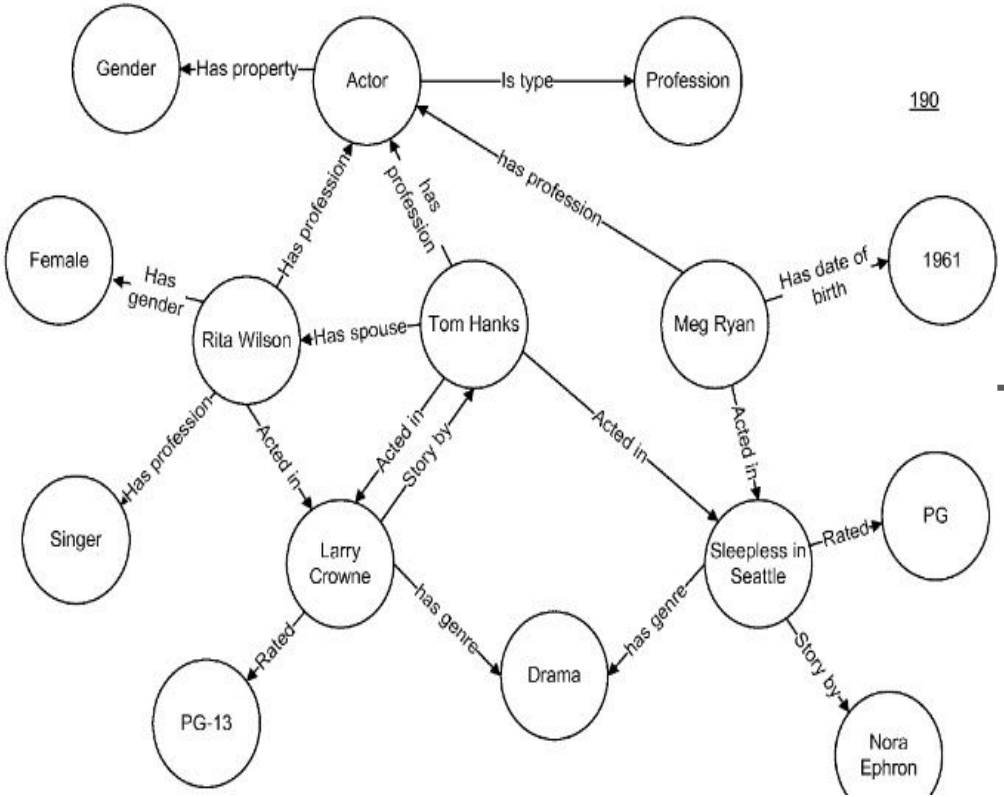
Archiving and versioning of Knowlegde Graphs

Digging into the Knowledge Graph
Univ. of Wisconsin, Nov. 2017

Contents

- **Defining** (limiting/scoping): Knowledge graphs as set of triples
- **Archiving**: copying or diffs?
- **Versioning**: getting insights into changes
- **Provenance of changes**: who, what, when and how
- **Implementing**: workflows and architecture
- **Insights**: using diffs and prov for statistics on usage, concept drift etc

- **Defining** (limiting/scoping): Knowledge graphs as set of triples



Subject	Predicate	Object
Actor	Is_type	Profession
Meg Ryan	has_prof	Actor
Meg Ryan	birth_date	"1961"

- **Archiving**: copying or diffs? (1/2)

V1

Subject	Predicate	Object
Actor	Is_type	Profession
Meg Ryan	has_prof	Actor
Meg Ryan	birth_date	"1961"
Meg Ryan	lives_in	New York

+

V2

Subject	Predicate	Object
Actor	Is_type	Profession
Meg Ryan	has_prof	Actor
Meg Ryan	birth_date	"1961"
Meg Ryan	lives_in	Paris

OR

Subject	Predicate	Object
Actor	Is_type	Profession
Meg Ryan	has_prof	Actor
Meg Ryan	birth_date	"1961"
Meg Ryan	lives_in	New York

+

ADD/DEL	Subject	Predicate	Object
DEL	Meg Ryan	lives_in	New York
ADD	Meg Ryan	lives_in	Paris

- **Archiving**: copying or diffs? (2/2)

CHALLENGES WITH MAKING DIFFS ON LINKED DATA TRIPLES

- Blank nodes
- Sorting

In other words, if two sets of triples result in the same knowledge graph, the canonical versions of both sets should be identical.

Related work: CARROLL, Jeremy J., et al. Named graphs, provenance and trust. In: *Proceedings of the 14th international conference on World Wide Web*. ACM, 2005. p. 613-622.
https://pdfs.semanticscholar.org/3190/6a2bc21c7871c72da219d712f78988eb73ed.pdf

- **Versioning**: getting insights into changes

"The object from predicate "lives_in" has a high frequency of changes"

"For two years source X adds many triples containing subjects of type 'book' with predicate 'has_ISBN' " (probably a publisher)

- **Provenance of changes**: who, what, when and how

In the archiving community, provenance is very important.

W3C PROV is a standard to describe the provenance information for data

Moreau, Luc, and Paul Groth. "Provenance: an introduction to prov." *Synthesis Lectures on the Semantic Web: Theory and Technology* 3.4 (2013): 1-129.

https://www.w3.org/TR/2013/NOTE-prov-primer-20130430/

Vyacheslav's work in the Parthenos project uses W3C prov to annotate their data

- **Implementing**: workflows and architecture

<<Nice diagrams and illustrations here on LOD Laundromat, Memento protocol, Linked-data fragments, and our idea on using LD-DIFFS for versioning>>

- **Insights**: using diffs and prov for statistics on usage, concept drift etc

Using Memento to develop micro processes that do fancy analysis on the things we are interested in.

- Concept drift
- Which graphs are used by other graphs
- Which graphs are stable
- Which graphs are often used
- Etc