

Adoption of rdf-toolkit for CASE normalization process


Date Submitted	31 Jul 2019
CP Number	003
Status	IMPLEMENTATION (PRODUCT MANAGER)
Impact	LOW
Issue	https://case.atlassian.net/browse/ONT-5
Level of Effort	LOW

CASE Version	0.1.0
Change Type	REVISION

Change Requestor

Requestor	@ Ryan Hohimer
Email	ryan@darklight.ai
Phone	(509) 430-6890
Organization	Darklight, Inc.
Member Class	For-Profit
Contributors	@ Alexander Nelson

Scheduling Track**

	REQUEST (CHECK ONE)	ASSIGNED (OC USE)
COMMITTEE REVIEW – CP IS EXPECTED TO HAVE A NON-TRIVIAL EFFECT ON CASE AND/OR UCO		
FAST-TRACK – CP IMPACT IS OF MINIMAL RISK TO THE CASE ONTOLOGY		
BUG FIX – CP IS A ZERO-RISK CHANGE TO THE CASE ONTOLOGY (E.G., CORRECTION TO SYNTAX OR SPELLING)		

** AT THE DISCRETION OF THE ONTOLOGY COMMITTEE CHAIR, ASSISTANT CHAIR, AND/OR PRODUCT MANAGER, A CHANGE PROPOSAL THAT REQUESTS BYPASSING FULL COMMITTEE REVIEW MAY BE DEEMED AS NEEDING MORE THOROUGH REVIEW AND BE PUT INTO THE FULL FORMAL COMMITTEE REVIEW PROCESS.

Justification for Change

Requirement	<ol style="list-style-type: none">1. CASE ontology files must be normalized to facilitate succinct, error-free merges.2. Instructions need to be developed to assist ontology change submitters with use of the suggested normalization tooling.
Impact: Benefit	Normalized style and syntax allow for less errors during merge processes. Normalization reduces the number of non-substantive changes, while not requiring ontology editing to be done by any particular tool.
Impact: Risk	

Description of Change

Details of Change	See Relevant data section
GitHub Proposal Encoding File	<p>This repository encodes a normalization process:</p> <p>https://github.com/ajnelson-nist/CASE-Examples-QC/</p> <p>Normalization, including by rdf-toolkit, runs with "make normalize" (documented in the repository's README).</p> <p>That repository shows how normalization can be performed, but does not necessarily need to be the exact implementation for ontology maintenance.</p> <p>Note that the referenced repository implements normalization for the CASE example data (JSON-LD) as well, whereas this Change Proposal is scoped to normalizing only the Turtle-formatted ontology definition files</p>
GitHub Supporting File(s)	
Other Supporting File(s)	

Relevant data

This Ontology Change Proposal requests a normalization process should be adopted to prevent errors and confusion during version control. This makes sure that the ontology file will reflect only substantive changes.

Clarifying the Problem

The CASE community will have multiple people contributing to the ontology using the CASE GitHub repository. Contributions will consist of edits that add new concepts, modify existing concepts, correct syntactic errors, and express logical axioms. Each edit will be submitted for review in a pull request. A pull request allows a reviewer to compare the new submission to the existing ontology to validate the new version.

However, the pull request and review process aren't simple if the changes are not substantive. Different versions (edits to the ontology) can contain the same substantive information in different languages (Turtle, RDF/XML, N-Triples, etc.) and different ontology editors may save the same substantive information with slightly different style and formatting (extra whitespace, tabs versus spaces, etc.). All changes in serialization are presented to the reviewer, both substantive and non-substantive. This requires the pull request reviewer to discern between the two.

Definitions to support "Normalization"

The "CASE normalization process" covers two data forms that can have independent normalization rules.

1. The ontology, encoded as Turtle, shall have its own normalization rules.
2. The example instance-data files, encoded as JSON-LD, shall have their own normalization rules.

This Change Proposal is scoped to the use of rdf-toolkit for normalization. However, experimentation has shown undesired behaviors of rdf-toolkit when used on JSON-LD, including confusing data re-structuring. Hence, this proposal will remain scoped to only normalization of the Turtle files that are the ontology serialization. A future proposal will address normalization of the examples.

For CASE, the "normalized" form of Turtle is the form that results from an application of rules to settle questions of Turtle style, such as:

- Definition sort order
- Whitespace usage
- Prefix usage
- Comma placement
- Blank node placement

The rules to be applied can satisfy CASE's needs for normalization as long as the rules have an unambiguous application mechanism. The CASE community can consider the output of a single tool, with all flags specified, to satisfy the function "normalizing the ontology serialization" – so long as that tool and its flags are agreed upon by the community.

The Solution

This change request proposes that the CASE ontology-development process should use the rdf-toolkit utility, which can be found here: <https://github.com/edmcouncil/rdf-toolkit>. The utility would be used before a pull request is submitted, which will remove non-substantive changes from the changed files.

Normalization command

This invocation of rdf-toolkit shows the flags to be used. "INPUT_TTL_FILE" is the CASE Turtle file that is not necessarily normalized. "OUTPUT_TTL_FILE" is the resulting Turtle file after normalization. This same command run on "OUTPUT_TTL_FILE" as input should produce a file with no further changes.

```
java -jar rdf-toolkit.jar \
  --infer-base-iri \
  --inline-blank-nodes \
  --source INPUT_TTL_FILE \
  --source-format turtle \
  --target OUTPUT_TTL_FILE \
  --target-format turtle
```

Why this Solution

The rdf-toolkit utility was created to help users submit Financial Industry Business Ontology (FIBO) ontology pull requests in RDF/XML format to the Object Management Group (OMG). This is a sizable effort that has created a best practice that we can adopt. It is reliable and is actively being updated with improvements. DarkLight, Inc. (CASE Adopter and member of the CASE community) uses the rdf-toolkit utility and is pleased with the consistency. The rdf-toolkit is stable and consistent.

The toolkit will ingest and output Turtle-format ontology files, and is also able to convert information-losslessly to RDF/XML and other ontology formats. This would allow us to output RDF/XML formatted files for documentation purposes and human legibility.

There is a webhook that can be set up to run the toolkit on commit which allows for automating the normalization process. The following notes are intended for the FIBO community, however, they are relevant to this proposal: (https://wiki.edmcouncil.org/display/FIBORDFKIT/Installing+and+Using+the+EDMC+RDF+Serializer?preview=/9013205/9013206/RDFSerializer_Notes.docx).

Related Links:

- <https://www.w3.org/RDF/>
- <https://www.w3.org/2013/data/>
- <https://github.com/edmcouncil/rdf-toolkit>

This Change Proposal was originally prepared by Ryan Hohimer with a broader scope. This reduced-scope version was prepared by Alex Nelson.

Vote

Topic	For	Against
Adoption of rdf-toolkit for CASE normalization process		

Action items

Outcome

Summary of Actions & Dates

(ONTOLOGY COMMITTEE USE ONLY)

Initially Received by OC Date	31 Jul 2019
Accepted as Complete Date	09 Sep 2019
Set Date for Requirements Review (OC)	19 Sep 2019
Distributed for OC Review Date	09 Sep 2019
Designated Special Reviewer(s)	
Review Comments Deadline**	09 Oct 2019
Link to Comment Matrix	https://docs.google.com/spreadsheets/d/1RdQhRQGIR0Rnb-DJ0zyYC5-Ca-aGgKzP/
Date & Result of Requirements Review (OC)	
(Y) Assigned for Development Date	14 Nov 2019
(N) Rejected (Reason)	
Solution Development Start Date	09 Aug 2019
Solution Development End Date	08 Apr 2020
Date & Result of Solutions Approval (OC)	
(Y) Assigned for Implementation Date	16 Apr 2020
(N-R) Returned for Revisions	
(N-X) Rejected	
Implementation Start Date	08 Apr 2020
Implementation Completion Date (publication of revised ontology and official documentation)	15 Jun 2020

** 30 DAYS BEFORE DISTRIBUTION