Understanding what you have

Content profiling with C3PO

Artur Kulmukhametov, Carl Wilson, Petar Petrov
Vienna University of Technology, Open Planets Foundation

Advanced Practitioner Course
Glasgow, 17 July 2013
Agenda

- Collection scale and heterogeneity
- An approach to getting control
- Characterisation tools
- C3PO, a tool for content profiling

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Heterogeneity of Data

- Personal
- Cultural Heritage
- Scientific Data
- Government Documents
- ... a huge variety of formats and information

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
What is It?

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Large Synoptic Survey Telescope

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

30 Terabytes of data nightly
What Happens in an Internet Minute?

- 20 New victims of identity theft
- 204 million Emails sent
- 47,000 App downloads
- 61,141 Hours of music
- 135 Botnet infections
- 277,000 Logins
- 30 Hours of video uploaded
- 1.3 million Video views
- 100,000 New Twitter accounts
- 320+ New Twitter accounts
- 100,000 New tweets
- 83,000 In sales
- 6,000 Photo views
- 3,000 Photo uploads
- 639,800 GB of global IP data transferred

And Future Growth is Staggering

Today, the number of networked devices = the global population
By 2015, the number of networked devices = 2x the global population
In 2015, it would take you 5 years to view all video crossing IP networks each second

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Conclusions?

..... that’s a lot of data ......
Do we know what it is?
Do we need to preserve it?
all of it??
Characterization

<table>
<thead>
<tr>
<th>Property</th>
<th>File A</th>
<th>File B</th>
<th>File C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>PDF 1.2</td>
<td>PDF 1.2</td>
<td>PDF 1.4</td>
</tr>
</tbody>
</table>

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
### Characterization

<table>
<thead>
<tr>
<th>Property</th>
<th>File A</th>
<th>File B</th>
<th>File C</th>
</tr>
</thead>
<tbody>
<tr>
<td>Format</td>
<td>PDF 1.2</td>
<td>PDF 1.2</td>
<td>PDF 1.4</td>
</tr>
<tr>
<td>Page Count</td>
<td>20</td>
<td>20.000</td>
<td>40</td>
</tr>
<tr>
<td>Encryption</td>
<td>Yes</td>
<td>No</td>
<td>Yes</td>
</tr>
<tr>
<td>File Size</td>
<td>1 MB</td>
<td>120 MB</td>
<td>2 MB</td>
</tr>
<tr>
<td>Valid</td>
<td>No</td>
<td>Yes</td>
<td>No</td>
</tr>
<tr>
<td>Well-formed</td>
<td>Yes</td>
<td>Yes</td>
<td>Yes</td>
</tr>
</tbody>
</table>
One size does not fit all
Scalability

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Many Collections, One View

- Global View of Content
  - distribution of file formats
  - distribution of characteristics

- Three Stages
  - collect metadata
  - combine and filter
  - analyse and reason

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Representative Samples

- Based upon metadata
- As few as possible, as many as necessary
- Stratification across file type, size, time or any other relevant characteristic for the use case
- Outliers identification

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Tools for Characterization

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
A few Problems...

- A lot of tools to manage and invoke
- Different output schemas
- Different configuration/environments
- No or bad higher level management
- Difficult to spot differences
• FITS identifies, validates, and extracts technical metadata for various file formats
• By Harvard University Library in 2009
• v0.6.2, LGPL
• Wraps other tools
• New version every 6-12 months

http://code.google.com/p/fits/
Main features:
• Consolidates output
• Can include raw output
• Configurable/Extendable

FITS includes:
• Droid
• Metadata Extra
• Jhove
• Exiftool
• FFident
• File Utility

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
FITS Output

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
FITS Output Conflict

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Advantages

• Only one output schema
• Basic QA hints
• Better type coverage (although...)

Disadvantages

• Consolidation is hard
• Performance and Scalability
• Conflicts

Yet Another?

This work was partially supported by the SCAPE Project. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
Clever, Crafty Content Profiling of Objects

C3PO is a tool for content profile generation.

Features:

• Uses characterization results,
• Deeper content analysis with nice visuals through the web-app,
• Generates content profiles (map/reduce)

Sometimes, I don’t understand human behavior?!

https://github.com/openplanets/c3po
Clever, Crafty Content Profiling of Objects

- **CLI-app**
  - Java
  - Parses and processes FITS files
  - Stores them in MongoDB
  - XML Profile + CSV

- **Web-app**
  - Play Framework
  - Overview and Browsing
  - Filtering
  - Representative Sample Set Generation

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
C3PO: Representative Samples

- **Size'o'Matic 3000**
  - The Size'o'Matic 3000 selects the smallest and the largest objects and fills the rest of the representative set with random objects near to the average objects size.

- **SysSampler**
  - This algorithm implements a common statistical approach, called systematic sampling. It divides the collection in bins and selects one element per bin at random. All elements have equal probability to be chosen.

- **DistSampler**
  - The distribution sampling algorithm takes a small number of properties as input and selects sample objects that together have (nearly) the same distribution as the whole collection or filter. Note, that if you select too many properties, or a special combination of properties, it could happen, that no representatives can be found.

---

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
C3PO: Performance

- Govdocs1
  - 945699 objects - 1h 48m
  - 112 different object properties
  - profile - 12 minutes
- Web Archive Data
  - 958638 objects - 2h 58m
  - 105 different object properties
  - profile - 13.5 minutes
- A single PC with 4GB RAM and 2.3GHz 2-core CPU.

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
C3PO: Performance

- SB (Denmark) dataset - 12 TB, 440M FITS files
- Test case 1 – Import
  - Linear ingest time of 0.65 ms for FITS file
- Test case 2 – GUI
  - 2.5 million FITS files limit
- Generate profile in command-line
  - 15 hours for 12M files

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).
C3PO: Challenges

- Conflict reduction
- Use the ontology from Planning and Watch for an alignment with other tools
- Generate a profile from SB dataset (0.4 billion FITS-files)
• Characterization is time consuming
• It can be faulty
• Know your tools
• A tool for content profiling? C3PO!

Summary

This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).