Specification for Creating and Using Terms

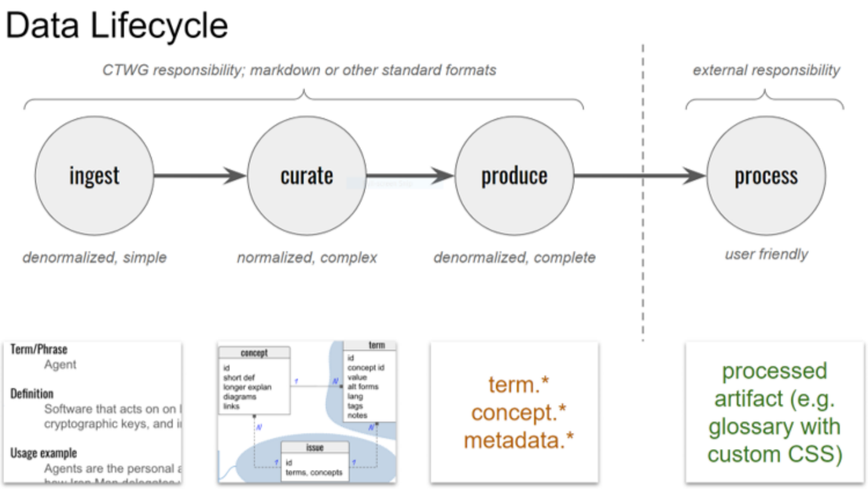
The contents of this page is work in progress and subject to discussion and changes.

Working with term(inologie)s comprises both the definition of terms, underlying concepts, and their interrelationships, as well as the use of such terms in documents, such as whitepapers, specifications and standards. The CTWG aims to facilitate different groups within ToIP to do both of these things, aiming to make life easier for authors and readers/audiences alike. The CTWG realizes that different (groups of) people typically have their own terminology (vocabulary, jargon, etc.) that they want to keep using as they author documents, and therefore makes it an objective to facilitate that. However, members of other groups are likely to understand such terms in their own way. To help them understand the intended meaning of such texts, CTWG:

- provides ToIP groups with easy ways to define the terminology of their group, that is the set of terms, the underlying concepts, the relations between such concepts, examples of various kinds, etc. (whatever turns out to be needed and used);
- provides authors (e.g. of whitepapers, specifications, standards etc.) with easy notations/syntax they can use in their documents to help their different audiences to better/actually understand the intended meaning; such notations/syntax may be references to definitions, the underlying concepts, examples and other properties that terms may have.
- provides ways to render such texts into various formats (e.g. web, pdf, 'digital' (JSON), thereby ensuring that the notations/syntax used to refer to some attribute of a term will be appropriately rendered for the kind of rendering that is being used. This means: provides a non-intrusive way for readers to better/actually understand the intended meaning, such as a popup for web-based documents.

CTWG does NOT provide the actual content of terminologies (meanings, documents). Its focus is on the process (and supporting tools) for allowing others to author and understand texts as they are intended. Participants of CTWG may however provide such content as part of another ToIP group. Also, CTWG (members) may actively help various ToIP groups to understand the various terminological issues they may want to address, teach them how to author texts (including texts that define terminology and their properties), how to use the notations/syntax to help their audiences, etc.

The below figure shows the lifecycle of texts from their raw authoring (INGEST stage) to the production of ToIP-style texts in various renderings (PRODUCE). It also shows how groups can create additional renderings, e.g. using their own styles and formats that are outside the CTWG scope.



Where the various stages are as follows:

INGEST

is a stage that contains contents of various kinds that authors can produce even if they have have limited or no technical working knowledge, such as your average business person, lawyer, etc. The kinds of contents that authors may submit would be limited, e.g. to a set of predefined (markdown) document templates, with or without a header construct (as e.g. in Docusaurus). An ingestible document may contain specific 'markers' (e.g. syntax similar to `[[text that shows|<reference>]]` in WikiPedia) for 'terminological enrichments' (further details in the 'PRODUCE' section). Any submission of a certain kind (template) will be processed according to the mechanisms specified for that template, unless the input cannot be processed (invalid input). The output of such processing typically is content that is added to the CURATE state, but other outputs could be created if that is beneficial.

To be discussed:

- the kinds of document templates that authors may use to submit contents; <Daniel: We already discussed this and created a single template for submitting individual terms and the concepts/definitions associated with them. I'm happy to discuss other templates (e.g., for patterns, for , but not ones for terms and concepts. Is it okay if we leave that particular decision alone for now?>
- for each of such kinds: the criterion for determining whether or not it valid for further processing; <Daniel: see this slide for my proposal: https://docs.google.com/presentation/d/1CzJ5G6qTQ06rohPHGUj9887XtB19EK-FUMhWlmgRvy8/edit#slide=id.gac44439c49_0_23>
- the 'marker' syntaxes we would need/like to support; <Daniel: this is a new topic and an interesting one. Does Rieks have a proposal? How to cross-link to other parts of the corpus is the main thing I have been thinking about. Footnotes or citations of references would also be good to clarify.>
- the specification of such processing, i.e. what parts of the contents go where in the curate stage (or elsewhere). <Daniel: I propose to write a script that preps the content for the curate stage. My first cut of the script would embody my proposal about the answer to this question. I've been saying I would write the script for 6 weeks now, but holidays and other pressing priorities have prevented my progress. Is anybody going to beat me to it, or should I press forward.>
- ...

CURATE

is a stage that contains all sorts of typed, normalized contents. 'Normalized' means that there is no duplication of content, and all content can be identified by a single identifier. 'Typed' means that files may contain different contents, e.g. a description of a concept, the definition of a term, etc.

Every such file will be owned/curated by (one or more designated participants of) a ToIP group. <Daniel: I still am not settled on this simplification. I had been expecting that any group could "claim" a piece of content, such that multiple groups would all say, "we endorse this definition of this term" – and 'this' in that sentence would use the same identifiers for both groups. But maybe I could be convinced. I understand why it would make things easier, and prevent debates and conflict... Let's schedule a discussion about it.> Curators make decisions about the contents of a file, whether or not to update it, etc. They are responsible for the quality of the contents. They need appropriate permissions to do so. A history of changes is kept that allows older content to be referenced continuously.

All curation content (the 'Corpus') is stored in a directory, with a subdirectory for each owner (curator) <Daniel: I don't agree with this. I think ownership should be tagged, with all owned files in a single directory tree, not segmented by owner.>, in which we will see further subdirectories for the various types of contents. Each file can thus be identified by the tuple (owner,type,filename), or by `https://toip.org/ctwg/corpus/owner/type/filename` or similar. A person that is a curator for an owner (ToIP group) is assigned appropriate permissions for everything under //toip.org/ctwg/corpus/owner/. We expect to use git (hub) to manage the corpus.

The purpose of curation content is for it to be used in the PRODUCE stage, which means that the structure of each kind of file must be defined such that it is fit for any/all of the purposes that this PRODUCE stage serves. This requirement may impose constraints on the internal structure of the different kinds of files in the corpus.

The CURATE stage should provide mechanisms for obtaining different attributes of the various typed contents for the purpose of creating different renderings. A mechanism could exist that takes `(<attributeid>,<owner>,<type>,<name>,<vsn>)` and returns a text for the referenced item. For example, when <attributeid> is `definition`, the returned text is the definition of the referenced item. Different mechanisms will needed for different rendering purposes.

A curated file has a header that specifies its owner, type, name and version (there may be a mapping between 'name' and the filename of the file, e.g. filename == 'name'+'_'+'version'), e.g. as:

```
==========
owner: "ToIP CTWG" <Daniel: I think this should be a many-to-many tag, not a one-to-many owner>
type: "concept" <Daniel: why do we need this? It would be redundant with directory structure?>
name: "actor" <Daniel: why do we need this? Wouldn't the name be derived from the value of item?>
vsn: "2" <Daniel: I don't agree that this is desirable. I proposed using github versions, which are not user-friendly but which require no extra maintenance.>
==========
```

Other header entries may be defined, as needed (e.g. an entry that specifies text to be used as a popover, or as a glossary entry, or ....).

Versioning needs to be discussed <Daniel: agreed>. While from a tech perspective it is beneficial to use github commits as a version, this may not be suitable for authors that want to refer to an element in the Corpus.

PRODUCE

is a stage in which all sorts of documentary artifacts are being created/generated/... We expect to see this stage produce glossaries in various renderings (web, pdf, ...), dumps of (parts of) the Corpus (e.g. a JSON or XML file), etc. The CTWG will cater for a minimal set of artifacts to be produced, such as a ToIP Dictionary, and a default glossary of terms for every owner. Depending on the needs we encounter, other artifacts may also be produced. Specifically, we expect one or two artifacts to be produced that allow third parties to do further processing on the contents of the Corpus. While the production of such artifacts is within CTWG scope, the further processing is an external responsibility.

Every artifact that is produced must have a 'definition document', i.e. a machine processable specification that allows the artifact to be automatically created, either on demand (of a user), or as a result of a trigger firing (such as the acceptance of a pull request). We may want to ponder the idea of including a file-type "artifact definition document" as one of the accepted file-types in the Corpus.