CODE

Commercially Empowered
Linked Open Data
Ecosystems in Research

Unleashing Semantics of Research Data

Florian Stegmaier

2nd Workshop on Big Data Benchmarking
18. December 2012
Pune, India

SEVENTH FRAMEWORK
PROGRAMME

supported by the European Commission
under the Seventh Framework Program (FP7)

UNIVERSITÄT
PASSAU

The dark side of research data

- Terrabytes of research data available, but
  - ... with varying quality
  - ... with contradicting facts
  - ... with missing data
  - ... labour intensive to compare

> *There is increasing concern that most current published research findings are false...* Ioannidis, 2005

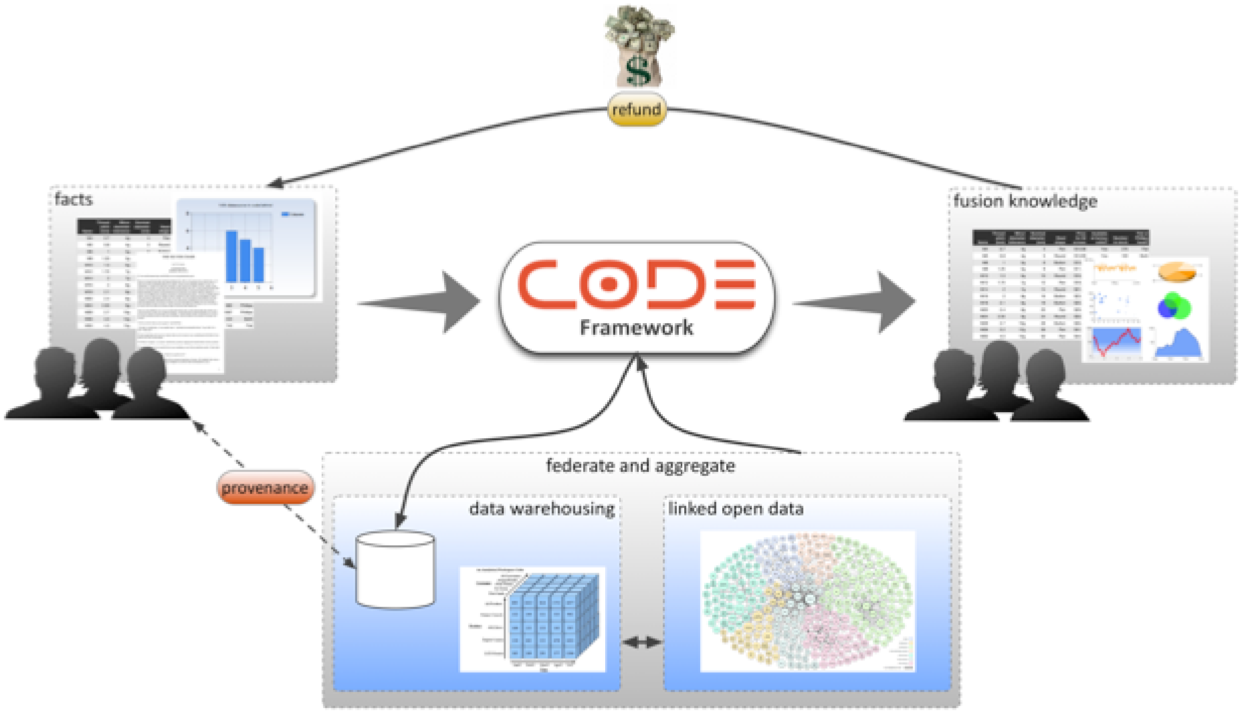
> *Dozens of individual published experiments report effectiveness improvements, and often claim statistical significance...* Armstrong et al., 2009

Global vision of CODE: „The Linked (Open) Science Cloud“

- Research data available in various formats:
  - Research paper (e.g., catalogues such as Mendeley)
  - Primary research data (e.g., evaluation campaigns)
  - Retrievable data (e.g., “ACM bubble” in Linked Data cloud)
  - Embedded data (e.g., exposed by microdata / -format)

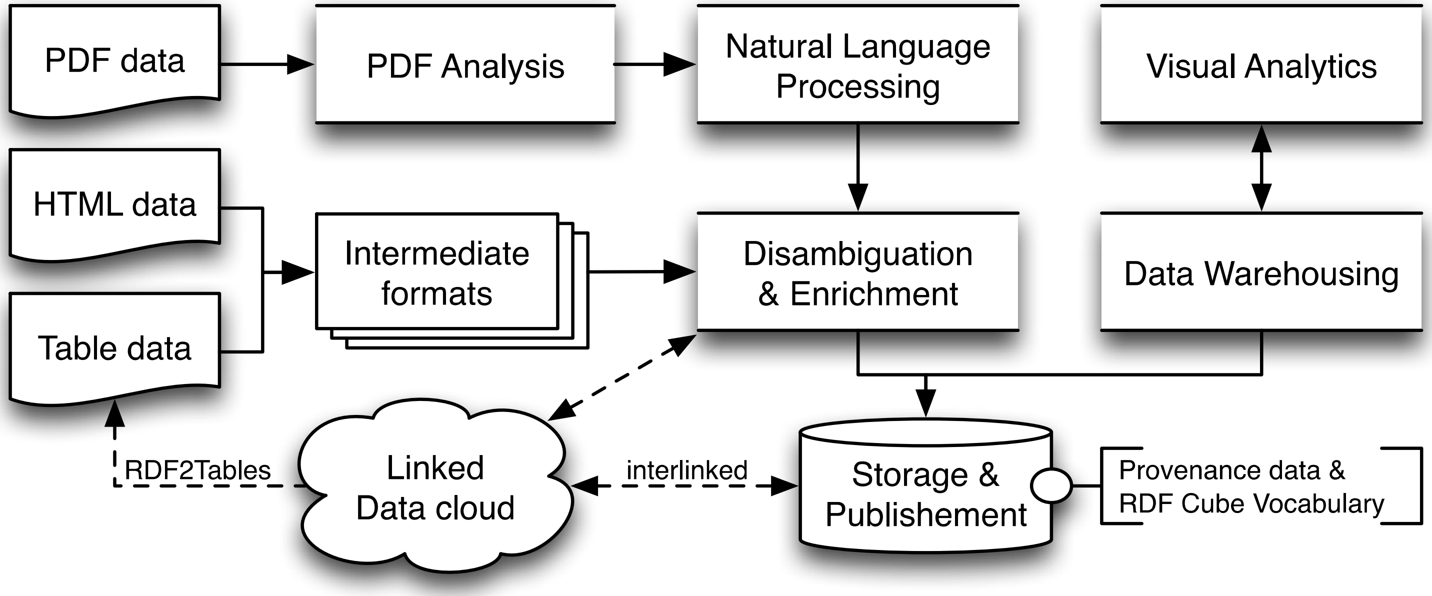
- Key-features of the ecosystem:
  - Knowledge extraction via atomic processing parts
  - Marketplace concepts lead to crowdsourcing
  - Integration of provenance fosters value-creation chains
  - Concepts of Linked Data enable a sophisticated data warehouse like retrieval

The long way to knowledge: The CODE view



13 TB of research data encapsulated in PDF, 3 M. users for crowdsourcing

Lifting of primary research data



...following stages of the „Big Data pipeline“ as well as observations of Labrinidis and Jagadish.

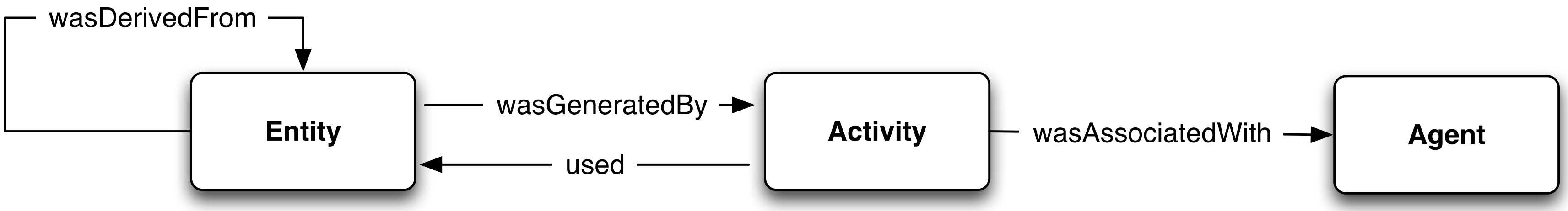
Really Big Data?
The classical "Vs" approach

- **Volume**
  - Explicit facts from research papers exposed as data warehouse
  - Interaction of peers with data (e.g., citing)
- **Velocity**
  - Real-time production (e.g., sensor data)
  - Batch-like production (e.g., conferences)
  - Single publication (e.g., white paper)
- **Variety**
  - Unstructured data (e.,g., PDF documents)
  - Semi-structured data (e.g., Excel spreadsheets)
  - Structured data (e.g., exposed in a Blog via Dublin Core)

Why provenance matters? The „semantic“ V(alue)

Mitchell and Wilson, 2012

- Every portions of data exposes indirect provenance
- Provenance chains enable mature interaction:
  - Tracing abilities
  - Quality estimation of the underlying data
  - “What interaction made the data worthy?”



Issues that we face...

- Exposing data portions via recent international W3C standards
  - Data warehousing: RDF Cube Vocabulary
  - Provenance: PROV-O Ontology

- Efficient internal storage:
  - (Big Data) Benchmarking must take place to ensure scalability
  - Interconnection between both data models requires efficient structures

How CODE could support Big Data Benchmarking

- CODE framework (already) offers services…
  - … to lift and interlink primary research / evaluation data
  - … to perform visual analysis on this data
  - … to manage time-dependent data

- Marketplace concepts for community engagement

> “In theory, there is no difference between theory and practice; in practice there is.
>
> - Chuck Reid, Yogi Berra

Thank you for your attention!

Online: www.code-research.eu
Twitter: @CODEresearchEU
Facebook: CODEresearchEU