Crowdsourcing semantic data management: challenges and opportunities

Elena Simperl
Karlsruhe Institute of Technology, Germany

Talk at WIMS 2012: International Conference on Web Intelligence, Mining and Semantics
Craiova, Romania; June 2012
Semantic technologies are all about automation

• Many tasks in semantic data management fundamentally rely on human input
  – Modeling a domain
  – Integrating data sources originating from different contexts
  – Producing semantic markup for various types of digital artifacts
  – ...
Great challenges

• Understand what drives users to participate in semantic data management tasks
• Design semantic systems reflecting this understanding to reach critical mass and sustained engagement
Great opportunities

If all US Mobile Internet time was condensed into one hour, how much time would be spent in the most heavily used sectors?

*Other refers to the 14 remaining online sectors visited from mobile phones.
Incentives and motivators

• What motivates people to engage with an application?
• Which rewards are effective and when?

• Motivation is the driving force that makes humans achieve their goals
• Incentives are ‘rewards’ assigned by an external ‘judge’ to a performer for undertaking a specific task
  – Common belief (among economists): incentives can be translated into a sum of money for all practical purposes
• Incentives can be related to extrinsic and intrinsic motivations
Incentives and motivators (2)

• Successful volunteer crowdsourcing is difficult to predict or replicate
  – Highly context-specific
  – Not applicable to arbitrary tasks

• Reward models often easier to study and control (if performance can be reliably measured)
  – Different models: pay-per-time, pay-per-unit, winner-takes-it-all, ...
  – Not always easy to abstract from social aspects (free-riding, social pressure)
  – May undermine intrinsic motivation
TURN WORK INTO PLAY
GWAPs and gamification

• **GWAPs**: human computation disguised as casual games
• **Gamification/game mechanics**: integrate game elements to applications
  – Accelerated feedback cycles
    • Annual performance appraisals vs immediate feedback to maintain engagement
  – Clear goals and rules of play
    • Players feel empowered to achieve goals vs fuzzy, complex system of rules in real-world
  – Compelling narrative
    • Gamification builds a narrative that engages players to participate and achieve the goals of the activity
  – **But in the end it’s about what tasks users want to get better at**
Example: ontology building
Example: relationship finding
Example: ontology alignment
Example: video annotation
Challenges

• Not all tasks are amenable to gamification
  – Work is decomposable into simpler (nested) tasks
  – Performance is measurable according to an obvious rewarding scheme
  – Skills can be arranged in a smooth learning curve
  – Player’s retention vs repetitive tasks

• Not all domains are equally appealing
  – Application domain needs to attract a large user base
  – Knowledge corpus has to be large-enough to avoid repetitions
  – Quality of automatically computed input may hamper game experience

• Attracting and retaining players
  – You need a critical mass of players to validate the results
  – Advertisement, building upon an existing user base
  – Continuous development
OUTSOURCING TO THE CROWD
Microtask crowdsourcing

- Work decomposed into small Human Intelligence Tasks (HITs) executed independently and in parallel in return for a monetary reward.
- Successfully applied to transcription, classification, and content generation, data collection, image tagging, website feedback, usability tests...
- Increasingly used by academia for evaluation purposes
- Extensions for quality assurance, complex workflows, resource management, vertical domains...
Examples

crowdSPRING

Need to get work done? Get Started

Want to earn money now? Learn More

Crowdsourcing ontology alignment

- Experiments using Amazon’s Mechanical Turk and CrowdFlower and established benchmarks
- Enhancing the results of automatic techniques
- Fast, accurate and cost-effective
Challenges

• Not all tasks can be addressed by microtask platforms
  – Routine work requiring common knowledge, decomposable into simpler, independent sub-tasks, performance easily measurable

• Ongoing research in task design, quality assurance (spam), estimated time of completion...
Give me the German names of all commercial airports in Baden-Württemberg, ordered by their most informative description.

„Retrieve the labels in German of commercial airports located in Baden-Württemberg, ordered by the better human-readable description of the airport given in the comment“.

• This query cannot be optimally answered automatically
  – Incorrect/missing classification of entities (e.g. classification as airports instead of commercial airports)
  – Missing information in data sets (e.g. German labels)
  – It is not possible to optimally perform subjective operations (e.g. comparisons of pictures or NL comments)
An integrated solution

- Integral part of Linked Data management platforms
  - At design time application developer specifies which data portions workers can process and via which types of HITs
  - At run time
    - The system materializes the data
    - Workers process it
    - Data and application are updated to reflect crowdsourcing results

- Formal, declarative description of the data and tasks using SPARQL patterns as a basis for the automatic design of HITs

- Reducing the number of tasks through automatic reasoning
Example using SPARQL

"Retrieve the labels in German of commercial airports located in Baden-Württemberg, ordered by the better human-readable description of the airport given in the comment."

SPARQL Query:

```
SELECT ?label WHERE {
  ?x a metar:CommercialHubAirport;
  rdfs:label ?label;
  rdfs:comment ?comment .
  ?x geonames:parentFeature ?z .
  FILTER (LANG(?label) = "de")
}
ORDER BY CROWD(?comment, "Better description of %x")
```
HITs design: Classification

- It is not always possible to automatically infer classification from the properties.
- **Example:** Retrieve the names (labels) of METAR station commercial airports.

```sparql
SELECT ?label WHERE {
  ?station a metar:CommercialHubAirport; rdfs:label ?label .}
```

**Input:**
```sparql
{?station a metar:Station; rdfs:label ?label; wgs84:lat ?lat; wgs84:long ?long}
```

**Output:**
```sparql
{?station a ?type. ?type rdfs:subClassOf metar:Station}
```
HITs design: Ordering

- Orderings defined via less straightforward built-ins; for instance, the ordering of pictorial representations of entities.
- SPARQL extension: ORDER BY CROWD
- Example: Retrieves all airports and their pictures, and the pictures should be ordered according to the more representative image of the given airport.

```sparql
SELECT ?airport ?picture WHERE {
    ?airport a metar:Airport;
    foaf:depiction ?picture .
} ORDER BY CROWD(?picture, "Most representative image for %airport")
```

Input: 

```sparql
{(?airport foaf:depiction ?x, ?y)
```

Output: 

```sparql
{((?x ?y) a rdf:List} UNION {(?y ?x) a rdf:List}}
```
Challenges

• Decomposition of queries
  – Query optimisation obfuscates what is used and should involve costs for human tasks

• Query execution and caching
  – Naively we can materialise HIT results into datasets
  – How to deal with partial coverage and dynamic datasets

• Appropriate level of granularity for HITs design for specific SPARQL constructs and typical functionality of Linked Data management components

• Optimal user interfaces of graph-like content
  – (Contextual) Rendering of LOD entities and tasks

• Pricing and workers’ assignment
  – Can we connect the end-users of an application and their wish for specific data to be consumed with the payment of workers and prioritization of HITs?
  – Dealing with spam / gaming
Thank you

e: elena.simperl@kit.edu, t: @esimperl

Publications available at www.insemtives.org

Team: Maribel Acosta, Barry Norton, Katharina Siorpaes, Stefan Thaler, Stephan Wölger and many others
Realizing the Semantic Web by encouraging millions of end-users to create semantic content.