University of California Curation Center

Merritt Atom-Based Submission

Rev. 0.2 – 2013-02-19

1 Introduction

It is desirable that the Merritt curation repository supports varied content submission protocols. Beyond its native protocol, as documented in [], other possibilities include SWORD and Atom. Atom is both an XML-based document syndication format [7] and an HTTP-based publishing protocol [6]. Merritt Atom-based submission can occur by translating an Atom feed into a set of conforming Merritt batch and object manifests are submitted to the Merritt Ingest micro-service.

2 Atom feed

A generic Atom feed has the following structure, where brackets "[" and "]" enclose optional elements, an ellipsis "..." indicates arbitrary repetition of the previous element, and variable strings are indicated by *underlined italics*:

```
<?xml version="1.0" encoding="utf-8"?>
<feed xmlns="http://www.w3.org/2005/Atom">
  feed-level-metadata ...
  <entry>
  [ <author>
      <name>author</name>
    </author>
    ... ]
  [ <contributor>
      <name>contributor</name>
    </contributor>
    ... ]
    <id>id</id>
  [ <link href="uri" [rel="relation"] ... />
    ... ]
  [ <published>date</published ]
    <title>title</title>
    <updated>date</updated>
  [ <content> ... </content> ]
  </entry>
```

```
  ...
</feed>
```

Feed-level metadata is not germane to Merritt submission. Each "<entry>" element is assumed to define a single coherent content object.

The "<author>" and "<contributor>" elements indicate a creator or contributor of the entry (and thus, object). The "<id>" element defines a permanent, universally unique identifier for the entry/object.

The "<link>" element defines an external reference to a web resource at the URI specified by the "href" attribute. The optional "rel" attribute indicates the link relation type. Atom defines five relations: "alternate", "enclosure", "related", "self", and "via". In addition, any relation type documented in the IANA registry may be used [2].

The "<published>" and "<updated>" elements define the publication date/times for the entry/object.

The optional "<content>" element is ignored in the context of translating an Atom feed into a Merritt submission package.

3 Atom-based submission

An Atom feed is translated into a Merritt object manifest.

1. Each "<entry>" element corresponds to a single object and a single object manifest.
2. Within the scope of a given "<entry>" element:
    a. The semi-colon-separated list of all optional "<author>" and "<contributor>" elements corresponds to the object-level ERC "who" (or creator) element. If no author or contributors are defined, set "who" to "(:unas)".
    b. The required <title> element corresponds to the object-level ERC "what" (title) element. If no title is defined, set "what" to "(:unas)".
    c. The required "<id>" element corresponds to the object-level ERC "where" (local identifier) element.

d. The required “<updated>” element corresponds to the object-level ERC “when” element. Atom dates are specified in the Internet timestamp format [3]:

    yyyy-*mm*-*dd*T*mm*:*ss*(Z|±*hh*:*mm*)

    The date value should be truncated to year/month/day granularity, if necessary.

e. The four ERC elements, “who”, “what”, “when”, “where”, are written to the “mrt-erc.txt” file, included as an object file-level component:

    ```
    erc:
    who: who
    what: what
    when: when
    where: where
    ```

f. Each optional “<link>” element defines it’s source content as an object file-level component and corresponds to a single file (and line) in the object manifest.

g. Within the scope of a given “<link>” element:

    i. If the “href” attribute is defined; and

    ii. The “rel” attribute is defined and is one of:

        1. “alternate”, or
        2. “archival”, or
        3. “enclosure”, or
        4. “http://purl.org/dc/terms/hasPart”; and

    iii. The URL is successfully resolvable, possibly after one or more 3xx redirects, then

    include the HTTP response body as an object file–level component with the filename defined by the filename, possibly without an extension, at the end of the URL path:

    http://*domain*[:*port*]/*path*/.../*filename*[.*ext*]

replacing any embedded escaped spaces,"%20", with underscores, "_". An additional suffix may need to be added to the filename to ensure the uniqueness of all filenames that comprise a given object.

filename[_*suffix*][.*ext*]

where "suffix" is drawn from a numeric sequence reinitialized to "2" for each entry/object and is separated from the filename by an underscore "_".

4 Implementation

The implementation must maintain a list of all filenames extracted from <link> elements for a given entry/object. If a filename/extension combination has already been used, a numeric suffix must be added to the filename. For example, it is common for multiple images to be specified for a given object. While these may distinguished in the path portion of the URL, they may share filenames:

```
.../meta/abcd
.../thumb/abcd.jpg
.../lowres/abcd.jpg
.../hires/abcd.jpg
```

The filenames used for the Merritt submission would be:

```
abcd
abcd.jpg
abcd_2.jpg
abcd_3.jpg
```

A new numeric sequence starting with "2" is established for each entry/object.

The implementation should support a configurable delay.

The implementation should support incremental submission. Since any new additions to the Atom feed will appear at the top of the feed, the translation code should check the Inventory micro-service for the local ID and timestamp to see if the object has already been submitted, i.e., the ERC "where" and "when" elements. New entries/ objects are processed (and submitted to Merritt) from the Atom feed until an entry is found that matches an object local ID and timestamp in the Inventory service.

References

[1] Gregorio, J. C., and B. de Hora, eds. (2007), *The Atom Publishing Protocol*, RFC 5023 <http://www.ietf.org/rfc/rfc5023.txt>.

[2] IANA (2013), *Atom Link Relations* <http://www.iana.org/assignments/link-relations>.

[3] Klyne, G., and C. Newman (2002), *Date and Time on the Internet: Timestamps*, RFC 3339 <http://www.ietf.org/rfc/rfc3339.txt>.

[4] Kunze, J., and A. Turner (2010), *Kernel Metadata and Electronic Resource Citations (ERCs)* <http://dublincore.org/groups/kernel/spec/>.

[5] Nottingham, M. (2007), *Feed Paging and Archiving*, RFC 5005 <http://www.ietf.org/rfc/rfc5005.txt>.

[6] Nottingham, M., and R. Sayre, eds. (2005), *The Atom Syndication Format*, RFC 4287 <http://www.ietf.org/rfc/rfc4287.txt>.

[7] UC Curation Center (2012), *Merritt Ingest Service* <https://confluence.ucop.edu/display/Curation/Ingest>.