Infrastructure - Story #8239

Additional formatIDs may be needed to support ADC content description

2018-01-09 14:49 - Dave Vieglais

Status:	Closed	**Start date:**	2018-01-09
Priority:	Normal	**Due date:**	
Assignee:	Jing Tao	**% Done:**	100%
Category:	Format ID	**Estimated time:**	0.00 hour
Target version:	CCI-2.3.10		
Story Points:	5		

Description

From Bryce via slack:

> Our team could use a new format ID or two. I was planning to file a ticket in redmine but I also wanted to know what the most productive way would be to get it on track for a CCI release. They need a bit of discussion, some level of agreement, etc.

Associated revisions

Revision 19080 - 2018-02-01 06:41 - Chris Jones

Add new formats for:
- Google Earth Keyhole Markup Language (KML) Compressed archive
- Mathworks MATLAB version 7.3 (R2006b or later) binary file - HDF5 compatible
- Mathworks MATLAB version 7 (R14 or later) binary file
- Mathworks MATLAB version 6 (R8 or later) binary file
- Mathworks MATLAB version 4 binary file
- JavaScript Object Notation (JSON) file
- JavaScript Object Notation (JSON) Linked Data file
- Markdown file (pandoc markdown compatible)
- R Markdown file (pandoc markdown compatible)
- RAW digital sensor image file

refs #8239

Revision 19080 - 2018-02-01 06:41 - Chris Jones

Add new formats for:
- Google Earth Keyhole Markup Language (KML) Compressed archive
- Mathworks MATLAB version 7.3 (R2006b or later) binary file - HDF5 compatible
- Mathworks MATLAB version 7 (R14 or later) binary file
- Mathworks MATLAB version 6 (R8 or later) binary file
- Mathworks MATLAB version 4 binary file
- JavaScript Object Notation (JSON) file
- JavaScript Object Notation (JSON) Linked Data file
- Markdown file (pandoc markdown compatible)
- R Markdown file (pandoc markdown compatible)
- RAW digital sensor image file

refs #8239

History

#1 - 2018-01-09 14:49 - Dave Vieglais

- Category set to Format ID

#2 - 2018-01-10 01:14 - Bryce Mecum

- Status changed from New to In Progress

- % Done changed from 0 to 30

From time to time on urn:node:ARCTIC we come across a file format not listed on the registered DataONE formats. It's not strictly required that we give each of these file formats its own DataONE format but, in an ideal world, we would like to have more specific formatIds than text/plain and application/octet-stream for objects with these file formats.

I have some high-level questions:

- Do file formats need to be sufficiently popular for us to include them in DataONE's list? Related: Do we care of the list of formats gets really long?
- What do we do about file formats that haven't gone through standardization, but optionally may, and thus, there media type is (1) non-standard and (2) subject to change? A good example is Markdown which was text/x-markdown and is now text/markdown (https://tools.ietf.org/html/rfc7763)
- How much do we care about open vs. proprietary formats? Do we want to avoid relatively obscure, proprietary formats? Can we come up with a good heuristic so as to reduce subjectivity here?

Here is the most up-to-date list of formats we could use:

KMZ

This is the zipped form of KMLs.

Reference: https://developers.google.com/kml/documentation/kmzarchives

- Should we just format them as application/vnd.google-earth.kml+xml or introduce another format? Or just use application/zip?
- I think we already discussed this one in another Redmine ticket. Haven't looked it up yet.

Proposal:

Type: DATA
Id: application/vnd.google-earth.kml+xml application/vnd.google-earth.kmz
Name: Google Earth Keyhole Markup Language (KML) archive (KMZ)
Media type: application/vnd.google-earth.kml+xml application/vnd.google-earth.kmz

RAW

This is a common file cameras produce but it is many formats. The closest thing is the DNG format which is highly related to the ISO format https://en.wikipedia.org/wiki/TIFF/EP which is not the same as a regular TIFF file but does often carry the same file extension.

- We could, instead, add format IDs for some or every RAW image file format. I don't like this option
- I see a potential problem in that the user may not actually know what file format their RAW file is in so they may end up providing overly-specific, incorrect information
- I'm overall not a big fan of adding this as a DataONE format

Proposal:

Type: DATA
Id: image/raw
Name: RAW image file formats (various)
Media type: There isn't one single good option

JSON-LD

JSON-LD is a JSON document with special data contained within it. The file extension is .json and any JSON parser/serializer can work with them.

Proposal:

Type: DATA
Id: application/ld+json
Name: JSON-LD
Media type: application/ld+json

RMarkdown

RMarkdown is a relatively new Markdown-like file format that adds some extra syntax around code blocks to help RMarkdown processors run the code inside the chunks.

- They're basically both text and executable.
- There isn't an official media type
- Jupyter notebooks appear to use application/vnd.jupyter (http://jupyter.readthedocs.io/en/latest/reference/mimetype.html) or some related media type

Proposal:

Type: DATA
Id: text/markdown | text/x-rmarkdown | application/x-rmarkdown | application/vnd.rmarkdown
Name: RAW image file formats (various)
Media type: text/markdown

MATLAB .mat files (binary)

Reference: https://www.mathworks.com/help/pdf_doc/matlab/matfile_format.pdf

- There are multiple versions of these files
- It appears as though older versions can still be read by newer versions so maybe it's not super important to have different format IDs for each version
- I think I've heard these files are highly related to HDF but I can't find a good ref right now

Proposal:

Type: DATA
Id: Not sure on this
Name: MATLAB data file
Media type: application/octet-stream

I'm still working on notes on these:

PROfile

RDI

GEOS-Chem

#3 - 2018-01-26 00:33 - Bryce Mecum

JSON

I think we all know what JSON is.

Proposal:

Type: DATA
Id: application/json
Name: JSON
Media type: application/json

#4 - 2018-01-31 19:41 - Bryce Mecum

Name for RMarkdown should be

Name: RMarkdown

(not what I put)

#5 - 2018-02-01 06:41 - Chris Jones

Hi Bryce, Jing,

I've added the formats to the objectFormatListV2.xml in the trunk of cn-buildout.

I changed image/raw to image/x-raw because the IANA has an image/raw listing but it points to the video/raw RFC on streaming raw uncompressed video over RTP, so I didn't want to confuse them. Otherwise, most of the rest are the same or similar to what Bryce proposed. For the RDI, PROfile, and GEOS-Chem formats, we need more discussion. Let's do that in another ticket.

Jing, will you please merge this file into the latest branch for the next CCI release, and also copy it over to the d1_libclient_java library where we cache a copy? I updated the total to 128, so we should be good. Also, will you then update the each CN environment with the new formats list? We should be able to close this then. Thanks!

Here they are:

```
<objectFormat>
    <formatId>application/vnd.google-earth.kmz</formatId>
    <formatName>Google Earth Keyhole Markup Language (KML) Compressed archive</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/vnd.google-earth.kmz"/>
    <extension>kmz</extension>
</objectFormat>
<objectFormat>
    <formatId>application/MATLAB-v7.3</formatId>
    <formatName>Mathworks MATLAB version 7.3 (R2006b or later) binary file - HDF5 compatible</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/octet-stream"/>
    <extension>mat</extension>
</objectFormat>
<objectFormat>
    <formatId>application/MATLAB-v7</formatId>
    <formatName>Mathworks MATLAB version 7 (R14 or later) binary file</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/octet-stream"/>
    <extension>mat</extension>
</objectFormat>
<objectFormat>
    <formatId>application/MATLAB-v6</formatId>
```

```
    <formatName>Mathworks MATLAB version 6 (R8 or later) binary file</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/octet-stream"/>
    <extension>mat</extension>
</objectFormat>
<objectFormat>
    <formatId>application/MATLAB-v4</formatId>
    <formatName>Mathworks MATLAB version 4 binary file</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/octet-stream"/>
    <extension>mat</extension>
</objectFormat>
<objectFormat>
    <formatId>application/json</formatId>
    <formatName>JavaScript Object Notation (JSON) file</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/json"/>
    <extension>json</extension>
</objectFormat>
<objectFormat>
    <formatId>application/json-ld</formatId>
    <formatName>JavaScript Object Notation (JSON) Linked Data file</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/ld+json"/>
    <extension>json</extension>
</objectFormat>
<objectFormat>
    <formatId>text/markdown</formatId>
    <formatName>Markdown file (pandoc markdown compatible)</formatName>
    <formatType>DATA</formatType>
    <mediaType name="text/markdown"/>
    <extension>md</extension>
</objectFormat>
<objectFormat>
    <formatId>text/x-rmarkdown</formatId>
    <formatName>R Markdown file (pandoc markdown compatible)</formatName>
    <formatType>DATA</formatType>
    <mediaType name="text/markdown"/>
    <extension>Rmd</extension>
</objectFormat>
<objectFormat>
    <formatId>image/x-raw</formatId>
    <formatName>RAW digital sensor image file</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/octet-stream"/>
    <extension>raw</extension>
</objectFormat>
```

#6 - 2018-02-01 06:42 - Chris Jones

- Assignee set to Jing Tao

#7 - 2018-05-04 01:24 - Matthew Jones

Without delving into checking whether the mime types are right, it looks good to me. All valuable additions.

#8 - 2018-05-08 23:18 - Jing Tao

- Status changed from In Progress to Closed

- Target version set to CCI-2.3.10

- % Done changed from 30 to 100

Those ids were deployed to cn-sandbox/stage/production. Close this ticket.
I also create a new ticket:

https://redmine.dataone.org/issues/8586

to handle another three format ids:
PROfile
RDI
GEOS-Chem