Characterization and Content Profiling

Artur Kulmukhametov

Department of Software Technology and Interactive Systems

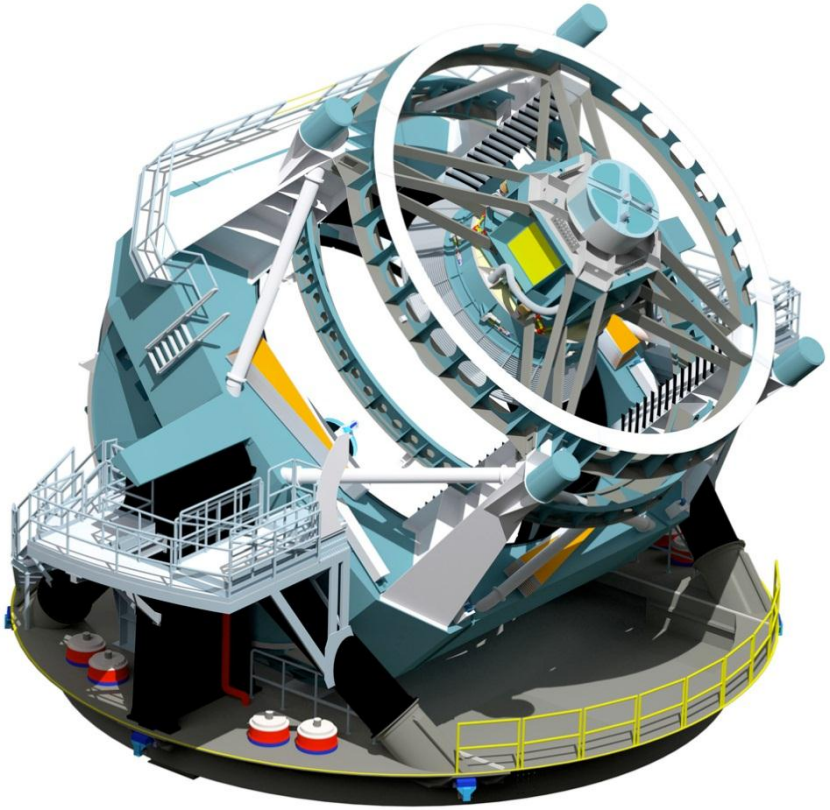
Vienna University of Technology

http://www.ifs.tuwien.ac.at/~kulmukhametov

Overview

- Motivation: collection scale and heterogeneity
- An approach to getting a control
- Characterisation tools
- Content profiling

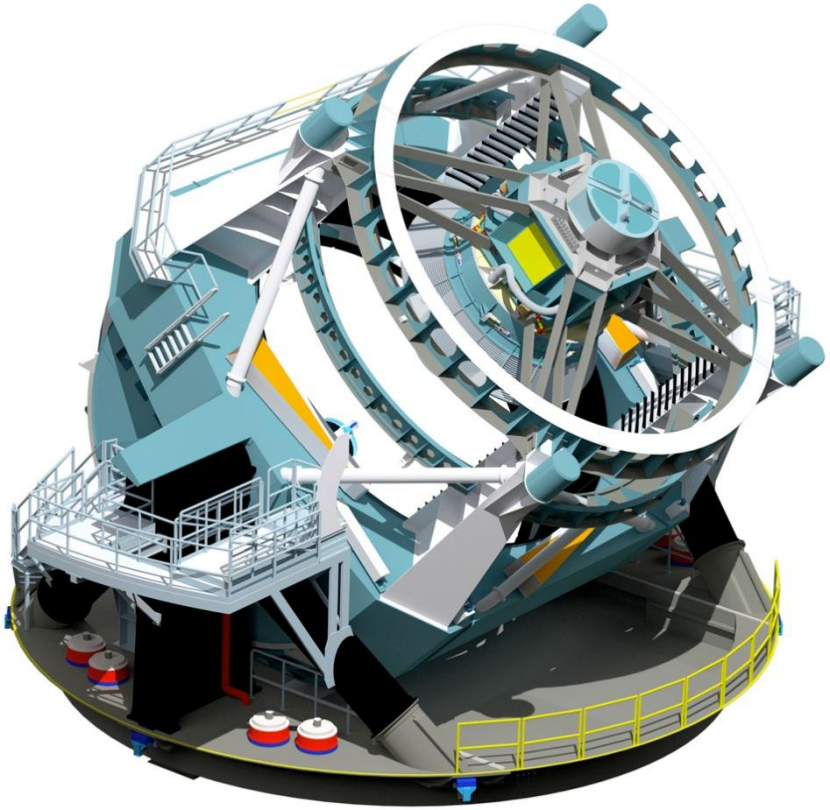
What is it?



*

- P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Large Synoptic Survey Telescope



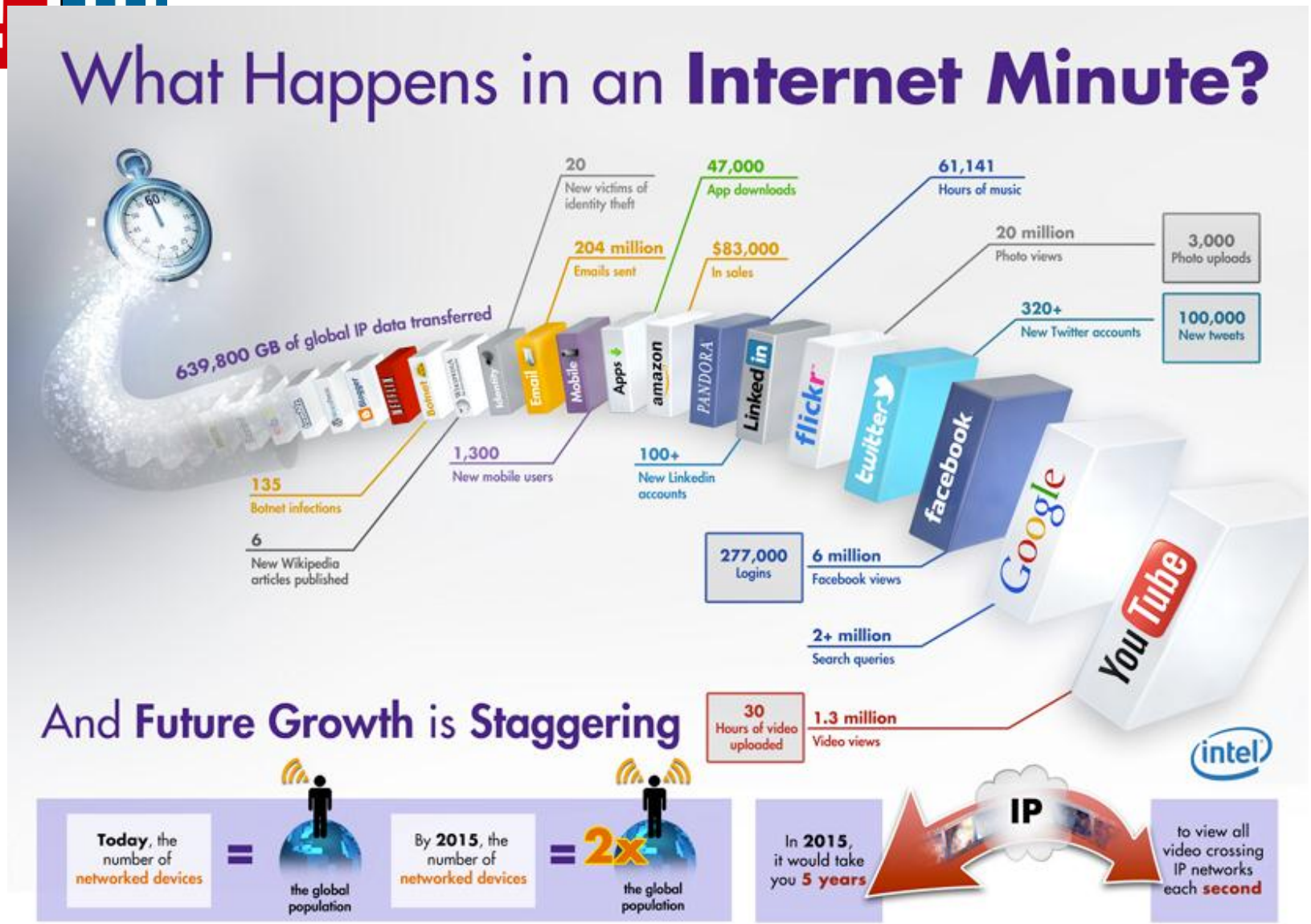
*

30 Terabytes of data nightly

- P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Variety of Data

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



* - P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Conclusions?

….. that’s a lot of data ……

Do you know what that data is?

Do you want to do something with it?

Characterization

the act of describing the character or qualities of someone or something

"Characterization." Merriam-Webster.com. Accessed March 10, 2014. http://www.merriam-webster.com/dictionary/characterization.

Characterization

*

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

* - P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Characterization

*

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

* - P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Characterization

*

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

* - P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Characterization

*

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

* - P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Characterization

*

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

! One size does not fit all !

* - P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Tools for Characterization

JHOVE — JSTOR/Harvard Object Validation Environment

JHOVE2

fido

Apache Tika

Exif

The National Archives

jpylyzer

ffident

Exiftool

METADATA XTRACTOR

TrID

Droid

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

File Information Tool Set

- Main features:
- Consolidates output
- Can include raw output
- Configurable/Extendable

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

Conflicts

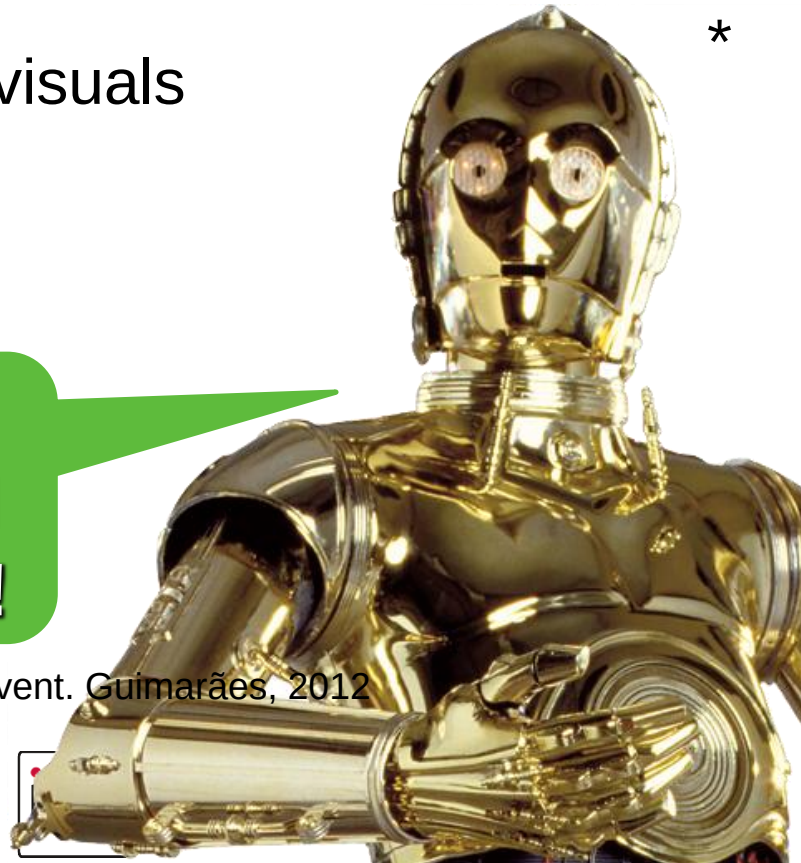
3 types of conflicts:

1. Inconsistent property naming, e.g: *image_width* and *imagewidth*
2. Competing characterisation results, e.g: tool1 identifies a file as *plain text*, but tool2 identifies the file as *PDF*
3. Close, but not the same property values, e.g: *application/xhtml+xml* vs. *application/xml*.

Clever, Crafty Content Profiling of Objects

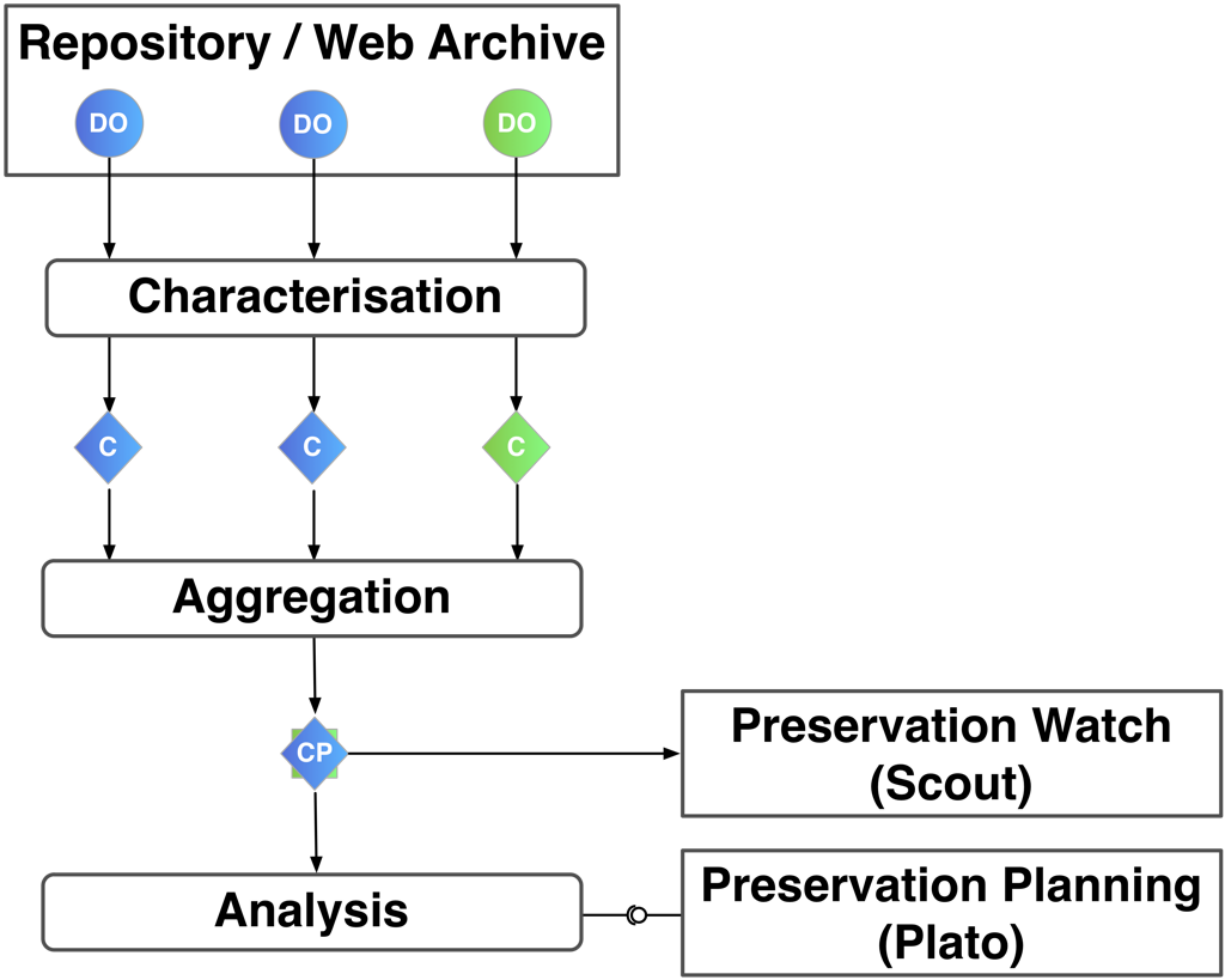
C3PO is a tool for content profile generation.

- Uses characterization results
- Deeper content analysis with nice visuals through the web-app
- Generates content profiles



* - P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Content Profiling



- Global View of Content
  - Distribution of characteristics
  - Statistics (size, min, max, …)
  - Sampling

* - P. Petrov, *Content Profiling and Planning,* SCAPE Training Event. Guimarães, 2012

Representative Sampling

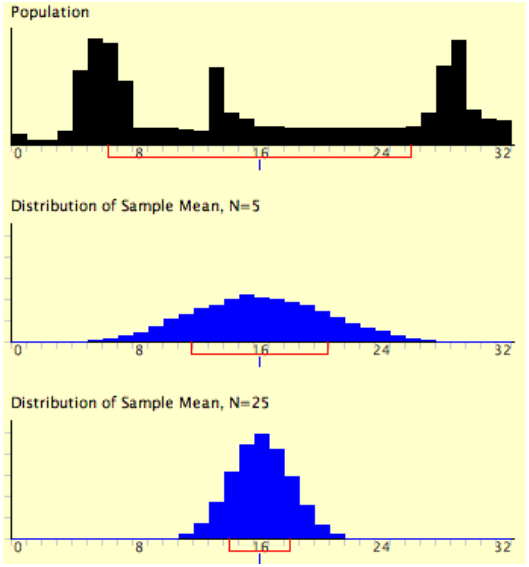
*

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

* - E. Poltorak, *Representative sampling,* Flickr, http://www.flickr.com/photos/44461316@N08/4110321514/, 2009

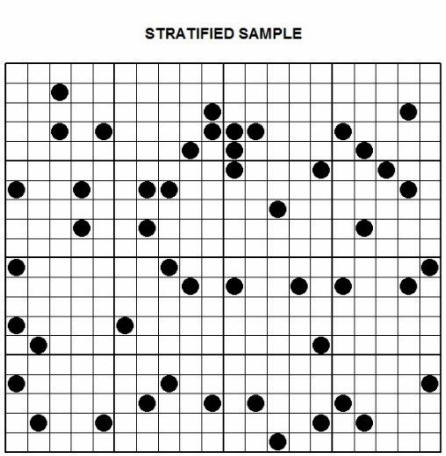
C3PO: Representative Samples

DistSampler



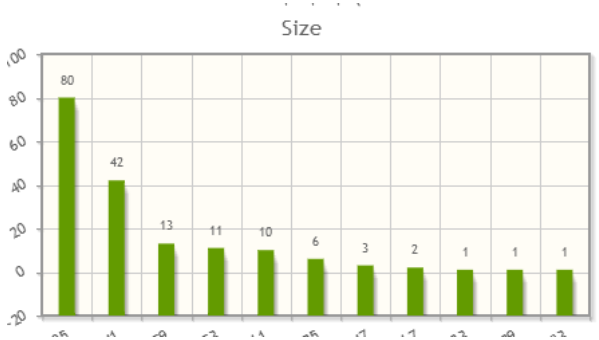
**

SysSampler



*

Size'o'Matic 3000



* - Statistical Consultants Ltd, http://www.statisticalconsultants.co.nz/weeklyfeatures/WF7.html, 2013

** - D. Lane, *Online Statistics Education*, http://onlinestatbook.com/2/sampling_distributions/samp_dist_mean.html, 2013

Summary

- Characterization is time consuming
- Characterisation can be faulty
- Know your tools
- Content profiling is a key to better understanding of your data

DP Exercise Task 1

- Get your data (at least 5000 files)
- Characterize the data with FITS
- Create and analyse a content profile using C3PO
- Write a report