Session: Social Semantics (9:30 – 11:10)

- Mark Greaves: *Crowdsourcing Semantics*
- Denny Vrandečić, Elena Simperl, Rudi Studer: *Diversity and the Semantic Web*
- Andreas Harth: *Beyond Privacy*
- Denny Vrandečić: *Shortipedia*

VULCAN

Crowdsourcing Semantic Information

Mark Greaves
Vulcan Inc.
markg@vulcan.com

Crowdsourcing Knowledge with the Semantic Web

- **Scalable knowledge acquisition (KA) is a grand challenge in AI**
- **There are several known factors which impact KA success**
  - The skill of the KA team and the amount of training/coordination required
  - The expressive power and ease of learning of the KR formalism
  - The usability and power of the KA tools
  - The formal complexity of an adequate domain model
- **Semantic web implies a KA strategy using crowdsourcing**
  - Encoder = DBA or webmaster SME, with minimal training and interaction
  - Huge numbers of authors and the web as a global publication fabric
  - KR with low expressive power = RDF or (sometimes) OWL
  - A set of tools and syntaxes
  - Modeling a domain of straightforward facts

Is crowdsourced KA via semantic web likely to succeed?
Is pay-as-you-go integration likely to work?

VULCAN

Is embedded semantic markup a promising KA strategy?

- **Can crowdsourced KA via page-embedded semantic markup succeed?**
  - The original Semantic Web use case
  - Combined structured and unstructured knowledge in the same place, with (hopefully) synchronized update
  - Machines could “read the web” without NLP
  - Best incentive ended up being to support SEO and publishing social data
- **The situation for powerful page-based semantic markup is not hopeful**
  - Despite 430M web pages with RDFa, Google said at SemTech that webmaster authoring in RDFa was too difficult, and this is probably right
  - Facebook said that >10% of OG markup is syntactically incorrect or incoherent
- **The KA answer so far: Schema.org and Facebook OG**
  - Data publishers are required to use a single common global ontology and vocabulary
  - Formal KR complexity is almost the lowest possible
  - Lowering the bar to achieve success

Clay Shirkey will be shown wrong…

… but it is difficult to take much satisfaction in this

Is Linked Data a Promising KA Crowdsourcing Strategy?

- **Can crowdsourced KA via Linked Data succeed?**
  - "Evolved" Semantic Web use case
  - Best incentive ended up being sharing and (government) data distribution
- **The situation for Linked Data markup is more hopeful**
  - Much more productive, yielding 10s of billions of semantic assertions
  - Many organizations are successfully publishing Linked Data
  - Overall semantic cohesion will increase as more data is mapped together
- **The KA answer so far: faith in Pay-As-You-Go (PAYGO)**
  - Building from a core set of known vocabularies and ontologies
  - PAYGO should yield a set of evolving, partial agreements on semantic models and terminology
  - What is the incentive to create and maintain high-quality PAYGO models?
  - There should be a business here… but no one has found it yet.

The community is better at publishing data than integrating it

How Can We Make PAYGO Succeed?

- **Is the PAYGO authorship just the familiar KA problem?**
  - PAYGO implies distributed authoring and managing a set of useful integration mappings
  - Data integration is much harder than just asserting links
- **PAYGO experiences**
  - How do companies succeed at “traditional” database PAYGO integration?
  - Powerful commercial reasons to differentiate products (brands, trademarks, etc.), so there are very high costs in creating product mappings
  - Achieving PAYGO in neuroscience has been exceptionally difficult
- **Mobile-social is next big tranche of Semantic Web data**
  - How can PAYGO work here given the fuzziness of the data collection?
  - Can we use machine learning for PAYGO in this area?

What PAYGO lessons can Semantic Web researchers learn from the database integration community?

VULCAN

Can Semantic Web Crowdsourced KA Yield Useful KBs?

- **The broad KA war is over and we won (sort of)**
  - RDF/OWL is the most important AI KR system on the planet
  - Semantic Web can create communities of engaged data publishers who take ownership of the integrated data
  - PAYGO and Linked Data is the vision for cost-effective KA and maintenance of specialized content
- **Is the semantic web KB technically useful?**
  - Data is often impossible to cache
  - Data at this scale always includes significant percentage of mistakes
  - Schema.org and Facebook OG markup has a very weak semantics
- **Responding to Watson: Most human questions are not precise**
  - Replacement of deduction by evidence gathering in large data sets
  - "Popular shopping areas in London"

Mark's Challenge to the EC: Continue on the LarKC Vision!

VULCAN

Thank You

Disclaimer: The preceding slides represent the views of the author only.
Any brands, logos and products are trademarks or registered trademarks of their respective companies.

VULCAN