On-the-fly Integration of Static and Dynamic Linked Data

Prof. Dr. Rudi Studer

Social Computing and Crisis in the New Information Age, Brussels
Oct 8, 2013
Election Monitoring

- Parliamentary elections in Albania
- Citizen portal ZaLart.al (based on Ushahidi platform)
- Citizens can report issues via mobile phones etc.

Reports by category

- Violent Incidents: 12
- Damages to Electoral Materials or Offices: 1
- Unethical Electoral Speech: 2
- Problems with Electoral Commissions: 9
- Breach of the Right to Vote and the Anonymity of the Ballot: 6
- Tension near CEAZ/VCC: 8
- Problems delivering electoral materials: 2
- Misuse of the State Police: 1
- Unbalanced Media Coverage: 3
- Other problems: 7
OpenStreetMap Project Haiti

http://vimeo.com/9182869
Issues in Complex Information Ecosystems

- The relevance of many types of data perishes or degrades over time (e.g., flood level, moving objects)
- A lot of websites provide access to their data via an API

- Data integration systems have to
  - Provide integrated access to live data
  - Be extendible for including new sources in an efficient manner

http://programmableweb.com
Semantic Web Technologies

- Enable the transition from the Web of Documents” to the “Web of Data”

- Technological basis: W3C Standards Resource Description Framework (RDF) and Web Ontology Language (OWL)

- Linked (Open) Data is “Semantic Web light”
  - Hitherto mostly focused on data publishing
  - Amount of available Linked Data is exploding

- Consuming and updating Linked Data is still to be improved
  - Semantics enable better means for data integration
Dynamic Data to Decisions (D3)

- Project in cooperation with USC/ISI
- Funded by ONR Global

Objectives:

- Real-time data integration for a variety of data sources in a range of data formats and access modalities
  - needed e.g. for election monitoring, crisis management and other quickly-evolving situations
- Ability to quickly add sources (via modeling), see KARMA
- Handling of geospatial data with temporal aspects
Wrapping Sources with KARMA

- System acts on RDF sources
- However, we need to interact with all kinds of tabular and tree like data

- KARMA is a visual data integration tool from ISI/USC
  - Maps all kinds of sources onto an ontology to generate Linked Data

![Diagram of KARMA interface](image-url)
Data-Fu - Overview

- Data-Fu is a **declarative rule language** to enable developers
  - to combine resources from several APIs
  - by flexibly defining the behaviour of the developed application (**dynamic reaction to states of resources at runtime**)
  - including the automated activation of links (i.e., that are discovered at runtime)
  - **scalable** execution

---

On-the-fly Integration of Static and Dynamic Linked Data
System Architecture

On-the-fly Integration of Static and Dynamic Linked Data
D3-Application: POIs, Events, Buses, … in L.A.
Enabled Actions

- Based on integrated data sources containing geospatial Regions, Points and georeferenced Events, Twitter Messages, Ushahidi, Epidemiological Records, Gnip data, ... you might realize application that ...

- Notify me if tick-borne encephalitis cases in certain regions are over a given threshold
- Notify me if there are online reports (in social media) of tick-borne encephalitis cases in high-risk areas
- Notify me if the people density in an area rises above a certain threshold
- Send me realtime reports about voter fraud from Ushahidi or from social media
Crowdsourcing of Tasks: Data Quality Improvement

- Acting on data depends on quality of data
  - E.g. resolving entity identities

- **Example:** *Retrieve the ICAO identifier of METAR stations*

```sparql
SELECT ?station ?icao WHERE {
  ?station a metar:Station ;
  owl:sameAs ?airport .
  ?airport a dbp-owl:Airport;
  dbp-ont:icaoLocationId ?icao .
}
```

- Split consolidation task into automated and human computation (humans verify algorithm output)

Source: http://www.slideshare.net/BarryNorton/crowdsourcing-tasks-in-linked-data-management
Identifiers for things

E.g., Fier region is identified via http://gadm.geovocab.org/id/1_71

Access to http://gadm.geovocab.org/id/1_71 returns RDF description (name, description, geometry…)

Other sources (e.g., DBpedia) can reuse the URI to reference unambiguously to Fier, facilitating querying over integrated data

Albania Electoral Districts (Qarks)

1. Berat (8 mandates)
2. Diber (6 mandates)
3. Durres (13 mandates)
4. Elbasan (14 mandates)
5. Fier (16 mandates)
6. Gjirokaster (5 mandates)
7. Korce (12 mandates)
8. Kukes (4 mandates)
9. Lezhe (7 mandates)
10. Shkoder (11 mandates)
11. Tirane (32 mandates)
12. Vlore (12 mandates)

Conclusion

Goal: Interoperation in Complex Information Ecosystems

- Realtime access to data and reactive behaviour for actions
- Ability to rapidly integrate new sources via Karma models and Data-Fu programs

Approach:
- Simplicity based on basic web architecture (URI, HTTP, REST)
- Simplicity based on basic data format (RDF)
- Simplicity based on basic logic dialect (Datalog Rules)
- Scalable interpreter suitable for parallelization

Future work
- Bandwidth issues: considering the “Disadvantaged Edge”
- Integration of crowdsourcing, e.g. for data cleansing