology* 91.11 (Nov. 2017), pp. 3697–3707. DOI: 10.1007/s00204-017-2020-z.

[6] Daniel L Villeneuve et al. "Adverse outcome pathway (AOP) development I: strategies and principles." In: *Toxicological sciences* 142.2 (Dec. 2014), pp. 312–20. DOI: 10.1093/toxsci/kfu199.

[7] Jaeseong Jeong and Jinhee Choi. "Use of adverse outcome pathways in chemical toxicity testing: potential advantages and limitations". In: *Environmental Health and Toxicology* 33.1 (Dec. 2017), e2018002. DOI: 10.5620/eht.e2018002.

[8] Noffisat O. Oki and Stephen W. Edwards. "An integrative data mining approach to identifying adverse outcome pathway signatures". In: *Toxicology* 350-352 (Mar. 2016), pp. 49–61. ISSN: 18793185. DOI: 10.1016/j.tox.2016.04.004.

[9] Antony J. Williams et al. "The CompTox Chemistry Dashboard: A community data resource for environmental chemistry". In: *Journal of Cheminformatics* 9.1 (Nov. 2017), p. 61. DOI: 10.1186/s13321-017-0247-6.

[10] Scott Federhen. "The NCBI Taxonomy database". In: *Nucleic Acids Research* 40 (D1). DOI: 10.1093/nar/gkr1178.

[11] Cataia Ives et al. "Creating a Structured AOP Knowledgebase via Ontology-Based Annotations." In: *Applied in vitro toxicology* 3.4 (Dec. 2017), pp. 298–311. DOI: 10.1089/aivt.2017.0017.

[12] Richard Cyganiak, David Wood, and Markus Lanthaler. *RDF 1.1 Concepts and Abstract Syntax*. 2014. URL: https://www.w3.org/TR/rdf11-concepts/ (visited on 12/08/2020).

[13] Mark D. Wilkinson et al. "Comment: The FAIR Guiding Principles for scientific data management and stewardship". In: *Scientific Data* 3.1 (Mar. 2016), pp. 1–9. DOI: 10.1038/sdata.2016.18.

[14] Christine Chichester et al. "Converting neXtProt into linked data and nanopublications". In: *Semantic Web* 6.2 (Jan. 2015), pp. 147–153. DOI: 10.3233/SW-140149.

[15] Monique Zahn-Zabal et al. "The neXtProt knowledgebase in 2020: Data, tools and usability improvements". In: *Nucleic Acids Research* 48.D1 (Jan. 2020), pp. D328–D334. DOI: 10.1093/nar/gkz995.

[16] Andra Waagmeester et al. "Using the Semantic Web for Rapid Integration of WikiPathways with Other Biological Online Data Resources". In: *PLOS Computational Biology* 12.6 (June 2016). Ed. by Christos A. Ouzounis, e1004989. DOI: 10.1371/journal.pcbi.1004989.

[17] Marvin Martens et al. "WikiPathways: connecting communities". In: *Nucleic Acids Research* 49.D1 (Nov. 2020), pp. 613–621. ISSN: 0305-1048. DOI: 10.1093/nar/gkaa1024.

[18] Fiona Sewell et al. "The future trajectory of adverse outcome pathways: a commentary". In: *Archives of Toxicology* 92.4 (Apr. 2018), pp. 1657–1661. DOI: 10.1007/s00204-018-2183-2.

[19] Lyle D. Burgoon. "The AOPOntology: A semantic artificial intelligence tool for predictive toxicology". In: *Applied In Vitro Toxicology* 3.3 (Sept. 2017), pp. 278–281. DOI: 10.1089/aivt.2017.0012.

[20] Clemens Wittwehr et al. "How Adverse Outcome Pathways Can Aid the Development and Use of Computational Prediction Models for Regulatory Toxicology." In: *Toxicological sciences : an official journal of the Society of Toxicology* 155.2 (2017), pp. 326–336. DOI: 10.1093/toxsci/kfw207.

[21] Marvin Martens. "marvinm2/AOPWikiRDF: Finished notebooks". In: (Nov. 2020). DOI: 10.5281/ZENODO.4292485.

[22] Nick Juty, Nicolas Le Novère, and Camille Laibe. "Identifiers.org and MIRIAM Registry: community resources to provide persistent identification". In: *Nucleic Acids Research* 40.D1 (Dec. 2011), pp. D580–D586. ISSN: 0305-1048. DOI: 10.1093/nar/gkr1097.

[23] Patricia L Whetzel et al. "BioPortal: enhanced functionality via new Web services from the National Center for Biomedical Ontology to access and use ontologies in software applications." In: *Nucleic acids research* 39.Web Server issue (July 2011), W541–5. DOI: 10.1093/nar/gkr469.

[24] *OWL Web Ontology Language Overview.* URL: https://www.w3.org/TR/owl-features/ (visited on 12/08/2020).

[25] Darren A Natale et al. "The Protein Ontology: a structured representation of protein forms and complexes." In: *Nucleic acids research* 39.Database issue (Jan. 2011), pp. D539–45. DOI: 10.1093/nar/gkq907.

[26] Bryony Braschi et al. "Genenames.org: the HGNC and VGNC resources in 2019". In: *Nucleic Acids Research* 47.D1 (Oct. 2018), pp. D786–D792. ISSN: 0305-1048. DOI: 10.1093/nar/gky930.

[27] Martijn P van Iersel et al. "The BridgeDb framework: standardized access to gene, protein and metabolite identifier mapping services". In: *BMC Bioinformatics* 11.1 (2010), p. 5. ISSN: 1471-2105. DOI: 10.1186/1471-2105-11-5.

[28] Kirill Degtyarenko et al. "ChEBI: a database and ontology for chemical entities of biological interest." In: *Nucleic acids research* 36.Database issue (Jan. 2008), pp. D344–50. ISSN: 1362-4962. DOI: 10.1093/nar/gkm791.

[29] Harry E. Pence and Antony Williams. "Chemspider: An online chemical information resource". In: *Journal of Chemical Education* 87.11 (Nov. 2010), pp. 1123–1124. DOI: 10.1021/ed100697w.

[30] Andra Waagmeester et al. "Wikidata as a knowledge graph for the life sciences". In: *eLife* 9 (Mar. 2020). ISSN: 2050084X. DOI: 10.7554/eLife.52614.

[31] Fredo Erxleben et al. "Introducing Wikidata to the Linked Data Web". In: Springer, Cham, Oct. 2014, pp. 50–65. DOI: 10.1007/978-3-319-11964-9_4.

[32] Anna Gaulton et al. "ChEMBL: a large-scale bioactivity database for drug discovery." In: *Nucleic acids research* 40.Database issue (Jan. 2012), pp. D1100–7. ISSN: 1362-4962. DOI: 10.1093/nar/gkr777.

[33] Sunghwan Kim et al. "PubChem substance and compound databases". In: *Nucleic Acids Research* 44.D1 (2016), pp. D1202–D1213. ISSN: 13624962. DOI: 10.1093/nar/gkv951.

[34] D. S. Wishart. "DrugBank: a comprehensive resource for in silico drug discovery and exploration". In: *Nucleic Acids Research* 34.90001 (Jan. 2006), pp. D668–D672. ISSN: 0305-1048. DOI: 10.1093/nar/gkj067.

[35] M. Kanehisa. "KEGG: Kyoto Encyclopedia of Genes and Genomes". In: *Nucleic Acids Research* 28.1 (Jan. 2000), pp. 27–30. ISSN: 0305-1048. DOI: 10.1093/nar/28.1.27.

[36] Eoin Fahy et al. "Update of the LIPID MAPS comprehensive classification system for lipids". In: *Journal of Lipid Research* 50.SUPPL. (Apr. 2009). ISSN: 00222275. DOI: 10.1194/jlr.R800095-JLR200.

[37] David S. Wishart et al. "HMDB: Database Statistics". In: (). ISSN: 03051048. DOI: 10.1093/nar/gkl923.

[38] Donna Maglott et al. "Entrez Gene: Gene-centered information at NCBI". In: *Nucleic Acids Research* 33.DATABASE ISS. (Jan. 2005). ISSN: 03051048. DOI: 10.1093/nar/gki031.

[39] Alex Bateman. "UniProt: A worldwide hub of protein knowledge". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D506–D515. ISSN: 13624962. DOI: 10.1093/nar/gky1049.

[40] T. Hubbard. "The Ensembl genome database project". In: *Nucleic Acids Research* 30.1 (2002), pp. 38–41. ISSN: 1362-4962. DOI: 10.1093/nar/30.1.38.

[41] *DCMI: Dublin Core^{TM} Metadata Element Set, Version 1.1: Reference Description.* URL: https://www.dublincore.org/specifications/dublin-core/dces/ (visited on 03/10/2020).

[42] *DCMI: DCMI Metadata Terms*. URL: https://www.dublincore.org/specifications/dublin-core/dcmi-terms/ (visited on 03/10/2020).

[43] Simon Cox et al. *Data Catalog Vocabulary (DCAT) - Version 2*. 2020. URL: https://www.w3.org/TR/2020/REC-vocab-dcat-2-20200204/ (visited on 12/08/2020).

[44] *FOAF Vocabulary Specification*. URL: http://xmlns.com/foaf/spec/ (visited on 03/10/2020).

[45] K Alexander et al. *Describing Linked Datasets with the VoID Vocabulary*. 2011. URL: http://scholar.google.com/scholar?cluster=14382548542971533685&hl=en&oi=scholarr#0 (visited on 12/08/2020).

[46] IDLab - Ghent University. *IDLabResearch/TurtleValidator: A Turtle validator on command line and in browser*. URL: https://github.com/IDLabResearch/TurtleValidator (visited on 02/27/2020).

[47] Janna Hastings et al. "The Chemical Information Ontology: Provenance and Disambiguation for Chemical Data on the Biological Semantic Web". In: *PLoS ONE* 6.10 (Oct. 2011). Ed. by Franca Fraternali, e25513. ISSN: 1932-6203. DOI: 10.1371/journal.pone.0025513.

[48] Janna Hastings et al. "eNanoMapper: Harnessing ontologies to enable data integration for nanomaterial risk assessment". In: *Journal of Biomedical Semantics* 6.1 (Mar. 2015), p. 10. ISSN: 20411480. DOI: 10.1186/s13326-015-0005-5.

[49] Sarala M Wimalaratne et al. "Uniform resolution of compact identifiers for biomedical data". In: (). DOI: 10.1038/sdata.2018.29.

[50] Julie A. McMurry et al. "Identifiers for the 21st century: How to design, provision, and reuse persistent identifiers to maximize utility and impact of life science data". In: *PLoS Biology* 15.6 (June 2017). ISSN: 15457885. DOI: 10.1371/journal.pbio.2001414.

[51] Hester M. Wain et al. "Guidelines for human gene nomenclature". In: *Genomics* 79.4 (Apr. 2002), pp. 464–470. ISSN: 08887543. DOI: 10.1006/geno.2002.6748.

[52] Susan Tweedie et al. "Genenames.org: the HGNC and VGNC resources in 2021". In: *Nucleic Acids Research* 1 (Nov. 2020). ISSN: 0305-1048. DOI: 10.1093/nar/gkaa980.

[53] Daniel Hanisch et al. "ProMiner: Rule-based protein and gene entity recognition". In: *BMC Bioinformatics* 6.SUPPL.1 (May 2005), S14. ISSN: 14712105. DOI: 10.1186/1471-2105-6-S1-S14.

[54] Yifeng Liu, Yongjie Liang, and David Wishart. "PolySearch2: a significantly improved text-mining system for discovering associations between human diseases, genes, drugs, metabolites, toxins and more". In: *Nucleic Acids Research* 43 (2015), pp. 535–542. DOI: 10.1093/nar/gkv383.

[55] Thierry Lombardot et al. "Updates in Rhea: SPARQLing biochemical reaction data". In: *Nucleic Acids Research* 47.D1 (Jan. 2019), pp. D596–D600. ISSN: 13624962. DOI: 10.1093/nar/gky876.

[56] Ethan G Cerami et al. "Pathway Commons, a web resource for biological pathway data". In: (). DOI: 10.1093/nar/gkq1039.

[57] Albert Meroño-Peñuela et al. "CLARIAH/grlc: January 2020 patch". In: (Jan. 2020). DOI: 10.5281/ZENODO.3606813.

[58] Holly M. Mortensen et al. "Leveraging human genetic and adverse outcome pathway (AOP) data to inform susceptibility in human health risk assessment". In: *Mammalian Genome* 29.1-2 (Feb. 2018), pp. 190–204. ISSN: 14321777. DOI: 10.1007/s00335-018-9738-7.

[59] Maureen E. Pittman et al. "AOP-DB: A database resource for the exploration of Adverse Outcome Pathways through integrated association networks". In: *Toxicology and Applied Pharmacology* 343 (Mar. 2018), pp. 71–83. ISSN: 10960333. DOI: 10.1016/j.taap.2018.02.006.

[60] Paul S. Price, Annie M. Jarabek, and Lyle D. Burgoon. "Organizing mechanism-related information on chemical interactions using a framework based on the aggregate exposure and adverse outcome pathways". In: *Environment International* 138 (May 2020), p. 105673. ISSN: 18736750. DOI: 10.1016/j.envint.2020.105673.

8 Annex

Table 1: **Prefixes in the RDF**

Ontology name	**Prefix in RDF**	**Base IRI**
Dublin Core	dc	`http://purl.org/dc/elements/1.1/`
DCMI Metadata Terms	dcterms	`http://purl.org/dc/terms/`
RDF Schema	rdfs	`http://www.w3.org/2000/01/rdf-schema#`
Friend Of A Friend	foaf	`http://xmlns.com/foaf/0.1/`
Adverse Outcome Pathway Ontology	aopo	`http://aopkb.org/aop_ontology#`
Phenotypic Quality Ontology	pato	`http://purl.obolibrary.org/obo/PATO_`
Chemical Information ontology	cheminf	`http://semanticscience.org/resource/CHEMINF_`
NCI Thesaurus	nci	`http://ncicb.nci.nih.gov/xml/owl/EVS/Thesaurus.owl#`
Measurement Method Ontology	mmo	`http://purl.obolibrary.org/obo/MMO_`
Simple Knowledge Organization System	skos	`http://www.w3.org/2004/02/skos/core#`
National Center for Biotechnology Information Organismal Classification	ncbitaxon	`http://purl.bioontology.org/ontology/NCBITAXON/`
Gene Ontology	go	`http://purl.obolibrary.org/obo/GO_`
EDAM bioinformatics operations, types of data, data formats, identifiers, and topics	edam	`http://edamontology.org/`
Provenance, Authoring and Versioning	pav	`http://purl.org/pav/`
Vocabulary of Interlinked Datasets	void	`http://rdfs.org/ns/void#`
Data Catalog Vocabulary	dcat	`http://www.w3.org/ns/dcat#`

Table 2: **Prefixes in the RDF for the Key Event Component annotations**

Ontology name	**Prefix in RDF**	**Base URI**
Cell Ontology	cl	`http://purl.obolibrary.org/obo/CL_`
Uber-anatomy ontology	uberon	`http://purl.obolibrary.org/obo/UBERON_`
Gene Ontology	go	`http://purl.obolibrary.org/obo/GO_`
Molecular Interactions Controlled Vocabulary	mi	`http://purl.obolibrary.org/obo/MI_`
Mammalian Phenotype Ontology	mp	`http://purl.obolibrary.org/obo/MP_`
Medical Subject Headings	mesh	`http://purl.bioontology.org/ontology/MESH/`
Human Phenotype Ontology	hp	`http://purl.obolibrary.org/obo/HP_`
Population and Community Ontology	pco	`http://purl.obolibrary.org/obo/PCO_`
Neuro Behavior Ontology	nbo	`http://purl.obolibrary.org/obo/NBO_`
Vertebrate trait ontology	vt	`http://purl.obolibrary.org/obo/VT_`
PRotein Ontology	pr	`http://purl.obolibrary.org/obo/PR_`
Chemical Entities of Biological Interest	chebio	`http://purl.obolibrary.org/obo/CHEBI_`
Foundational Model of Anatomy Ontology	fma	`http://purl.org/sig/ont/fma/fma`