the Large Knowledge Collider

Frank van Harmelen
Vrije Universiteit Amsterdam
The Vision

“a configurable platform for infinitely scalable semantic web reasoning”
Why we need
The Large Knowledge Collider

Gartner (May 2007):
"By 2012,
70% of public Web pages will have some level of semantic markup,
20% will use more extensive Semantic Web-based ontologies"

• Semantic Technologies at Web Scale?
  – 20% of 30 billion pages @ 1000 triples per page = 6 trillion triples
  – 30 billion and 1000 are underestimates, imagine in 6 years from now…
  – data-integration and semantic search at web-scale?
1 triple:
Distance Sun – Pluto

Fensel / Harmelen estimate
$10^{14}$ Triples

$\sim 10^{14}$ Triples
Sun

~$10^{11}$ Triples

Earth

Jupiter

Pluto
Infinitely scalable (1/2)

- by giving up 100% correctness:
  - trading quality for size
  - often completeness is not needed
  - sometimes even correctness is not needed
Infinitely scalable (2/2)

- by parallelisation:
  - cluster computing

- wide area distribution
  “Thinking@home”,
  “self-computing semantic Web”

- cloud computing
Why “LarKC”?

• The Large Knowledge Collider

A configurable platform for experimentation by others
“Configurable platform”

“a configurable platform for infinitely scalable semantic web reasoning”
The consortium

50 people present
Use case: Drug Discovery

- Problem: pharmaceutical R&D in early clinical development is stagnating

FDA white paper Innovation or Stagnation (March 2004):
“developers have no choice but to use the tools of the last century to assess this century's candidate solutions.”

“industry scientists often lack cross-cutting information about an entire product area, or information about techniques that may be used in areas other than theirs.”

“Show me any potential liver toxicity associated with the compound's drug class, target, structure and disease.”

Show me all liver toxicity associated with the target or the pathway.

Genetics

Chemistry

LITERATURE

Current NCBI: linking but no inference
Use Case: City on-line

- Our cities face many challenges
- Urban Computing is the ICT way to address them

How can we redevelop existing neighborhoods to improve the quality of life?
How can we create more choices in housing, accommodating diverse lifestyles and all income levels?
How can we reduce traffic congestion yet stay connected?
How can we include citizens in planning their communities rather than limiting input to only those affected by the next project?
How can we fund schools, bridges, roads, and clean water while meeting short-term costs of increased security?
• **MaRVIN**: cluster-based P2P inference engine:
  – Divide data-set between peers
  – Do local reasoning on single peers
  – Forward results to peers
  – Repeat

• **WebPIE**: MapReduce-based inference engine:
  – Currently scales to **100 billion triples**

• General architecture for **workflows of plugins**
  – Identify + Select + Reason + Decide
  – Abstracts over local/remote deployent