HTTPreserve

Auditing document-based hyperlinks

Ross Spencer **@beet_keeper**
@httpreserve, https://github.com/httpreserve

Extract

```
goatslayer@goatslayer:~/git/go/src/github.com/httpreserve/tikalinkextract$ ./tikalinkextract -file archives-nz-demo/
2017/06/17 13:33:11 INFO: 'E2_070910 - LIANZA Conference.DOC' being processed.
2017/06/17 13:33:11 INFO: 'E1_Heartlands speech.doc' being processed.
2017/06/17 13:33:11 INFO: 'E1_Te Waihora Signing.doc' being processed.
2017/06/17 13:33:11 INFO: 'E1_National War Memorial - Reappointments.doc' being processed.
2017/06/17 13:33:11 INFO: 'E2_$559mAuckland roads.30june2006.v2.doc' being processed.
2017/06/17 13:33:11 INFO: 'E1_Information Pack.pdf' being processed.
2017/06/17 13:33:11 INFO: 'E1_Backup of 05-12-08 TeWaihora plan.wbk' being processed.
2017/06/17 13:33:11 INFO: 'E1_Backup of Protected Objects Bill.wbk' being processed.
```

- Extract hyperlinks from documents using regular expression

*- Apache Tika 1.15 supports **333** formats*

- Easily extensible for other tools and regex

Request

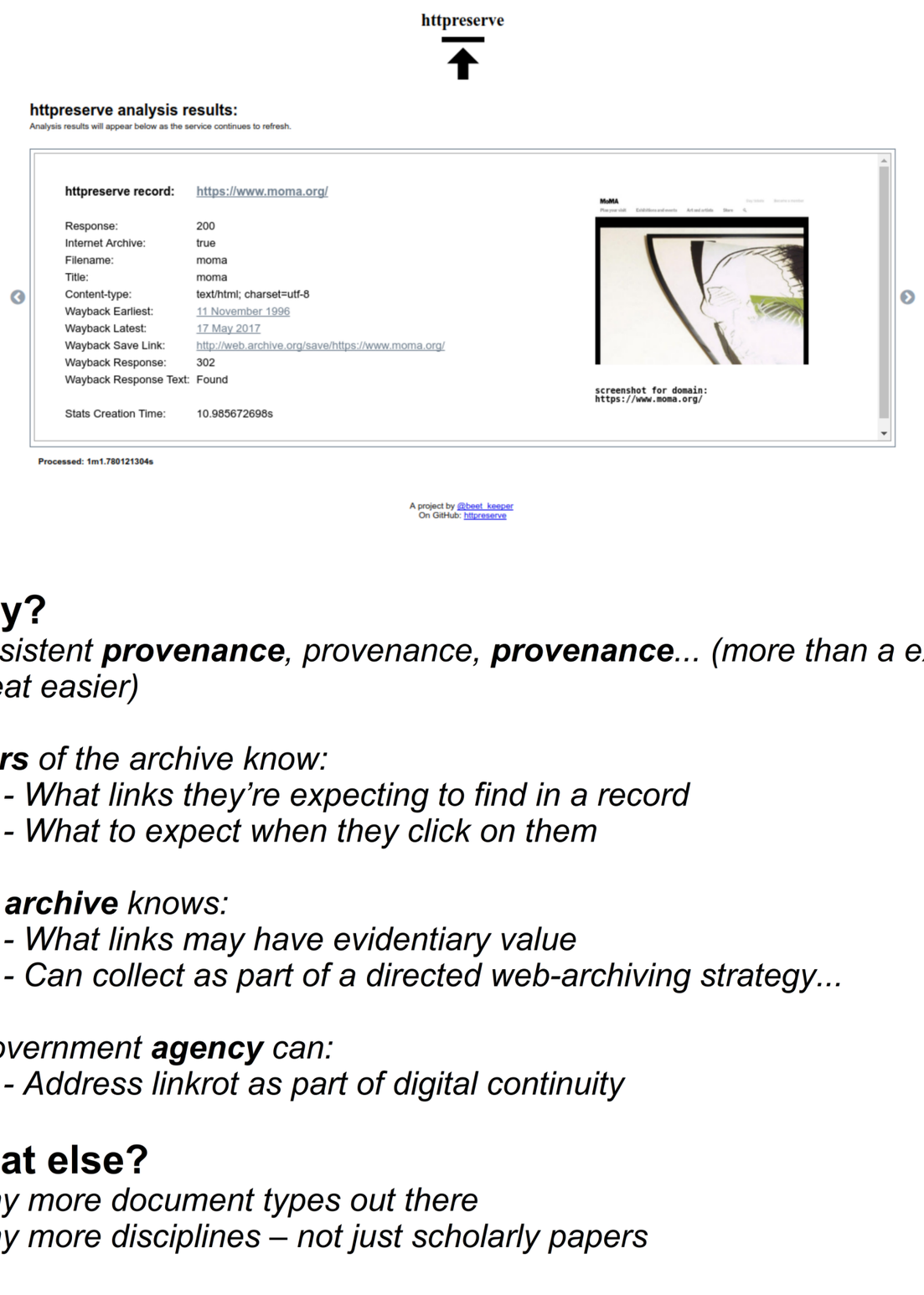
*- **HTTPreserve Core**, RESTFul, plus, CLI, plus, Web App*

- IA Memento used to retrieve earliest and latest links

*- Indication of links being gone, plus screenshot and HTML **<title>** to observe content drift*

- Other web archive URI-base could be used instead of IA or as well as...

Record and Describe



*- Foundation to report per **your** needs*

*- JSON, BoltDB, CSV, **(CLI)**; Webapp*

- Screenshots rendered enable scanning and auditing by archivists

Why?

*Consistent **provenance**, provenance, **provenance**... (more than a experiment. Let others repeat easier)*

***Users** of the archive know:*
- *What links they're expecting to find in a record*
- *What to expect when they click on them*

*The **archive** knows:*
- *What links may have evidentiary value*
- *Can collect as part of a directed web-archiving strategy...*

*A government **agency** can:*
- *Address linkrot as part of digital continuity*

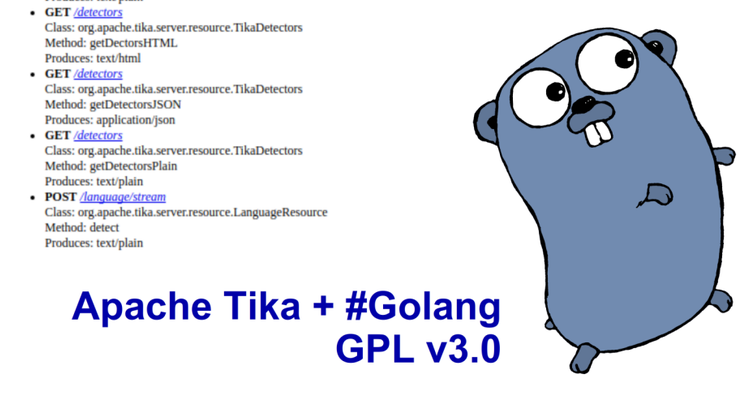
*OPF **Blog** Link:*
https://perma.cc/7JSG-STMN

What else?

Many more document types out there
Many more disciplines – not just scholarly papers

In summary...

*As part of **description** and as part of **preservation**, report everything, use as required*



Apache Tika + #Golang
GPL v3.0