How Does Knowledge Evolve in Open Knowledge Graphs?

Axel Polleres

Vienna University of Economics and Business
wish I had more time to prepare for the talk... … but I was hit by dynamic events
wish I had more time to prepare for the talk...
... but I was hit by dynamic events

Instance data changing in a highly dynamic fashion:

Not only the route, but also temporal information:
wish I had more time to prepare for the talk...  
... but I was hit by dynamic events

Instance data changing in a highly dynamic fashion:

<table>
<thead>
<tr>
<th>Time</th>
<th>Departure</th>
<th>Platform</th>
</tr>
</thead>
<tbody>
<tr>
<td>14:43</td>
<td>Budapest-Keleti</td>
<td>3 A-C</td>
</tr>
<tr>
<td>14:43</td>
<td>Flughafen Wien</td>
<td>3 A-C</td>
</tr>
<tr>
<td>15:19</td>
<td>Innsbruck Hbf</td>
<td>4 B-D</td>
</tr>
<tr>
<td>15:45</td>
<td>Flughafen Wien</td>
<td>3 B-D</td>
</tr>
<tr>
<td>15:45</td>
<td>Bratislava hl.st.</td>
<td>3</td>
</tr>
<tr>
<td>16:43</td>
<td>Flughafen Wien</td>
<td>3 A-C</td>
</tr>
<tr>
<td>16:46</td>
<td>Ausfall/Cancelled</td>
<td>1 D</td>
</tr>
<tr>
<td>16:50</td>
<td>Schwarzach-St Veit</td>
<td>1 D</td>
</tr>
</tbody>
</table>

Information: SCOTTY mobil I Twitterer.  
All changes between Kufstein and Salzburg Hbf
wish I had more time to prepare for the talk...  
... but I was hit by dynamic events

Instance data changing in a highly dynamic fashion:

Not only the route, but also temporal information:
wish I had more time to prepare for the talk...  
... but I was hit by dynamic events

Instance data changing in a highly dynamic fashion:

Not only the route, but also temporal information:
wish I had more time to prepare for the talk...  
... but I was hit by dynamic events.

Instance data changing in a highly dynamic fashion:

**Dynamic constraints violated**
(inconcistency via dynamics!):
wish I had more time to prepare for the talk... 
... but I was hit by dynamic events

Even seemingly static data no longer valid:
wish I had more time to prepare for the talk...  
... but I was hit by dynamic events

Even seemingly static data no longer valid...
wish I had more time to prepare for the talk...
... but I was hit by dynamic events

... working on the train as planned not possible :-S
wish I had more time to prepare for the talk…
… but I was hit by dynamic events…

...after a 5hrs delay arrival …

...just to realize the next morning, this:

Your Austrian flight has been cancelled

Dear Mr. Polleres,

Your Austrian flight OS405 from Vienna to Lyon on 03.12.2023 has been cancelled.

Please accept our sincere apologies.

...which perfectly motivates a talk about dynamics and evolution in KGs!
Thanks to my co-authors and to the new TGDK journal!

Transactions on Graph Data and Knowledge (TGDK)

Publications
All documents published in this journal are available open access on ORSIP.

Aims and Scope
Transactions on Graph Data and Knowledge (TGDK) is an Open Access journal that publishes original research articles and survey articles on graph-based abstractions for data and knowledge, and the techniques that such abstractions enable with respect to integration, querying, reasoning and learning. The scope of the journal thus intersects with such research areas as Conformal Graph Algorithms, Graph Databases, Graph Representation Learning, Knowledge Graphs, Knowledge Representation, Linked Data and the Semantic Web. Also in scope for the journal are research investigating graph-based abstractions of data and knowledge in the context of Data Integration, Data Science, Information Extraction, Information Retrieval, Machine Learning, Natural Language Processing, and the Web.

The journal is Open Access without fees for readers nor for authors (also known as Diamond Open Access).

Open Access Policy
TGDK articles are peer-reviewed and published according to the principle of OpenAccess, i.e., they are available online and free of charge. The authors retain their copyright.
What did I want to talk about?

- Open Knowledge Graphs?
- Perspectives on evolution: dimensions of time and temporality
- Observability of evolution
- Metrics for evolution
- How does evolution affect downstream tasks and resp. techniques?
- Recent own work...
Open, collaborively edited KGs

• E.g.
  - DBPedia
  - Wikidata

• Maintain collaboratively edited, curated, reusable knowledge

• Serve as a backbone for various applications!
Main questions:

• Which publicly accessible, open KGs are observable in a manner that would allow a longitudinal analysis of their evolution and how?
  ➔ What dimensions does evolution have at all?

• Do we have the right metrics to analyse KGs’ evolution?

• Do we have the right techniques to process evolving KGs?
What dimensions does evolution have at all?

• Temporal KGs: Time as Data ("valid time")

• Dates/Timestamps
What dimensions does evolution have at all?

- Temporal KGs: Time as Data (“valid time”)
  - Dates/Timestamps
  - Intervals
  - (start/end events)

Challenge: Needs reification!
The statement that Picasso created Guernica was only created in Wikidata in March 2013

What dimensions does evolution have at all?

- **Time-varying KGs**: Time as meta-data/log data ("transaction time")

- **Edit events**(Dates/Timestamps)

"Guernica" (Q175036) was created on 28 November 2012 in Wikidata

The statement that Picasso created Guernica was only created in Wikidata in March 2013
What dimensions does evolution have at all?

- Time-varying KGs: Time as meta-data/log data ("transaction time")
  - Edit events (Dates/Timestamps)
  - Metadata also contains additional provenance information
  - ... yet, often not part of the RDF/Graph model itself!
What dimensions does evolution have at all?

• Time-varying KGs: Granularity of dynamicity (observability):
  • “Versioned KG” ("snapshot") of particular materialisations
  vs.
  • “Dynamic KG”
  down to single Edit events
What dimensions does evolution have at all?

- Lines are blurry...
- Location could also be modelled as data:
What dimensions does evolution have at all?

- Instance evolution
- Schema evolution
- Collaboration models

(\textit{single edge, node changes})
Continuous Change

Temporal KG

Dynamic KG

Versioned KG

Time-Varying KG

"Time as data" (\textit{valid time})

"Time as metadata/log data" (\textit{transaction time})

Discrete Changes (\textit{full graph snapshots/dumps})
What dimensions does evolution have at all?

- Instance evolution
- Schema evolution
- Collaboration evolution
- Structural evolution
- Dynamics (change frequencies, etc.)
- Timeliness (recency of temporal information, delays)
- Monotonicity ("growth" vs. "deletions")
### Availability of Dbpedia and Wikidata:

<table>
<thead>
<tr>
<th>Level</th>
<th>Queryable</th>
<th>Collaborative</th>
<th>Formats</th>
<th>Protocol</th>
<th>Metadata</th>
<th>Temporality</th>
<th>Timeliness</th>
</tr>
</thead>
<tbody>
<tr>
<td>V</td>
<td>Yes</td>
<td>Yes</td>
<td>NT, TTL, HDT, JSON</td>
<td>HTTP, SPARQL</td>
<td>schema.org</td>
<td>No</td>
<td>2-3 Days</td>
</tr>
<tr>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
<td>NT, TTL, JSON</td>
<td>HTTP, SPARQL</td>
<td>schema.org</td>
<td>No</td>
<td>2-3 Days</td>
</tr>
<tr>
<td>CL</td>
<td>Yes</td>
<td>Yes</td>
<td>JSON</td>
<td>SSE</td>
<td>No</td>
<td>Event TS</td>
<td>Seconds</td>
</tr>
<tr>
<td>V</td>
<td>Yes</td>
<td>Partial</td>
<td>NT</td>
<td>HTTP, SPARQL</td>
<td>No</td>
<td>No</td>
<td>Quarterly</td>
</tr>
<tr>
<td>S</td>
<td>Yes</td>
<td>Yes</td>
<td>RDF</td>
<td>HTTP, SPARQL</td>
<td>No</td>
<td>No</td>
<td>Daily</td>
</tr>
<tr>
<td>CL</td>
<td>Yes</td>
<td>Yes</td>
<td>RDF</td>
<td>HTTP</td>
<td>No</td>
<td>Graph TS</td>
<td>Daily</td>
</tr>
</tbody>
</table>

*Availability of Open KG Versions (V), Schema (S), and Change logs (CL), find more in the paper!*
Note: additional Caveat:

• Historic versions are hard to maintain and host:

https://www.rdfhdt.org/datasets/

Important Note (12 April 2022): We are experiencing some technical problems on our “gaia” server, so unfortunately some datasets could be unavailable (e.g. Wikidata). We hope to resolve this issue soon, thanks for your understanding!

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Size</th>
<th>Triples</th>
<th>Details</th>
<th>Provenance</th>
</tr>
</thead>
<tbody>
<tr>
<td>Latest Wikidata (3rd march 2021)</td>
<td>53GB uncomp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Latest Wikidata (9th march 2020)</td>
<td>50GB uncomp</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>DBPedia 2016-10 English</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**What is the “DBpedia Snapshot” Release?**

Historically, this release has been associated with many names: “DBpedia Core”, “EN DBpedia”, and — most confusingly — just “DBpedia”. In fact, it is a combination of —

• EN Wikipedia data — A small, but very useful, subset (~ 1 Billion triples or 14%) of the whole DBpedia extraction using the DBpedia Information Extraction Framework (DIEF), comprising structured information extracted from the English Wikipedia plus some enrichments from other Wikipedia language editions, notably multilingual abstracts in ar, ca, cs, de, el, eo, es, eu, fr, ga, id, it, ja, ko, nl, pl, pt, sv, uk, ru, zh.

• Links — 62 million community-contributed cross-references and owl:sameAs links to other linked data sets on the Linked Open Data (LOD) Cloud that allow to effectively find and retrieve further information from the largest, decentral, change-sensitive knowledge graph on earth that has formed around DBpedia since 2007.

https://www.dbpedia.org/resources/snapshot-release/
Another approach (for dbpedia):

- Dbpedia “Wayback machine”
  
  https://wayback.cluster.ai.wu.ac.at/

The DBpedia Wayback Machine * SEMANTiCS2015

Javier D. Fernández  
Vienna University of Economics and Business, Vienna, Austria  
javier.fernandez@wu.ac.at

Patrik Schneider  
Vienna University of Economics and Business, Vienna, Austria  
patrik.schneider@wu.ac.at

Jürgen Umbirch  
Vienna University of Economics and Business, Vienna, Austria  
juergen.umbrich@wu.ac.at

Challenge: keep these services running...
Another Open Question:

• How representative are big Open Collaborative KGs like Dbpedia & Wikidata at all?

... What about Enterprise KGs?

Challenge: not accessible 😞

Bachelor Thesis

SMW Cloud: A Corpus of Domain-Specific Knowledge Graphs from Semantic MediaWikis
Daniil Dobriy

Wikis

- Powered by MediaWiki software.
- We know of 60527 currently active wikis.
  (It’s a lot 😊)
How many **Semantic MediaWikis**?

SMW Cloud (1458 wikis)

Crawled RDF data available at [semantic-data.cluster.ai.wu.ac.at/smwcloud/](http://semantic-data.cluster.ai.wu.ac.at/smwcloud/)

Currently ongoing work/next steps:
- also crawl historic data (Semantic MediaWiki edit history)
- also crawl Wikiba.se instances!
SMW Cloud:

- Very different KGs than Dbpedia and Wikidata
- Small, narrow scheme
- ...very likely, very different evolution!
What dimensions does evolution have at all?

- Instance evolution
- Schema evolution
- Collaboration evolution

- Structural evolution
- Dynamics (change frequencies, etc.)
- Timeliness (recency of temporal information, delays)
- Monotonicity (“growth” vs. “deletions”)

![Diagram showing the relationship between different types of knowledge graphs and the time dimensions of evolution. The diagram illustrates the concepts of temporal KG, dynamic KG, and versioned KG, with axes representing continuous change and discrete changes. The time dimensions are indicated with "Time as data" (valid time) and "Time as metadata/log data" (transaction time).]
Underlying Collaborative KG-Creation Processes

<table>
<thead>
<tr>
<th>KG</th>
<th>Expert-driven</th>
<th>Crowd-sourced</th>
<th>Resource-dependent</th>
<th>Community-driven</th>
<th>Bot-assisted</th>
</tr>
</thead>
<tbody>
<tr>
<td>Wikidata [244]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>DBpedia [148]</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

• **Question**: How can we observe and analyse KG collaboration models?

Following:
Possible further directions to analyse collaboration in more detail:

- **Question:** How does the schema evolve in relation to the data?
- **Question:** How is the use of the schema related to specific user communities?
Possible further directions to analyse collaboration in more detail:

- **Challenge:**
  - *Again: User/collaboration data not “readily available”*

---

<table>
<thead>
<tr>
<th>index</th>
<th>username</th>
<th>userID</th>
<th>vocabulary</th>
<th>entity</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>&quot;David&quot;</td>
<td>1234</td>
<td>P31:0128207,</td>
<td>Q2</td>
</tr>
<tr>
<td>2</td>
<td>&quot;David&quot;</td>
<td>1234</td>
<td>P2067,</td>
<td>Q2</td>
</tr>
<tr>
<td>3</td>
<td>&quot;David&quot;</td>
<td>u1234</td>
<td>P138,</td>
<td>Q2</td>
</tr>
<tr>
<td>4</td>
<td>&quot;David&quot;</td>
<td>u1234</td>
<td>P2579,</td>
<td>Q3</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Towards analysing the evolution of community-driven (sub-)schemas within Wikidata**

Nicola Pascal Krenn

[http://polleres.net/supervised_theses/Nicola_Krenn_MSc_2023.pdf](http://polleres.net/supervised_theses/Nicola_Krenn_MSc_2023.pdf)
Main questions:

• Which publicly accessible, open KGs are observable in a manner that would allow a longitudinal analysis of their evolution and how?

• Do we have the right metrics to analyse KGs’ evolution?

• Do we have the right techniques to process evolving KGs?
From static metrics to dynamic metrics

- Basic (static) **Graph metrics**, e.g.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Description</th>
<th>Used/Defined in</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute depth</td>
<td>$d_a = \text{sum over the cardinality of each path in a set of paths in graph}$</td>
<td>IsA graph [11, 91, 143, 252]</td>
</tr>
<tr>
<td>Average depth</td>
<td>$d_a /</td>
<td>\text{paths}</td>
</tr>
<tr>
<td>Maximal depth</td>
<td>longest path</td>
<td>IsA graph [11, 91, 143], graph [37]</td>
</tr>
<tr>
<td>Number of paths</td>
<td>$</td>
<td>\text{paths}</td>
</tr>
<tr>
<td>Tangledness</td>
<td>$\frac{n_G}{t}$, $n_G = \text{cardinality of G, } t = \text{cardinality of the set of nodes with more than one ingoing IsA arc in G}$</td>
<td>IsA graph [11, 91, 143]</td>
</tr>
<tr>
<td>Degree Distribution</td>
<td>mean-square deviation of the degree of graph nodes</td>
<td>graph [37, 67, 143]</td>
</tr>
</tbody>
</table>
From static metrics to dynamic metrics

- Basic (static) **Knowledge Graph metrics**, e.g.

<table>
<thead>
<tr>
<th>Primitives</th>
<th>Definition</th>
<th>Knowledge Graph T-Box/Schema</th>
<th>Knowledge Graph A-Box/Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Entities</td>
<td>number of entities, classes and instances</td>
<td>graph [37, 143], IsA graph [91], OWL [215], DAG [252]</td>
<td>number of connected components</td>
</tr>
<tr>
<td>Properties</td>
<td>number of unique properties or relations</td>
<td>OWL schema [174], OWL [229, 231, 233], DAG [254]</td>
<td>number of connected components</td>
</tr>
<tr>
<td>Classes</td>
<td>$</td>
<td>C</td>
<td>= \text{number of classes (concepts)}$</td>
</tr>
<tr>
<td>Instances</td>
<td>$</td>
<td>I</td>
<td>= \text{number of individuals}$</td>
</tr>
<tr>
<td>Object properties</td>
<td>$P_o = \text{number of object properties (non-inheritance)}$</td>
<td>Schema [143], OWL [215, 231]</td>
<td>number of connected components</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge Graph</th>
<th>T-Box/Schema</th>
<th>A-Box/Data</th>
</tr>
</thead>
<tbody>
<tr>
<td>Depth of Inheritance Tree</td>
<td>Tree [174], OWL [73, 207, 229], DAG [252]</td>
<td>number of connected components</td>
</tr>
<tr>
<td>Property Class Ratio</td>
<td>OWL [231, 174, 73], DAG [252]</td>
<td>number of connected components</td>
</tr>
<tr>
<td>Inheritance Richness</td>
<td>OWL [71, 73, 207, 229], Schema [143]</td>
<td>number of connected components</td>
</tr>
<tr>
<td>Attribute Richness</td>
<td>OWL [71, 229], Schema [143]</td>
<td>number of connected components</td>
</tr>
<tr>
<td>Class Property Ratio</td>
<td>Onto [11, 91, 143]</td>
<td>number of connected components</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Knowledge Graph</th>
<th>T-Box/Schema</th>
</tr>
</thead>
<tbody>
<tr>
<td>Average Population</td>
<td>OWL [73, 229], Onto [91]</td>
</tr>
</tbody>
</table>
Challenge: What do these metrics tell us over time?

- **Bottomline/Challenges:**
  - These metrics are not sufficient to track *patterns of evolution*...
  - **We need to track changes on a finer granularity level**
  - **We need new metrics (from other fields):**
    - *Time series analyses (change frequencies, seasonality)*
    - *Network science (dynamics of networks)*
    - etc.
More Open question(s)

• How does consistency evolve over time (and why is this important)?

• Challenge:
  • Our “classical” tools (OWL, SHACL?) are not really useful here directly
Reasoning + Constraints in Wikidata: OWL?

- Challenges:
  - Wikidata does not use OWL!
  - Big Open KGs are all quite inconsistent!

<table>
<thead>
<tr>
<th>RDFS/OWL property</th>
<th>Equivalence established through Wikidata property</th>
</tr>
</thead>
<tbody>
<tr>
<td>rdfs:type</td>
<td>equivalent property (P1628)</td>
</tr>
<tr>
<td>rdfs:subClassOf</td>
<td>equivalent property (P1628)</td>
</tr>
<tr>
<td>rdfs:subPropertyOf</td>
<td>equivalent property (P1628)</td>
</tr>
<tr>
<td>rdfs:range</td>
<td>equivalent property (P1628)</td>
</tr>
<tr>
<td>rdfs:domain</td>
<td>equivalent property (P1628)</td>
</tr>
<tr>
<td>rdfs:label</td>
<td>documented as matching</td>
</tr>
<tr>
<td>rdfs:comment</td>
<td>documented as matching</td>
</tr>
<tr>
<td>rdfs:first</td>
<td>documented as matching</td>
</tr>
<tr>
<td>rdfs:rest</td>
<td>documented as matching</td>
</tr>
<tr>
<td>rdfs:member</td>
<td>documented as matching</td>
</tr>
<tr>
<td>rdfs:member</td>
<td>inverse property (P1696) of part of (P361)</td>
</tr>
<tr>
<td>owl:equivalentProperty</td>
<td>equivalent property (P1628)</td>
</tr>
<tr>
<td>owl:equivalentClass</td>
<td>equivalent property (P1628)</td>
</tr>
<tr>
<td>owl:inverseOf</td>
<td>equivalent property (P1628)</td>
</tr>
<tr>
<td>owl:disjointUnionOf</td>
<td>equivalence intended</td>
</tr>
<tr>
<td>owl:disjointWith</td>
<td>no documented equivalence</td>
</tr>
<tr>
<td>owl:sameAs</td>
<td>no documented equivalence</td>
</tr>
<tr>
<td>owl:disjointWith</td>
<td>no documented equivalence</td>
</tr>
<tr>
<td>owl:propertyDisjointWith</td>
<td>no documented equivalence</td>
</tr>
<tr>
<td>owl:propertyChainAxiom</td>
<td>no documented equivalence</td>
</tr>
<tr>
<td>owl:assertionProperty</td>
<td>no documented equivalence</td>
</tr>
</tbody>
</table>
Reasoning+Constraints in Wikidata: SHACL?

Challenges:
• Wikidata does not use SHACL either!
  • Formalization
  • Analysis of violations over time!

Formalizing and Validating Wikidata’s Property Constraints using SHACL and SPARQL

Nicolas Ferranti a, Jairo Francisco De Souza b, Shqiponja Ahmetaj c and Axel Polleres

https://www.semantic-web-journal.net/content/formalizing-and-validating-wikidatas-property-constraints-using-shacl-and-sparql
Reasoning + Constraints in Wikidata: SHACL?

• Challenges:
• Wikidata does not use SHACL either!
  • Formalization - what makes it challenging? ---> Reification galore! ;(-)

![Diagram of Wikidata relationships with SHACL constraints](image-url)
Main questions:

• Which publicly accessible, open KGs are observable in a manner that would allow a longitudinal analysis of their evolution and how?

• Do we have the right metrics to analyse KGs’ evolution?

• Do we have the right techniques to process evolving KGs?
Do we have the right techniques to process evolving KGs?

- What else will you find in our paper? Survey of ...
  - Storage techniques for evolving KGs
  - Reasoning & Querying techniques for evolving KGs
  - Learning & Embeddings for evolving KGs

- Challenges:
  - Again: How do we make these methods scale to large-scale, evolving, collaborative KGs?
  - E.g. How to reason and query over evolving KGs?
    - Need to extend our techniques to deal with reification?
    - What’s the “right” reification?
      - Labelled property Graphs?
      - RDF-*?
      - Wikidata’s proprietary reification mechanism?
    - How to scale and modularize existing techniques over highly reified KGs?
Some take home messages:

- There’s a lot to learn about the dynamics of (Open) KGs!
  - Understanding the evolution of knowledge (graphs) is a hot topic!

Yet, there are some major challenges:

- Data Availability:
  - (fine-grained) data about their evolution it not available (Streams!)
  - We need more long-tail data!
  - It’s hard to sustain efforts to sustain data about evolution!

- Metrics:
  - We need metrics and techniques to analyse KG dynamics and evolution:
    - New metrics, look into other fields!
    - Adaptations and extensions of existing metrics

- Techniques: Storage/Querying/Learning
  - How to scale and modularize existing techniques over highly reified KGs?
  - Dynamic Embeddings/Model Dynamics?

- P.S.: Submit to TGDK 😊

  ... would have loved to discuss these challenges with you in person 😊