Semantic Publishing Benchmark

Second Workshop on Graph-based Technologies and Applications
21.February.2014, Barcelona

Problem

- Publishing Domain
  - Why create a benchmark for that domain?
    - Constantly generating new content
    - Constantly updating existing content
    - Constantly consuming content
  - Semantic technologies in the publication pipeline
    - Annotation of content
    - Content multi-purposing

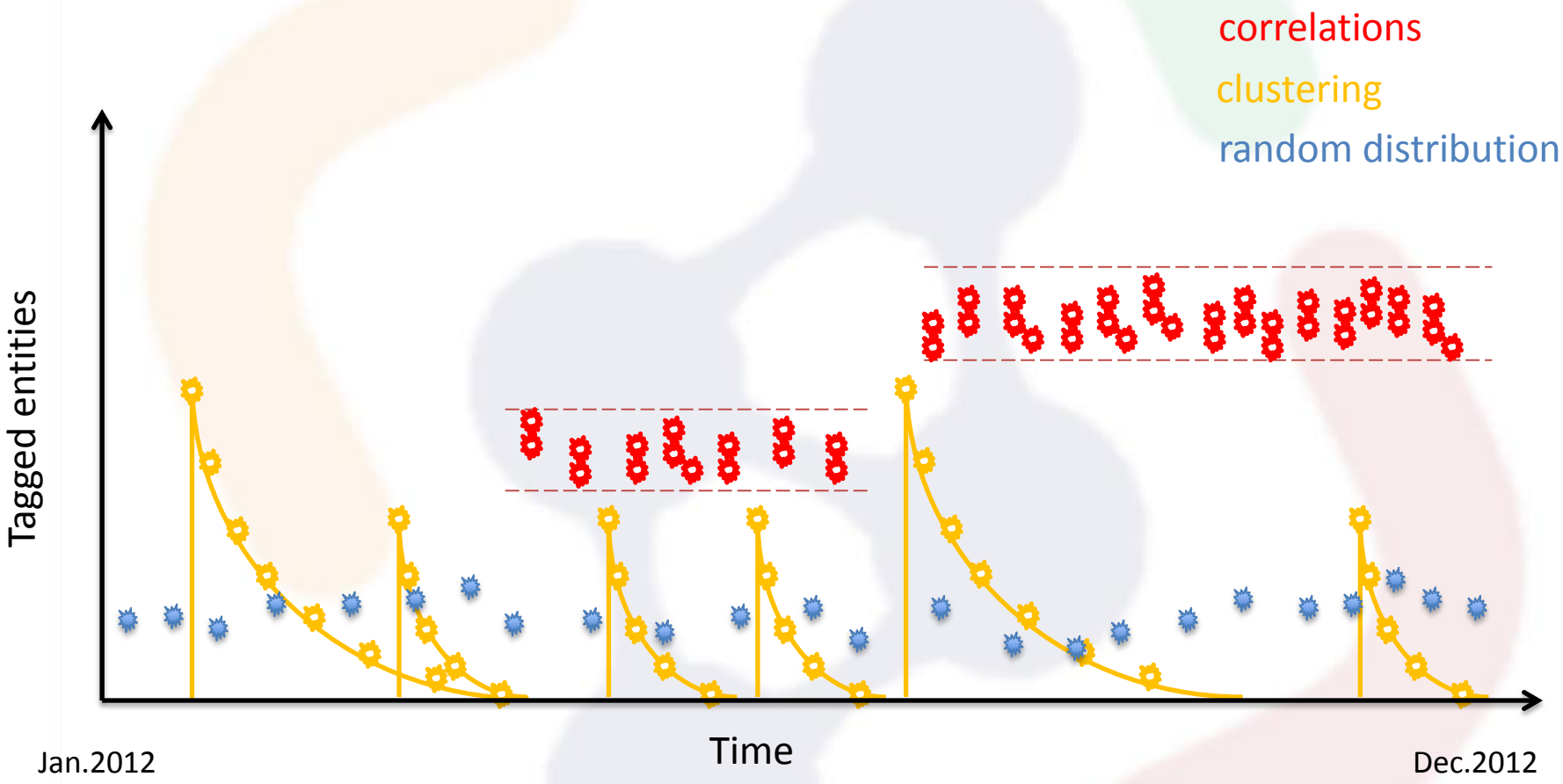
Solution

- LDBC Semantic Publishing Benchmark
- A benchmark for ***RDF Databases, SPARQL 1.1***
- Scenario : a media organization which maintains a catalogue of meta-data (Creative Works) for its assets : News, Articles, Blogs, Journals
- The benchmark simulates :
  - Consumption of that meta-data
  - Management of that meta-data

Solution (2) - Features

- Features of the benchmark
  - Uses real reference data provided by the **BBC** and **DBPedia**
  - Constantly evolving data-generator
    - Started with random distributions of entities
    - Added ***clustering*** of data – e.g. modeling major and minor events
    - Currently implementing modeling ***correlations*** between entities

Solution (3)
Data Generator - evolution



LDBC
Linked Data Benchmark Council

Solution (4) - Features

- Queries
  - Aggregation, Geo-spatial, Time range, Full-text search, Drill-down, Faceted search
- Choke points
  - Choose the optimal query plan
  - Correct estimation of cardinalities
- Online replication and Backup

Results

- Query performance rate
  - Editorial operations, Aggregation operations
  - Total QPS
- Benefits
  - Using the benchmark as a part of the release procedure for OWLIM RDF Store
  - Detect performance issues

Interested ?

Thank you!