Infrastructure - Task #8820

Add new DataONE Object format for HDF4/5 file formats

2019-06-13 19:54 - Bryce Mecum

Status:	Closed	**Start date:**	2019-06-13
Priority:	Normal	**Due date:**	
Assignee:	Jing Tao	**% Done:**	100%
Category:		**Estimated time:**	0.00 hour
Target version:			
Milestone:	None	**Story Points:**	
Product Version:	*		

Description

HDF 4 and 5 are efficient binary formats for data commonly used in science: https://en.wikipedia.org/wiki/Hierarchical_Data_Format, https://www.hdfgroup.org/solutions/hdf5/. I don't think we have a lot of content in this format, if any, but it's a pretty common format and a good one at that.

I did some research on MIME types and file extensions:

Re: MIME type:

> The recommended content is application/x-hdf5 for data in HDF5 or application/x-hdf for data in earlier versions.

https://www.hdfgroup.org/2018/06/citations-for-hdf-data-and-software/

Re: Extension,

- https://en.wikipedia.org/wiki/Hierarchical_Data_Format lists a few of the variants I've seen
- The R hdf5 package uses the h5 extension (https://github.com/grimbough/rhdf5/tree/master/inst/testfiles) so I went with this
- .hdf4 and .hdf5 seem common too but h4/h5 seems a tad more common

Here are the details for each of the new formats:

HDF4

- formatId: application/x-hdf
- formatName: Hierarchical Data Format version 4 (HDF4)
- mediaType: application/octet-stream
- extension: h4

HDF5

- formatId: application/x-hdf5
- formatName: Hierarchical Data Format version 5 (HDF5)
- mediaType: application/octet-stream
- extension: h5

History

#1 - 2019-06-13 22:39 - Matthew Jones

This looks great. I would suggest that the proper media types should be application/x-hdf annd application/x-hdf5 rather than application/octet-stream. Best to be specific. Thoughts?

#2 - 2019-06-13 22:49 - Bryce Mecum

> This looks great. I would suggest that the proper media types should be application/x-hdf annd application/x-hdf5 rather than

application/octet-stream. Best to be specific. Thoughts?

I thought this too, but based my decision off of how other formats were done. e.g., the MATLAB formats (including 7.3 which is HDF5) and RAW are typed as octet-stream. I'm cool with either way. And now that I look, I think your suggestion is probably the better one.

Let's go with with mediaType values that match the formatId values.

#3 - 2019-06-17 16:12 - Jing Tao

They will look like:

```
<objectFormat>
    <formatId>application/x-hdf</formatId>
    <formatName>Hierarchical Data Format version 4 (HDF4)</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/x-hdf"/>
    <extension>h4</extension>
</objectFormat>

<objectFormat>
    <formatId>application/x-hdf5</formatId>
    <formatName>Hierarchical Data Format version 5 (HDF5)</formatName>
    <formatType>DATA</formatType>
    <mediaType name="application/x-hdf5"/>
    <extension>h5</extension>
</objectFormat>
```

#4 - 2019-06-17 18:18 - Bryce Mecum

Looks good to me.

#5 - 2019-06-17 21:20 - Jing Tao

- % Done changed from 0 to 100

- Status changed from New to Closed

The formats have been added to the code (dataone-cn-metacat and d1_common_java). The formats also were added to cn-dev, cn-dev-2, cn-sandbox, cn-sandbox-2, cn-stage, cn-stage-2 and production cns.