Automatic Preservation Watch

Using Information Extraction on the Web

Luis Faria  lfaria@keep.pt
KEEP SOLUTIONS  www.keep-solutions.com

Alan Akbik, Barbara Sierman, Marcel Ras, Miguel Ferreira, José Carlos Ramalho

iPRES 2013
Lisbon, September 2, 2013
Why do we need monitoring?

- Organisation mission
- Organisation policies
- Bit rot
- Resource capability
- System availability
- Security breach
- Format obsolescence
- New standards
- Emerging technology
- Producer trends
- Consumer trends
- Economical limitations
- Social and political factors
Why do we need monitoring?

Risks
- Bit rot
- Resource capability
- System availability
- Security breach

Opportunities
- Organisation mission
- Organisation policies
- Producer trends
- Consumer trends
- Economical limitations
- Social and political factors
Risk Assessment

- Yes but manual and adhoc: 60%
- None: 40%

Survey on:
Scout: a preservation watch system

Monitors aspects of the world to detect preservation risks and opportunities

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).
Information Sources

• Format registries & software catalogues
• Digital repositories & web archives
• Organizational objectives
• Experiments
• Simulation
• Human knowledge
Currently supported information sources

- PRONOM
- Repository content and events
- Web archive content
- Web archive renderability experiments
- SCAPE Policy model
Define triggers

- Notify me when there are tools that can render the format X.
Define triggers
Simple query with templates

Query

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

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

Search
Create trigger
Receive notifications

- Email
- HTTP Push API

There are tools that can render format X.
Automatic Watch Limitations

Machine readable data

• Explicit and formal specified information
  • Controlled vocabulary
  • Ontology

• All instances use same structure and set of values
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).
e-journal coverage questions

- Which **publisher** provides which **journal titles**
- Publisher changes:
  - **Ceases** to provide journal
  - **Transfers** journal to other publisher(s)
  - Publishers **merge**
- Journal changes:
  - **Name** changes
  - **ISSN** changes
  - **Ceased to exist**
“In 1991, two years before the merger with Reed, Elsevier acquired Pergamon Press in the UK.”

“The Asia-Europe Foundation (ASEF) sold the Asia Europe Journal and transferred the copyright to its long-time partner Springer.”

“Acta Chirurgica Iugoslavica is available free of charge as an Open Access journal on the Internet.”
“In 1991, two years before the merger with Reed, Elsevier acquired Pergamon.”

“The Asia-Europe Foundation (ASEF) sold the Asia Europe Journal and transferred the copyright to its long-time partner Springer.”

“Acta Chirurgica Iugoslavica is available free of charge as an Open Access journal.”

Where is this information?

In the publisher website!

Not machine readable!
Information Extraction

- **Extract structural** information from unstructured data
- **Pattern-based** information extraction

“[X] acquired [Y]”

- Some training and supervision may be needed
1. Data acquisition and pre-processing

2. Relation discovery

3. Information extraction

4. Validation of results
1. Data acquisition and pre-processing

- Focused crawler with seed words (12,000 entries)
  - Publisher names
  - Journal titles
  ➔ 500,000 Web pages

- Pre-process with NLP tools
  ➔ 18 million sentences
  ➔ 8 GB
2. Relation discovery

<table>
<thead>
<tr>
<th>Prominent pattern</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>[X] journal of [Y]</td>
<td>1</td>
</tr>
<tr>
<td>[X] published by [Y]</td>
<td>2</td>
</tr>
<tr>
<td>[X] journal on [Y]</td>
<td>3</td>
</tr>
<tr>
<td>[X] journal published by [Y]</td>
<td>4</td>
</tr>
<tr>
<td>[X] available as [Y] journal</td>
<td>5</td>
</tr>
<tr>
<td>PubMed [X] [Y]</td>
<td>9</td>
</tr>
<tr>
<td>[X] science proceedings of [Y]</td>
<td>25</td>
</tr>
<tr>
<td>[X] subscription available to [Y]</td>
<td>30</td>
</tr>
</tbody>
</table>
2. Relation discovery

<table>
<thead>
<tr>
<th>Prominent pattern</th>
<th>Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>[X] journal of [Y]</td>
<td>1</td>
</tr>
<tr>
<td>[X] published by [Y]</td>
<td>2</td>
</tr>
<tr>
<td>[X] journal on [Y]</td>
<td>3</td>
</tr>
<tr>
<td>[X] journal published by [Y]</td>
<td>4</td>
</tr>
<tr>
<td>[X] available as [Y] journal</td>
<td>5</td>
</tr>
<tr>
<td>PubMed [X] [Y]</td>
<td>9</td>
</tr>
<tr>
<td>[X] science proceedings of [Y]</td>
<td>25</td>
</tr>
<tr>
<td>[X] subscription available to [Y]</td>
<td>30</td>
</tr>
</tbody>
</table>
3. Information extraction

2,000 journal titles

500 journal-publisher attributions
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).

4. Validation of results

**Journal titles in eDepot**
- 86%
- 10%
- 4%

**Title-publisher in the Keepers registry**
- 50%
- 35%
- 15%

- **Should add**
- **False-positives**
- **Existing**
False-positives

- **Detecting boundaries** of titles and publisher names
  
  “European Journal of Nuclear Medicine and Molecular Imaging”

- **Using abbreviations** on titles and publisher names
  
  IAAE - “International Association of Agricultural Economists”

- **Technical problems like encoding**
  
  ”ňőťřą+buda University”
Conclusions

• **We need data** to support digital preservation

• **Explicit and formal specified** for automation

• **Registries** tend to be incomplete and outdated

• **Information Extraction Technologies** can help

• Still, **some supervision** may be needed
Send us your use cases!

Preservation Watch
What risks to monitor?

Information Extraction
What to extract from the web?

Luis Faria
lfaria@keep.pt

Alan Akbik
alan.akbik@tu-berlin.de

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).
Thank you, questions?

- Scout - a preservation watch system
  - Site: http://openplanets.github.io/scout/
  - Demo: http://scout.scape.keep.pt

- SCAPE Planning and Watch suite iPRES poster

- SCAPE
  - http://www.scape-project.eu