

# LOD-LAUNDROMAT

**PUBLISHING OTHER PEOPLE'S DIRTY DATA**

Wouter Beek, Laurens Rietveld, Hamid Bazoobandi, Jan  
Wielemaker, Stefan Schlobach

VU University Amsterdam

<http://wouterbeek.github.io/ll.html>

# Freebase 'Monkey' (< 10% syntactically correct)

```
Activities Termin... Thu 23 Oct, 13:29 wbeek@localhost:~  
wbeek@localhost:~ 111x18  
40\u00E19\u00E37\u00E48\u00E2D\u00E07\u00E08\u00E32\u00E01\u00E21\u00E35\u00E1A\u00E23\u00E23\u00E1E\u00E1A\u00E38\u00E23\u00E38\u00E29\u00E23\u00E48\u00E27\u00E21\u00E01\u00E31\u00E19\u00E21\u00E32 \u00E01\u00E48\u00E2D\u00E19\u00E17\u00E35\u00E48\u00E08\u00E30\u00E21\u00E35\u00E27\u00E34\u00E27\u00E31\u00E12\u00E19\u00E32\u00E01\u00E32\u00E23\u00E41\u00E22\u00E01\u00E15\u00E31\u00E27\u00E2D\u00E2D\u00E01\u00E08\u00E32\u00E01\u00E01\u00E31\u00E19 \u00E40\u00E21\u00E37\u00E48\u00E2D 2-3 \u00E25\u00E49\u00E32\u00E19\u00E18\u00E35\u00E01\u00E48\u00E2D\u00E19\u00E25\u00E34\u00E07\u00E21\u00E35\u00E15\u00E35\u00E19\u00E28\u00E19\u00E49\u00E32\u00E41\u00E25\u00E30\u00E15\u00E35\u00E19\u00E28\u00E25\u00E31\u00E07\u00E43\u00E0A\u00E49\u00E08\u00E31\u00E1A\u00E40\u00E01\u00E32\u00E30\u00E44\u00E14\u00E49 \u00E21\u00E31\u00E1B\u00E35\u00E19\u00E28\u00E32\u00E01\u00E34\u00E19\u00E1A\u00E19\u00E15\u00E49\u00E19\u00E44\u00E21\u00E49\u00E44\u00E14\u00E49\u00E2D\u00E22\u00E48\u00E32\u00E07\u00E04\u00E25\u00E48\u00E2D\u00E07\u00E41\u00E04\u00E25\u00E48\u00E27 \u00E1E\u00E27\u00E01\u00E17\u00E35\u00E48\u00E21\u00E35\u00E28\u00E32\u00E07 \u00E40\u00E0A\u00E48\u00E19 \u00E25\u00E34\u00E27\u00E2D\u00E01 \u00E41\u00E25\u00E30\u00E1E\u00E27\u00E01\u00E17\u00E35\u00E48\u00E44\u00E21\u00E48\u00E21\u00E35\u00E28\u00E32\u00E07 \u00E40\u00E48\u00E19 \u00E01\u00E2D\u00E23\u00E1\u00E34\u00E25\u00E25\u00E32 \u00E40\u00E1B\u00E47\u00E19\u00E15\u00E49\u00E19"@th .  
<http://rdf.freebase.com/ns/m.08pbxl> <http://rdf.freebase.com/ns/common.topic.description> "Abe er, i snefver forstand, betegnelsen pe5 en primat der ikke er en halvabe, en spf8gelsesabe eller et menneske. Det vil sige al le arter i Abe-infraordenen med undtagelsen af mennesket. I bred forstand kan " .  
rapper: Failed to parse file 08pbxl turtle content  
rapper: Parsing returned 33 triples  
[wbeek@localhost ~]$ █  
wbeek@localhost:~ 111x18  
Warning: /home/wbeek/label.nt:241250:0: Syntax error: subject expected  
Warning: /home/wbeek/label.nt:241252: Syntax error: newline in string  
Warning: /home/wbeek/label.nt:241254:0: Syntax error: subject expected  
Warning: /home/wbeek/label.nt:241263: Syntax error: newline in string  
Warning: /home/wbeek/label.nt:241265:0: Syntax error: subject expected  
Warning: /home/wbeek/label.nt:241267:0: Syntax error: subject expected  
Warning: /home/wbeek/label.nt:241268: Syntax error: newline in string  
Warning: /home/wbeek/label.nt:241270:0: Syntax error: subject expected  
Warning: /home/wbeek/label.nt:241271: Syntax error: newline in string  
Warning: /home/wbeek/label.nt:241273:0: Syntax error: subject expected  
Warning: /home/wbeek/label.nt:241274: Syntax error: newline in string  
Warning: /home/wbeek/label.nt:241276:0: Syntax error: subject expected  
Warning: /home/wbeek/label.nt:241277: Syntax error: newline in string  
Warning: /home/wbeek/label.nt:241279:0: Syntax error: subject expected  
% Parsed "label.nt" in 2.96 sec; 214,722 triples  
true.  
?- █
```

# DIRTY DATA

- Character encoding issues
- Socket errors
- Protocol errors
- Corrupted archives
- Authentication problems
- Syntax errors
- Wrong metadata

# PROBLEM STATEMENT

After 10+ years of SW evangelization data quality is still not as high as it should be.

... therefore the SW is not generally machine-processable today.

Data preparation / data cleaning tasks take 10-40% of research time.

Existing solutions for cleaning data (standards, guidelines, best practices, tools) are targeted towards human data creators, who can (and do) choose not to use them.

# GOALS

- Automate the data preprocessing phase
- Disseminate *all* LOD in a standards-compliant / machine-processable way, *right now*:
  - Scale: billions of triples
  - Days not decades
- Support common uses cases:
  - Splitting/combining data
  - Streamed processing
  - Non-SW tooling: regex, GNU tools (e.g., grep), Pig, etc.

# LOD LAUNDROMAT



WEB SERVICE: [HTTPS://LODLAUNDROMAT.ORG](https://lodlaundromat.org)

# METADATA

- Duplicate triples
- Most occurring errors

# CURRENT USE CASES

- Automated load balancing:
  - Use *reliable* metadata for determining sizes/skews
  - Multi-node computing cluster
- Streamed data processing:
  - Streaming window: 1 triple
  - 10B+ triples processed on discount hardware
  - PrefLabel
- Improve evaluations:
  - "We tested our algorithm on the English version of DBpedia"
  - We are currently optimizing all our algorithms for <5 datasets!
  - Evaluate an algorithm against 15,000 datasets in 10-40 lines of code.



# FUTURE USE CASES

- LOD Observatory
- Feedback to dataset publishers
- Meta-data dataset
- Algorithm heuristics