SCAPE
Scalable Preservation Environments

Preservation Watch

What to monitor and how Scout can help

Luis Faria lfaria@keep.pt

KEEP SOLUTIONS www.keep-solutions.com

Digital Preservation Advanced Practitioner Course
Glasgow, 15th-19th July 2013

KEEP SOLUTIONS

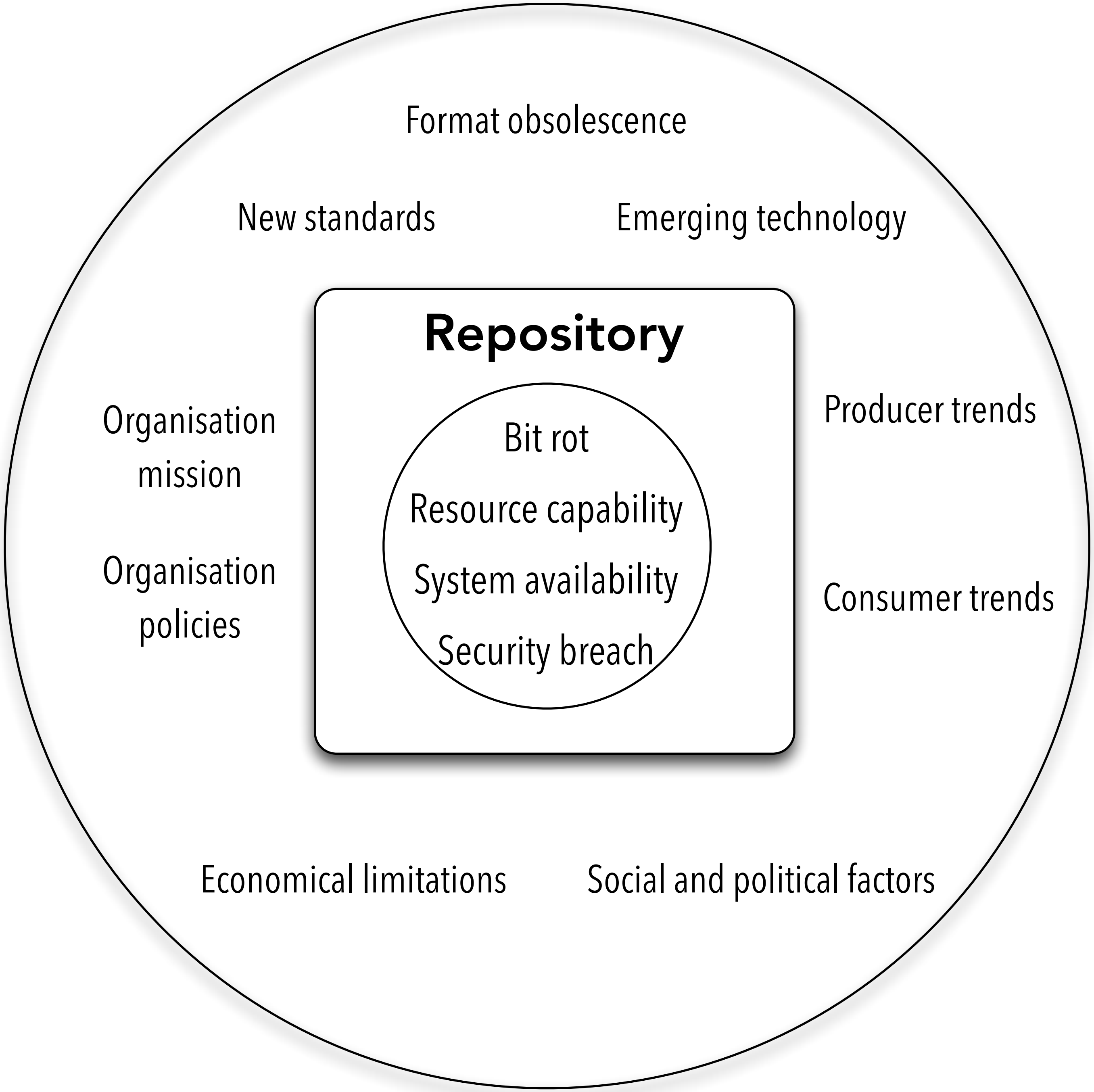
- Company specialized in information management
- Digital preservation experts
- Open source: RODA, KOHA, DSpace, Moodle, etc.
- Scientific research
  - **SCAPE**: large-scale digital preservation environments
  - **4C**: digital preservation cost modeling

http://www.keep-solutions.com

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).

Preservation monitoring

Why do we need monitoring?



Format obsolescence

New standards

Emerging technology

Repository

Organisation mission

Bit rot

Resource capability

System availability

Security breach

Producer trends

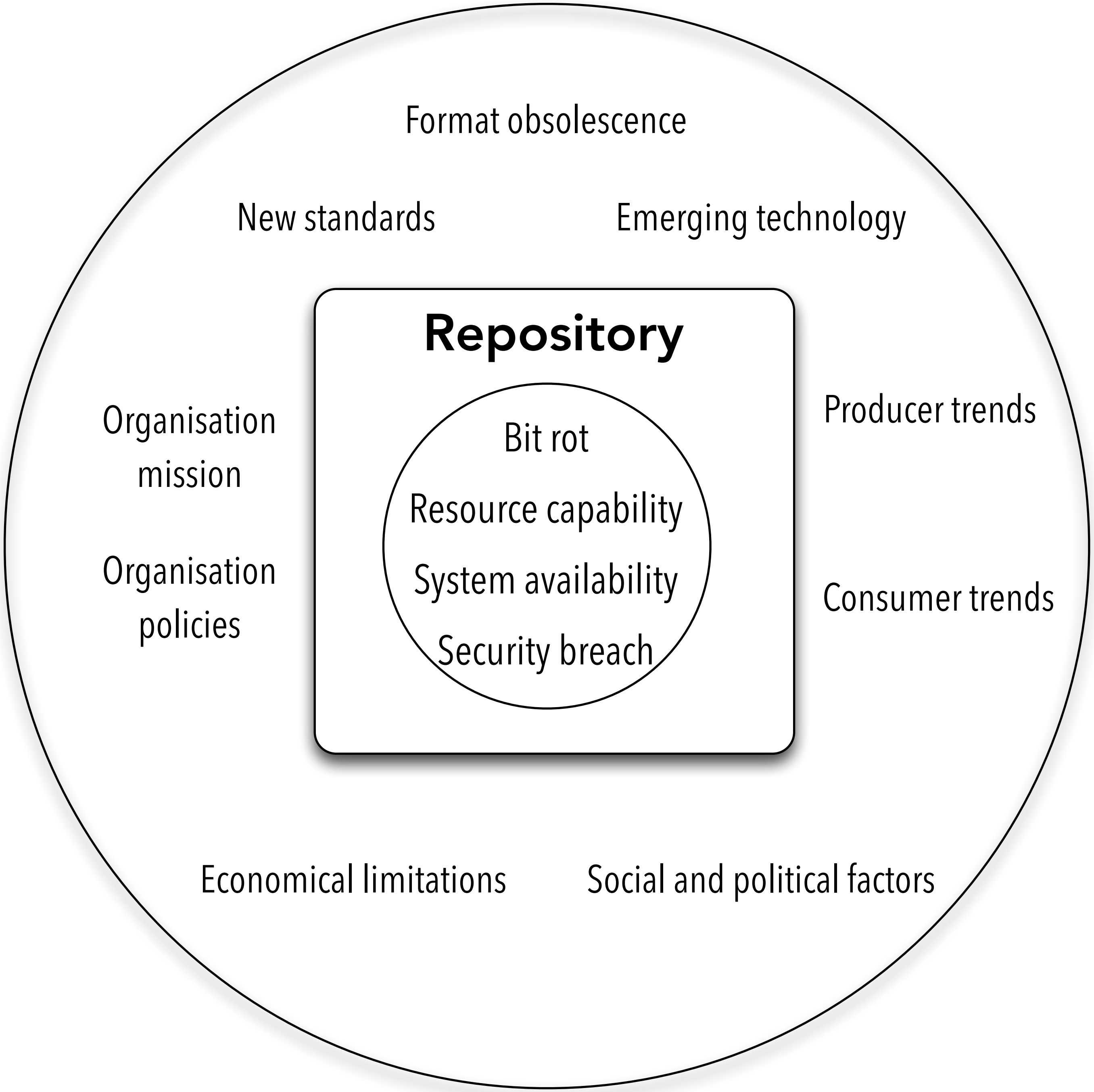
Organisation policies

Consumer trends

Economical limitations

Social and political factors

Why do we need monitoring?


Format obsolescence

New standards

Emerging technology

Repository

Organisation mission

Organisation policies

Bit rot

Resource capability

System availability

Security breach

Producer trends

Consumer trends

Economical limitations

Social and political factors

Risks

Opportunities

State of the Art

- Digital Format Registries
- Automatic Obsolescence Notification System (AONS)
- Technology watch reports

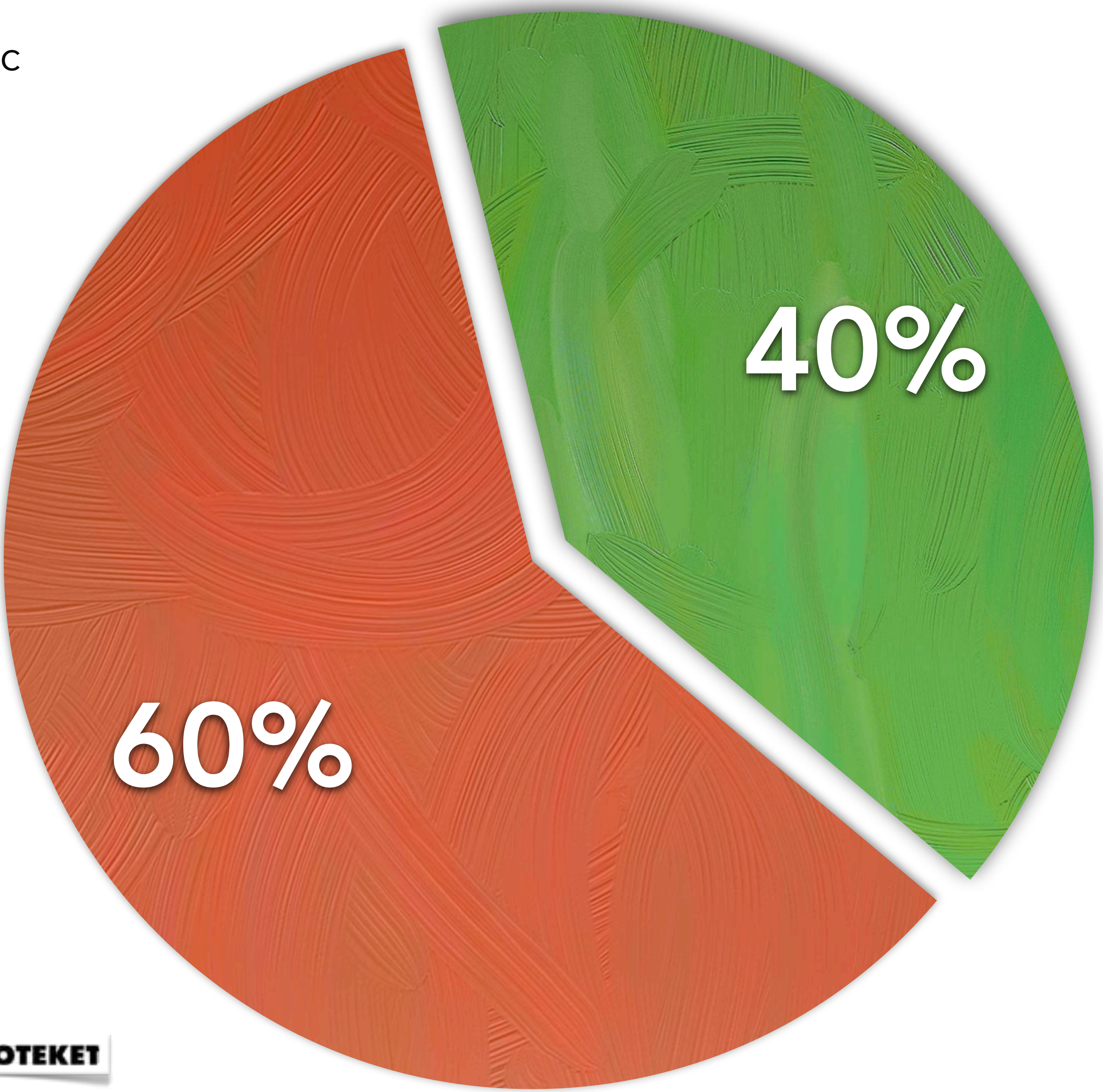
State of the Art

- Digital Format Registries
  - Lack of coverage
  - Statically-defined generic risks
  - Lack of structure in risks
  - Focus on format obsolescence
- AONS
  - Total dependency on format registries
- Technology watch reports
  - Machine unreadable

Risk Assessment

🟩 Yes but manual and adhoc

🟥 None



Survey on:

Science & Technology Facilities Council

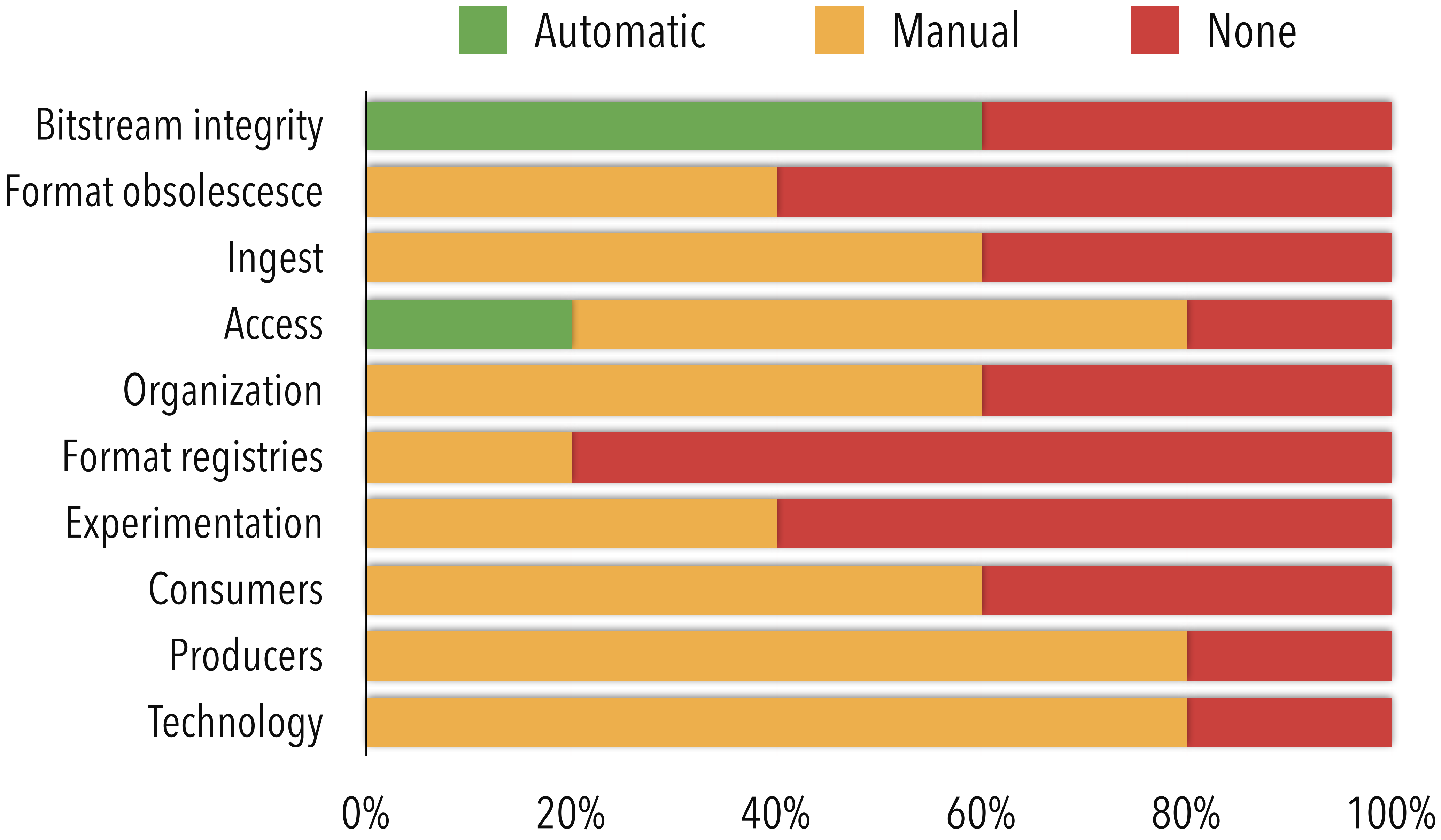
iM INTERNET MEMORY FOUNDATION

BRITISH LIBRARY

STATSBIBLIOTEKET

KB Koninklijke Bibliotheek National Library of the Netherlands

Monitoring

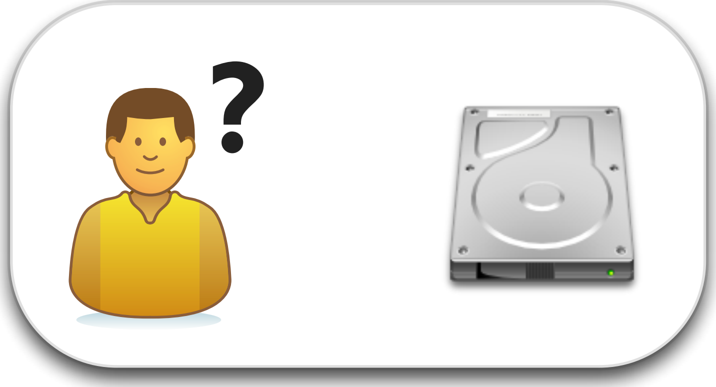

Automatic
Manual
None
Bitstream integrity
Format obsolescesce
Ingest
Access
Organization
Format registries
Experimentation
Consumers
Producers
Technology
0%
20%
40%
60%
80%
100%

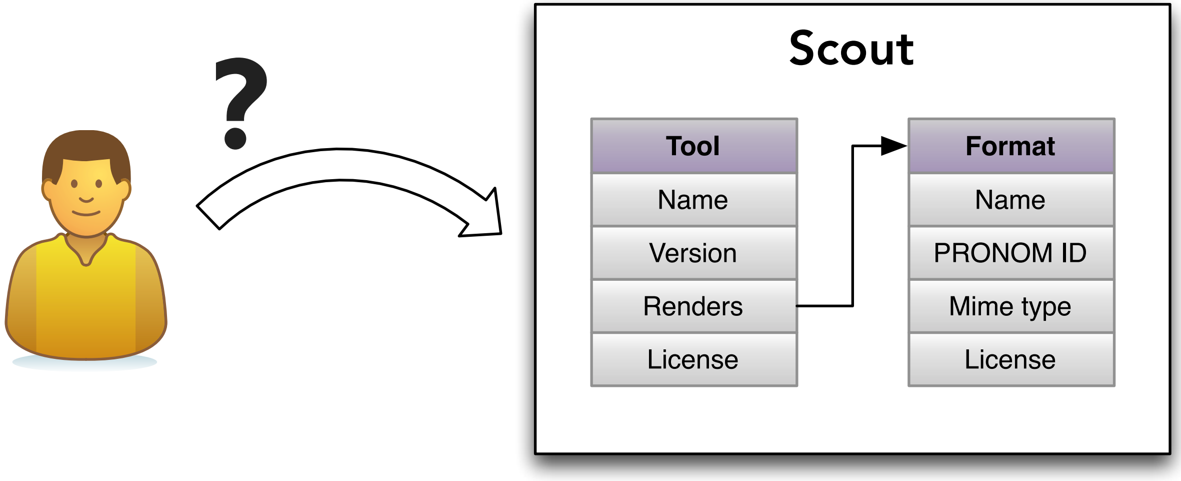
What is needed?

- We need data!
  - From anywhere and everywhere
  - Sharing
- Usability & Scalability
  - Structured data
  - Controlled vocabulary

Scout

A novel approach

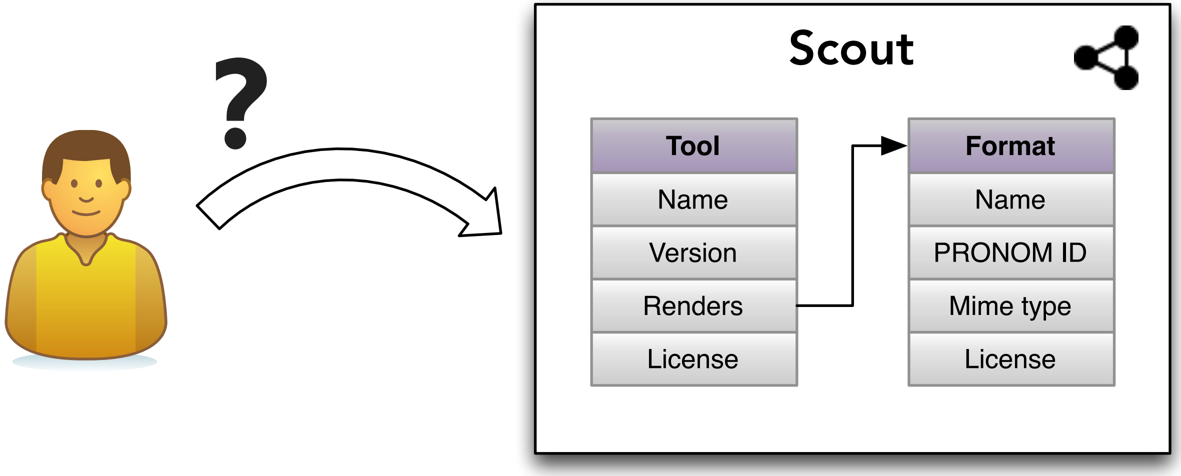




Scout

Tool
Name
Version
Renders
License

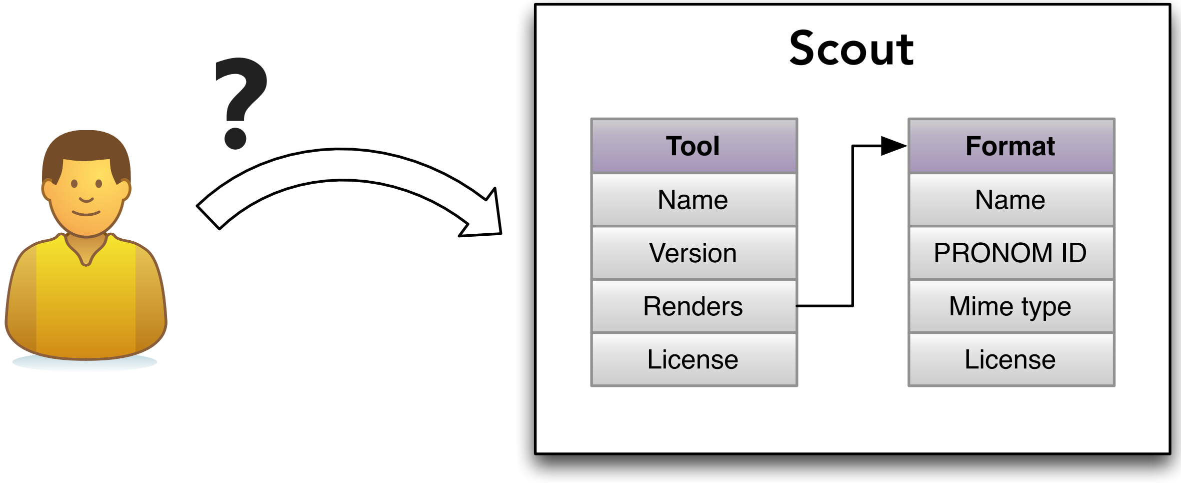
Format
Name
PRONOM ID
Mime type
License



Scout

Tool
Name
Version
Renders
License

Format
Name
PRONOM ID
Mime type
License



Scout

Tool

Name

Version

Renders

License

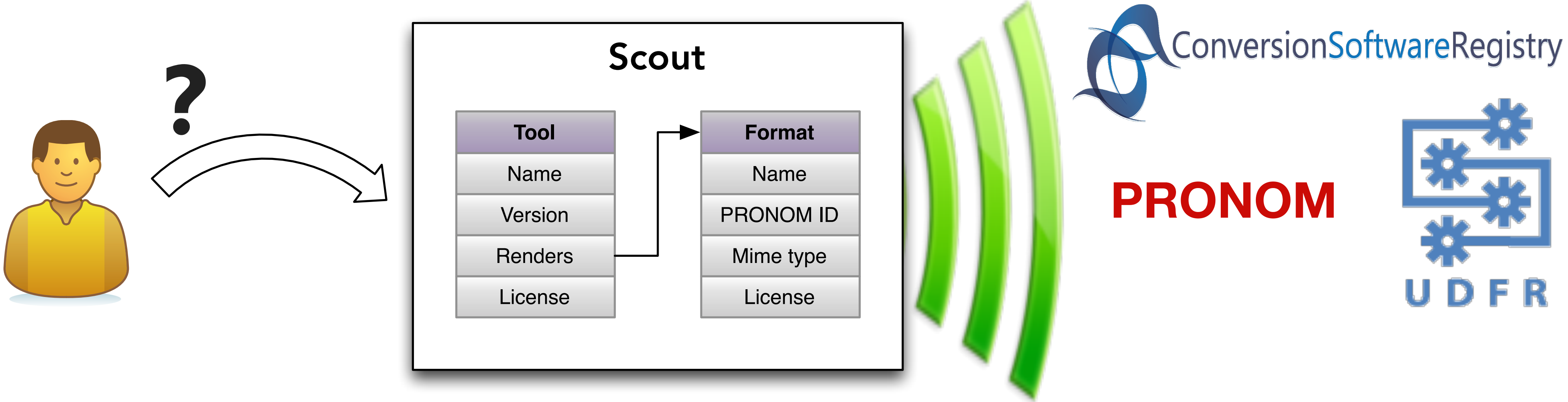
Format

Name

PRONOM ID

Mime type

License



?

Scout

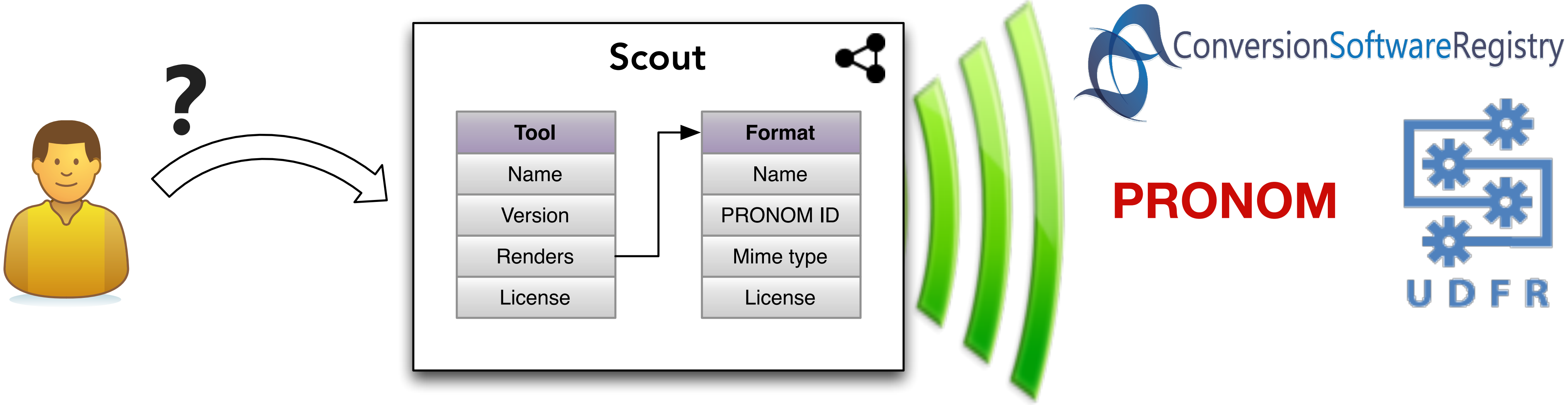
Tool
- Name
- Version
- Renders
- License

Format
- Name
- PRONOM ID
- Mime type
- License

ConversionSoftwareRegistry

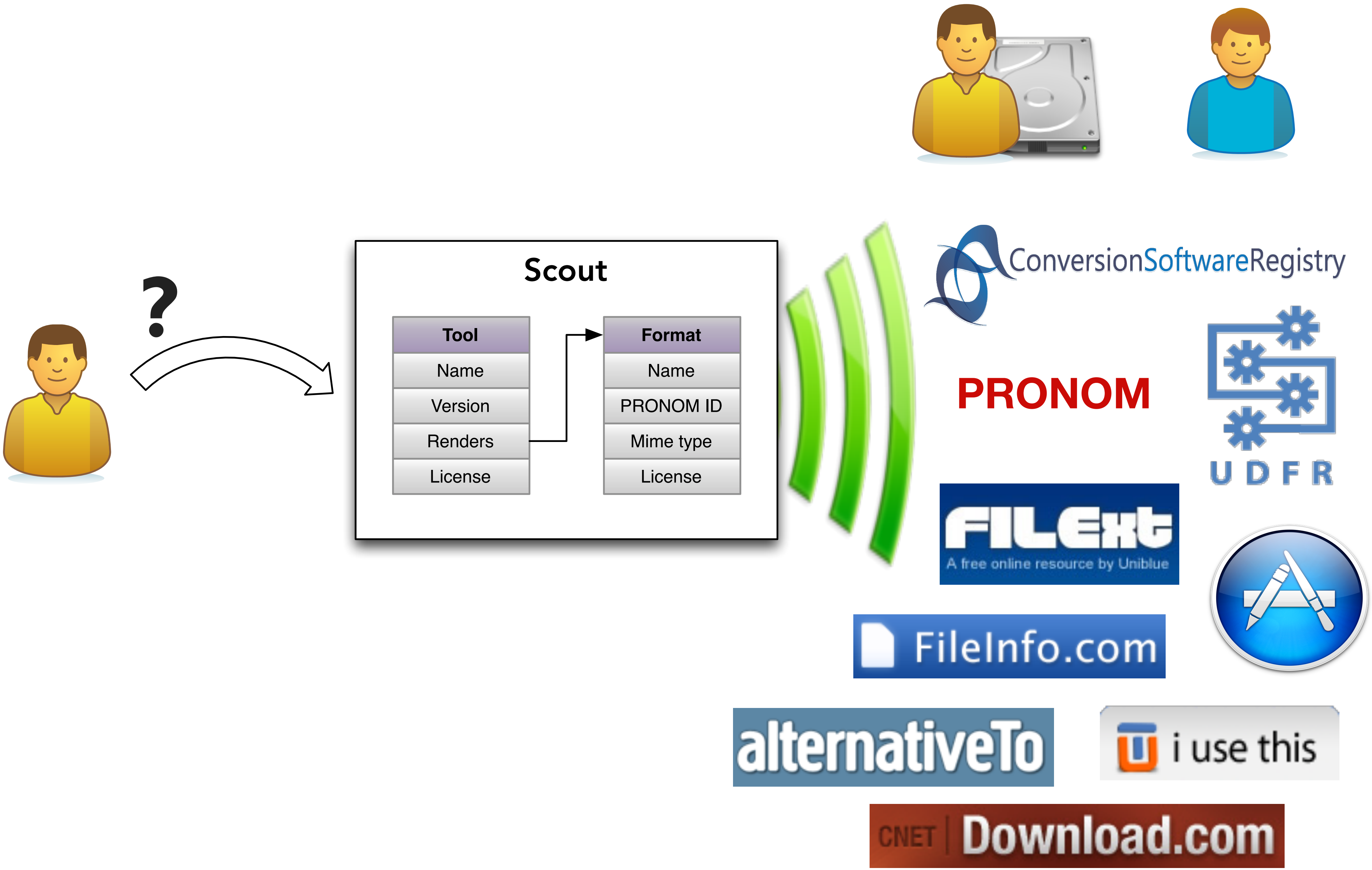
PRONOM

UDFR



Tool
Name
Version
Renders
License

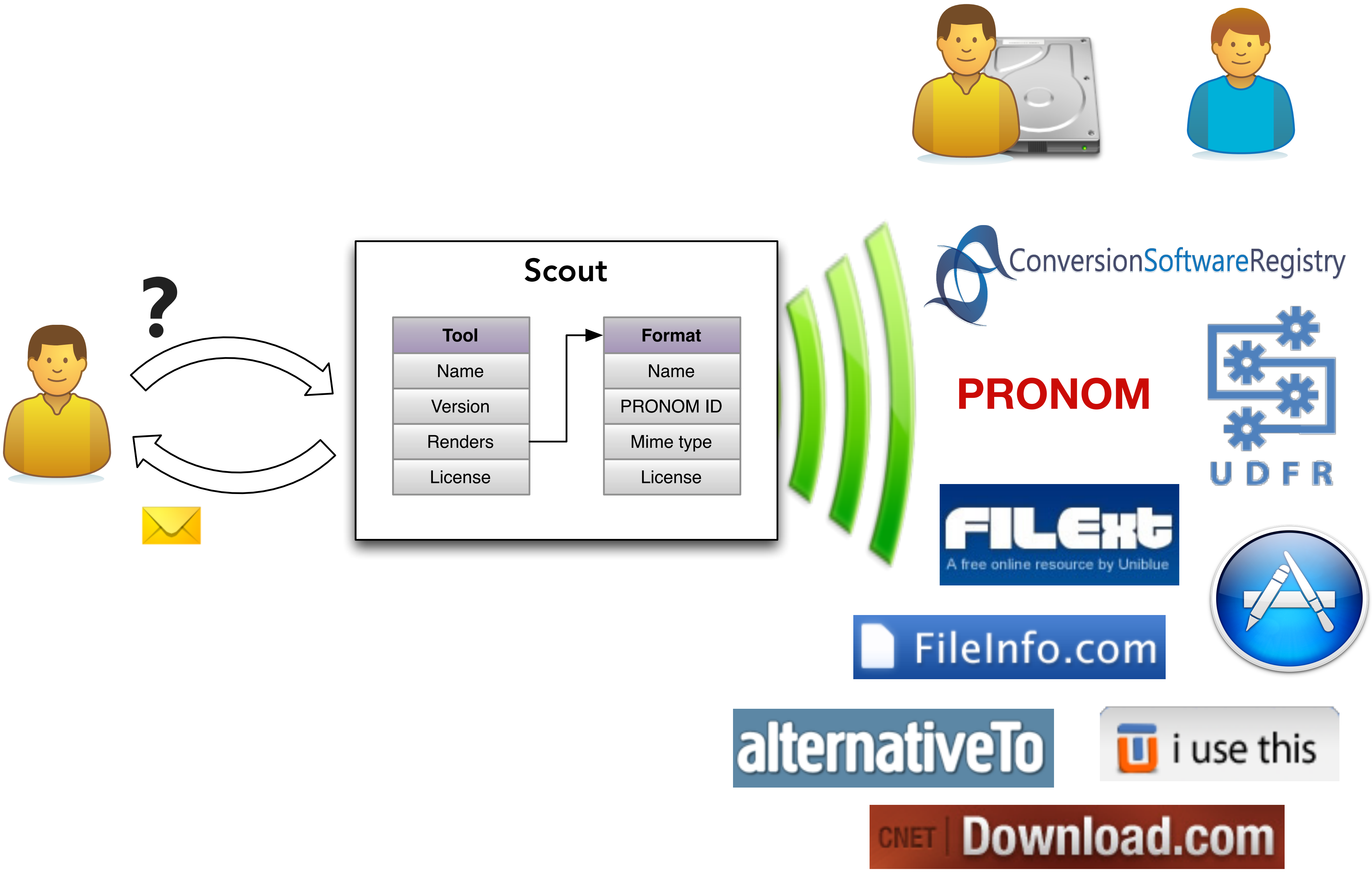
Format
Name
PRONOM ID
Mime type
License



Scout

Tool
- Name
- Version
- Renders
- License

Format
- Name
- PRONOM ID
- Mime type
- License



Scout

Tool
- Name
- Version
- Renders
- License

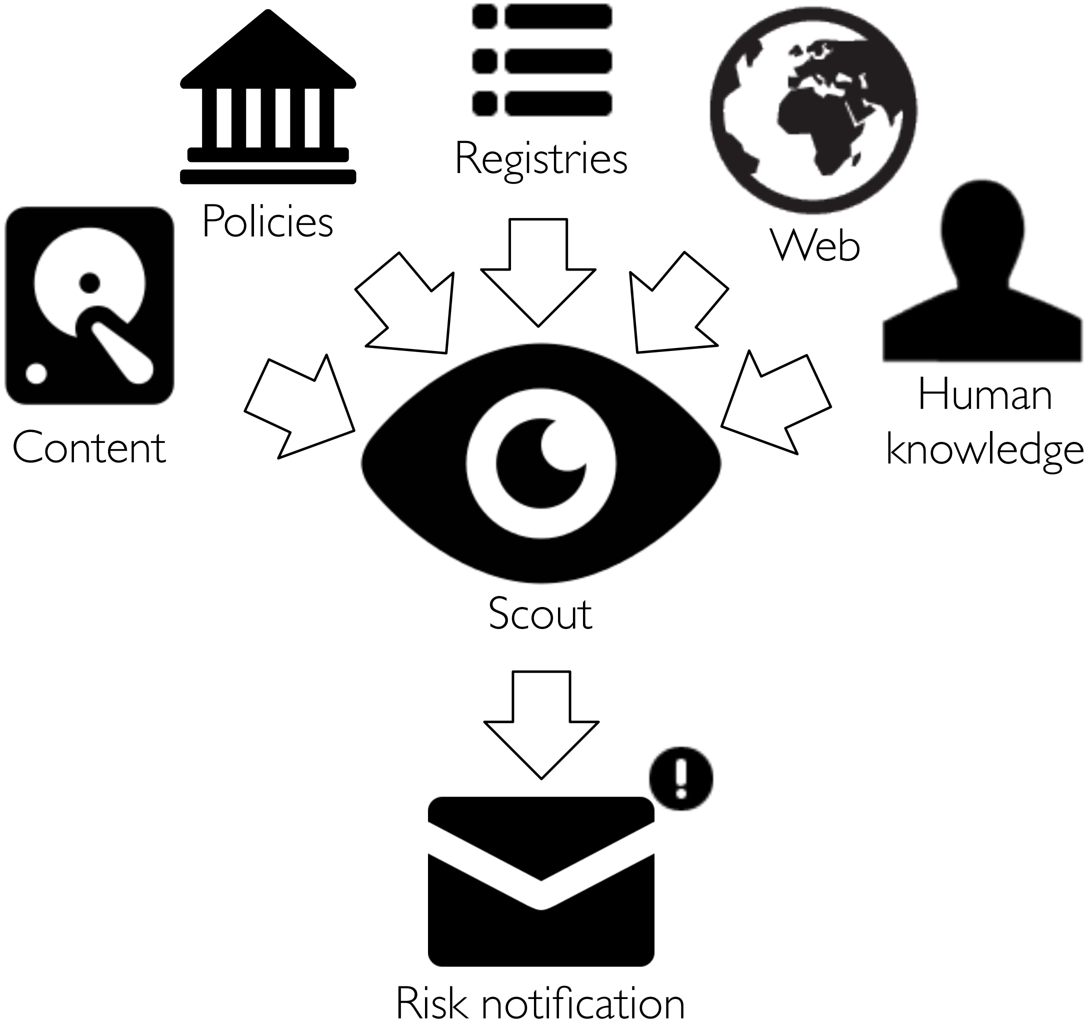
Format
- Name
- PRONOM ID
- Mime type
- License

Goals

- Collect information from different sources
- Enable human input of data
- Central knowledge base for digital preservation
- Enable users to pose questions
- Notify users of significant events and plan validity
- Easily support for new sources and questions

Scout: a preservation watch system

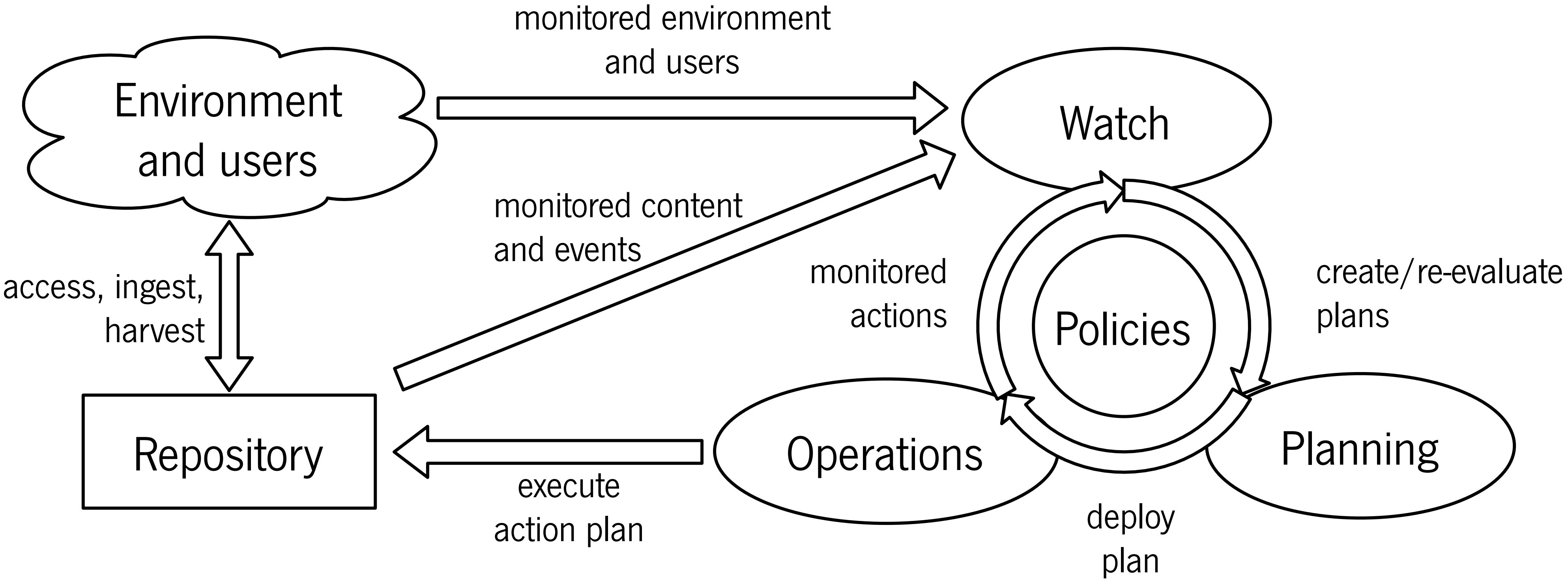
- Monitors aspects of the world to detect preservation risks and opportunities
- Triple store
- Adaptors
  - Data Connector & Report API
  - SCAPE Policy model
  - PRONOM
  - Web semantic extraction
  - Renderability experiments
- Web interface
- Triggers: templates and SPARQL
- Email notifications
- Demo: http://scout.scape.keep.pt



http://openplanets.github.io/scout/

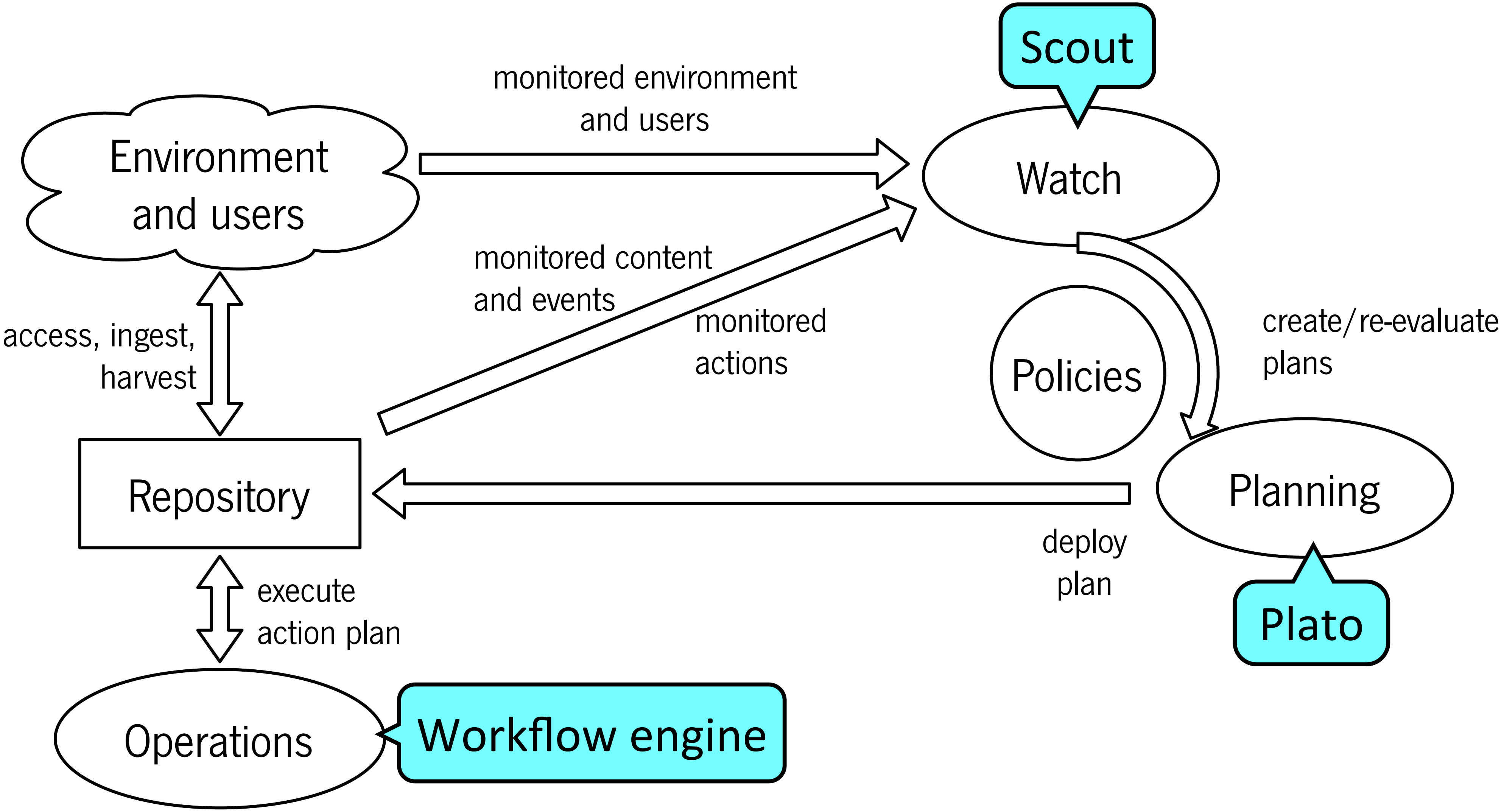
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Preservation lifecycle



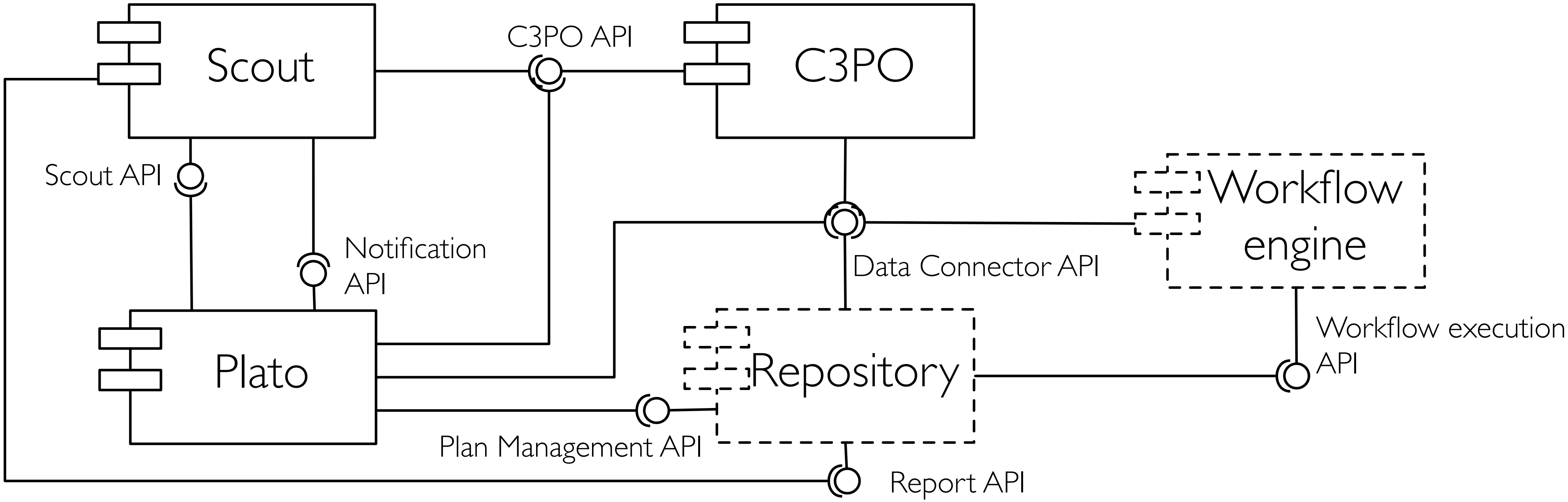
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Preservation lifecycle (in practice)



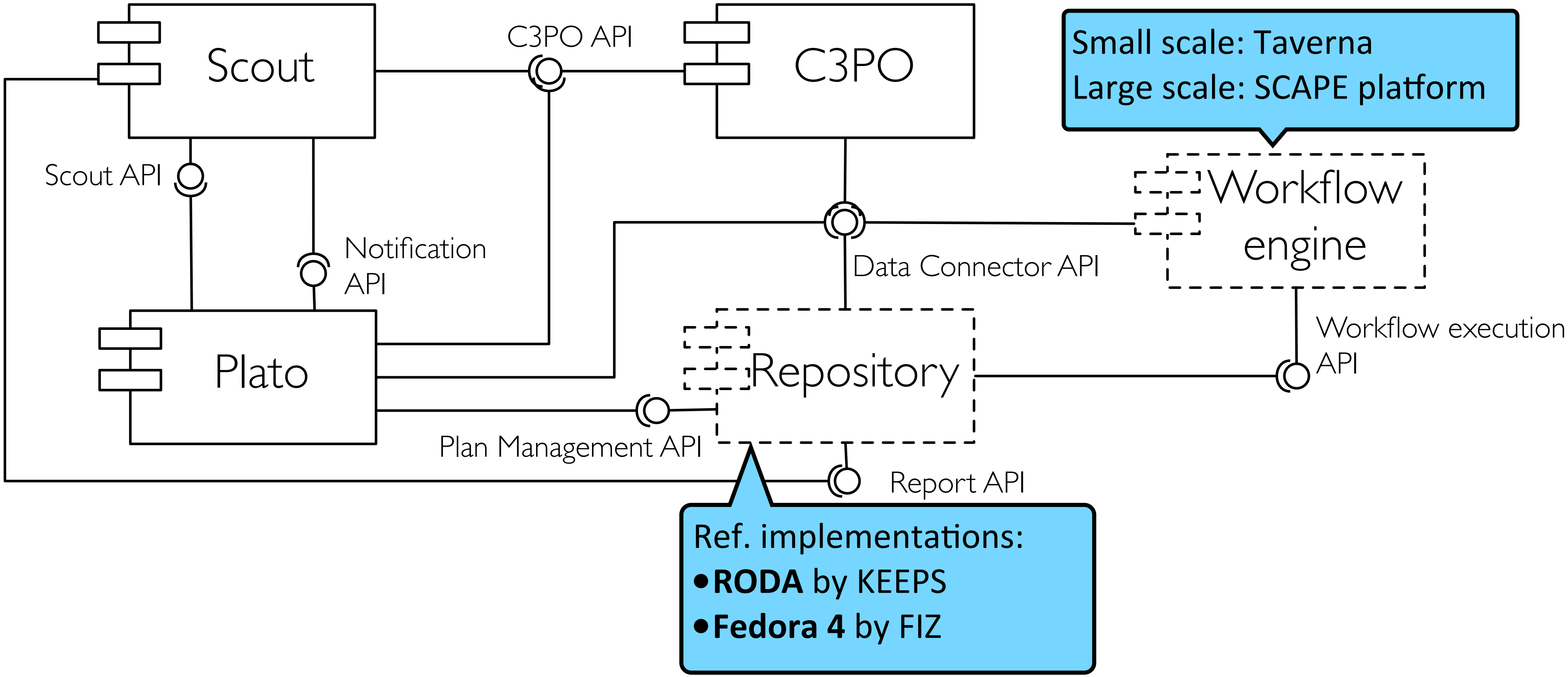
This work was partially supported by the SCAPE Project.
The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

Architecture



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).

Architecture

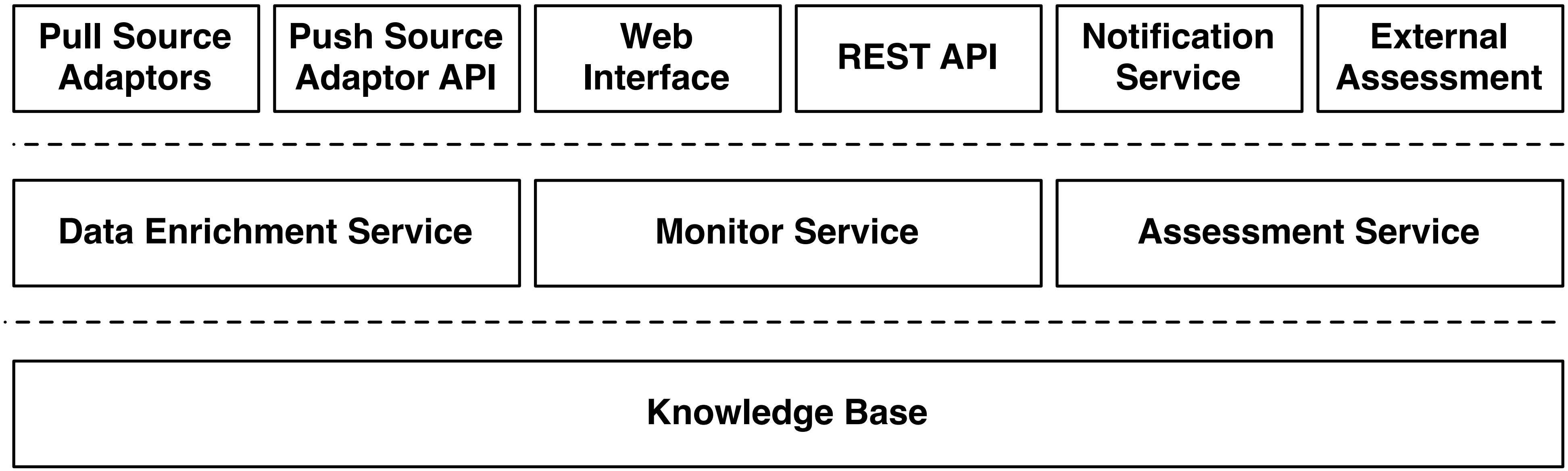


Small scale: Taverna
Large scale: SCAPE platform

Ref. implementations:
- **RODA** by KEEPS
- **Fedora 4** by FIZ

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).

Scout Architecture



Pull Source Adaptors

Push Source Adaptor API

Web Interface

REST API

Notification Service

External Assessment

Data Enrichment Service

Monitor Service

Assessment Service

Knowledge Base

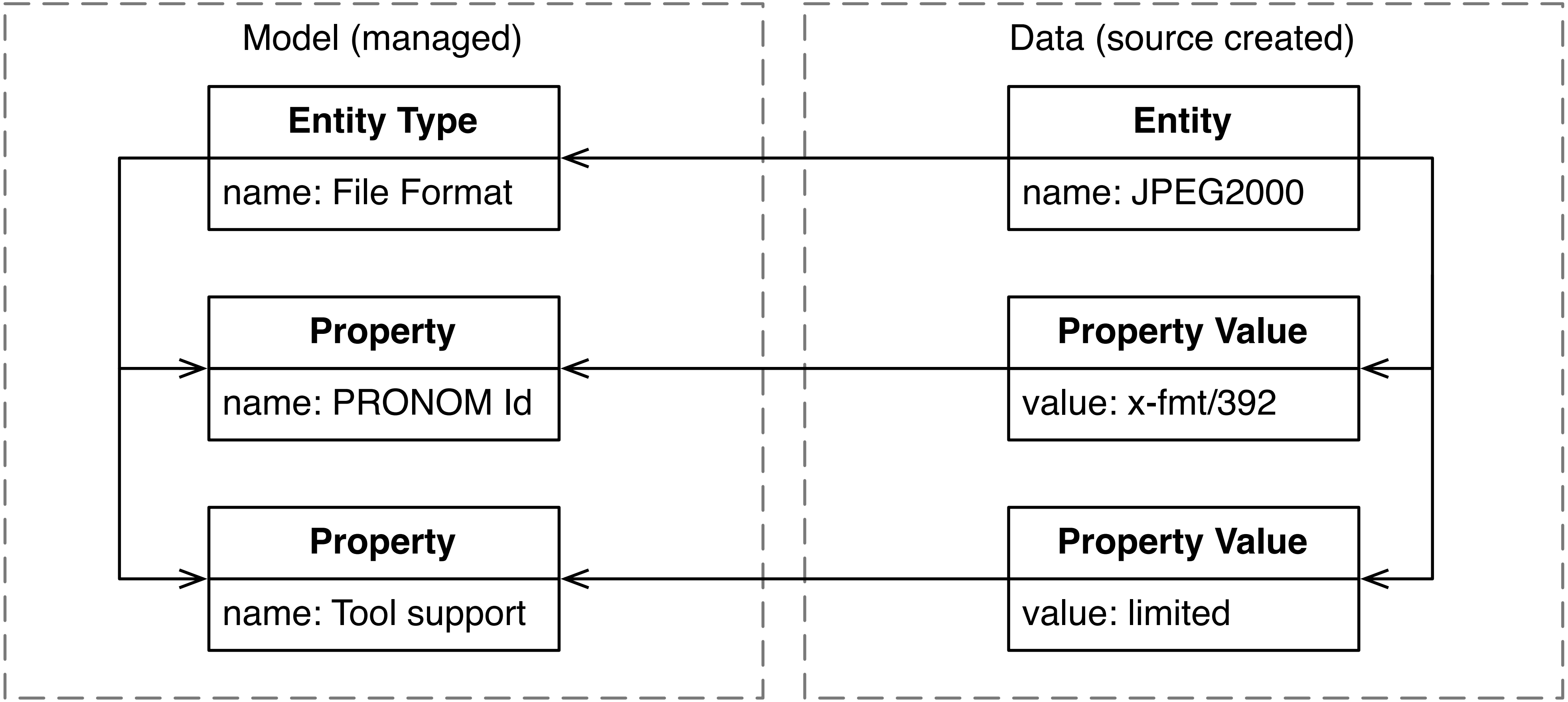
Example questions

- Are there any tools that can render the format X?
- Is my repository the only one that has format Y?
- Are my preservation plans still valid?
- Are my repository policies being enforced?

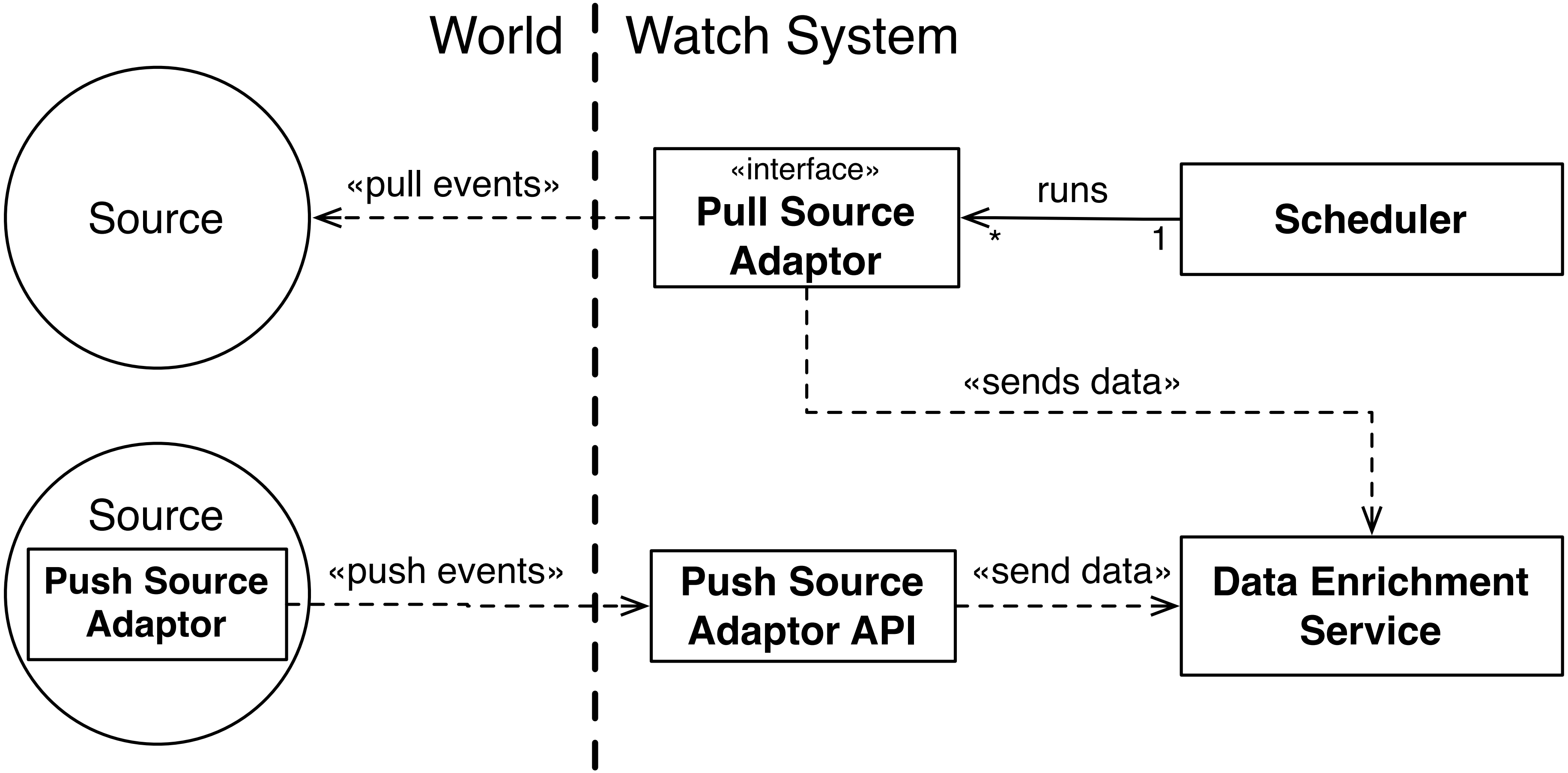
Information Sources

- Format registries & software catalogues
- Digital repositories & web archives
- Organizational objectives
- Experiments
- Simulation
- Human knowledge

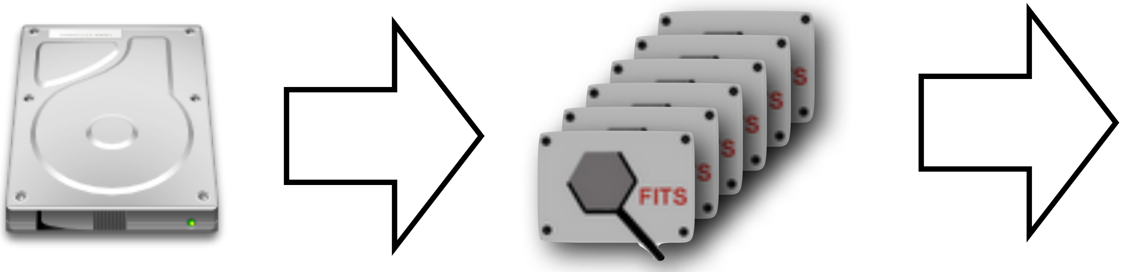
Normalized data model



Information source adaptor



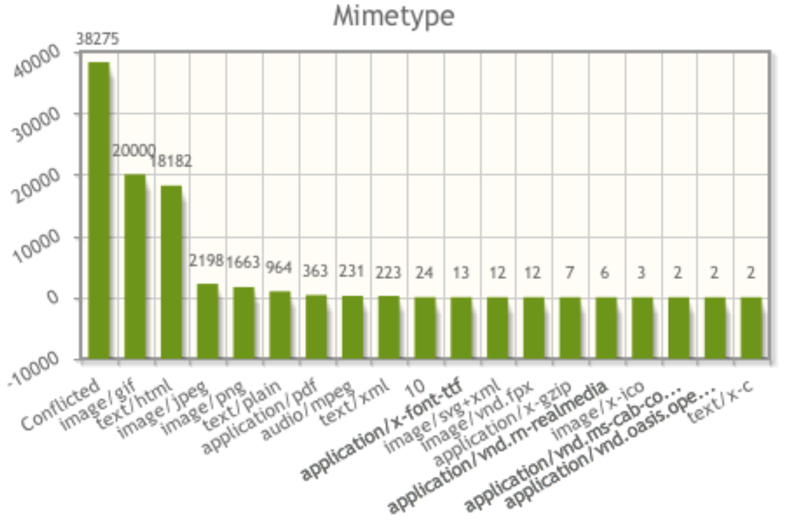
C3PO Content profile tool

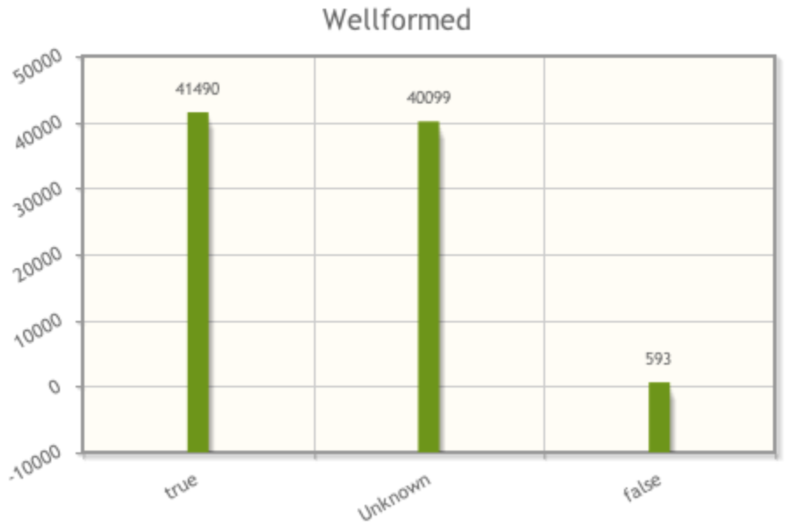


Characterization

Reports:

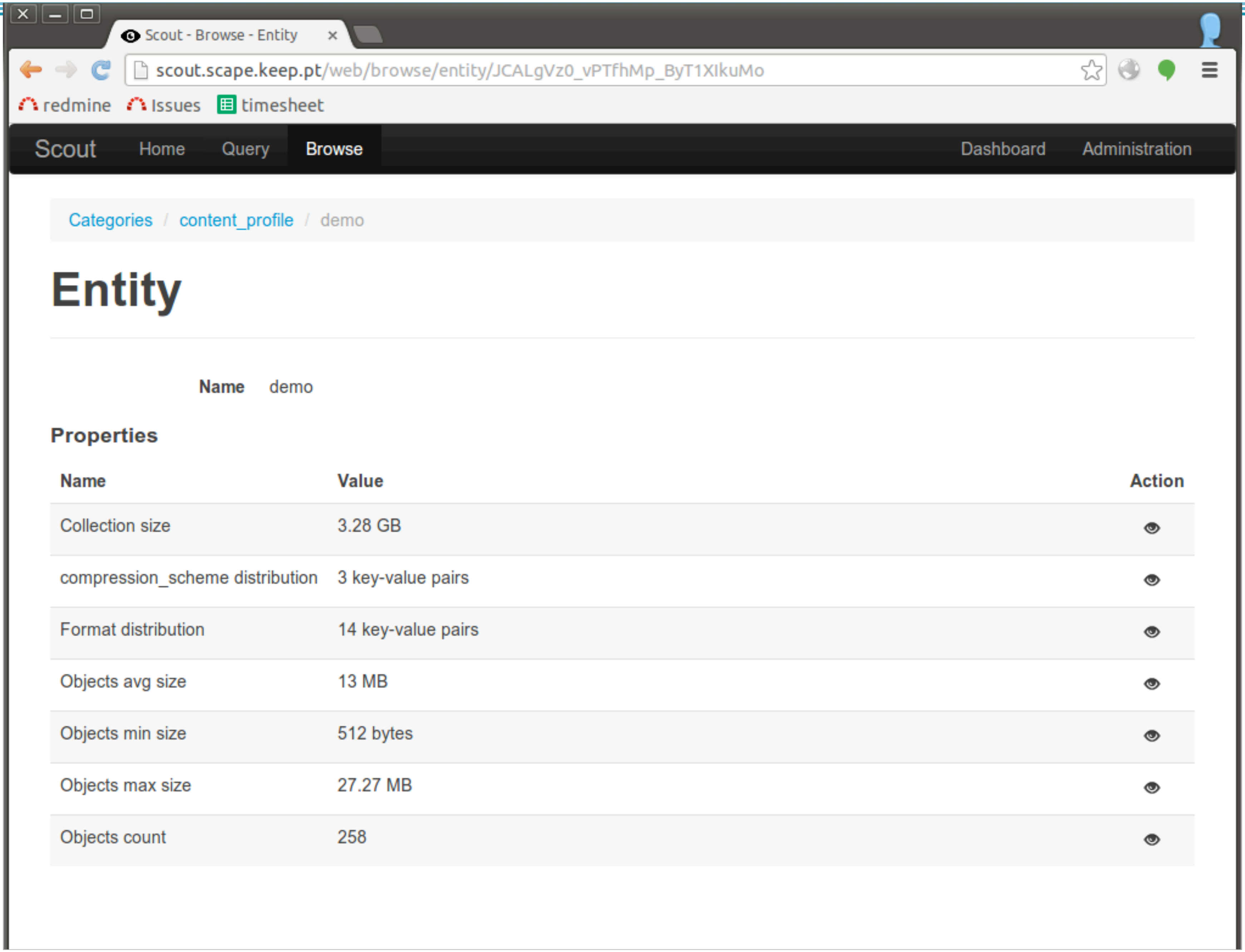
- Aggregation
- Analysis
- Representative datasets





https://github.com/peshkira/c3po

Petar Petrov <petrov@ifs.tuwien.ac.at>



Categories / content_profile / demo

Entity

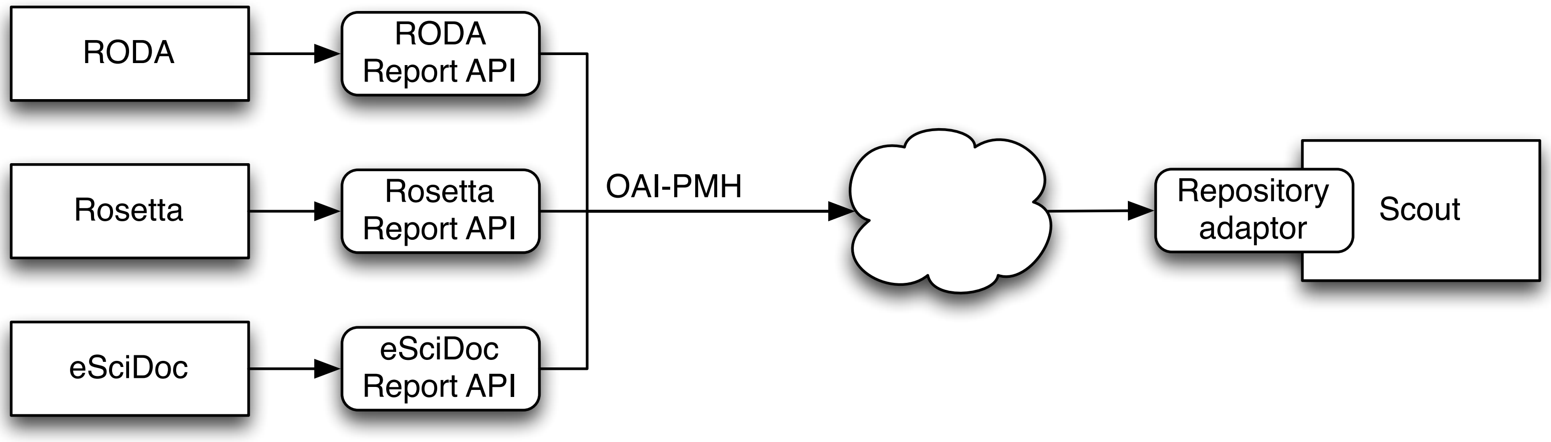
Name demo

Properties

Name	Value	Action
Collection size	3.28 GB	👁
compression_scheme distribution	3 key-value pairs	👁
Format distribution	14 key-value pairs	👁
Objects avg size	13 MB	👁
Objects min size	512 bytes	👁
Objects max size	27.27 MB	👁
Objects count	258	👁

Repository content

Repository events API and adaptor



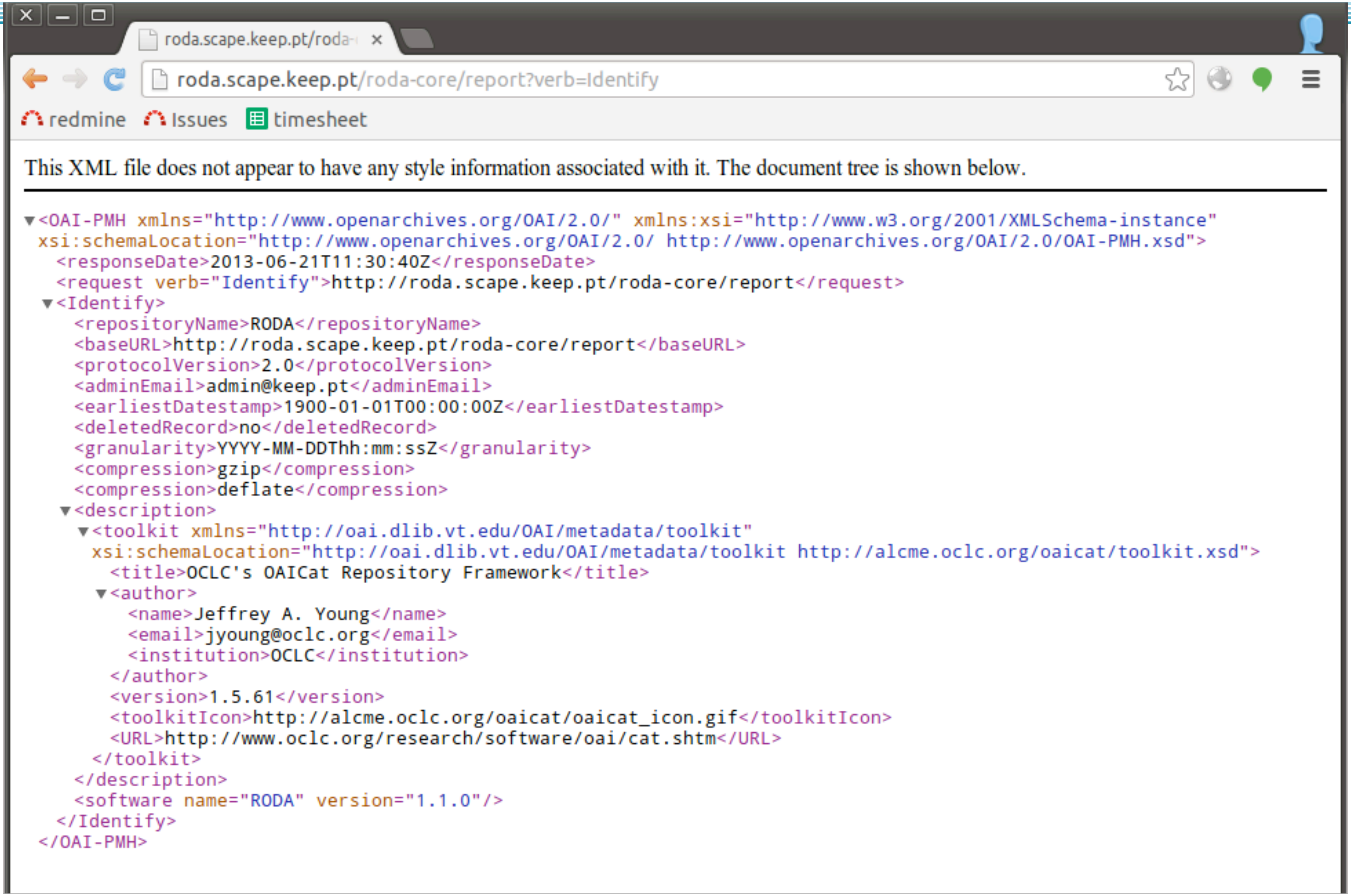
- OAI-PMH with PREMIS
- Normalize events
- Fine-grain events
- History

- Events example
  - Ingest started/ended
  - Representation downloaded
  - Plan executed

Repository Events API (Report API)

- Provides access to repository events
- Events:
  - **Ingest** started and finished
  - **Viewed** or **downloaded** descriptive metadata or representation
  - Preservation **plan executed**
- OAI-PMH data provider
- PREMIS events metadata
  - Agent: **who** triggered the event
  - Date/time: **when** did the event occur
  - Details: **what** happened
- API specification: https://github.com/openplanets/scape-platform-api
- Ref. implementation: https://github.com/openplanets/roda

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).

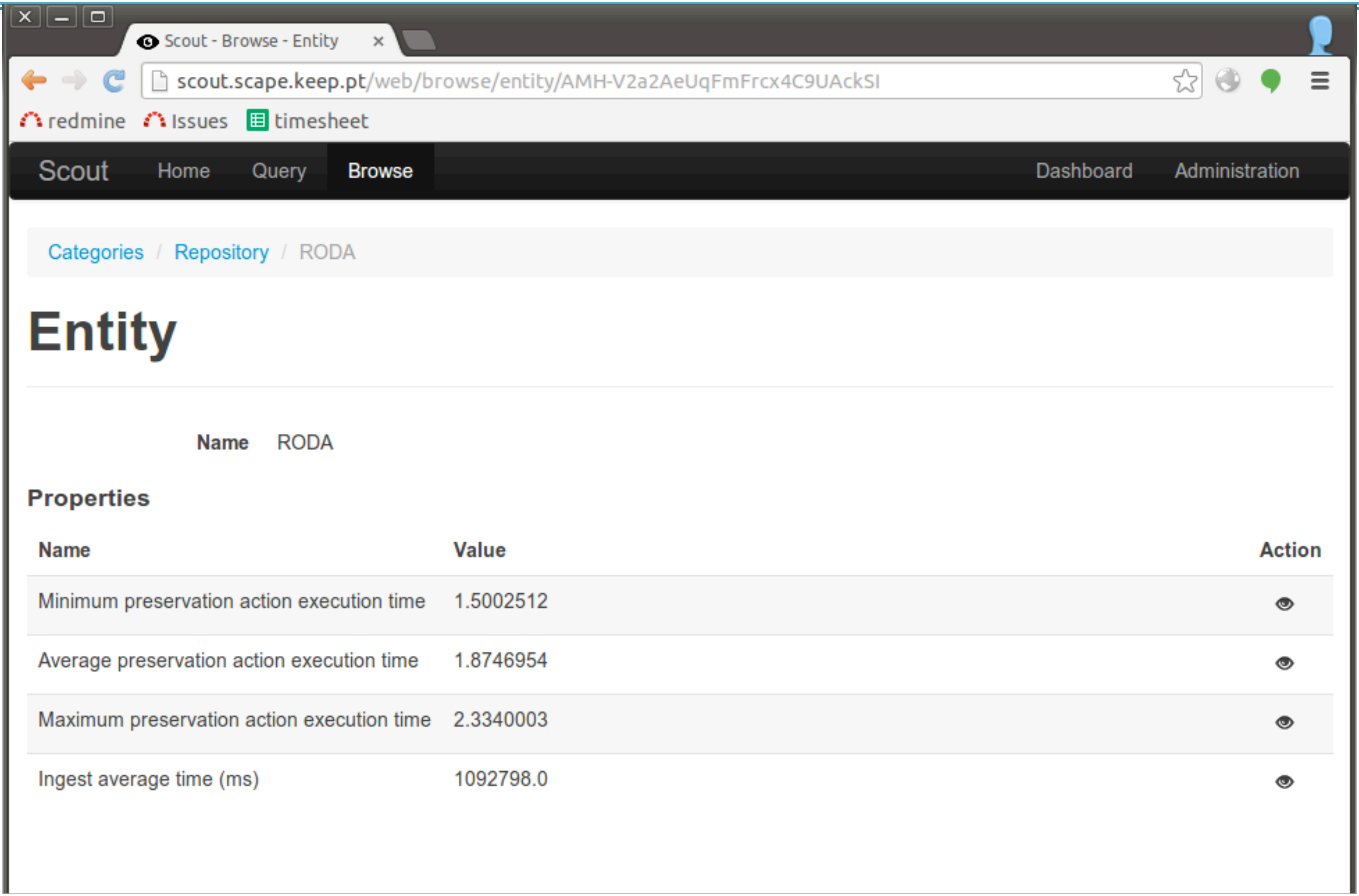


This XML file does not appear to have any style information associated with it. The document tree is shown below.

```
▼<OAI-PMH xmlns="http://www.openarchives.org/OAI/2.0/" xmlns:xsi="http://www.w3.org/2001/XMLSchema-instance"
 xsi:schemaLocation="http://www.openarchives.org/OAI/2.0/ http://www.openarchives.org/OAI/2.0/OAI-PMH.xsd">
   <responseDate>2013-06-21T11:30:40Z</responseDate>
   <request verb="Identify">http://roda.scape.keep.pt/roda-core/report</request>
  ▼<Identify>
     <repositoryName>RODA</repositoryName>
     <baseURL>http://roda.scape.keep.pt/roda-core/report</baseURL>
     <protocolVersion>2.0</protocolVersion>
     <adminEmail>admin@keep.pt</adminEmail>
     <earliestDatestamp>1900-01-01T00:00:00Z</earliestDatestamp>
     <deletedRecord>no</deletedRecord>
     <granularity>YYYY-MM-DDThh:mm:ssZ</granularity>
     <compression>gzip</compression>
     <compression>deflate</compression>
    ▼<description>
      ▼<toolkit xmlns="http://oai.dlib.vt.edu/OAI/metadata/toolkit"
       xsi:schemaLocation="http://oai.dlib.vt.edu/OAI/metadata/toolkit http://alcme.oclc.org/oaicat/toolkit.xsd">
         <title>OCLC's OAICat Repository Framework</title>
        ▼<author>
           <name>Jeffrey A. Young</name>
           <email>jyoung@oclc.org</email>
           <institution>OCLC</institution>
         </author>
         <version>1.5.61</version>
         <toolkitIcon>http://alcme.oclc.org/oaicat/oaicat_icon.gif</toolkitIcon>
         <URL>http://www.oclc.org/research/software/oai/cat.shtm</URL>
       </toolkit>
     </description>
     <software name="RODA" version="1.1.0"/>
   </Identify>
 </OAI-PMH>
```

Report API (RODA reference implementation)

https://github.com/openplanets/roda/tree/master/roda-core/roda-core-services/src/main/java/eu/scape_project/roda/report



Categories / Repository / RODA

Entity

Name RODA

Properties

Name	Value	Action
Minimum preservation action execution time	1.5002512	👁
Average preservation action execution time	1.8746954	👁
Maximum preservation action execution time	2.3340003	👁
Ingest average time (ms)	1092798.0	👁

Repository events

Web archive adaptor

Content via C3PO

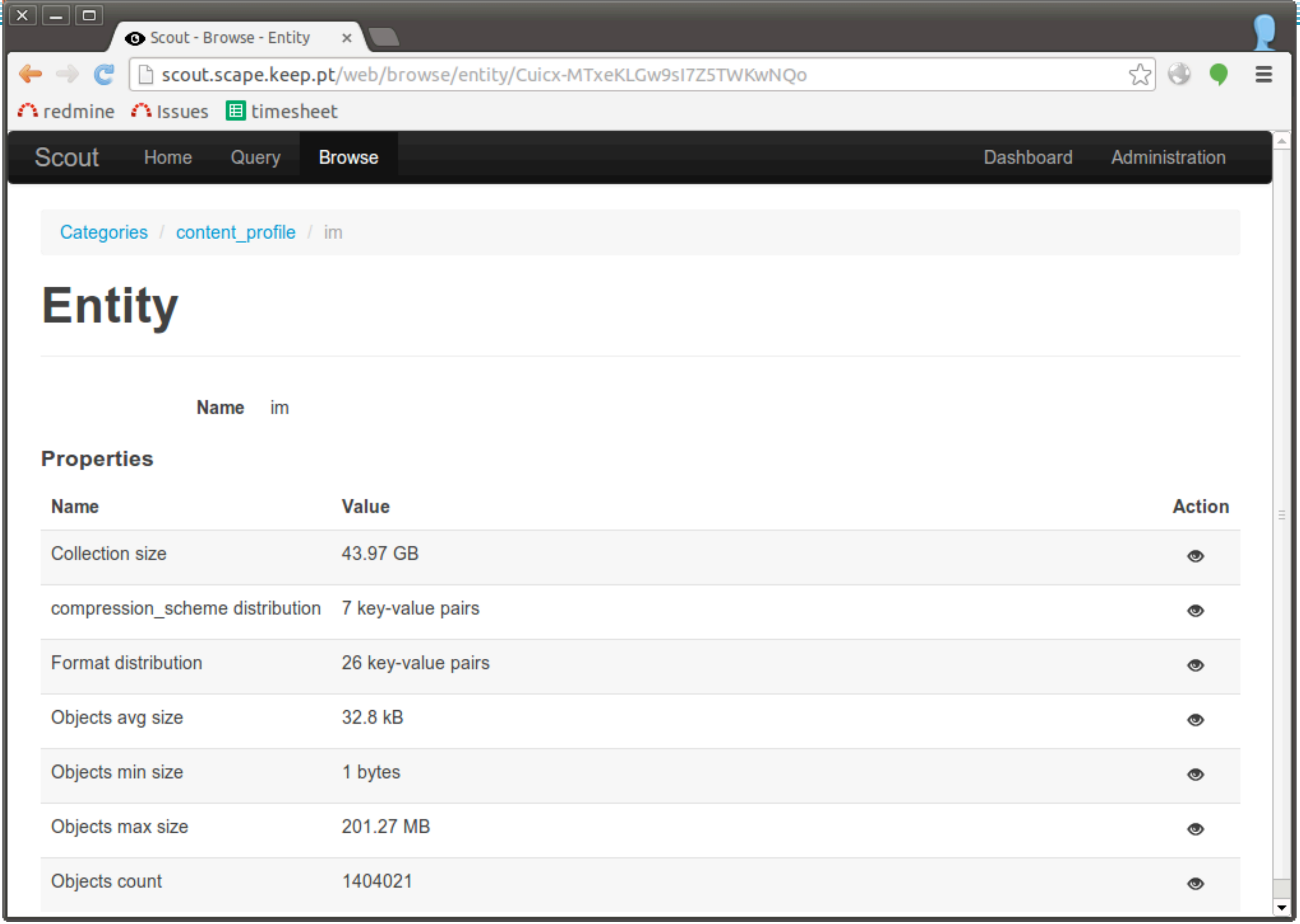
- IM-C3PO prototype integration (2010-2012)
- SB-C3PO prototype integration (large-scale: 300 TB)

Renderability analysis experiments:

- Browser snapshots comparison large-scale platform prototype
- C3PO adaptor for experiment results
- Scout C3PO adaptor profile support

Other web characterization info:

- ARC header extractor (ongoing)



Categories / content_profile / im

Entity

Name im

Properties

Name	Value	Action
Collection size	43.97 GB	👁
compression_scheme distribution	7 key-value pairs	👁
Format distribution	26 key-value pairs	👁
Objects avg size	32.8 kB	👁
Objects min size	1 bytes	👁
Objects max size	201.27 MB	👁
Objects count	1404021	👁

Web archive content (IM)

Web archive content (SB)

12 TB, 440M FITS files

Test case 1 - Import

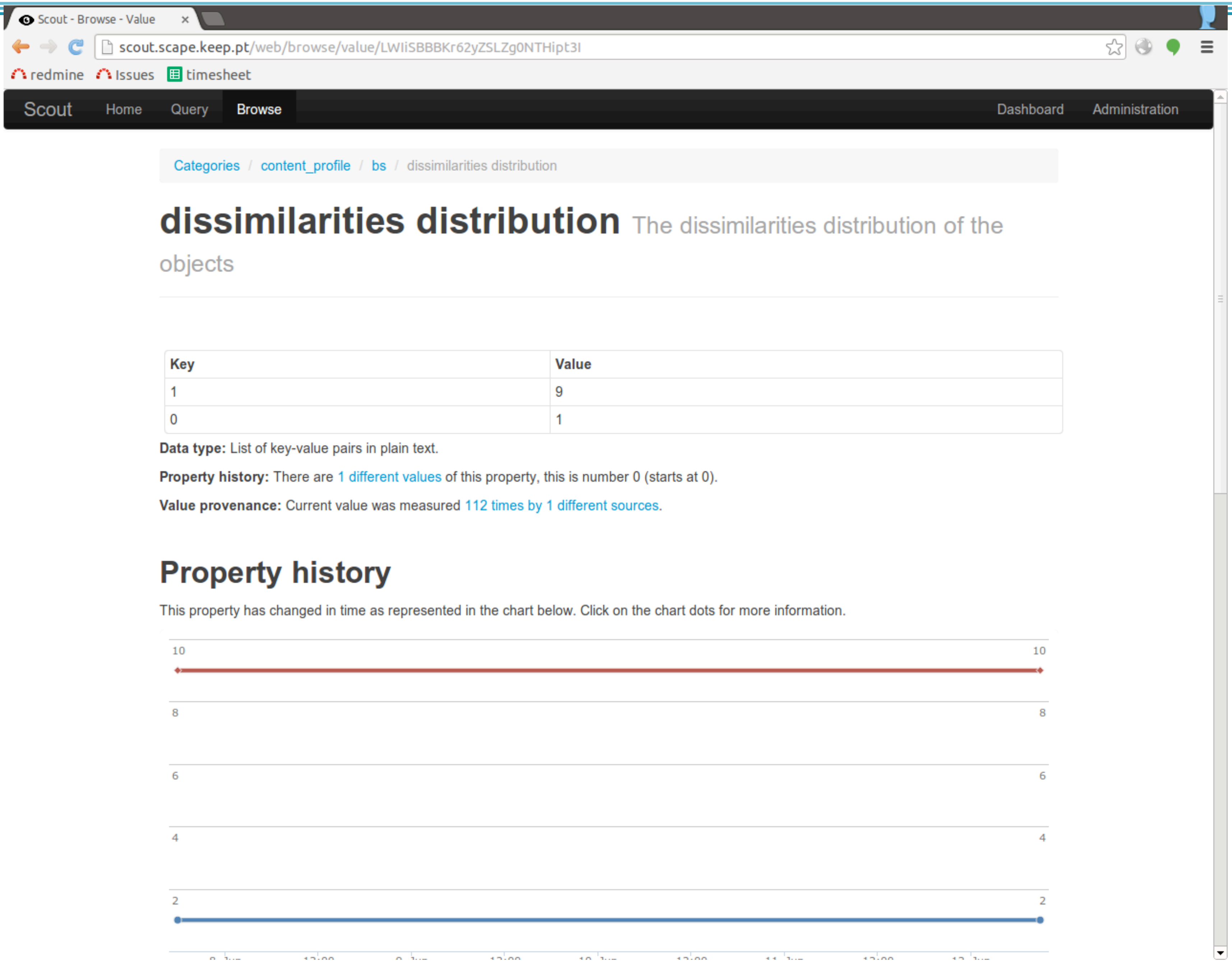
- Linear ingest time of 0.65 ms for FITS file

Test case 2 - GUI

- 2.5 million FITS files limit

Generate profile in command-line

- 15 hours for 12M files



Categories / content_profile / bs / dissimilarities distribution

dissimilarities distribution The dissimilarities distribution of the objects

Key	Value
1	9
0	1

Data type: List of key-value pairs in plain text.

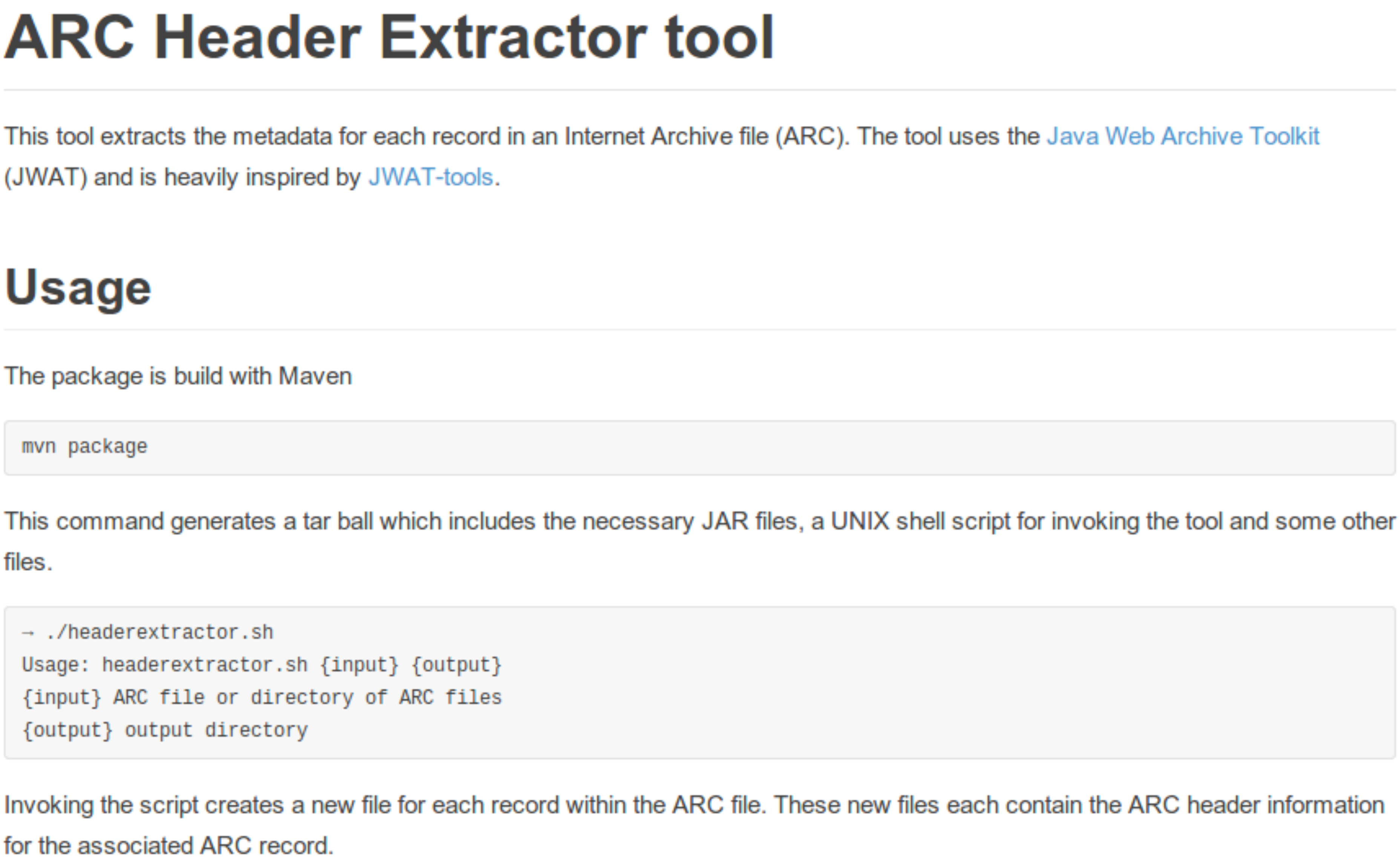
Property history: There are 1 different values of this property, this is number 0 (starts at 0).

Value provenance: Current value was measured 112 times by 1 different sources.

Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.

Web archive renderability analysis



ARC Header Extractor tool

This tool extracts the metadata for each record in an Internet Archive file (ARC). The tool uses the Java Web Archive Toolkit (JWAT) and is heavily inspired by JWAT-tools.

Usage

The package is build with Maven

```
mvn package
```

This command generates a tar ball which includes the necessary JAR files, a UNIX shell script for invoking the tool and some other files.

```
→ ./headerextractor.sh
Usage: headerextractor.sh {input} {output}
{input} ARC file or directory of ARC files
{output} output directory
```

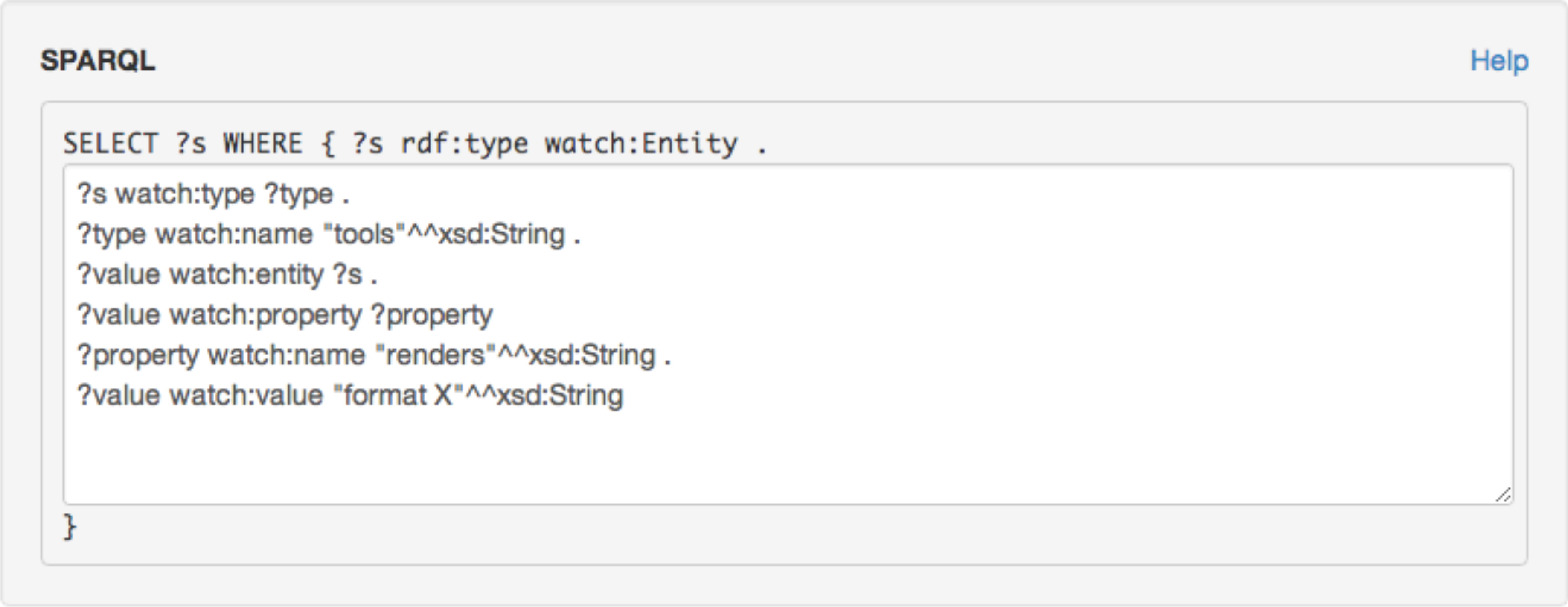
Invoking the script creates a new file for each record within the ARC file. These new files each contain the ARC header information for the associated ARC record.

ARC Header extractor tool

https://github.com/statsbiblioteket/arc-header-extractor

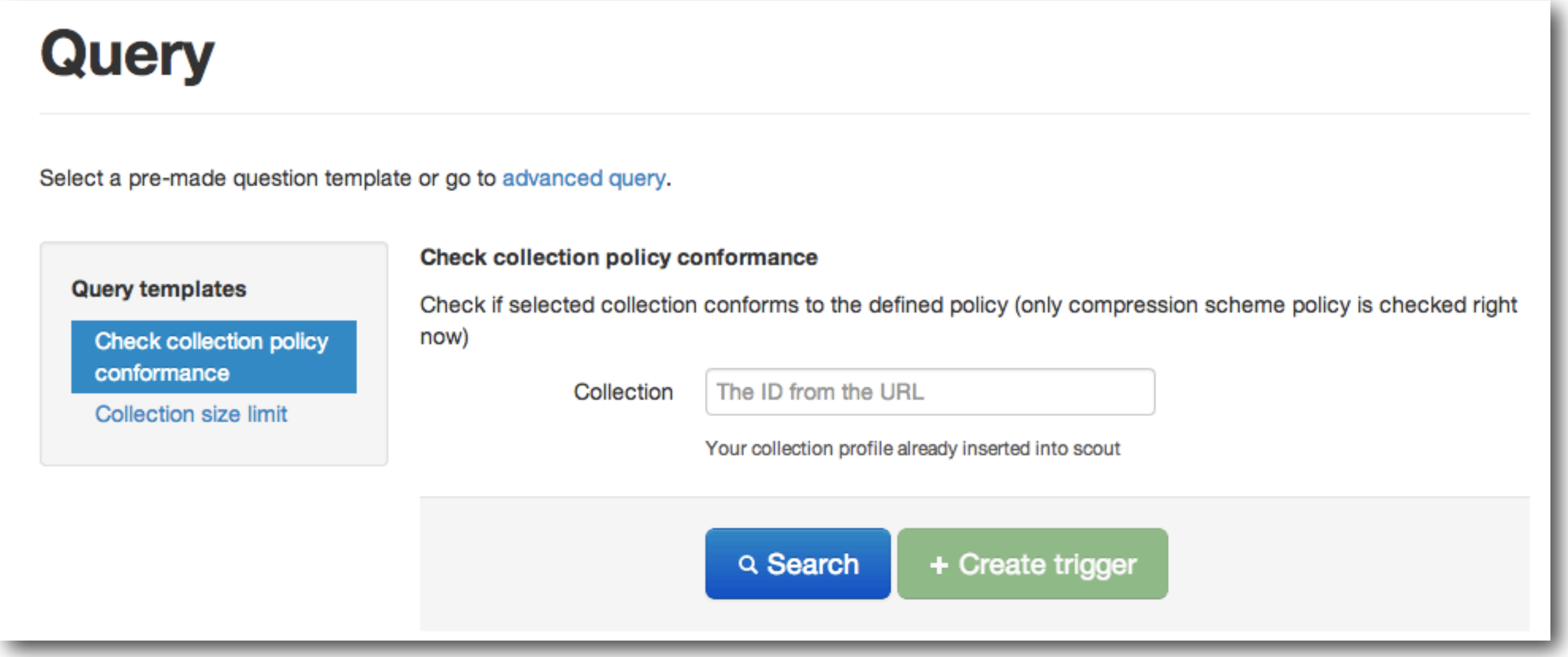
Define triggers

- Notify me when there are tools that can render the format X.



```
SELECT ?s WHERE { ?s rdf:type watch:Entity .
?s watch:type ?type .
?type watch:name "tools"^^xsd:String .
?value watch:entity ?s .
?value watch:property ?property
?property watch:name "renders"^^xsd:String .
?value watch:value "format X"^^xsd:String
}
```

Define triggers
Simple query with templates



Query

Select a pre-made question template or go to advanced query.

Query templates

Check collection policy conformance

Collection size limit

Check collection policy conformance

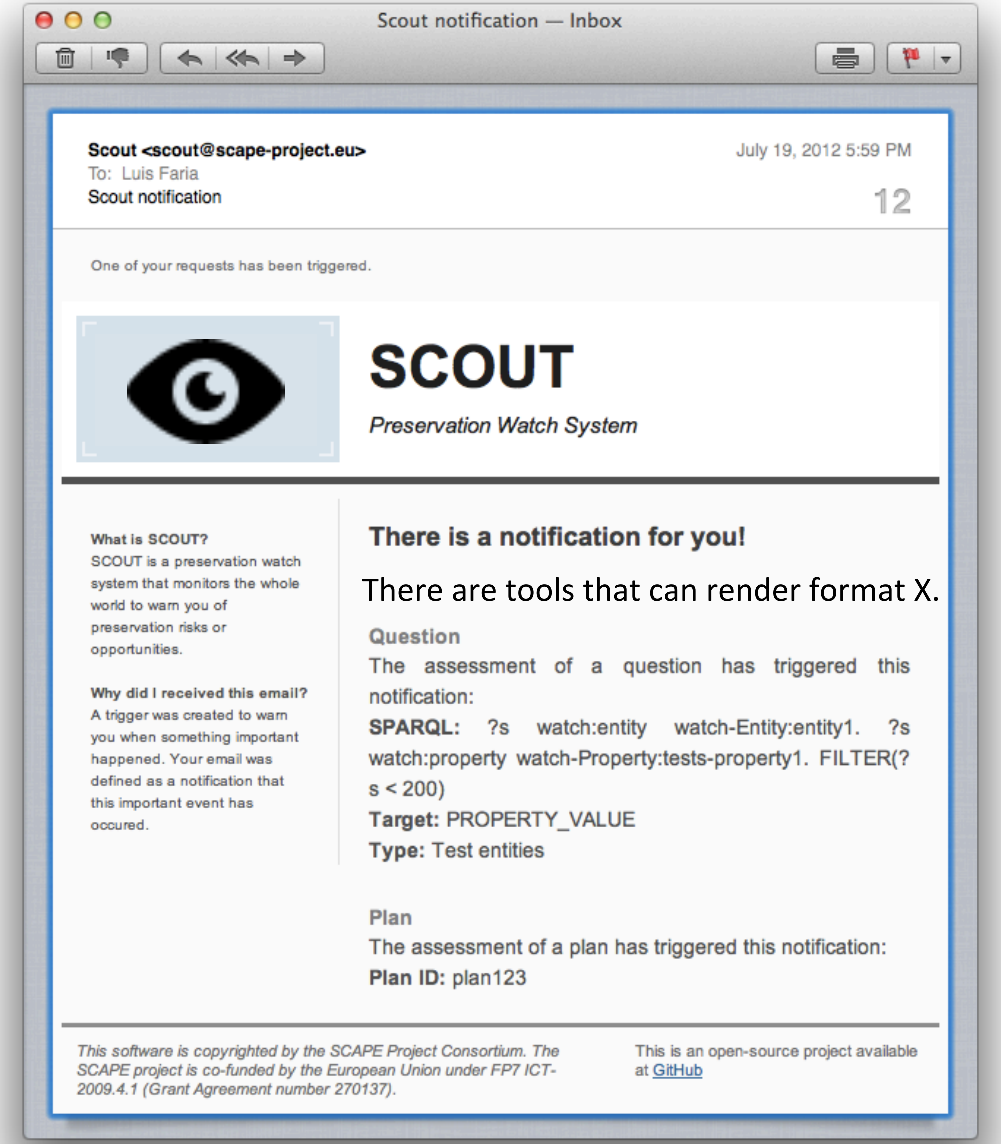
Check if selected collection conforms to the defined policy (only compression scheme policy is checked right now)

Collection: The ID from the URL

Your collection profile already inserted into scout

Receive notifications

Email



Scout <scout@scape-project.eu>
To: Luis Faria
Scout notification

July 19, 2012 5:59 PM
12

One of your requests has been triggered.

SCOUT

Preservation Watch System

What is SCOUT?
SCOUT is a preservation watch system that monitors the whole world to warn you of preservation risks or opportunities.

Why did I received this email?
A trigger was created to warn you when something important happened. Your email was defined as a notification that this important event has occured.

There is a notification for you!

There are tools that can render format X.

Question
The assessment of a question has triggered this notification:
SPARQL: ?s watch:entity watch-Entity:entity1. ?s watch:property watch-Property:tests-property1. FILTER(?s < 200)
Target: PROPERTY_VALUE
Type: Test entities

Plan
The assessment of a plan has triggered this notification:
Plan ID: plan123

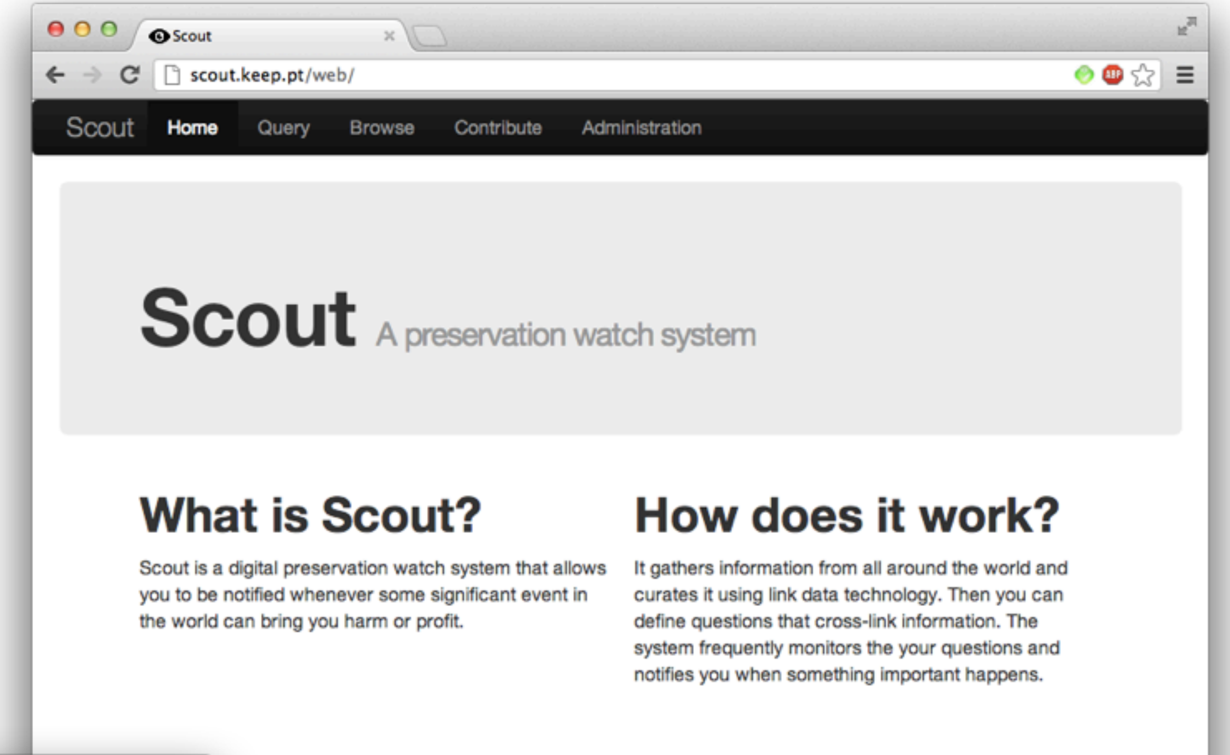
This software is copyrighted by the SCAPE Project Consortium. The SCAPE project is co-funded by the European Union under FP7 ICT-2009.4.1 (Grant Agreement number 270137).

This is an open-source project available at GitHub

HTTP Push API

Interfaces

Web page



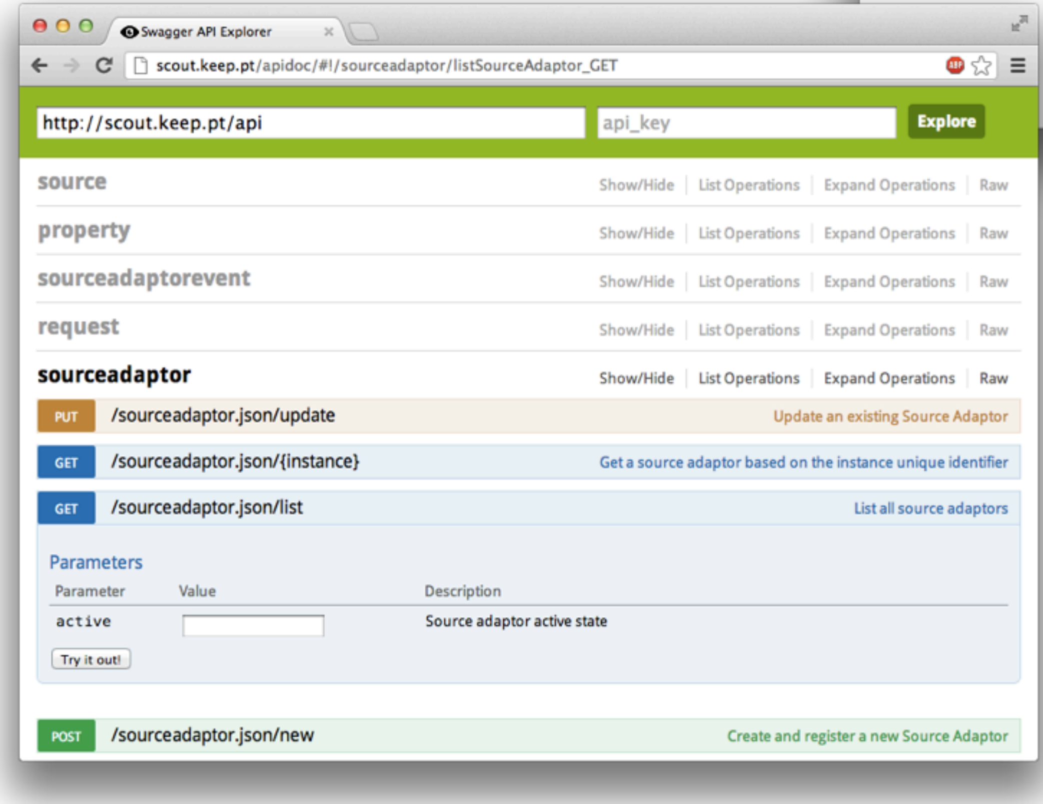
Scout A preservation watch system

What is Scout?

Scout is a digital preservation watch system that allows you to be notified whenever some significant event in the world can bring you harm or profit.

How does it work?

It gathers information from all around the world and curates it using link data technology. Then you can define questions that cross-link information. The system frequently monitors the your questions and notifies you when something important happens.



source Show/Hide | List Operations | Expand Operations | Raw

property Show/Hide | List Operations | Expand Operations | Raw

sourceadaptorevent Show/Hide | List Operations | Expand Operations | Raw

request Show/Hide | List Operations | Expand Operations | Raw

sourceadaptor Show/Hide | List Operations | Expand Operations | Raw

PUT /sourceadaptor.json/update — Update an existing Source Adaptor

GET /sourceadaptor.json/{instance} — Get a source adaptor based on the instance unique identifier

GET /sourceadaptor.json/list — List all source adaptors

Parameters

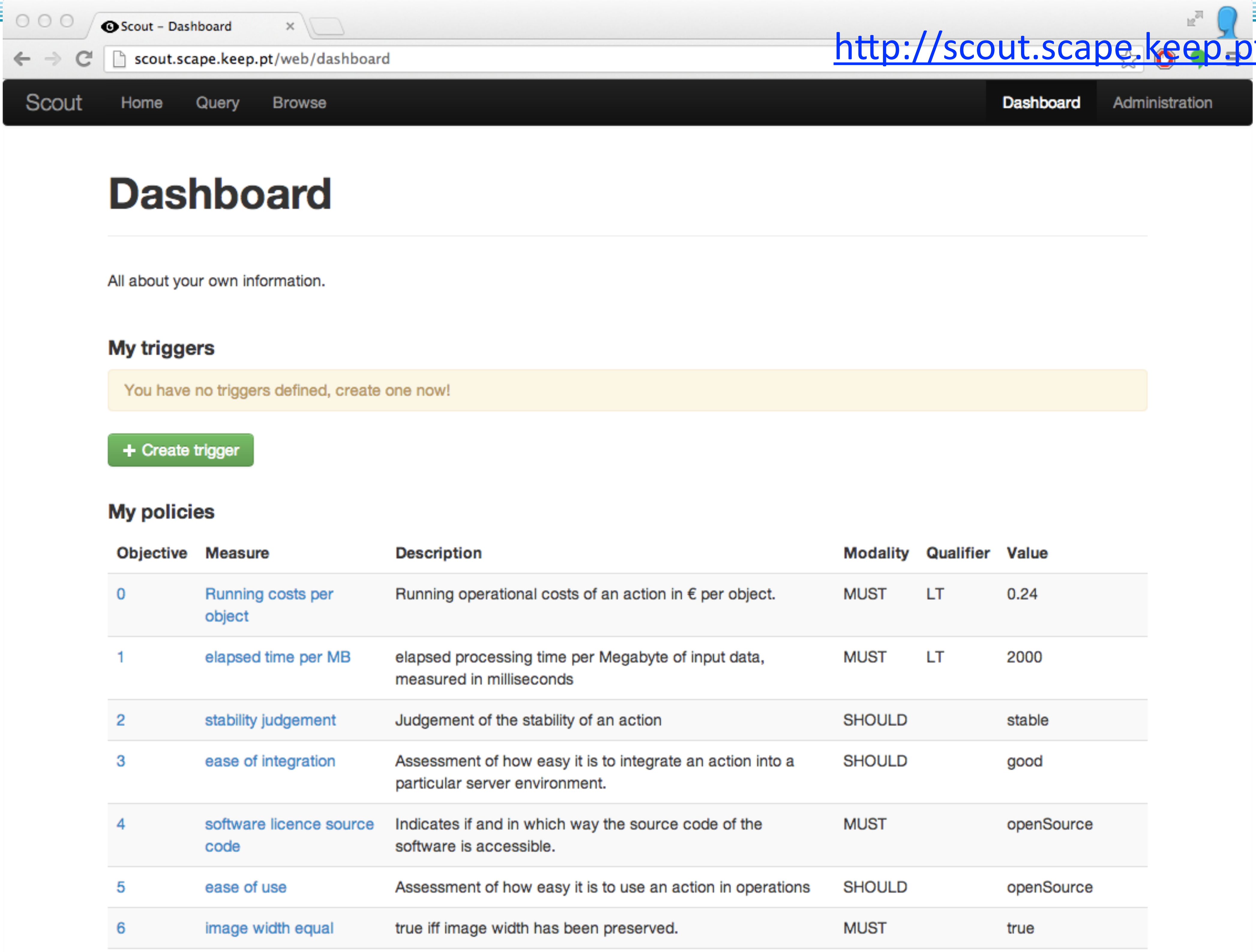
Parameter	Value	Description
active		Source adaptor active state

Try it out!

POST /sourceadaptor.json/new — Create and register a new Source Adaptor

REST API

http://scout.scape.keep.pt



Dashboard

All about your own information.

My triggers

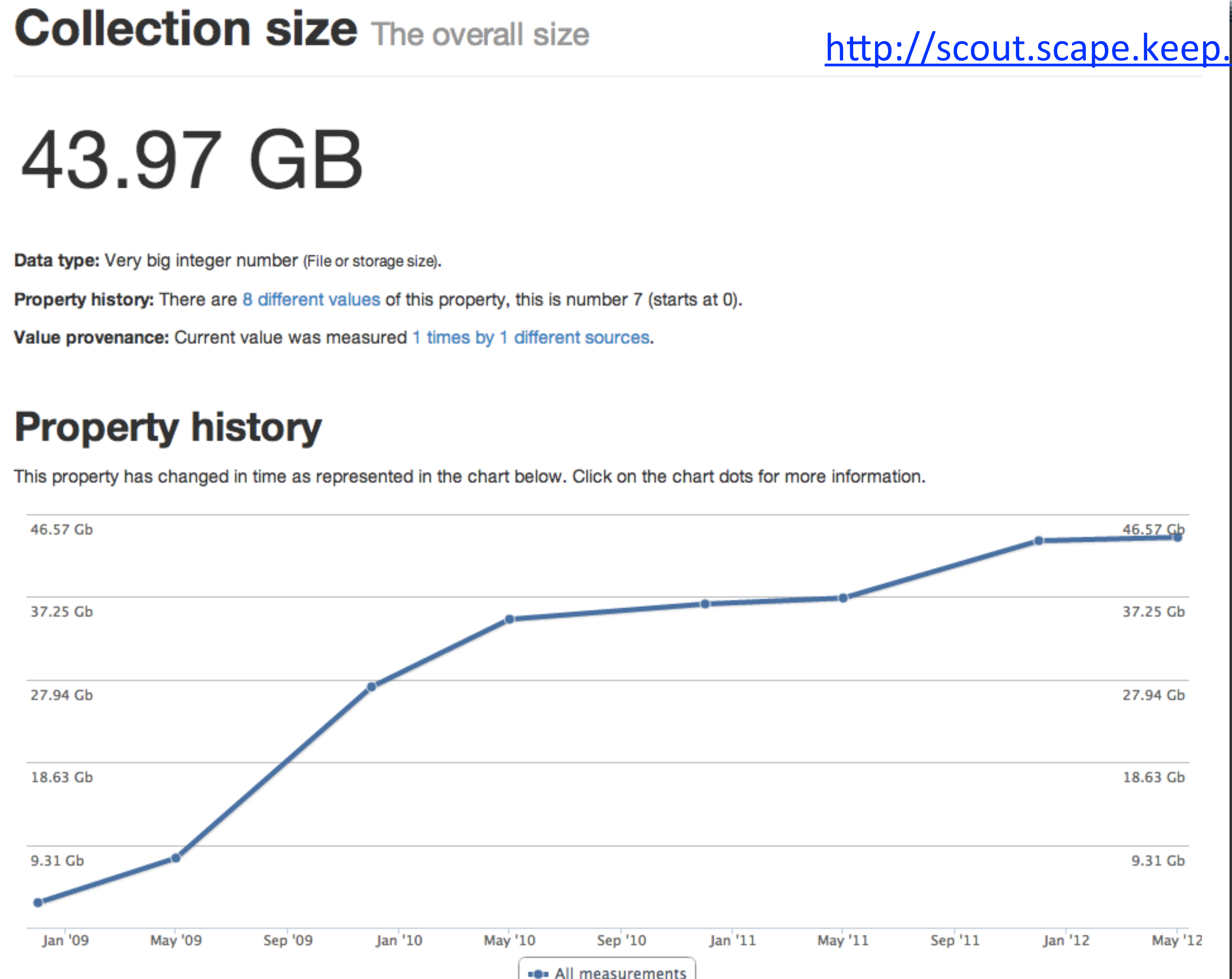
You have no triggers defined, create one now!

+ Create trigger

My policies

Objective	Measure	Description	Modality	Qualifier	Value
0	Running costs per object	Running operational costs of an action in € per object.	MUST	LT	0.24
1	elapsed time per MB	elapsed processing time per Megabyte of input data, measured in milliseconds	MUST	LT	2000
2	stability judgement	Judgement of the stability of an action	SHOULD		stable
3	ease of integration	Assessment of how easy it is to integrate an action into a particular server environment.	SHOULD		good
4	software licence source code	Indicates if and in which way the source code of the software is accessible.	MUST		openSource
5	ease of use	Assessment of how easy it is to use an action in operations	SHOULD		openSource
6	image width equal	true iff image width has been preserved.	MUST		true

http://scout.scape.keep.pt



Collection size The overall size

43.97 GB

Data type: Very big integer number (File or storage size).

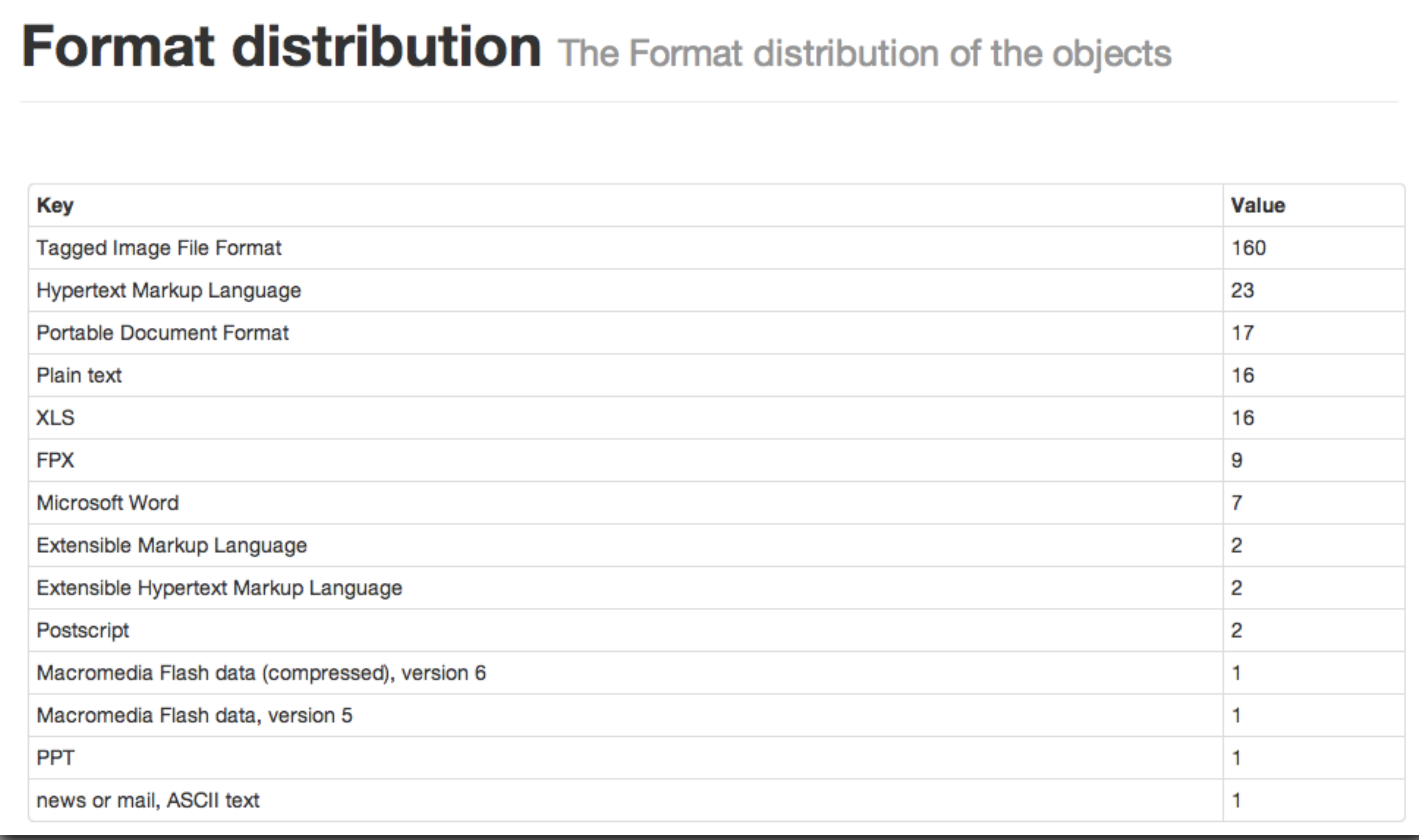
Property history: There are 8 different values of this property, this is number 7 (starts at 0).

Value provenance: Current value was measured 1 times by 1 different sources.

Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.

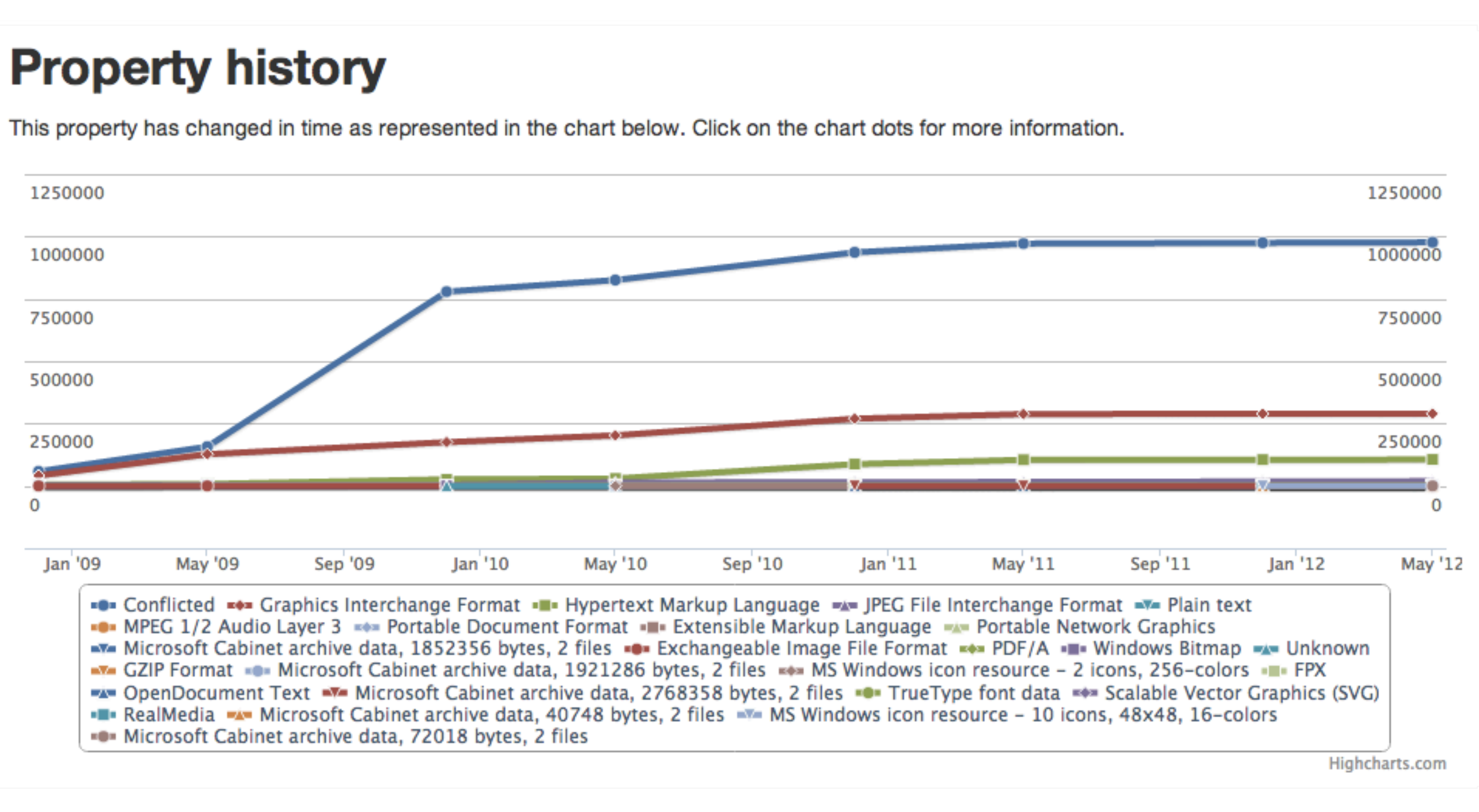
http://scout.scape.keep.pt



Format distribution The Format distribution of the objects

Key	Value
Tagged Image File Format	160
Hypertext Markup Language	23
Portable Document Format	17
Plain text	16
XLS	16
FPX	9
Microsoft Word	7
Extensible Markup Language	2
Extensible Hypertext Markup Language	2
Postscript	2
Macromedia Flash data (compressed), version 6	1
Macromedia Flash data, version 5	1
PPT	1
news or mail, ASCII text	1

http://scout.scape.keep.pt

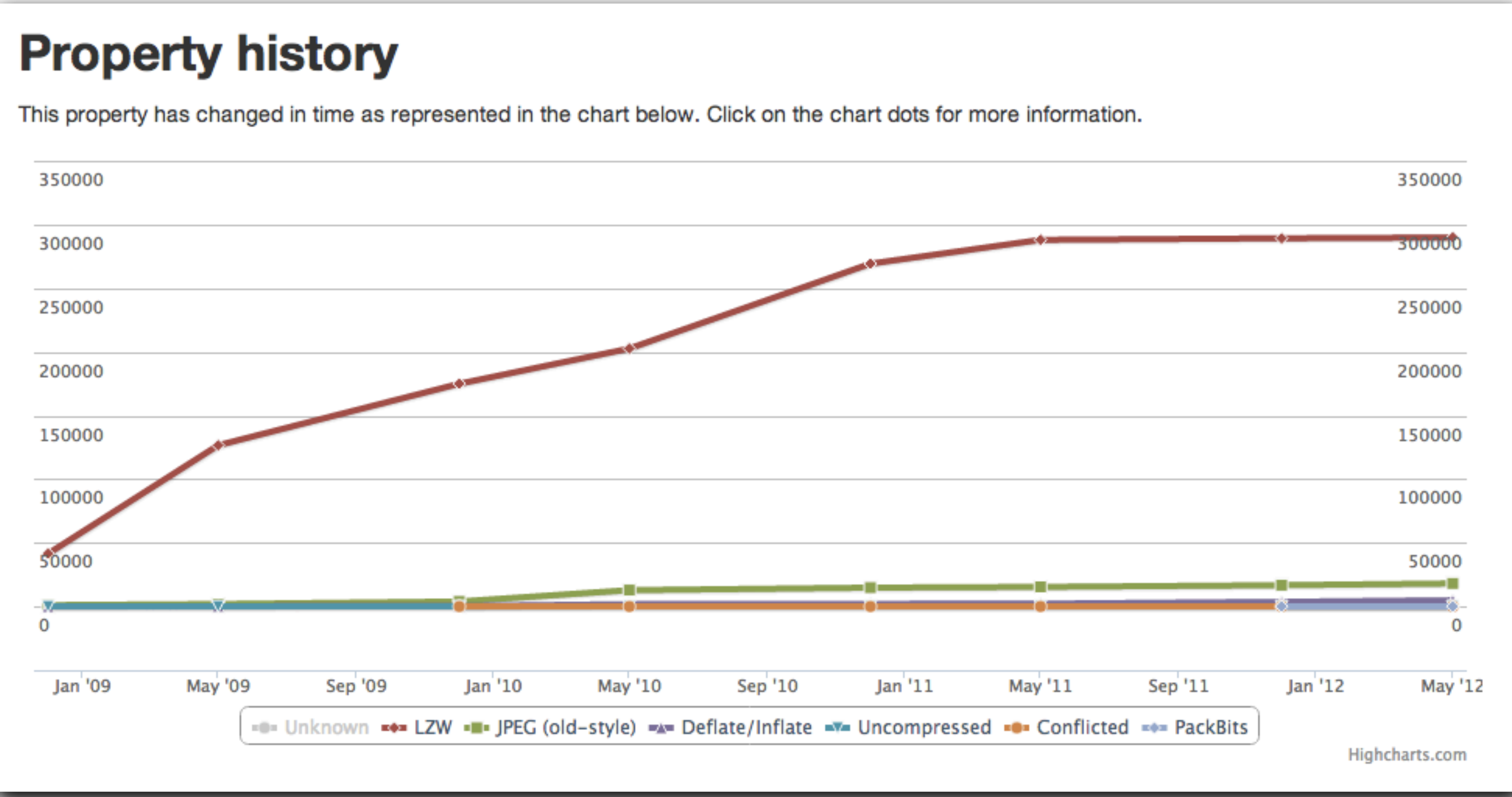


Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.

1250000
1000000
750000
500000
250000
0
Jan '09
May '09
Sep '09
Jan '10
May '10
Sep '10
Jan '11
May '11
Sep '11
Jan '12
May '12
Conflicted
Graphics Interchange Format
Hypertext Markup Language
JPEG File Interchange Format
Plain text
MPEG 1/2 Audio Layer 3
Portable Document Format
Extensible Markup Language
Portable Network Graphics
Microsoft Cabinet archive data, 1852356 bytes, 2 files
Exchangeable Image File Format
PDF/A
Windows Bitmap
Unknown
GZIP Format
Microsoft Cabinet archive data, 1921286 bytes, 2 files
MS Windows icon resource - 2 icons, 256-colors
FPX
OpenDocument Text
Microsoft Cabinet archive data, 2768358 bytes, 2 files
TrueType font data
Scalable Vector Graphics (SVG)
RealMedia
Microsoft Cabinet archive data, 40748 bytes, 2 files
MS Windows icon resource - 10 icons, 48x48, 16-colors
Microsoft Cabinet archive data, 72018 bytes, 2 files
Highcharts.com

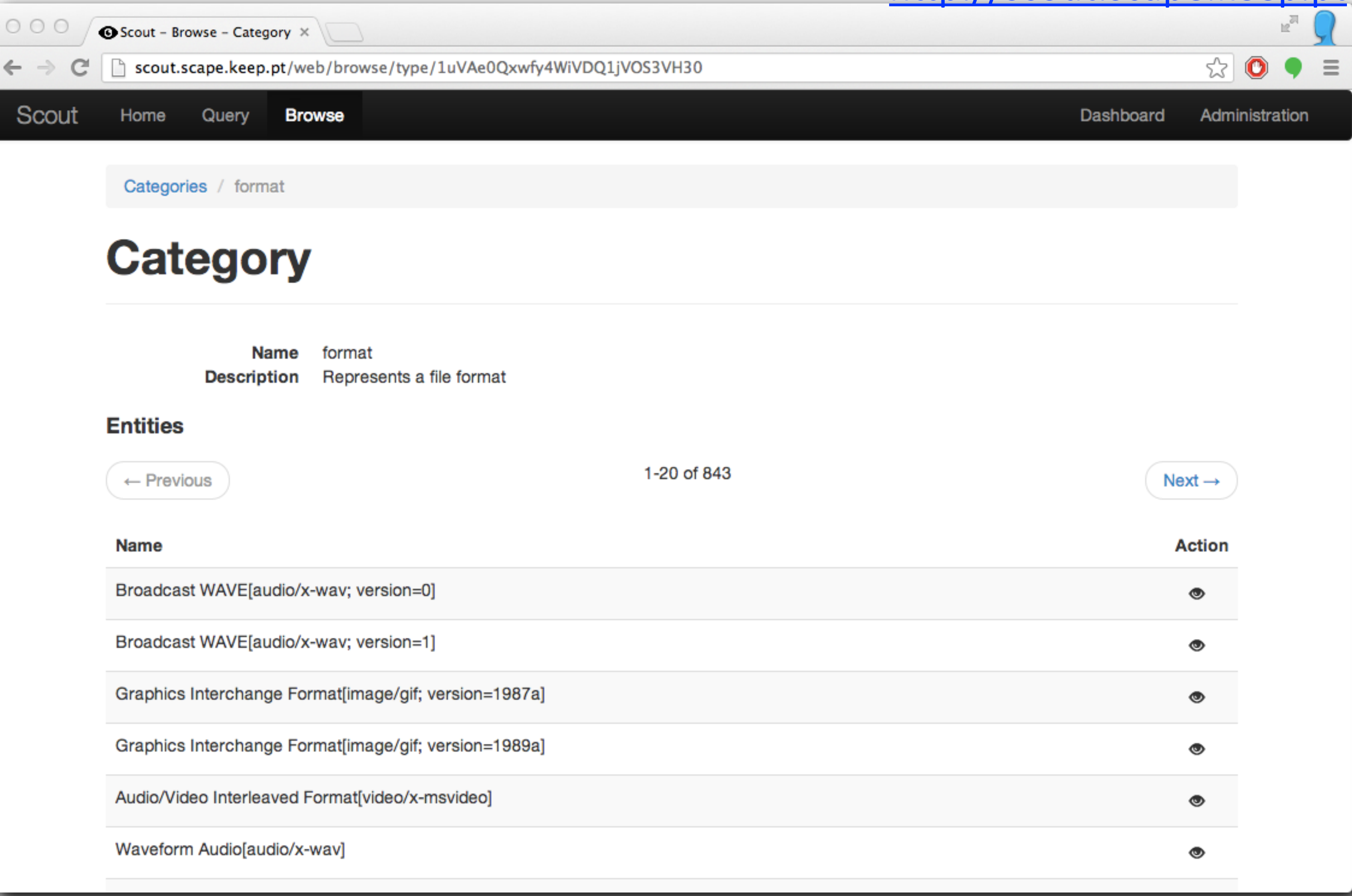
http://scout.scape.keep.pt



Property history

This property has changed in time as represented in the chart below. Click on the chart dots for more information.

http://scout.scape.keep.pt


Scout - Browse - Category
scout.scape.keep.pt/web/browse/type/1uVAe0Qxwfy4WiVDQ1jVOS3VH30
Scout
Home
Query
Browse
Dashboard
Administration

Categories / format

Category

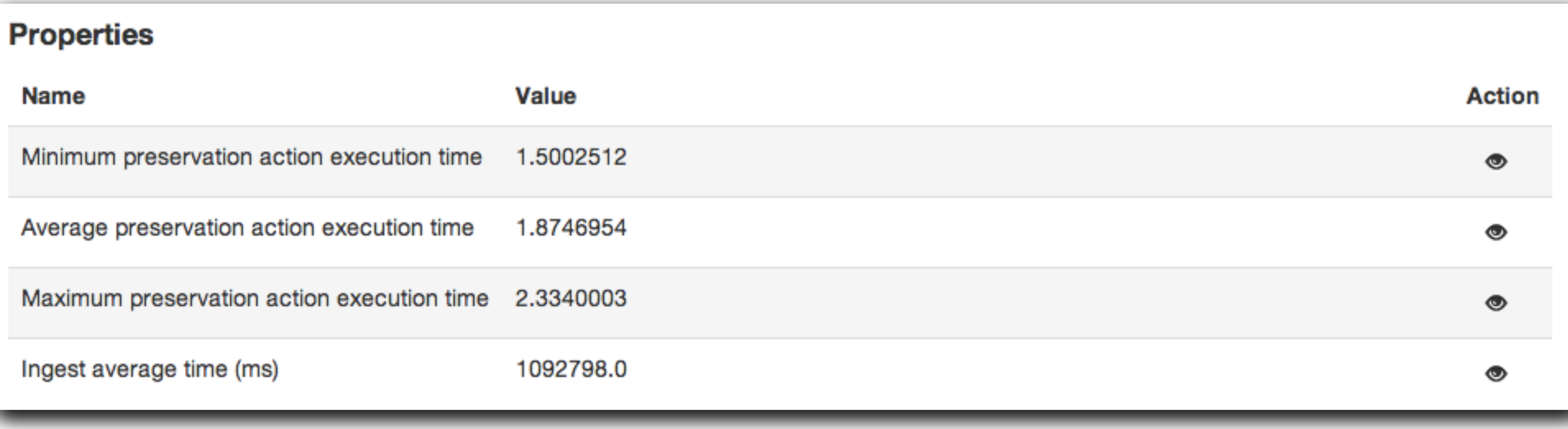
Name format
Description Represents a file format

Entities

← Previous 1-20 of 843 Next →

Name	Action
Broadcast WAVE[audio/x-wav; version=0]	👁
Broadcast WAVE[audio/x-wav; version=1]	👁
Graphics Interchange Format[image/gif; version=1987a]	👁
Graphics Interchange Format[image/gif; version=1989a]	👁
Audio/Video Interleaved Format[video/x-msvideo]	👁
Waveform Audio[audio/x-wav]	👁

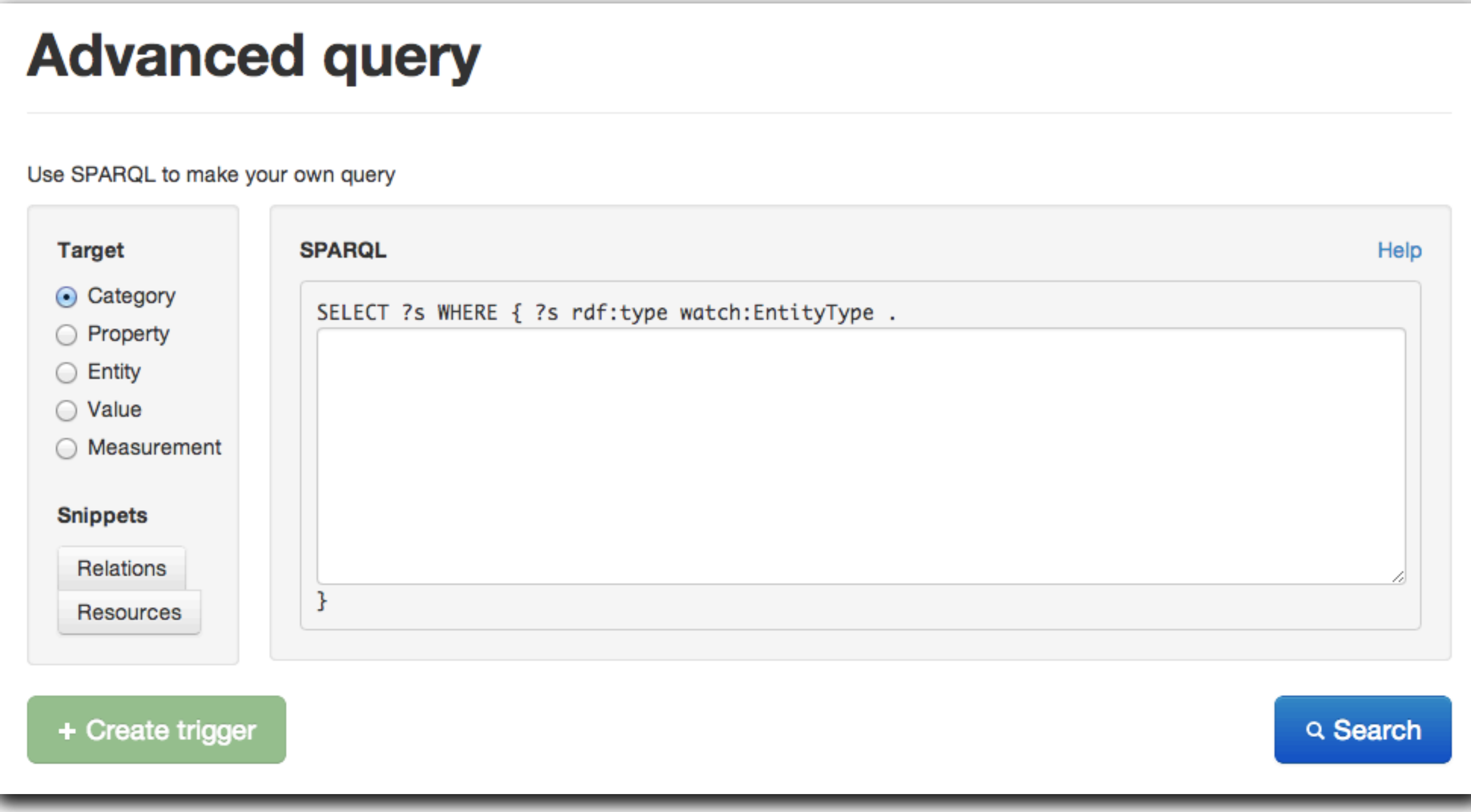
http://scout.scape.keep.pt



Properties

Name	Value	Action
Minimum preservation action execution time	1.5002512	👁
Average preservation action execution time	1.8746954	👁
Maximum preservation action execution time	2.3340003	👁
Ingest average time (ms)	1092798.0	👁

http://scout.scape.keep.pt



Advanced query

Use SPARQL to make your own query

Target

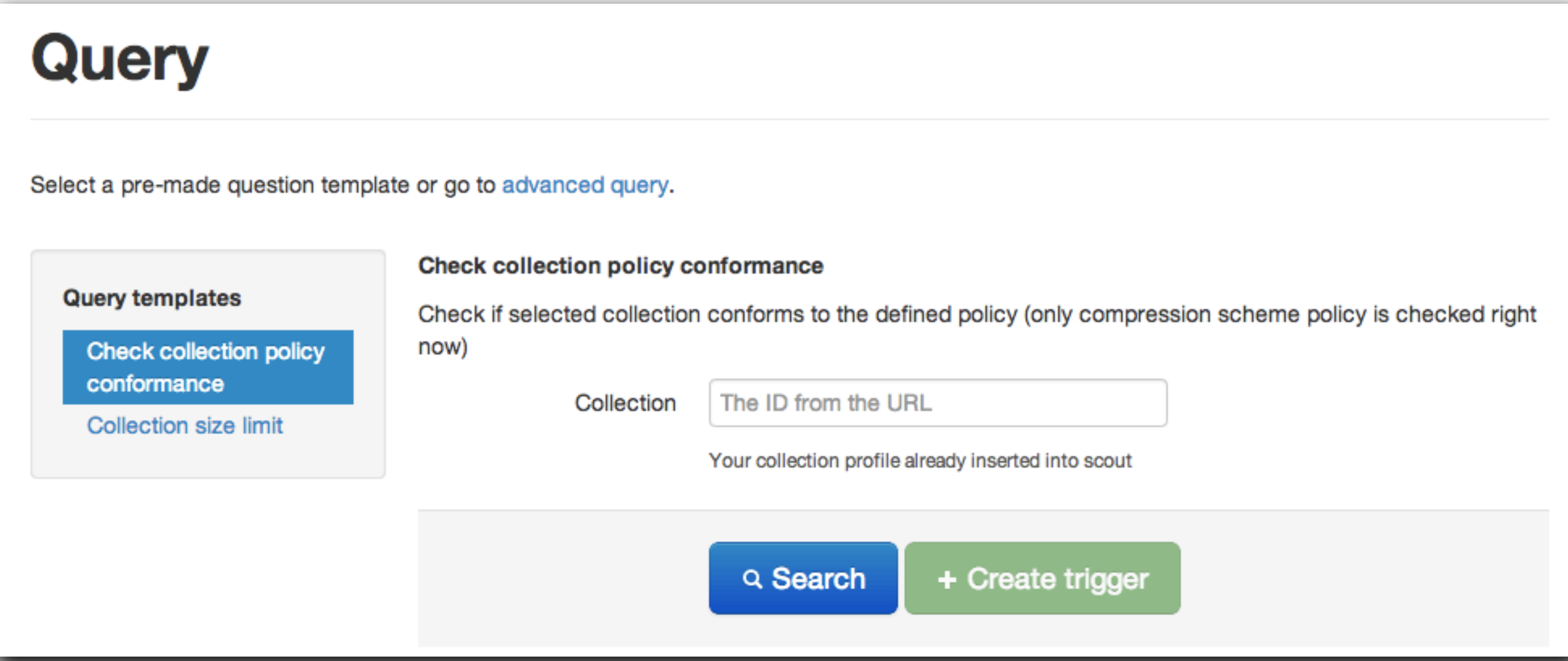
● Category
○ Property
○ Entity
○ Value
○ Measurement

SPARQL

```
SELECT ?s WHERE { ?s rdf:type watch:EntityType .

}
```

http://scout.scape.keep.pt



Query

Select a pre-made question template or go to advanced query.

Query templates

- Check collection policy conformance
- Collection size limit

Check collection policy conformance

Check if selected collection conforms to the defined policy (only compression scheme policy is checked right now)

Collection: The ID from the URL

Your collection profile already inserted into scout

How to be a part of Scout

- Join the surveys
  - Send me your email address <lfaria@keep.pt>
- Integrate your content
  - Send your content profile with C3PO
  - Send repository events with Report API
- Contribute with information (soon)
  - Use Scout form for manual input of knowledge

Benefits

- **Synergy:** together we can do more
- **Sharing:** know about your peers
- **Centralize knowledge:** holistic view of influencers
- **Traceability:** record the inputs to decision-making
- **Find opportunities:** reduce costs and optimize

Roadmap

- User support
- More trigger templates
- More adaptors
  - KrakeN
  - Software catalogues
  - Other format registries
  - Other experiments information sources
  - Manual input (human knowledge)
  - Simulation

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).

SCAPE

Scalable Preservation Environments

Preservation Watch

What to monitor and how Scout can help

Luis Faria lfaria@keep.pt

KEEP SOLUTIONS www.keep-solutions.com

Digital Preservation Advanced Practitioner Course
Glasgow, 15th-19th July 2013