# Identification of Science Resources & Tool for Extracting Standard Metadata Properties

**Pratik Shrivastava[1], Dave Vieglais[2]**
[1]University of Illinois at Urbana-Champaign, [2]DataONE

## Introduction

- Identifying the correct file format is imperative for processing its contents.
- Many metadata standards are serialized as XML requires additional details of namespace information for processing.
- Packaging data into data packages requires metadata identification and parsing of the files.
- A tool for reliable identification makes processing easier.

## Aim

- Determine the scientific resources using the Linux file command and Apache Tika which are excellent tools for file format identification.
- Use Apache Tika for parsing the metadata contents of the resources.
- Extraction of standard set of properties from the metadata.

## File Command

- File command performs several additional tests for determining the file format instead of using the file extensions.
- Uses the format signatures, known as magic numbers for identifying the file format.
- The magic directory contains the files, these files consist of the magic numbers. File command uses a compiled binary file containing the magic files.

## Apache Tika

- It is an open source toolkit for detecting and extracting metadata and contents of the files.
- Its ability to detect and parse file formats from over a 1000 different formats makes it a useful tool for search engine indexing, content analysis, translation etc.
- The new file types can be detected by creating a custom XML file containing the information.
- New parsers can be easily created and integrated into the application for fresh file formats.

## DataONE Magic file

- Gathered a Test corpus for the known DataONE file formats.
- Define rules for DataONE file format Identification.
- Create Magic files for identifying DataONE file formats.

```
# EML (Ecological Metadata Language Format)
0 string <?xml
>&0 regex (eml)-[0-9].[0-9].[0-9]+
formatid="eml://ecoinformatics.org/%s”

# onedcx (DataONE Dublin Core Extended v1.0)
>&0 regex (onedcx/v)[0-9].[0-9]+
formatid="http://ns.dataone.org/metadata/schema/%s"

# ISOTC211 (Geographic MetaData (GMD) Extensible Markup
Language)
>&0 regex  isotc211
>>&0 regex eng;USA formatid=http://www.isotc211.org/2005/gmd
```

- Compile magic files for the libmagic library of the file command.
- Tested magic file using unittest library in python.

```
$ file -m dataone.mgc 00_eml-211.xml
00_eml-211.xml: formatid="eml://ecoinformatics.org/eml-2.1.1"
```

## Custom File Detector using Tika

- Create custom-mimetypes.xml and a jar file for identifying new file format .
- The xml supports magic numbers for file Identification.
- Tika app with custom-mimtypes.jar is used for file detection.
- It uses regex for matching patterns defined in value attribute.

```
<mime-type
  type="text/xml;formatid=eml://ecoinformatics.org/eml-2.0.0">
  <magic priority="60">
    <match value="eml://ecoinformatics.org/eml-2.0.0”
           type="string” offset="50:1000"/>
  </magic>
</mime-type>
<mime-type
  type='text/xml;formatid=http://www.isotc211.org/2005/gmd-
noaa’>
  <magic priority="75">
    <match value="gov.noaa.nodc" type="string"
offset="50:1000"/>
  </magic>
</mime-type>
<mime-type
type='application/rdf+xml;formatid="http://www.openarchives.org/
ore/terms"'>
  <magic priority="75">
    <match value="openarchives.org/ore" type="string"/>
  </magic>
</mime-type>
```

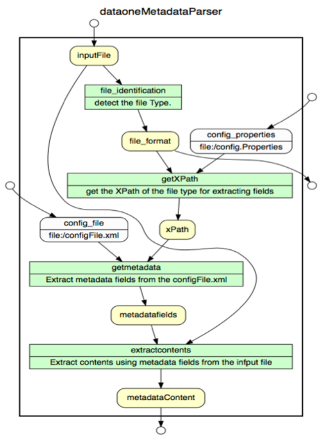
## DataONE Metadata Extraction Tool Using Tika

- A command line tool for detecting and parsing  standard metadata properties for science resources.

```
---------------------------------------------------------------------------
FILE FORMAT :text/xml; formatid="http://www.isotc211.org/2005/gmd"
---------------------------------------------------------------------------
Title: Herring Infection Prevalence Data, 2007-2016, EVOS Herring Program
Creator: Paul Hershberger; Paul Hershberger; Paul Hershberger; Paul Hershberger
Subject: Clupea pallasii; Pacific herring, arenque del Pacífico; Prince William
Sound; Exxon Valdez; oil spill; Exxon Valdez Oil Spill Trustee Council; EVOSTC; EVOS
Herring Survey; EVOS Herring; EVOS Herring Research and Monitoring; infection; viral
hemorrhagic Publisher: Carol Janzen
Publisher: Alaska Ocean Observing System
Date: 20170101
```

- It uses Tika detector for identification of the file type.
- It is a custom namespace aware parsers for extraction of the metadata content from different file formats
- Uses a configuration file for extracting the metadata properties either by specifying the element tag or using an XPath values.

```
<FileFormat name="isotc211-gmd-noaa">
     <namespaces>
       <namespace prefix="gco" uri="http://www.isotc211.org/2005/gco" />
       <namespace prefix="gmd" uri="http://www.isotc211.org/2005/gmd" />
       <namespace prefix="gmi" uri="http://www.isotc211.org/2005/gmi"/>
     </namespaces>
     <metadataFields>
       <field>
         //gmi:MI_Metadata/gmd:contact
       </field>
     </metadataFields>
 </FileFormat>
```



```
 ## XPath for DublinCore.
http\://ns.dataone.org/metadata/schema/onedcx/v1.0=//FileFormats/FileFormat[@name='onedcx']/*

## XPath for rdf
http\://www.openarchives.org/ore/terms=//FileFormats/FileFormat[@name='openarchives']/*

## XPath for Mercury
http\://purl.org/ornl/schema/mercury/terms/v1.0=//FileFormats/FileFormat[@name='mercury']/*

## XPath for the datadryad
http\://datadryad.org/profile/v3.1=//FileFormats/FileFormat[@name='datadryad’]/*

## XPath for the FGDC
FGDC-STD-001-1998=//FileFormats/FileFormat[@name='fgdc-001-1998']/*

## XPath for isotc211
http\://www.isotc211.org/2005/gmd=//FileFormats/FileFormat[@name='isotc211']/*
```

## Conclusion:

- Successful identification of file format using Libmagic and Apache Tika
- Easy to add support for new file metadata properties for file extraction using the configuration file.
- The output can be exported to JSON, CSV format.
- Useful in searching and indexing metadata content.

## References

- http://tika.apache.org
- https://github.com/apache/tika
- https://github.com/file/file
- http://openpreservation.org/blog/2012/08/09/magic-editing-and-creation-primer
- https://linux.die.net/man/1/file
- https://filemagic.readthedocs.io/en/latest/guide.html

## Github

- https://github.com/DataONEorg/dataone-tika-parser
- https://github.com/DataONEorg/file_identification

## Acknowledgments

Supported by NSF under Grant Numbers 0830944 and 1430508.

DataONE