Tools for Characterisation

Deploying tools and understanding the results

Carl Wilson

Open Planets Foundation

File Format Day of Action
Wellcome Trust Centre, 2013-01-28

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).

Who Am I?

Carl Wilson

Open Planets Foundation

www: http://openplanetsfoundation.org

email: carl@openplanetsfoundation.org

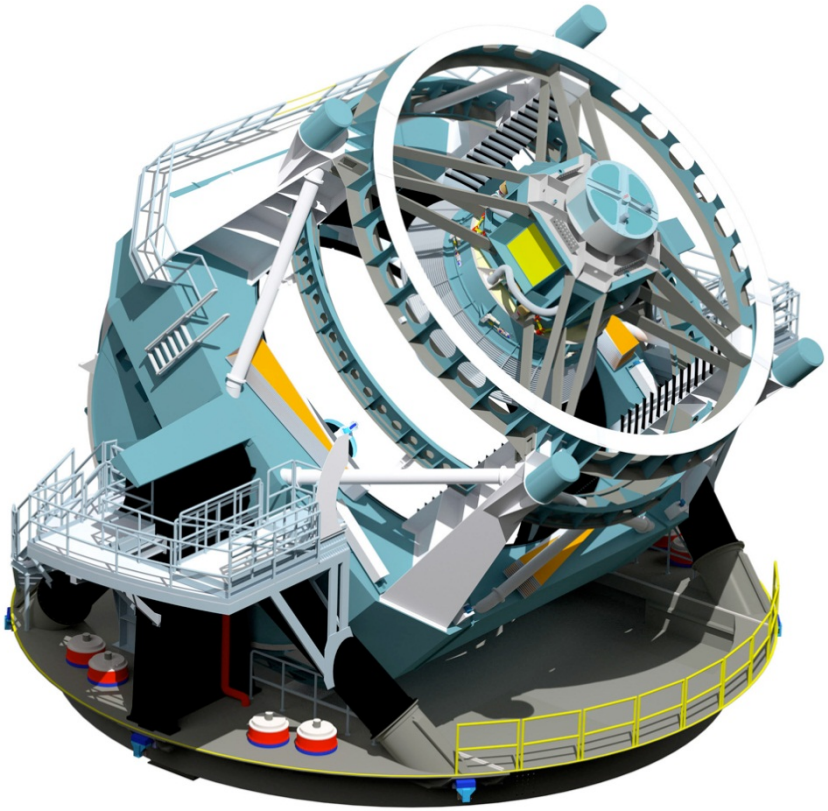
Agenda

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

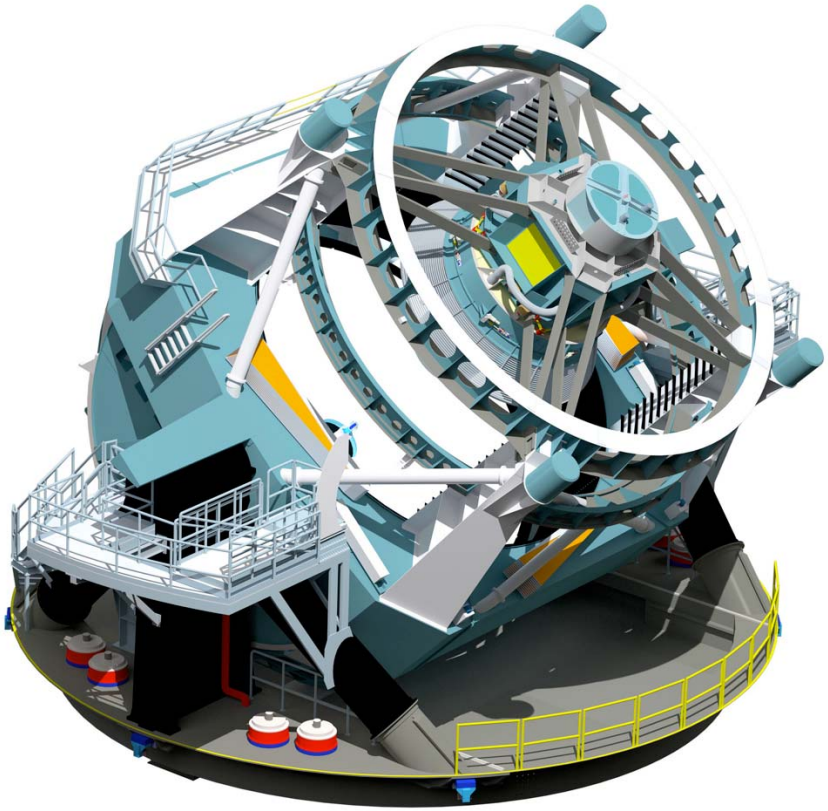
Heterogeneous Content

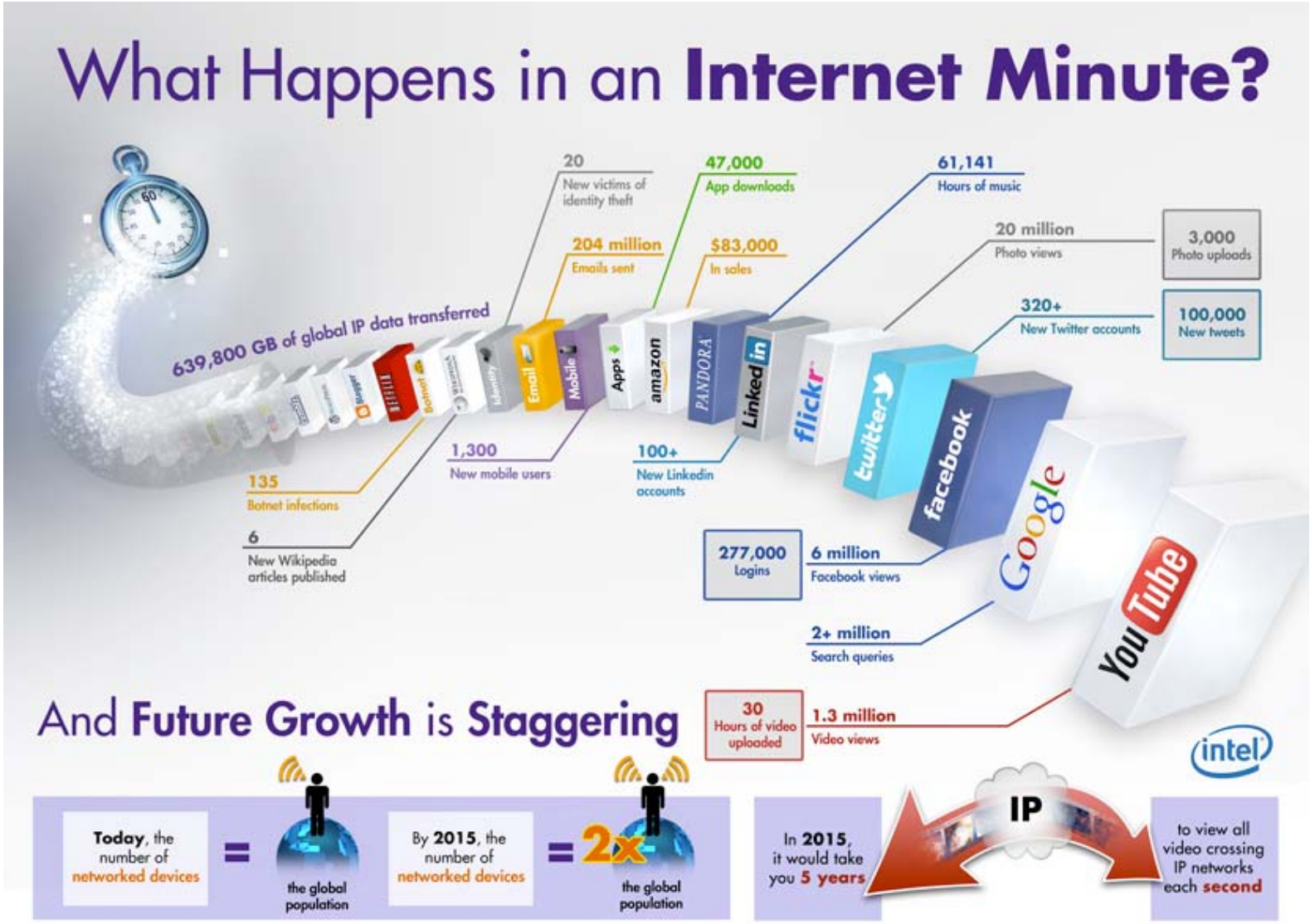
- Personal
- Cultural Heritage
- Scientific Data
- Government Documents
- …. a huge variety of formats and information

What is this?



Large Synoptic Survey Telescope




What Happens in an Internet Minute?
639,800 GB of global IP data transferred
20
New victims of identity theft
204 million
Emails sent
47,000
App downloads
$83,000
In sales
61,141
Hours of music
20 million
Photo views
3,000
Photo uploads
320+
New Twitter accounts
100,000
New tweets
1,300
New mobile users
100+
New Linkedin accounts
135
Botnet infections
6
New Wikipedia articles published
277,000
Logins
6 million
Facebook views
2+ million
Search queries
30
Hours of video uploaded
1.3 million
Video views
Email
Mobile
Apps
amazon
PANDORA
LinkedIn
flickr
twitter
facebook
Google
YouTube
intel
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
IP
to view all video crossing IP networks each second

Conclusions?

….. that’s a lot of data ……

Do we know what it is?

Do we need to preserve it?

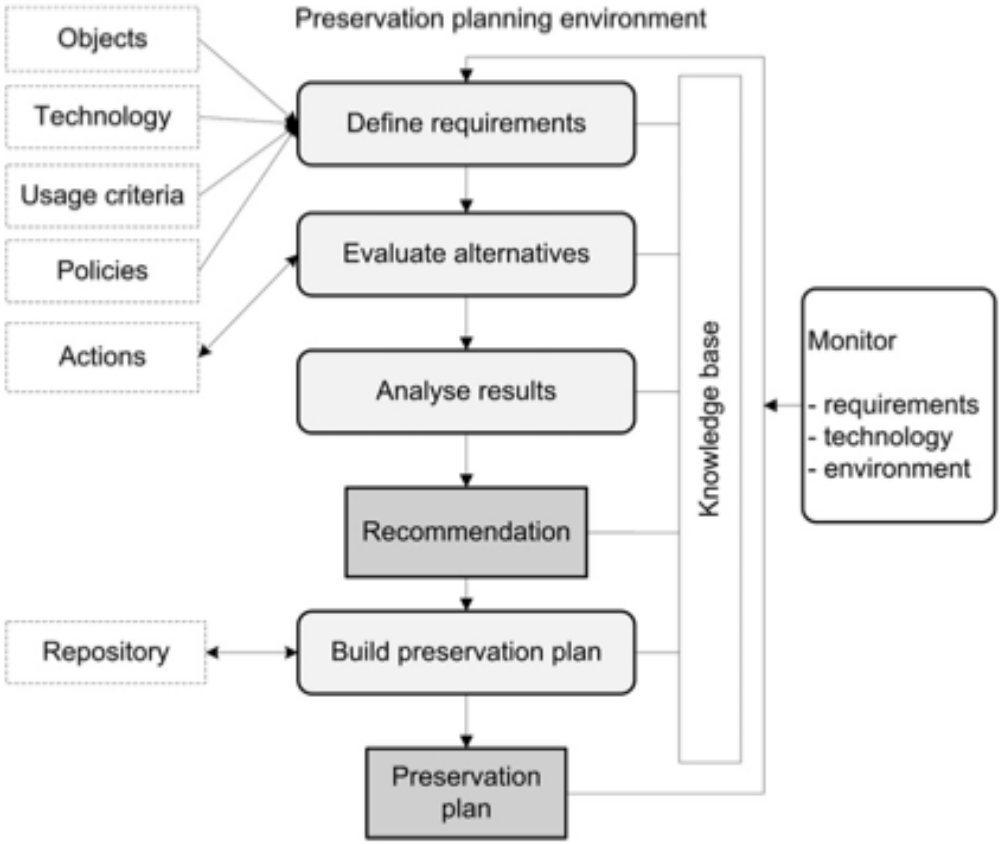
all of it??

What Can We Do?

A Plan of Action

- Identification
- Characterisation
- Risk Assessment
- Planning
- Action

What is Preservation Planning?



Preservation Plan

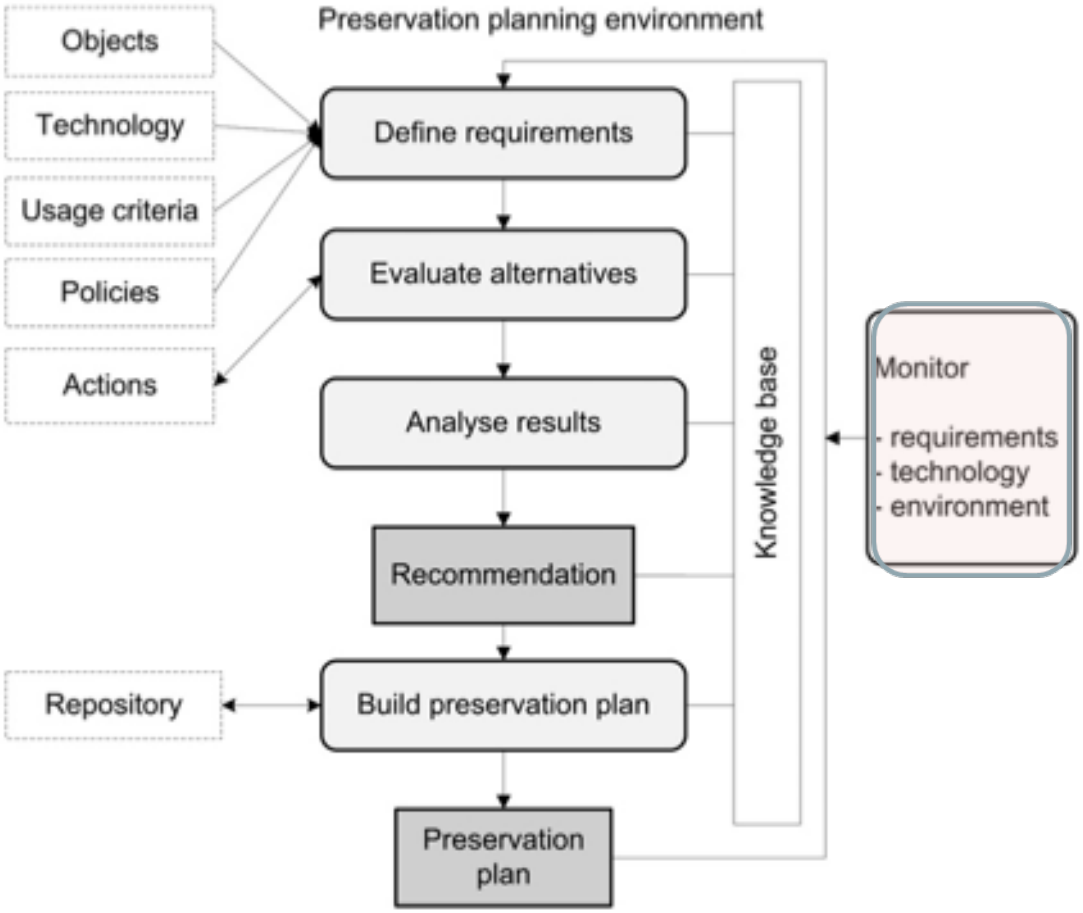
A preservation plan defines a series of preservation actions to be taken by responsible institution to address an identified risk for a given set of digital objects or records (called collection).

Plato

- Supports the Preservation Planning Workflow
- Helps the user through the process
- Enables the creation of a real preservation plan

https://github.com/openplanets/plato

Preservation Watch

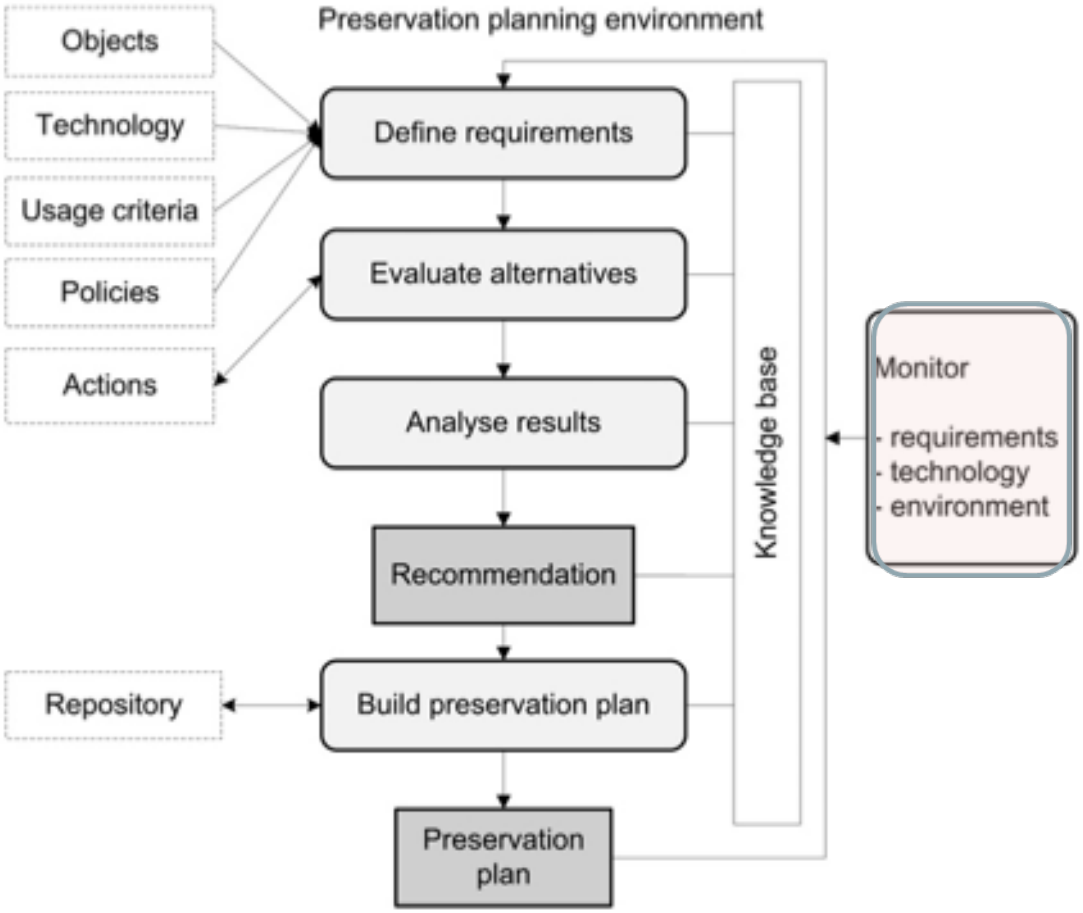


Scout

- Monitors interesting aspects of the world
- Notifies you when certain events occur
- Helps you to realise when you should re-evaluate or create a new plan

https://github.com/openplanets/scout

Characterising Large Collections



Format Profile

Property	File A	File B	File C
Format	PDF 1.2	PDF 1.2	PDF 1.4

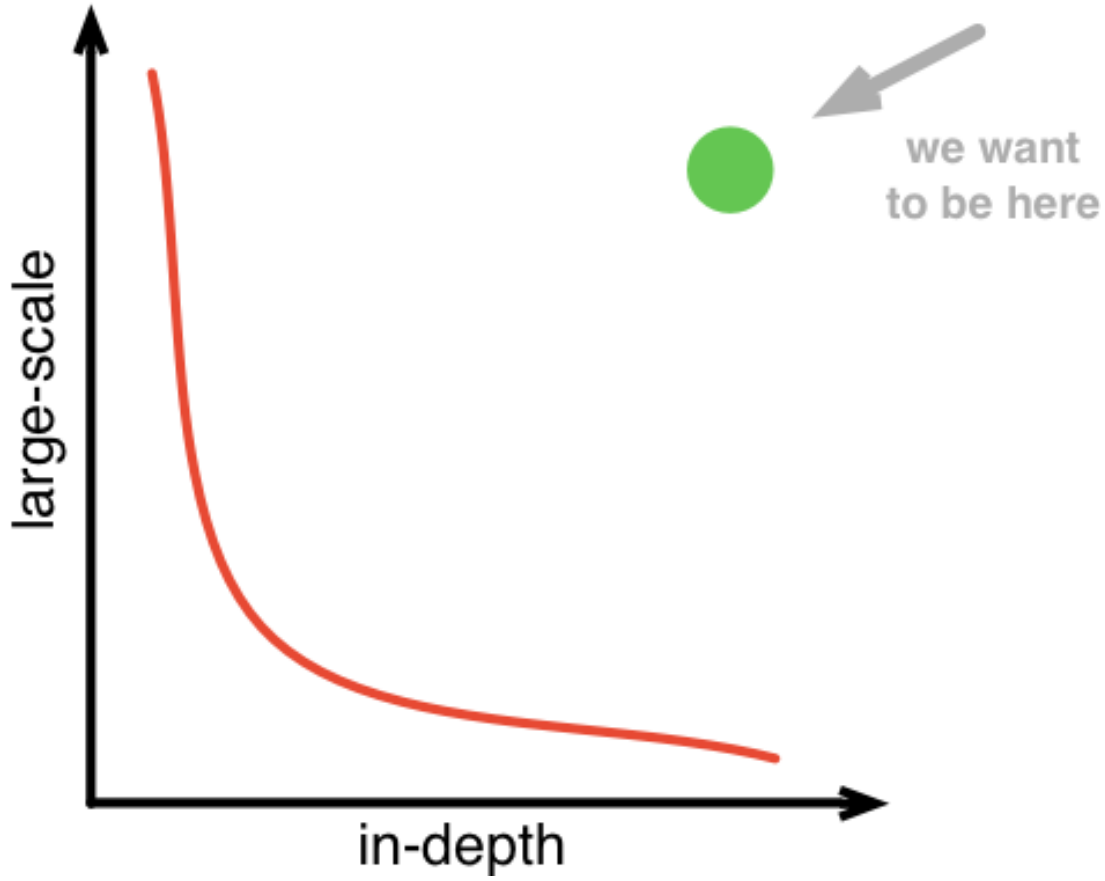
Content Profile

Property	File A	File B	File C
Format	PDF 1.2	PDF 1.2	PDF 1.4
Page Count	20	20.000	40
Encryption	Yes	No	Yes
File Size	1 MB	120 MB	2 MB
Valid	No	Yes	No
Well-formed	Yes	Yes	Yes

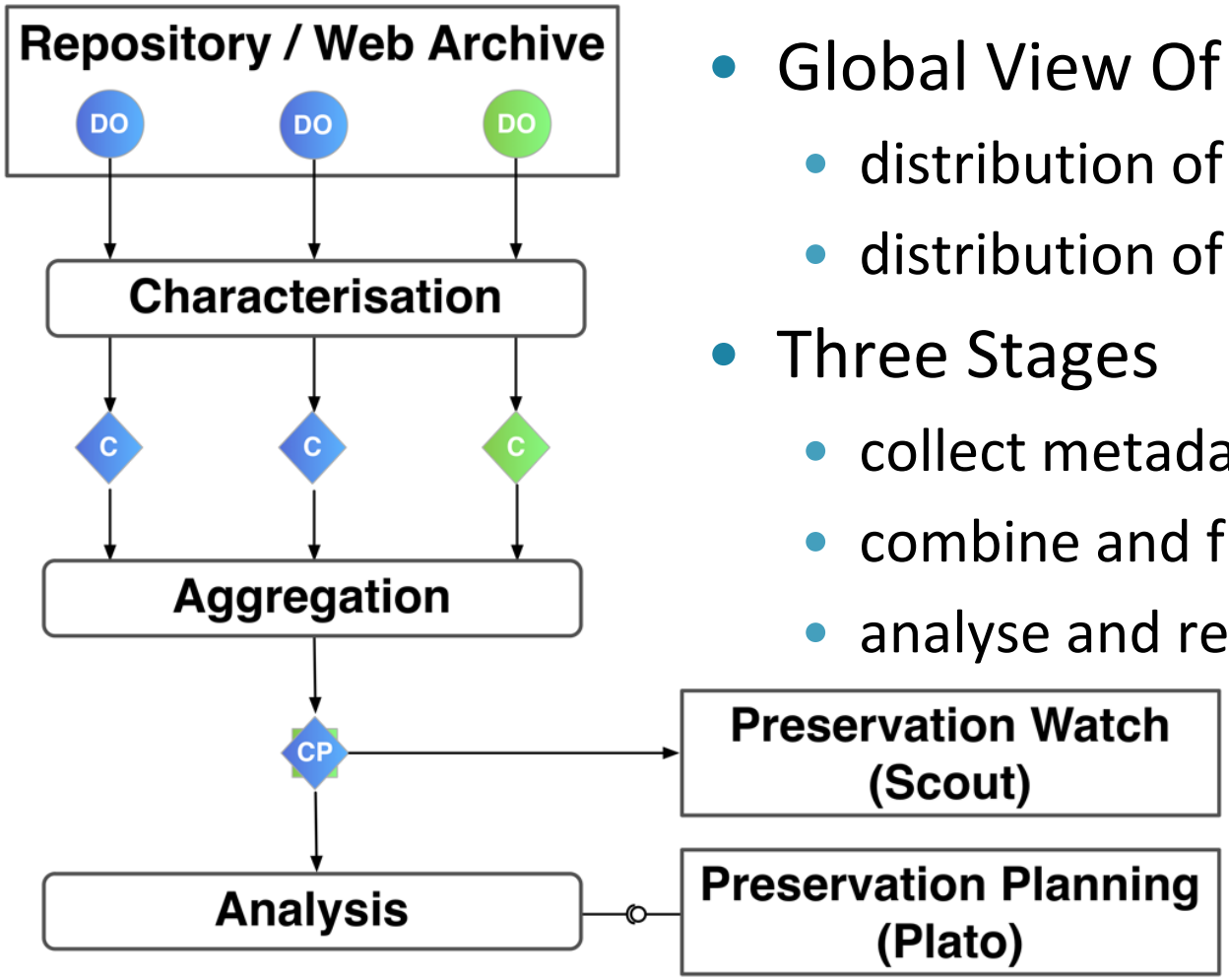
Heterogeneity

One Size Does Not Fit All

Scalability



Many Collections, One View

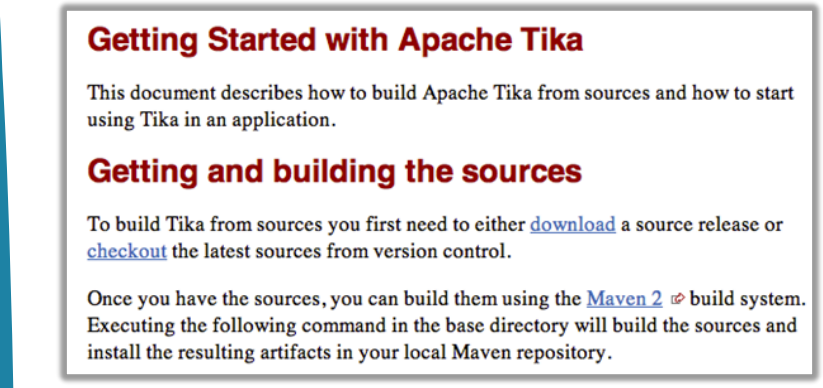


- Global View Of Content
  - distribution of file formats
  - distribution of characteristics
- Three Stages
  - collect metadata
  - combine and filter
  - analyse and reason

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

Digital Preservation Tools for Developers



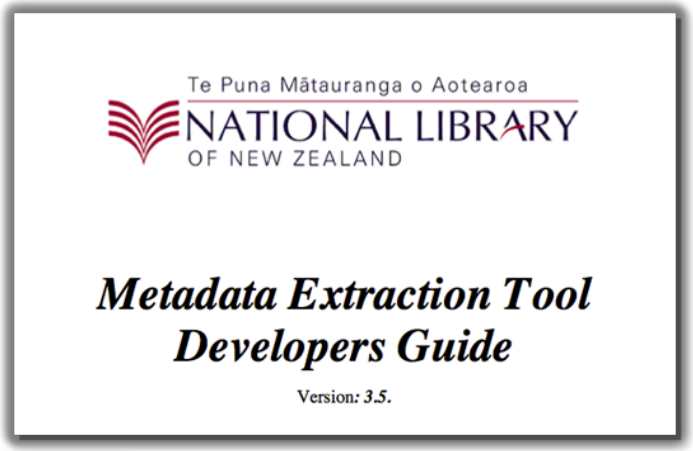
Getting Started with Apache Tika

This document describes how to build Apache Tika from sources and how to start using Tika in an application.

Getting and building the sources

To build Tika from sources you first need to either download a source release or checkout the latest sources from version control.

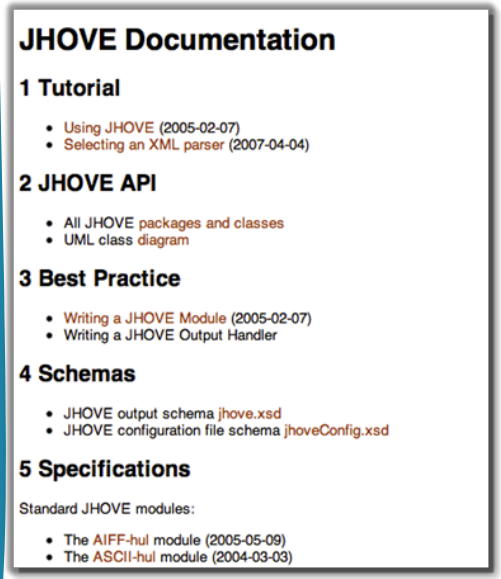
Once you have the sources, you can build them using the Maven 2 build system. Executing the following command in the base directory will build the sources and install the resulting artifacts in your local Maven repository.



Te Puna Mātauranga o Aotearoa
NATIONAL LIBRARY OF NEW ZEALAND

Metadata Extraction Tool Developers Guide

Version: *3.5.*



JHOVE Documentation

1 Tutorial

- Using JHOVE (2005-02-07)
- Selecting an XML parser (2007-04-04)

2 JHOVE API

- All JHOVE packages and classes
- UML class diagram

3 Best Practice

- Writing a JHOVE Module (2005-02-07)
- Writing a JHOVE Output Handler

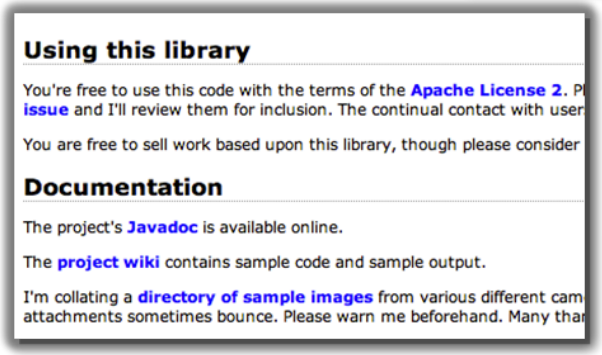
4 Schemas

- JHOVE output schema jhove.xsd
- JHOVE configuration file schema jhoveConfig.xsd

5 Specifications

Standard JHOVE modules:

- The AIFF-hul module (2005-05-09)
- The ASCII-hul module (2004-03-03)



Using this library

You're free to use this code with the terms of the **Apache License 2.** P
issue and I'll review them for inclusion. The continual contact with user

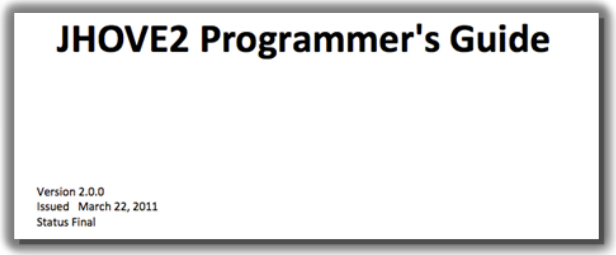
You are free to sell work based upon this library, though please consider

Documentation

The project's **Javadoc** is available online.

The **project wiki** contains sample code and sample output.

I'm collating a **directory of sample images** from various different cam
attachments sometimes bounce. Please warn me beforehand. Many thar



JHOVE2 Programmer's Guide

Version 2.0.0
Issued March 22, 2011
Status Final

Digital Preservation Tools for Developers

```
<?xml version="1.0"?>
<jhoveConfig version="1.0"
 xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xmlns="http://hul.harvard.edu/ois/xml/ns/jhove/jhoveConfig"
 xsi:schemaLocation="http://hul.harvard.edu/ois/xml/ns/jhove/jhoveConfig
                     http://hul.harvard.edu/ois/xml/xsd/jhove/jhoveConfig.xsd">
  <jhoveHome>jhove-home-directory</jhoveHome>
[ <defaultEncoding>encoding</defaultEncoding> ]
[ <tempDirectory>directory</tempDirectory> ]
[ <bufferSize>buffer</bufferSize> ]
[ <mixVersion>version</mixVersion> ]
[ <sigBytes>n</sigBytes> ]
  <module>
    <class>module-class-name</class>
  [ <init>optional-module-init-argument</init> ]
  [ <param>optional-module-parameter</param> ]
    ...
  </module>
  ...
  <outputHandler>
    <class>output-handler-class-name</class>
  </outputHandler>
  ...
[ <logLevel>logging-level</logLevel> ]
</jhoveConfig>
```

```
Options:
    -?  or --help           Print this usage message
    -v  or --verbose        Print debug level messages
    -V  or --version        Print the Apache Tika version n

    -g  or --gui            Start the Apache Tika GUI
    -s  or --server         Start the Apache Tika server
    -f  or --fork           Use Fork Mode for out-of-proces

    -x  or --xml            Output XHTML content (default)
    -h  or --html           Output HTML content
    -t  or --text           Output plain text content
    -T  or --text-main      Output plain text content (main
    -m  or --metadata       Output only metadata
    -j  or --json           Output metadata in JSON
    -y  or --xmp            Output metadata in XMP
    -l  or --language       Output only language
    -d  or --detect         Detect document type
    -eX or --encoding=X     Use output encoding X
    -pX or --password=X     Use document password X
    -z  or --extract        Extract all attachements into c
    --extract-dir=<dir>     Specify target directory for -z
```

file(1) - Linux man page

Name

file - determine file type

Synopsis

file [**-bchikLNnprsvz0**] [**--apple**] [**--mime-encoding**] [**--mime-type**] [**-e** *testname*] [**-F** *separator*] [**-f** *namefile*] [**-m** *magicfiles*] *file* ...

file -C [**-m** *magicfiles*]

file [**--help**]

Description

DP Tools for Preservation Experts


```
<jhove
xmlns:xsi="http://www.w3.org/2001/XMLSchema-xmlns="http://hul.harvard.edu/ois/xml/ns/jhove"
xsi:schemaLocation="http://hul.harvard.edu/ois/xmlhul.harvard.edu/ois/xml/xsd/jhove/1.6/jhove.xsd"
name="Jhove" release="1.6" date="2011-01-04">
 <date>2011-03-23T10:03:54+01:00</date>
 <repInfo uri="what_am_i.jpg">
 <reportingModule release="1.2" date="2007-02-13">JPreportingModule>
 <lastModified>2009-09-16T15:23:14+02:00</lastM <size>83672</size>
<format>JPEG</format>
<version>1.01</version>
<status>Well-Formed and valid</status>
<sigMatch>
<module>JPEG-hul</module>
</sigMatch>
<mimeType>image/jpeg</mimeType>
<profiles>
  <profile>JFIF</profile>
</profiles>
<properties>
  <property>
   <name>JPEGMetadata</name>
   <values arity="List" type="Property">
 <property>
  <name>CompressionType</name>
....
</jhove>
```

DP Tools for Preservation Experts

ExifToolVersion 7.74
FileNametest_file.pdf
FileSize 2.7 MB
FileModifyDate2012:11:02 18:17:28+01:00
FileType PDF
MIMEType application/pdf
PageCount 2
PDFVersion 1.4

<jhove>
<repInfo xmlns="" uri="test_file.pdf">
<reportingModule release="1.8" date="2009-05-22">PDFhul</reportingModule>
<lastModified>2012-11-02T18:17:28+01:00</lastModified>
<size>2846900</size>
<format>PDF</format>
<version>1.3</version>
<status>Well-Formed and valid</status>
<sigMatch>
<module>PDF-hul</module>
</sigMatch>
<mimeType>application/pdf</mimeType>
<properties>
<property>
<name>PDFMetadata</name>
</jhove>

To Sum Up…..

JHOVE JSTOR/Harvard Object Validation Environment

JHOVE2

Apache Tika

fido

The National Archives

jpylyzer

ffident

METADATA XTRACTOR

TrID

A few problems….

- A lot of tools to manage and invoke
- Harder for developers and administrators
- Different output schemas
- Different configuration/environments
- No or bad higher level management
- Difficult to spot differences

FITS

- File Information Tool Set
- Harvard University Library
- 2009
- v0.6.1, LGPL
- Wraps other tools
- 6m - 1y for a new version

http://code.google.com/p/fits/

http://github.com/gmcgath/openfits

Tools (inside)

- Droid
- Metadata Extra
- Jhove
- Exiftool
- FFident
- File Utility

FITS

- Consolidates output
- Can include raw output
- Configurable/Extendable

FITS Output

```
<?xml version="1.0" encoding="UTF-8"?>
<fits xmlns="http://hul.harvard.edu/ois/xml/ns/fits/fits_output" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
xsi:schemaLocation="http://hul.harvard.edu/
ois/xml/ns/fits/fits_output http://hul.harvard.edu/ois/xml/xsd/fits/fits_output.xsd" version="0.6.0" timestamp="12/27/11 10:49 AM">
  <identification>
   <identity format="Portable Document Format" mimetype="application/pdf" toolname="FITS" toolversion="0.6.0">
     <tool toolname="Jhove" toolversion="1.5" />
     <tool toolname="file utility" toolversion="5.03" />
     <tool toolname="Exiftool" toolversion="7.74" />
     <tool toolname="Droid" toolversion="3.0" />
     <tool toolname="NLNZ Metadata Extractor" toolversion="3.4GA" />
     <tool toolname="ffident" toolversion="0.2" />
     <version toolname="Jhove" toolversion="1.5">1.4</version>
     <externalIdentifier toolname="Droid" toolversion="3.0" type="puid">fmt/18</externalIdentifier>
   </identity>
  </identification>
  <fileinfo>
    <size toolname="Jhove" toolversion="1.5">39586</size>
    <creatingApplicationName toolname="NLNZ Metadata Extractor" toolversion="3.4GA"
status="SINGLE_RESULT">/XPP</creatingApplicationName>
    <lastmodified toolname="Exiftool" toolversion="7.74" status="SINGLE_RESULT">2011:12:27 10:44:28+01:00</lastmodified>
    <created toolname="Exiftool" toolversion="7.74" status="SINGLE_RESULT">2002:04:25 13:02:24Z</created>
    <filepath toolname="OIS File Information" toolversion="0.1"
status="SINGLE_RESULT">/home/petrov/taverna/tmp/000/000009.pdf</filepath>
```

Advantages

- I care only about one output schema
- I get (basic) QA hints
- Better type coverage (although...)

Disadvantages

- Consolidation is hard
- Flexibility
- Performance and Scalability
- Stability

C3PO

- 0.2.0 – alpha
- Command line application
- Web application

https://github.com/peshkira/c3po

https://github.com/openplanets/c3po

{name: “mongo”, type:”DB”}

- Widely used
- Stores documents as BSON
- Native support for Map/Reduce
- Sharding and auto balancing out of the box
- NoSQL != schemaless

CLI

- Java
- Parses and processes FITS output
- Stores the results in the document store
- XML Profile + CSV

WebApp

- Overview
- Browsing
- Filtering
- Samples

- No Java EE stack
- Supports MVC and REST
- Scala and Java versions
- Templating engine, etc.