TrainiG towards a society of data-saVvy inforMation prOfessionals to enable open leadership INnovation

Analyzing the Evolution of Vocabulary Terms and their Impact on the LOD Cloud

Mohammad Abdel-Qader, Ansgar Scherp, and Iacopo Vagliano
ZBW – Leibniz Information Centre for Economics
Christian-Albrechts-University in Kiel, Germany

June 6th , 2018, 15th ESWC 2018,
03.06.2018- 07.06.2018, Heraklion, Crete, Greece.

www.moving-project.eu

Introduction

- Vocabulary terms define the schema of Knowledge Graphs (KGs) such as the Linked Open Data (LOD) cloud or Wikidata.
- But vocabularies are subject to change.
- So far it is unknown how such vocabulary changes are reflected by the KGs.
- Explicitly triggering data publishers to update their model is also challenging due to the distributed nature of KGs.
- Ontology engineers are not aware of who uses their vocabularies

Research Questions

1. What is the use rate of terms for a set of vocabularies in each dataset? And by whom?
2. When are the newly created terms adopted in Knowledge Graphs?
3. What is the rate of using the deprecated terms in Knowledge Graphs?

Analysis Method

- Two steps:

1. We determined vocabularies that have two or more published versions on the web.
   - We selected a set of vocabularies that:
     a. have at least two published versions,
     b. at least two versions are covered by the dataset that we investigate,
     c. and the vocabulary terms occur in at least one triple in the published dataset.

Analysis Method

2. We investigated how the changed terms of vocabularies are adopted and used in the evolving Knowledge Graphs.

   - Extract all the Pay Level Domains (PLDs) from the subject of the crawled triples that use any of the terms.

   - A PLD is any web domain that requires payment at a Top Level Domain (TLD) or country code TLD (cc-TLD) registrar [1]

1. H.-T. Lee, D. Leonard, X. Wang, and D. Loguinov, “Irlbot: scaling to 6 billion pages and beyond,” ACM Transactions on the Web (TWEB), vol. 3, no. 3, p. 8, 2009.

Datasets

- We applied our analysis approach on three large-scale KGs:
  - Dynamic Linked Data Observatory (DyLDO)[1].
    - 242 snapshots.
  - Billion Triples Challenge (BTC)[2].
    - 2009 to 2012, and 2014.
  - Wikidata[3].
    - 25 RDF dump files.

[1] http://km.aifb.kit.edu/projects/dyldo/
[2] http://challenge.semanticweb.org/
[3] https://www.wikidata.org

Selecting the vocabularies

134 Most used vocabularies

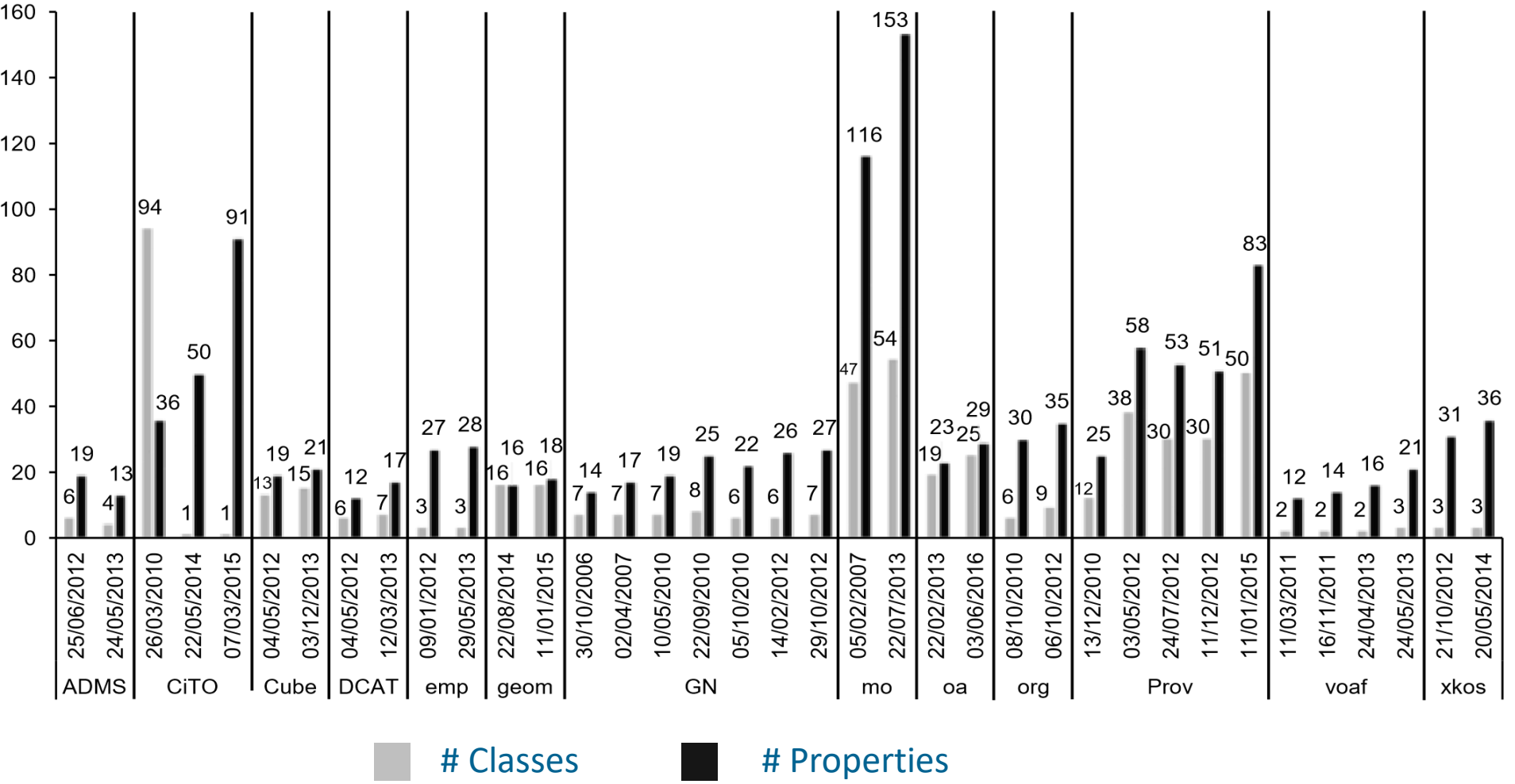
18 Have more than one version after the timeframe of datasets

13 Have changes

Resulted vocabularies

Vocabulary	Prefix	Versions	Term Changes
Asset Description Metadata Schema	ADMS	2	18
Citation Typing Ontology	CiTO	3	218
The data cube vocabulary	Cube	2	6
Data Catalog Vocabulary	DCAT	2	13
A vocabulary for jobs	EMP	2	1
Ontology for geometry	GEOM	2	2
The Geonames ontology	GN	7	31
The music ontology	MO	2	46
Open Annotation Data Model	OA	2	31
Core organization ontology	ORG	2	8
W3C PROVenance Interchange	PROV	5	168
Vocabulary of a Friend	VOAF	4	8
An extension of SKOS for representation of nomenclatures	XKOS	2	1

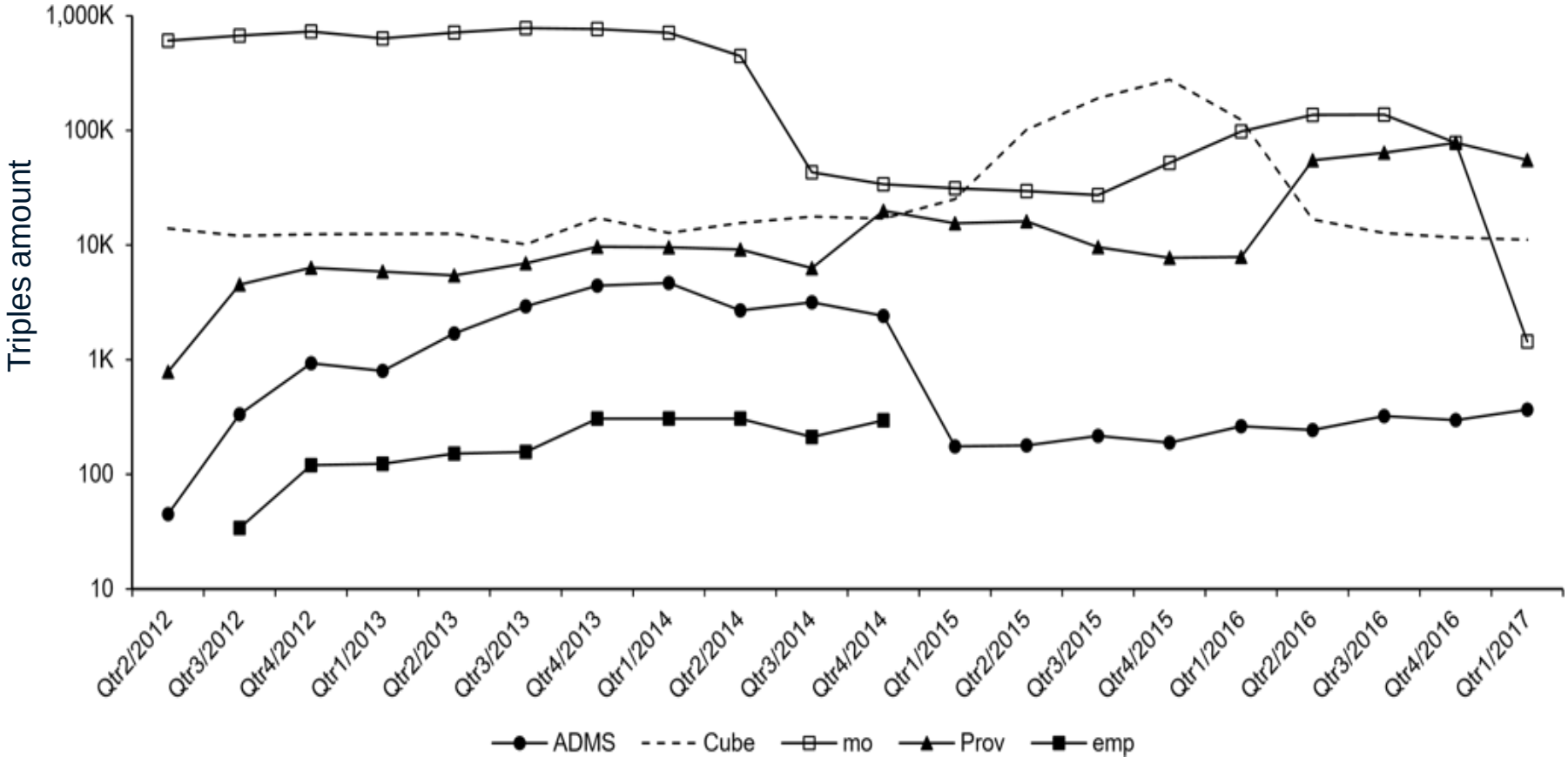
Overview on vocabulary versions



RQ1- Results- PLDs

Dataset	PLD	Triples
BTC 2009	geonames.org	81M
BTC 2010	geonames.org	7M
BTC 2011	zitgist.com	2.6M
BTC 2012	rdfize.com	3.8M
BTC 2014	dbtune.org	81.5M
DyLDO	dbtune.org	160M

RQ1- Results- Vocabularies use in DyLDO



RQ2- Results- New terms adoption in DyLDO

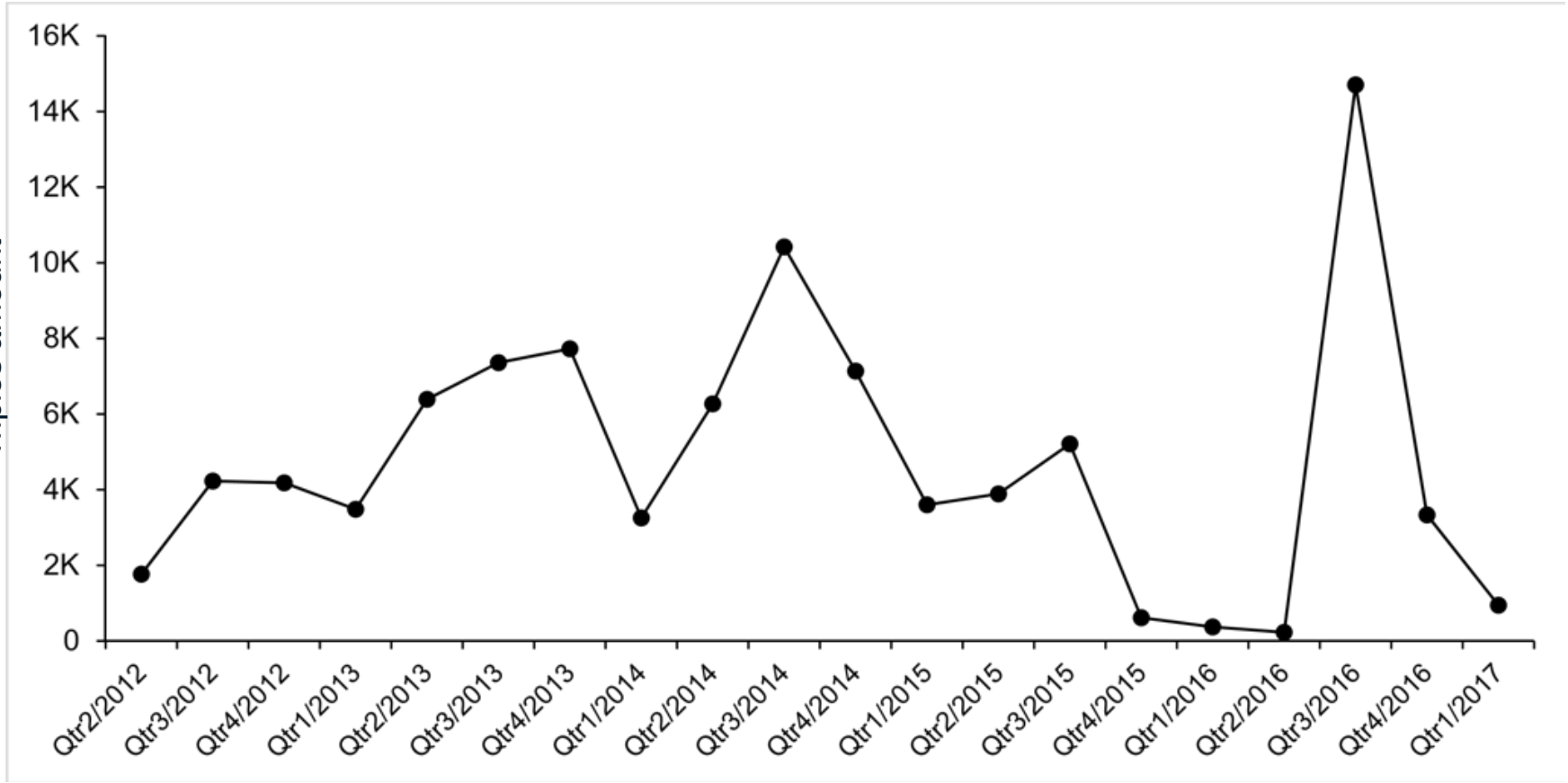
Vocabulary	New terms	Adopted terms	Triples	μ (Days)	σ
ADMS	6	100%	31K	7	0
CiTO	80	100%	281K	7	0
Cube	5	100%	15K	7	0
DCAT	5	100%	104K	8.4	3.13
EMP	1	100%	4K	7	0
GEOM	2	100%	16K	420	0
GN	21	100%	160M	127.76	255.33
MO	44	100%	45M	8.75	9.68
OA	21	0%	-	-	-
ORG	8	100%	173K	7	0
PROV	106	85%	121M	30.15	37.49
VOAF	10	100%	75K	43.33	68.58
XKOS	1	0%	-	-	-

RQ2- Results- New terms adoption in DyLDO

Vocabulary	New terms	Adopted terms	Triples	μ (Days)	σ
ADMS	6	100%	31K	7	0
CiTO	80	100%	281K	7	0
Cube	5	100%	15K	7	0
DCAT				.4	3.13
EMP				7	0
GEOM				20	0
GN				7.76	255.33
MO	44	100%	45M	8.75	9.68
OA	21	0%	-	-	-
ORG	8	100%	173K	7	0
PROV	106	85%	121M	30.15	37.49
VOAF	10	100%	75K	43.33	68.58
XKOS	1	0%	-	-	-

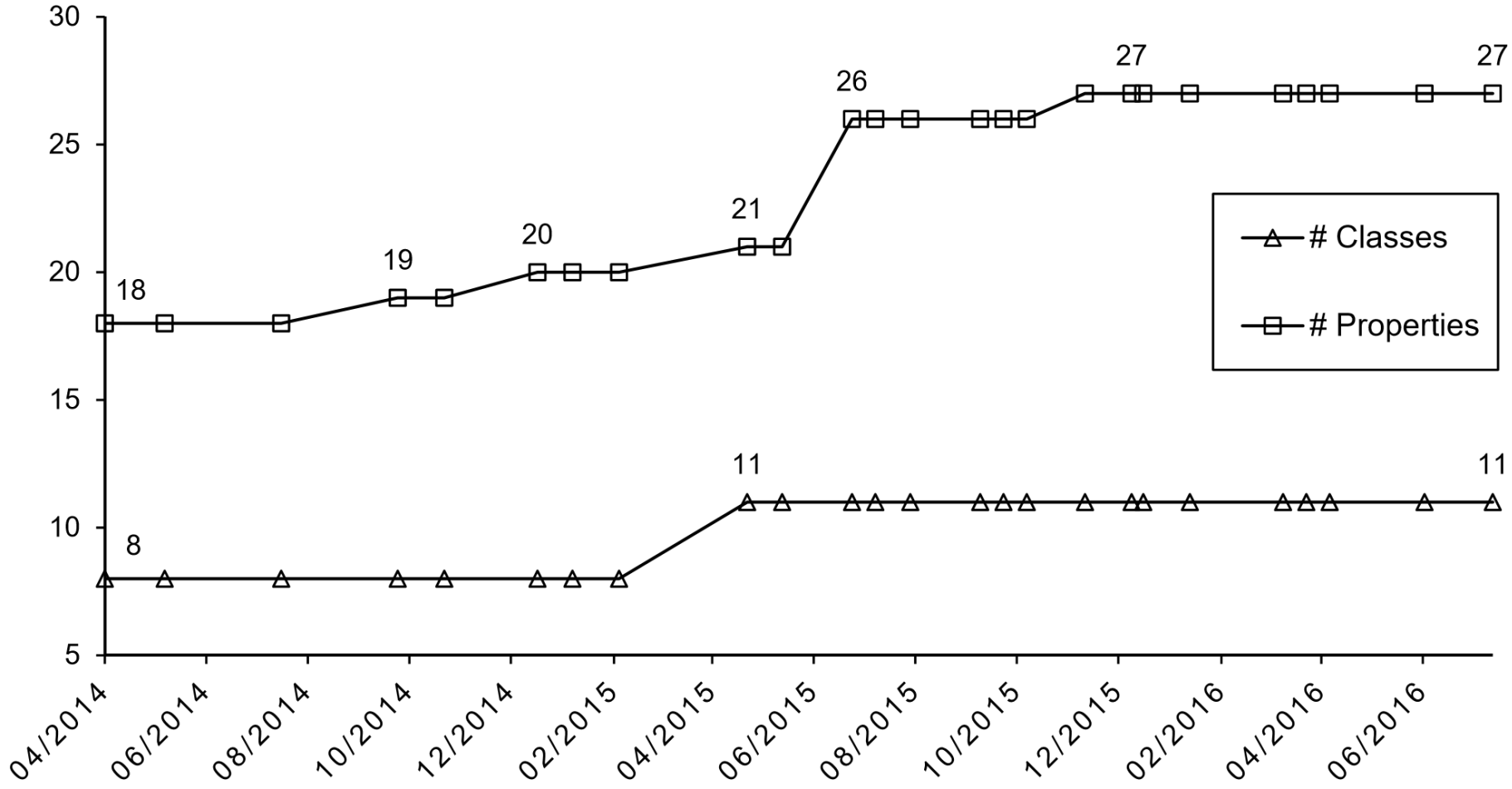
We found that some terms were adopted before their official publishing date!

RQ3- Results- gn:Country use in DyLDO

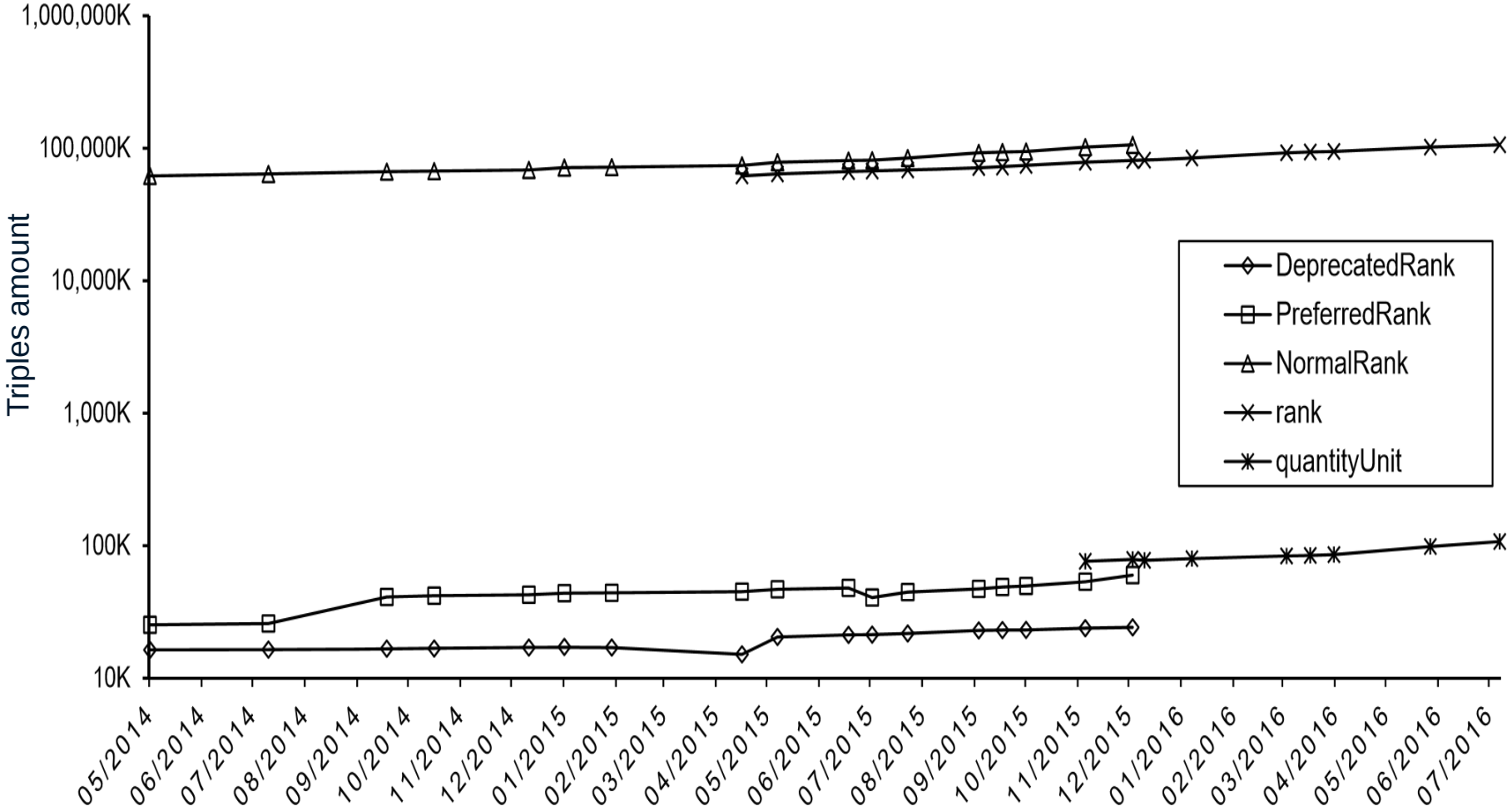


Note: `gn:Country` was deprecated in September 2010

Results- Wikidata Change



Results- Wikidata new terms adoption



Conclusions

- Even small changes of vocabulary terms can have a deep impact on the real data that use those terms.
- Most of newly coined terms are adopted immediately afterward.
- Most of the deprecated terms are still used.
- Foster a better understanding what is the impact of their changes and how they are adopted.

Thank you for your attention!

Any questions?

Project consortium and funding agency

MOVING is funded by the EU Horizon 2020 Programme under the project number INSO-4-2015: 693092

References

1. Abdel-Qader, M., Scherp, A.: Towards understanding the evolution of vocabulary terms in knowledge graphs. ArXiv e-prints (Sep 2017), https://arxiv.org/pdf/1710.00232.pdf
2. Abdel-Qader, M., Scherp, A.: Qualitative analysis of vocabulary evolution on the linked open data cloud. In: PROFILES Workshop co-located with ESWC, Volume 1598. CEUR-WS. org (2016)
3. Chawuthai, R., Takeda, H., Wuwongse, V., Jinbo, U.: Presenting and preserving the change in taxonomic knowledge for linked data. Semantic Web 7(6), 589{616 (2016)
4. Dividino, R., Scherp, A., Gröner, G., Grotton, T.: Change-a-LOD: does the schema on the linked data cloud change or not? In: Consuming Linked Data Workshop colocated with ISWC, Volume 1034. pp. 87{98. CEUR-WS. org (2013)
5. Gottron, T., Knauf, M., Scherp, A.: Analysis of schema structures in the linked open data graph based on unique subject URIs, pay-level domains, and vocabulary usage. Distributed and Parallel Databases 33(4), 515{553 (2015)
6. Guha, R.V., Brickley, D., Macbeth, S.: Schema.org: Evolution of structured data on the web. Communications of the ACM 59(2), 44{51 (2016)
7. Käfer, T., Abdelrahman, A., Umbrich, J., O'Byrne, P., Hogan, A.: Observing linked data dynamics. In: ESWC. pp. 213{227. Springer (2013)
8. Käfer, T., Umbrich, J., Hogan, A., Polleres, A.: Towards a dynamic linked data observatory. LDOW co-located with WWW (2012)
9. Meusel, R., Bizer, C., Paulheim, H.: A web-scale study of the adoption and evolution of the schema.org vocabulary over time. In: International Conference on Web Intelligence, Mining and Semantics. p. 15. ACM (2015)
10. Mihindukulasooriya, N., Poveda-Villalon, M., Garca-Castro, R., Gomez-Perez, A.: Collaborative ontology evolution and data quality-an empirical analysis. In: International Experiences and Directions Workshop on OWL. pp. 95{114. Springer (2016)