Crawl Me Maybe: An Iterative Linked Dataset Crawler

Besnik Fetahu, Ujwal Gadiraju, and Stefan Dietze
{lastname}@L3S.de
L3S Research Center, Leibniz University of Hannover

Iterative Dataset Crawler
1) Metadata crawler
2) Instance-level crawler

What’s the story?
- Evolution of Linked Datasets
- Flexible Crawl Definitions
- Iterative Crawls
- Capturing and Storing Diffs on-the-fly
- Dataset Crawler Web Interface

Online diff computation
1) Insertions: Additions introduced in the dataset correspond to insertions.
2) Deletions: Over time, triples are deleted from a dataset due to various reasons ranging from persisting correctness to detection of errors.
3) Updates: Updates correspond to the update of one element of a triple <s,p,o>.

Dataset crawler schema

Key Features
- Capture dataset evolution
- Online diff computation based on three logging operators (insertions/updates/deletions)
- Analyze the state of a dataset at any given time point
- Web interface which allows the setup of the crawler, and facilitates simple query functionalities over the crawled data