Identification of Science Resources & Tool for Extracting Standard Metadata Properties

Pratik Shrivastava[1], Dave Vieglais[2]

[1]University of Illinois at Urbana-Champaign, 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 it 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

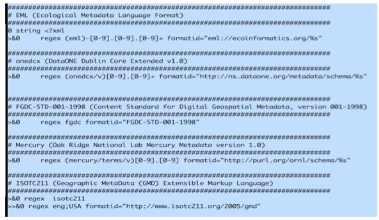
- File command performs several additional tests for determining the file format instead of using the file extensions.
- Uses the format signatures, known as magic numbers for identifying the file format.
- The magic directory contains the files, these files consist of the magic numbers. File command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

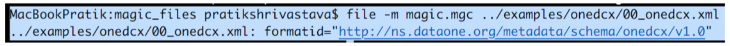
DataONE Magic file

- 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"


##########################################################################
# FGDC-STD-001-1998 (Content Standard for Digital Geospatial Metadata, version 001-1998)
##########################################################################
>&0     regex fgdc formatid="FGDC-STD-001-1998"

##########################################################################
# Mercury (Oak Ridge National Lab Mercury Metadata version 1.0)
##########################################################################
>&0     regex (mercury/terms/v)[0-9].[0-9] formatid="http://purl.org/ornl/schema/%s"

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

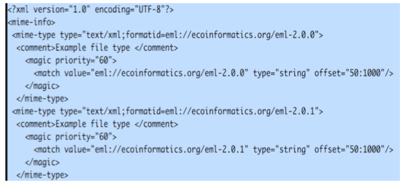
- Compile magic files for the libmagic library of the file command.
- Tested the magic file using unittest library in python.



```
MacBookPratik:magic_files pratikshrivastava$ file -m magic.mgc ../examples/onedcx/00_onedcx.xml
../examples/onedcx/00_onedcx.xml: formatid="http://ns.dataone.org/metadata/schema/onedcx/v1.0"
```

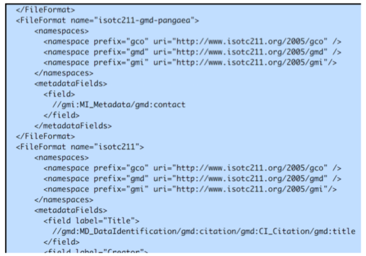
Custom File Detector using Tika

- 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.



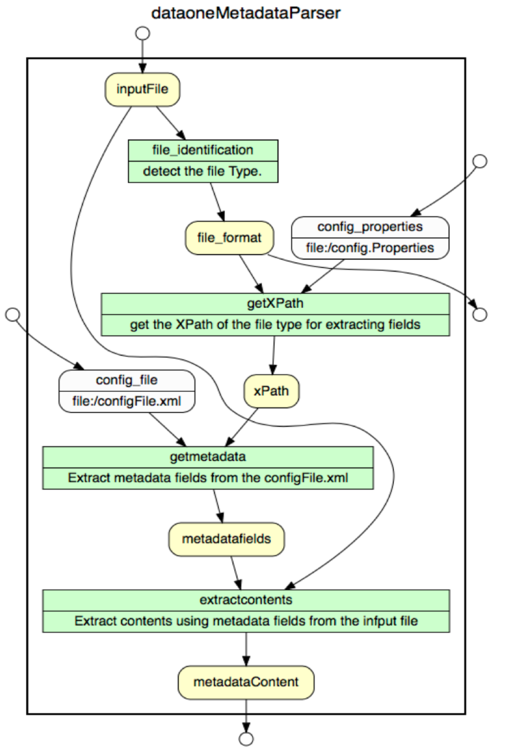
```
<?xml version="1.0" encoding="UTF-8"?>
<mime-info>
 <mime-type type="text/xml;formatid=eml://ecoinformatics.org/eml-2.0.0">
  <comment>Example file type </comment>
    <magic priority="60">
      <match value="eml://ecoinformatics.org/eml-2.0.0" type="string" offset="50:1000"/>
    </magic>
  </mime-type>
 <mime-type type="text/xml;formatid=eml://ecoinformatics.org/eml-2.0.1">
  <comment>Example file type </comment>
    <magic priority="60">
      <match value="eml://ecoinformatics.org/eml-2.0.1" type="string" offset="50:1000"/>
    </magic>
  </mime-type>
```

DataONE Metadata Extraction Tool



```
</FileFormat>
<FileFormat name="isotc211-gmd-pangaea">
    <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>
<FileFormat name="isotc211">
    <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 label="Title">
        //gmd:MD_DataIdentification/gmd:citation/gmd:CI_Citation/gmd:title
      </field>
```

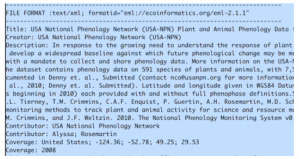
- A configurable command line tool for extracting standard metadata properties for science resources.
- It uses custom detector for identification of the file type.
- It is a custom namespace aware parsers for extraction of the metadata content from different file formats.
- Uses a configuration file for extracting the metadata properties from a science resource.



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

Results / Conclusion:



```
--------------------------------------------------------------------------------
FILE FORMAT :text/xml; formatid="eml://ecoinformatics.org/eml-2.1.1"
--------------------------------------------------------------------------------
Title: USA National Phenology Network (USA-NPN) Plant and Animal Phenology Data
Creator: USA National Phenology Network (USA-NPN)
Description: In response to the growing need to understand the response of plant
 develop a widespread baseline against which future phenological change may be me
with a mandate to collect and share phenology data. More information on the USA-
he dataset contains phenology data on 591 species of plants and animals, with 7,
cumented in Denny et. al., Submitted (contact nco@usanpn.org for more informatio
 al., 2010; Denny et. al. Submitted). Latitude and longitude given in WGS84 Datu
s beginning in 2010) each provided with and without full phenophase definitions.
.L. Tierney, T.M. Crimmins, C.A.F. Enquist, P. Guertin, A.H. Rosemartin, M.D. Sc
monitoring methods to track plant and animal activity for science and resource m
M. Crimmins, and J.F. Weltzin. 2010. The National Phenology Monitoring System v0
Contributor: USA National Phenology Network
Contributor: Alyssa; Rosemartin
Coverage: United States; -124.36; -52.78; 49.25; 29.53
Coverage: 2008
```

- An easily configurable tool for adding new file format.
- Easier to add and remove metadata properties for file extraction.
- The output can be exported to JSON, CSV format.
- Highly usable in searching and indexing metadata contents.

Acknowledgments

Supported by NSF under Grant Numbers 0830944 and 1430508.

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 it 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.
- It 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.

Method

- Create Magic files for identifying DataONE file formats.
- Gathered a Test corpus for the known DataONE file formats.
- Define rules for DataONE file format Identification.
- Compile magic files for the Libmagic library used by the file command.
- Create custom-mimetypes.xml file for identification using Tika.
- It uses magic numbers as well for identification.
- Tika performs detection of the file type and based on that uses parsers for metadata extraction.
- Created custom namespace aware parsers for extraction of the metadata content from different file formats.
- Created a command line application based on Tika, which uses a configuration file for extracting standard set of metadata fields based on the file type.
- It takes file as input and identifies the file format and extract metadata properties.

Results

- Successful identification of the file types using Libmagic and Apache Tika.
- Used Python for unittest and the latest file version will contain the changes.
- Configurable command line application for detecting file types.
- Configurable tool for extracting desired set of metadata fields from the input file type.
- Configuration file helps in addition of new file formats and the respective metadata fields for extraction.