WIRTSCHAFTS
UNIVERSITÄT
WIEN VIENNA
UNIVERSITY OF
ECONOMICS
AND BUSINESS

EFMD
EQUIS
ACCREDITED

Measures for assessing the data freshness in Open Data portals

Sebastian Neumaier and Jürgen Umbrich

Vienna University of Economics and Business

Problem/Challenge

- How up-to-date are resources in Open Data portals?

- Required information for such a metric:
  - **Change history** of documents in portals

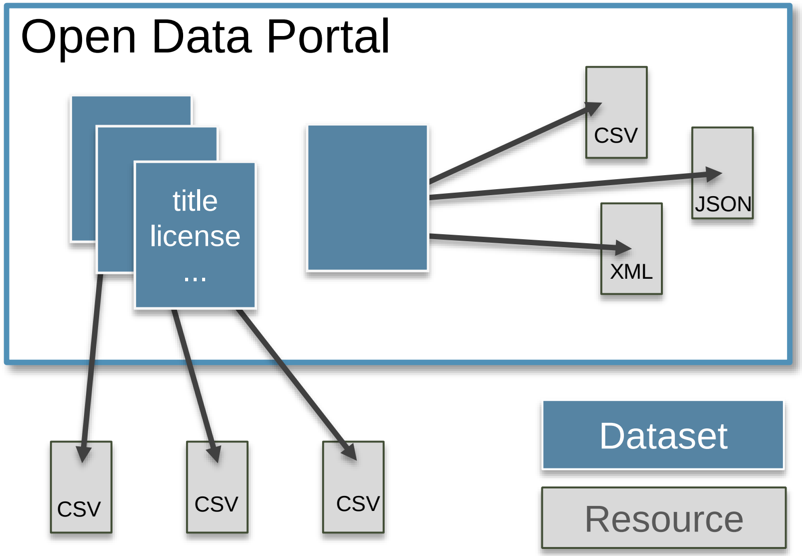
- Challenge:
  1. Collect available change history
  2. Estimated next change time to assess up-to-dateness

- Two scenarios:
  - *Portal provider*: wants to add freshness measure to metadata
  - *Data consumer*: updating of application, DB, etc..

Open Data Portals

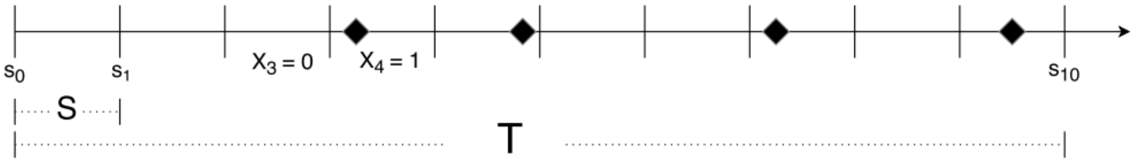
- Single point of access
- Local and external resources
- Meta data
  - Title
  - **Modification date**
  - …

- Typical software:

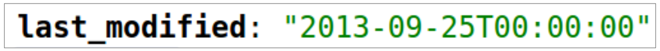


ckan Socrata OpenDataSoft

Sources of change information in OD portals

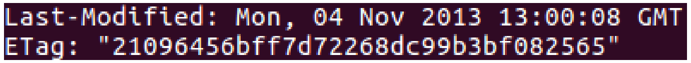


- **Push-based** history:
  - Data provider push change information to portal
    - If *local*, by uploading new version
    - If *external*, by updating a specific metadata field



```
last_modified: "2013-09-25T00:00:00"
```

- Pull-based history:
  - **Age sampling**:
    - Access to latest change time of a resource (i.e., last-modified timestamp in *HTTP Header*)



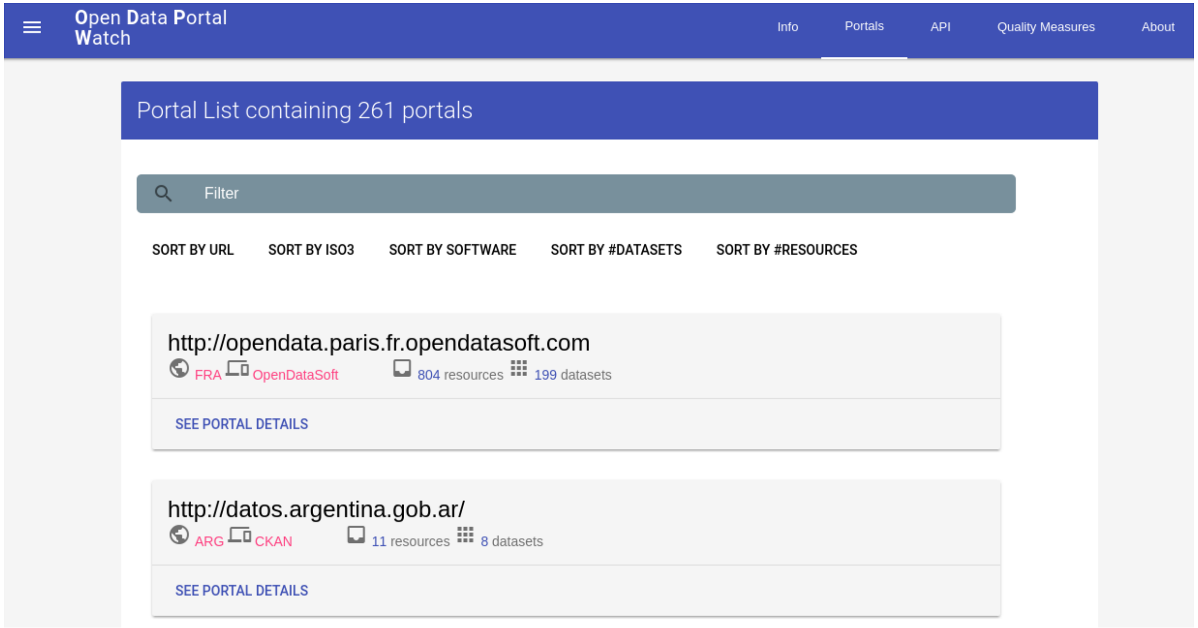
```
Last-Modified: Mon, 04 Nov 2013 13:00:08 GMT
ETag: "21096456bff7d72268dc99b3bf082565"
```

  - **Comparison sampling**:
    - Detect changes by monitoring and comparing the resources

Open Data Portal Watch

http://data.wu.ac.at/portalwatch/

- Periodically monitoring over 260 Open Data portals
- Metadata quality assessment
  - Uniform handling of metadata (using DCAT mapping)
- Evolution tracking & archiving
  - Meta data
  - Data

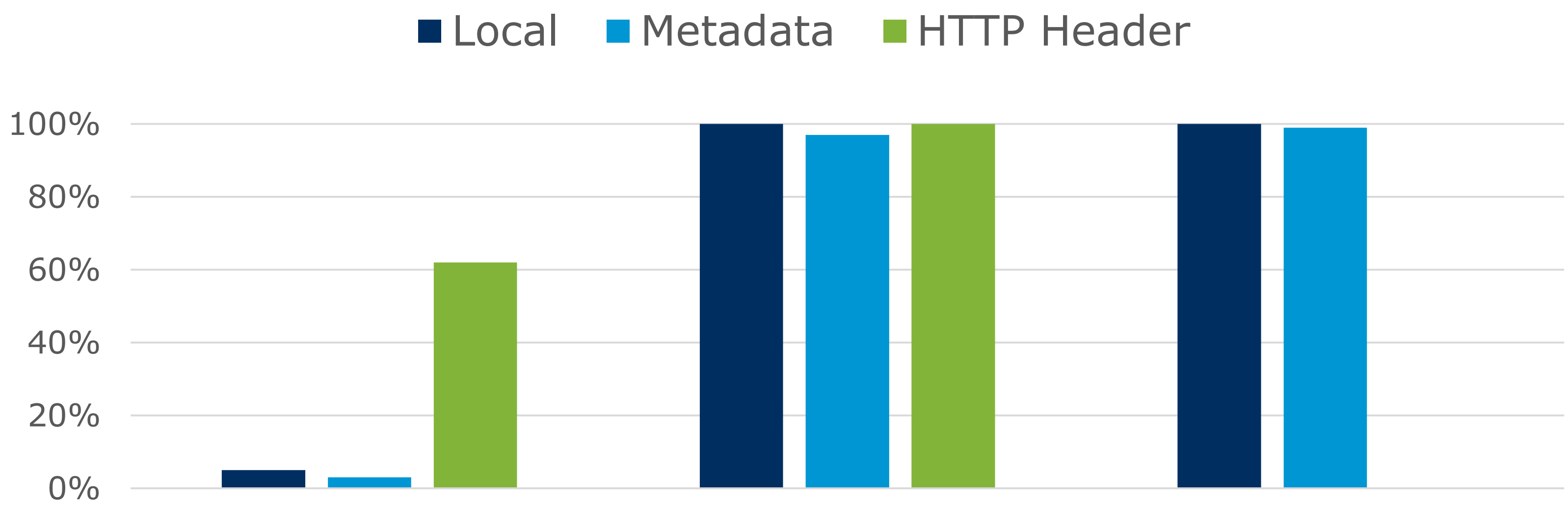


Portal List containing 261 portals

http://opendata.paris.fr.opendatasoft.com
FRA OpenDataSoft 804 resources 199 datasets

http://datos.argentina.gob.ar/
ARG CKAN 11 resources 8 datasets

Available change information

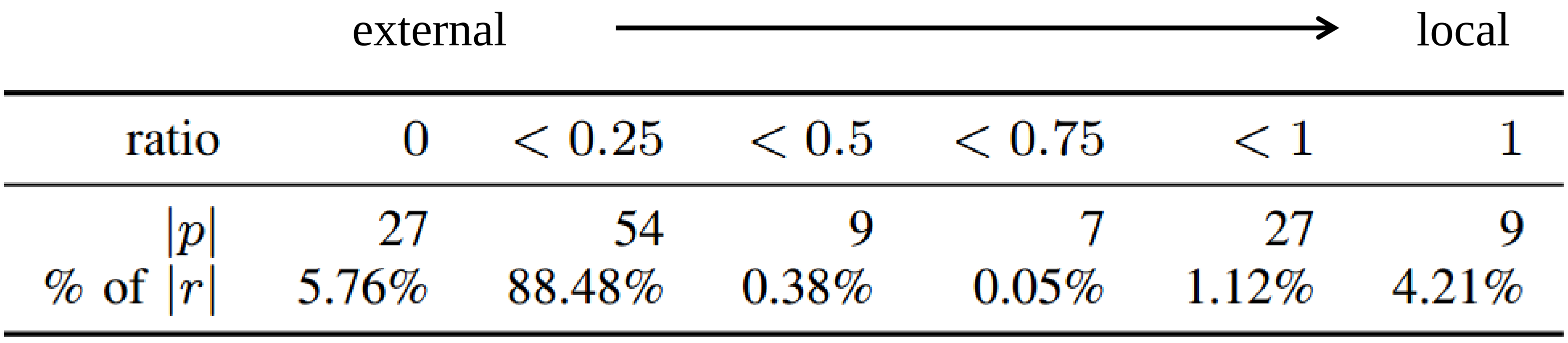


	CKAN	Socrata	OpenDataSoft
URLs	**4,049,851**	181,548	8757
distinct	**2,116,940**	165,966	8757

- *CKAN*: age- and comparison-sampling required
- *Socrata* & *OpenDataSoft*: push-based possible

Local vs external resources on CKAN

- 130 CKAN portals:
  - 27 portals host all resources externally, 9 all locally
  - Majority of all URLs (~88%) belong to 54 portals with <25% local resources
  - **HDX portal**: 9574 URLs, 8833 distinct, 2114 **local** (~**24**%)



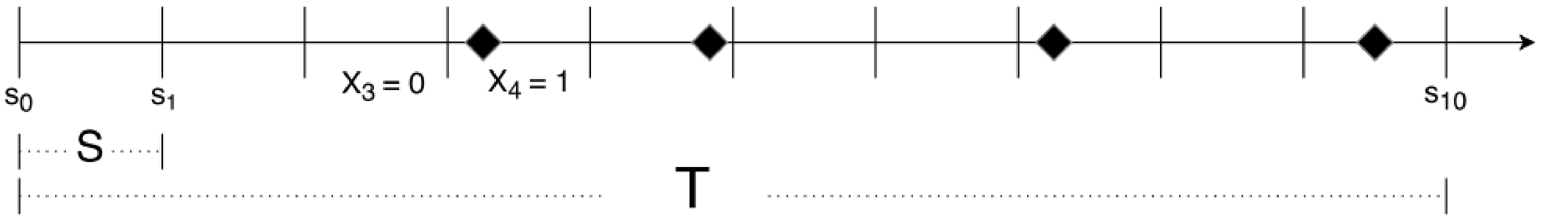
external → local

ratio	0	< 0.25	< 0.5	< 0.75	< 1	1
$\|p\|$	27	54	9	7	27	9
% of $\|r\|$	5.76%	88.48%	0.38%	0.05%	1.12%	4.21%

Estimation of next updates

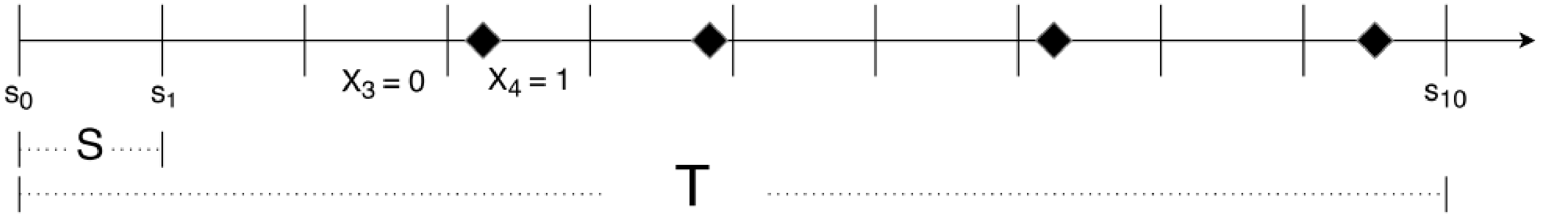
- Evaluating three change estimation heuristics:
  - *Poisson process*
    - Cho and Garcia-Molina (2003) propose Poisson process model to estimate updates in the context of Web sites
  - *Markov chain approach*
    - Umbrich et al. (2015) use Markov chains to schedule next crawl times for URLs based on previous observed changes
  - *Empirical distribution*
    - Build empirical distribution of changes based on intervals

Estimation of next updates (cont'd)



- **Age sampling** *(last-modified timestamp in HTTP Header)*
  - *Poisson distribution*
    - X/T $\left(= \frac{number\ of\ changes}{monitoring\ period}\right)$ as estimator for Poisson parameter
    - Compute next change time by considering *p*-quantiles
  - *Empirical distribution*
    - Use intervals between the observed last-modified times
    - *p*-quantiles of empirical distribution

Estimation of next updates (cont'd)

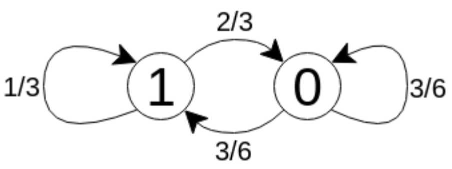


- **Comparison sampling** *(comparing the actual content)*
  - Only binary information/states available:

0	0	0	1	1	0	0	1	0	1

- *Markov chain approach*
  - Probability of next change based on previous state, e.g.:

i \i+1	1	0	TOTAL
0	3	3	6
1	1	2	3



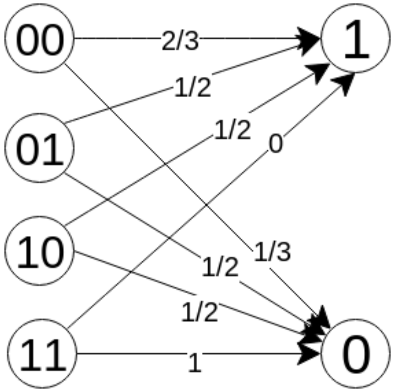
$P(1|0) = 3/6$

Extending the Markov chain approach

H:	0	0	0	1	1	0	0	1	0	1

- Extend approach by considering the last *k* states for computing the probabilities:

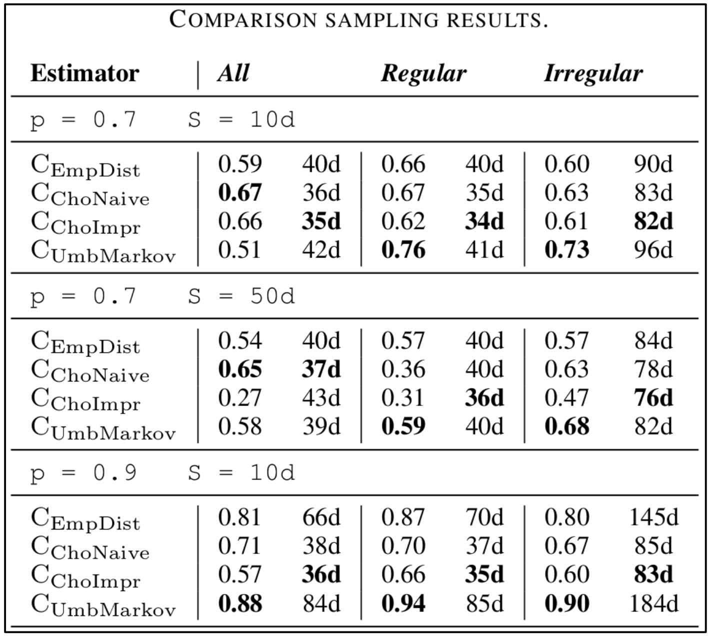
i \i+1	1	0	TOTAL
00	2	1	3
01	1	1	2
10	1	1	2
11	0	1	1



$$P(1|00) = 2/3$$

Evaluation Summary

- Controlled environment:
  - Evaluation using revision histories of Wikipedia articles
    - 1562 randomly Wiki articles with >3 years history and >30 revisions
    - Wiki change history does not follow Poisson distribution
  - Different confidence values:
    - *For fixed p, we report the ratio of successfully predicted updates*
- Conclusion:
  - *Markov chain approach* best for comparison-based sampling
  - *Empirical distribution* best for push-based and age-based sampling



COMPARISON SAMPLING RESULTS.

Estimator	*All*		*Regular*		*Irregular*	
`p = 0.7   S = 10d`						
C_{EmpDist}	0.59	40d	0.66	40d	0.60	90d
C_{ChoNaive}	**0.67**	36d	0.67	35d	0.63	83d
C_{ChoImpr}	0.66	**35d**	0.62	**34d**	0.61	**82d**
$C_{\text{UmbMarkov}}$	0.51	42d	**0.76**	41d	**0.73**	96d
`p = 0.7   S = 50d`						
C_{EmpDist}	0.54	40d	0.57	40d	0.57	84d
C_{ChoNaive}	**0.65**	**37d**	0.36	40d	0.63	78d
C_{ChoImpr}	0.27	43d	0.31	**36d**	0.47	**76d**
$C_{\text{UmbMarkov}}$	0.58	39d	**0.59**	40d	**0.68**	82d
`p = 0.9   S = 10d`						
C_{EmpDist}	0.81	66d	0.87	70d	0.80	145d
C_{ChoNaive}	0.71	38d	0.70	37d	0.67	85d
C_{ChoImpr}	0.57	**36d**	0.66	**35d**	0.60	**83d**
$C_{\text{UmbMarkov}}$	**0.88**	84d	**0.94**	85d	**0.90**	184d

Thank you for your attention

- *Goal*
  - Data Freshness estimation in Open Data
- *Challenge*
  - Collecting change history (push vs pull)
- *Approach*
  - Estimators for different scenarios
  - Empirical evaluation

Sebastian Neumaier
WU Vienna, Institute for Information Business

email: sebastian.neumaier@wu.ac.at
url: https://sebneumaier.wordpress.com/
twitter: @sebneum