LOD-LAUNDROMAT

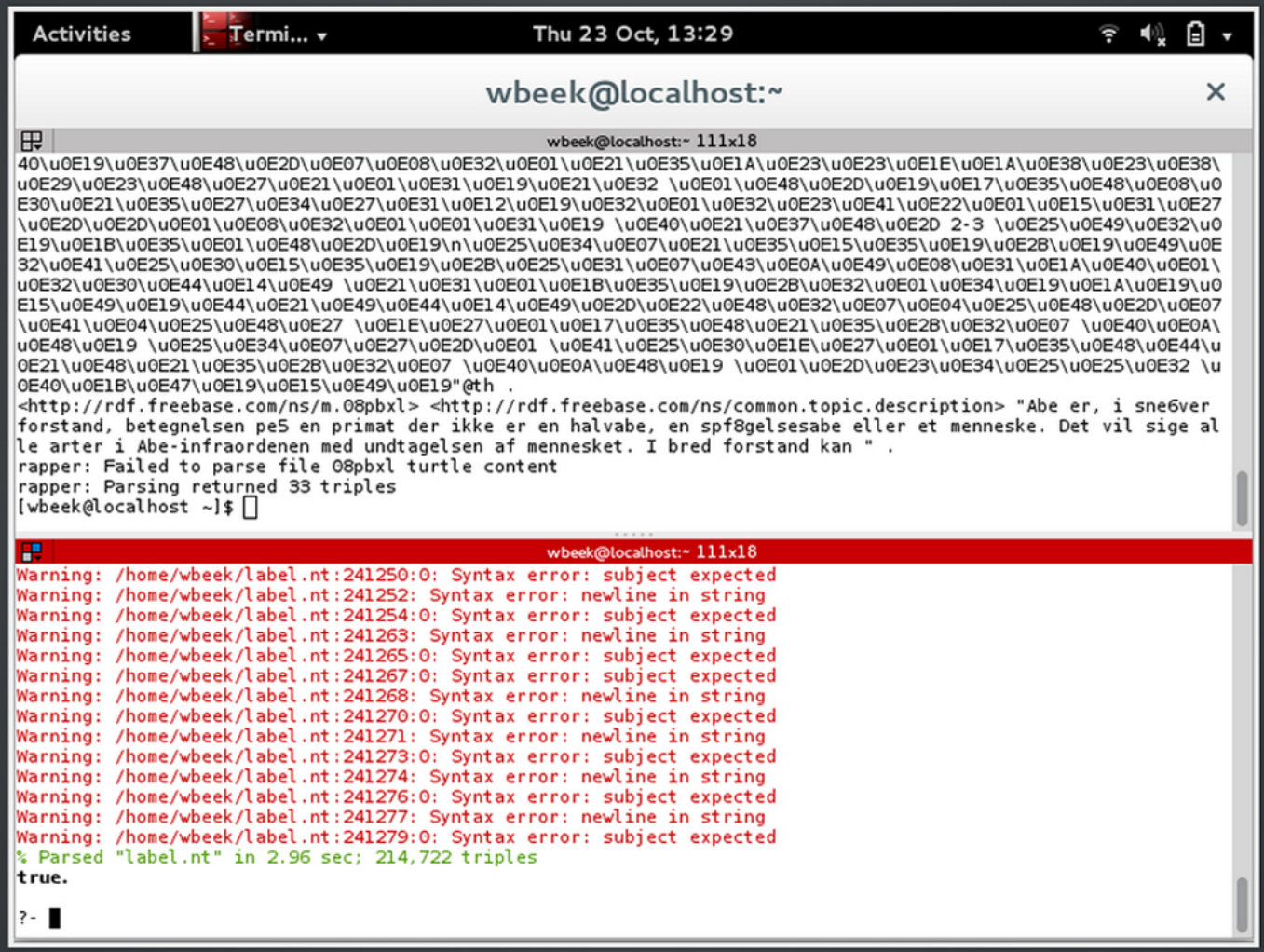
PUBLISHING OTHER PEOPLE'S DIRTY DATA

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

VU University Amsterdam

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

Freebase 'Monkey' (< 10% syntactically correct)



LinkedBrainz

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: reges, GNU tools (e.g., grep), Pig, etc.

LOD LAUNDROMAT

WEB SERVICE: 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