Annotations

EDRLab

Context

- Thorium Reader will integrate annotations by the end of the year.
- The format will be based on W3C annotations, model and json-ld serialization.
- Annotations can be detached, or embedded in an EPUB (and WP package).
- We'll make annotations accessible, and shareable.
- **First step -> export / import of sets of annotations, as detached files.**
- Second step -> embedding of annotation in an EPUB or Packaged Web Publication.
- Third step -> implementation of an open Web Protocol (à la WebDav).
- We'll also study interop with Hypothesis.

Annotations

- Modeled in the W3C Annotation Model spec (3.1)
    - Has an id, type ("Annotation"), body (the information), and target (the annotated publication)
    - The body may be embedded or remote; it may also be a choice between multiple resources (?).
    - There may be alternative ways to present the target (good).

Our profile

- "created" and "modified" will be added, as extended properties.
- The body will only be plain text, and embedded in the annotation.
- "textDirection" and "language" will be usable.
- A "color" property will be added; its value will be CSS friendly (#01E3F6).

Example 1

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "urn:uuid:123-123-123-123",

  "created": "2023-09-10T15:13:28Z",
  "modified": "2023-00-12T09:00:00Z",

  "type": "Annotation",
  "body": {
    "type" : "TextualBody",
    "value" : "j'adore !",
    "format" : "text/plain",
    "color" : "#01E3F6",     # css-friendly
    "textDirection" : "ltr", # optionnal
    "language" : "fr"        # optionnal
  },
  "target": {
   …
   }
}
```

Location of the annotation

- A target is a locator inside the publication. Several will be generated.
- A "source" locating a resource in the publication. For EPUBs, it is the path to the resource, from the root of the OCF. For Web Publications, this is the absolute URL of the resource.
- meta "headings", each with a "level" and "text", which act as breadcrumbs for helping visualize the position of the annotation.
- A meta "page", indicates the page number on which the annotation is found. This will usually be a visual indicator only.
- An epub-cfi target if the publication is an EPUB, and if the annotation is on a spine item. It will then be mandatory and will present both the left-hand part (resource location) and the right-hand part (location in the resource).
- A TextQuoteSelector target, which helps locate the annotation in a fuzzy way. It will be mandatory for Web publications, and highly recommended for EPUBs.
- A DomRangeSelector (new) target, which is the way Thorium Reader (and maybe others) will locate the annotation. Optional.
- A ProgressionSelector (new) target, which represents % of progression in the resource. Optional, fuzzy.

Example 2

```
"target": {
  "source": "OEBPS/text/chapter1.html",
  "meta": {
    "headings": [
      {
        "level": 1,
        "txt": "Section 10",
      },
      {
        "level": 1,
        "txt": "Section 11",
      },
      {
        "level": 2,
        "txt": "Sub Section 1",
      }
     ],
     "page": "XI", # label or content

   },
  "selector": [
       ...
   ]
}
```

Exemple 3

```
"target": {
   "source": "OEBPS/text/chapter1.html",
   "meta": {
     ...
   },
   "selector": [
      {
       "type": "FragmentSelector",
       "conformsTo": "http://www.idpf.org/epub/linking/cfi/epub-cfi.html",
       "value": "epubcfi(/6/4!/4[body01]/10[para05]/3:/10[para05]/10)"
      }
     {
       "type": "TextQuoteSelector",
       "exact": "Combien de fois \n\n\    ne m’avait-il", #raw
      "prefix": "ouver quelqu’un    \n      comme vous. ",
      "suffix": " pas \n\n\      reproché de travailler ma"
     },
     {
      "type": "ProgressionSelector",
      "value": 0.53423425
     },
     {
      "type": "DomRangeSelector",
      "startContainerElementCssSelector": ".calibre_3",
      "startContainerChildTextNodeIndex": 0,
      "startOffset": 1066,
      "endContainerElementCssSelector": ".calibre_3",
      "endContainerChildTextNodeIndex": 0,
      "endOffset: 1095
      },
   ]
}
```

Sets of annotations

- Collections of Annotations modeled in the W3C Rec (5.1)
    - Has an id and specific type ("AnnotationCollection")
    - Can be extended with profile-specific properties.
- Burdens:
    - MUST contain pages, which MUST have an id, type, startIndex, partOf. Cumbersome.
    - MUST reference or embed the first page (= json property), and SHOULD reference the last page.
    - SHOULD contain an indication of the total number of annotations. Cumbersome.
    - The json schema we found doesn't have any metadata extensibility point, and total is required.
- Solution: create another container, a simpler "AnnotationSet". And develop a json schema for it.

Identifying the source publication

- Only for detached annotations.
- We'll use the extended metadata allowed in a collection.
- Create an "about" property, which contains the set of properties useful to identify the publication associated with the annotations.
    - dc:identifier (array of URIs), dc:format (content type), dc:title, dc:publisher, dc:creator, dc:date, dc:source (if present in the EPUB).
- It will be the task of the reading system to associate a collection of annotations with a publication, on the import of the annotation set.
- If several annotation sets are imported for the same publication, a reading system should offer a choice to the user, or even present the different sets on the same screen (but the UX would be tricky).

Identifying the generator of the annotations

- An Annotation (not AnnotationCollection) can have a "generator" and "generated" as date of generation (spec 3.3.1). We'll use them at the level of the AnnotationSet.

- "generator" will only be a set of properties defined in the spec 3.3.2 (an URI won't be allowed):

  - Properties are: id, type, name, homepage

Example

```
{
  "@context": "http://www.w3.org/ns/anno.jsonld",
  "id": "urn:uuid:123-123-123-123",
  "type": "AnnotationSet",
  "generator": {
    "id": "https://github.com/edrlab/thorium-reader/releases/tag/v2.3.0",
    "type": "Software",
    "name": "v2.3.0",
    "homepage": "https://thorium.edrlab.org"
  },
  "generated": "2023-09-01T10:00:00Z",
  "label": "Annotations Mme Prof, La Peste, cours 1ere B",
  "about": {                     # only if the annotations are detached
    "dc:identifier": [
    "urn:isbn:1234567890",
    "...",
    ],
    "dc:format": "application/epub+zip",
    "dc:title": "Alice in Wonderland",
    "dc:publisher": "Example Publisher",
    "dc:creator": "Anne O'Tater",
    "dc:date": "1865",
    "dc:source": "urn:isbn:1234567891", # if present in the EPUB
  }
  "items": [     # list of annotations
    ]
}
```

References

- Web Annotation Data Model, W3C Rec, 2017
- json schema for collections
- EPUB CFUI specification (2017)
- Open Annotations in EPUB, CFI pros and cons (2014)
- epub-cfi cannot reference content in non-spine items (2013)