ProvQL: Querying about the Provenance of your Data

Argyro Avgoustaki, Giorgos Flouris,
Irini Fundulaki, Dimitris Plexousakis

ΕΛΙΔΕΚ
Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας
ΓΓΕΤ
ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ
ΕΡΕΥΝΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
FORTH
INSTITUTE OF COMPUTER SCIENCE

Provenance in the World

- A large volume of heterogeneous data are published on the Web
- Everyone is a potential publisher of data that are available for free access and use
- Need for assessing the quality and reliability of data

  ⇨ Recording the *provenance* of data

- Provenance is the history of data and can be used to support applications related to:
  - ✓ Data Quality
  - ✓ Reliability
  - ✓ Trustworthiness
  - ✓ Copyrights
  - ✓ Access Control
  - ✓ Accountability

ΕΛΙΔΕΚ Ελληνικό Ίδρυμα Έρευνας & Καινοτομίας
ΓΓΕΤ ΓΕΝΙΚΗ ΓΡΑΜΜΑΤΕΙΑ ΕΡΕΥΝΑΣ ΚΑΙ ΤΕΧΝΟΛΟΓΙΑΣ
FORTH INSTITUTE OF COMPUTER SCIENCE

ProvQL Query Language

- A great number of provenance models has been proposed
- **ProvQL** is a high-level structured query language suitable for seeking information related to data provenance
- ProvQL can answer provenance queries such as:
  1. Identify the different ways to derive a specific data record?

     ```
     SELECT SPROV(?id) WHERE QUADS(?id) = (<a>, <b>, <c>, <n>)
     ```

  2. Identify all data records whose provenance includes a specific data source

     ```
     SELECT QUADS(?id) WHERE PROV(?id) CONTAINS c1
     ```

  3. Identify all sources that originate from a data item referring to “Donald_Trump”

     ```
     SELECT PROV(?id) WHERE QUADS(?id) = (?a, ?b, “Donald Trump”, ?n)
     ```