Large Scale Reasoning on the Semantic Web

or:

When success is becoming a problem

Frank van Harmelen

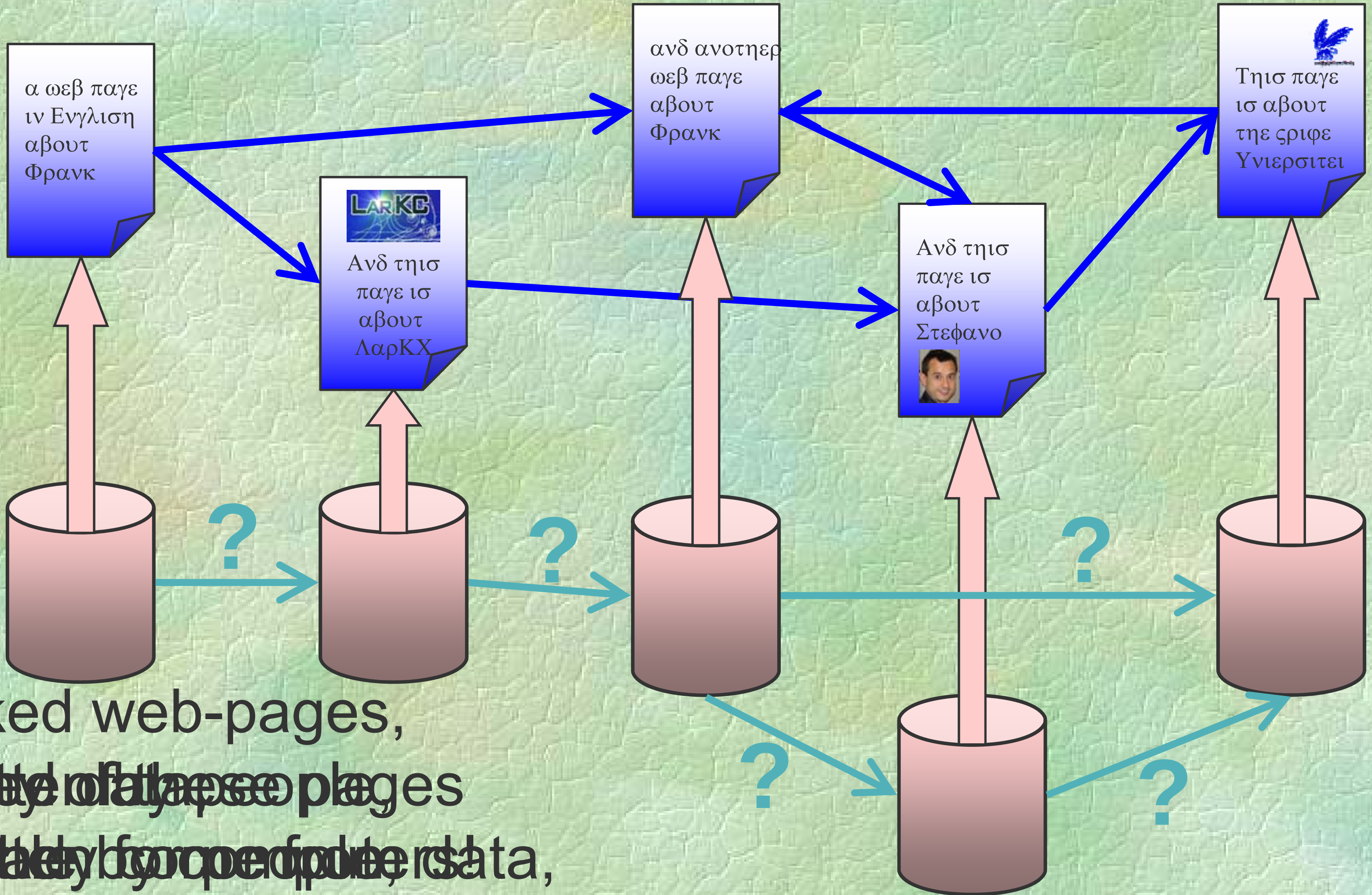
Vrije Universiteit Amsterdam

vrije Universiteit

What is
the Semantic Web

The Current Web of text and pictures

The Future Web of Data



α ωεβ παγε ιν Ενγλιση αβουτ Φρανκ
Ανδ τηισ παγε ισ αβουτ ΛαρΚΧ
ανδ ανοτηερ ωεβ παγε αβουτ Φρανκ
Ανδ τηισ παγε ισ αβουτ Στεφανο
Τηισ παγε ισ αβουτ τηε ςριφε Υνιερσιτει
?
?
?
?
?

linked web-pages,

Many of these pages
Written for people,

already computed from data,
written by people,

Best if also readable by computers!
But that's only readable by people

General idea of Semantic Web

Make current web more machine accessible
(currently all the intelligence is in the user)

Do this by:

1. Making **data and meta-data** available on the Web in machine-understandable form (**formalised**)
2. Structure the data and meta-data in **ontologies**

These are non-trivial design decisions. Alternative would be:

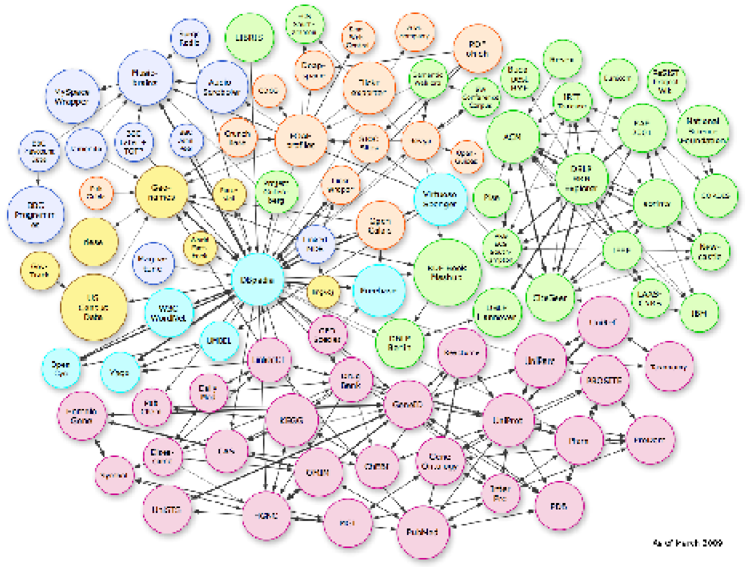
GIGA Google

~~Semantic Web~~

"Web of Data" (TBL)

1. expose data on the web ("**facts**") in interoperable form (RDF)
2. expose **knowledge** on the web with interoperable semantics (ontologies, RDF Schema, OWL)
3. Apply lightweight inference for
   - Interoperability
   - Query answering
   - Search
   - Unexpected reuse
   - ...

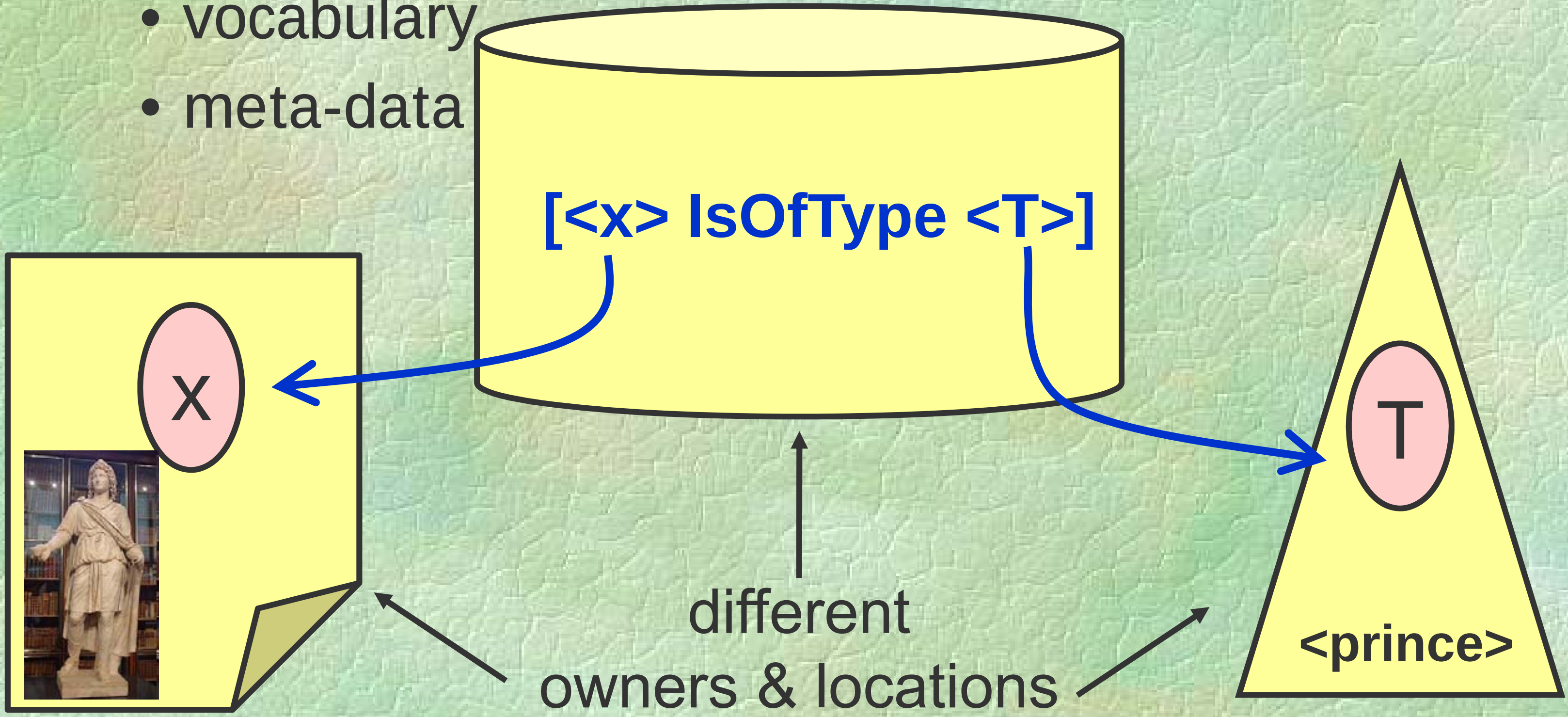
Not just data, also knowledge



- All of this:
  - Low expressivity logic (**RDF**)
  - That allows some inference: Property inheritance, domain/range inference
- Some of this:
  - Medium expressive logic (**OWL**)
  - That allows more inference: (in)equality, number restrictions, datatypes

Web of Data: anybody can say anything about anything

- All identifiers are URL's (= on the Web)
  - Allows total decoupling of
    - data
    - vocabulary
    - meta-data



Are you getting anywhere?

Linked Open Data cloud

already many b

any CD ever recorded (almost)

life science databases

basic facts on every country on the planet

common sense rules & facts (100.000's)

(UK, FR, NL)

scientific bibliographies

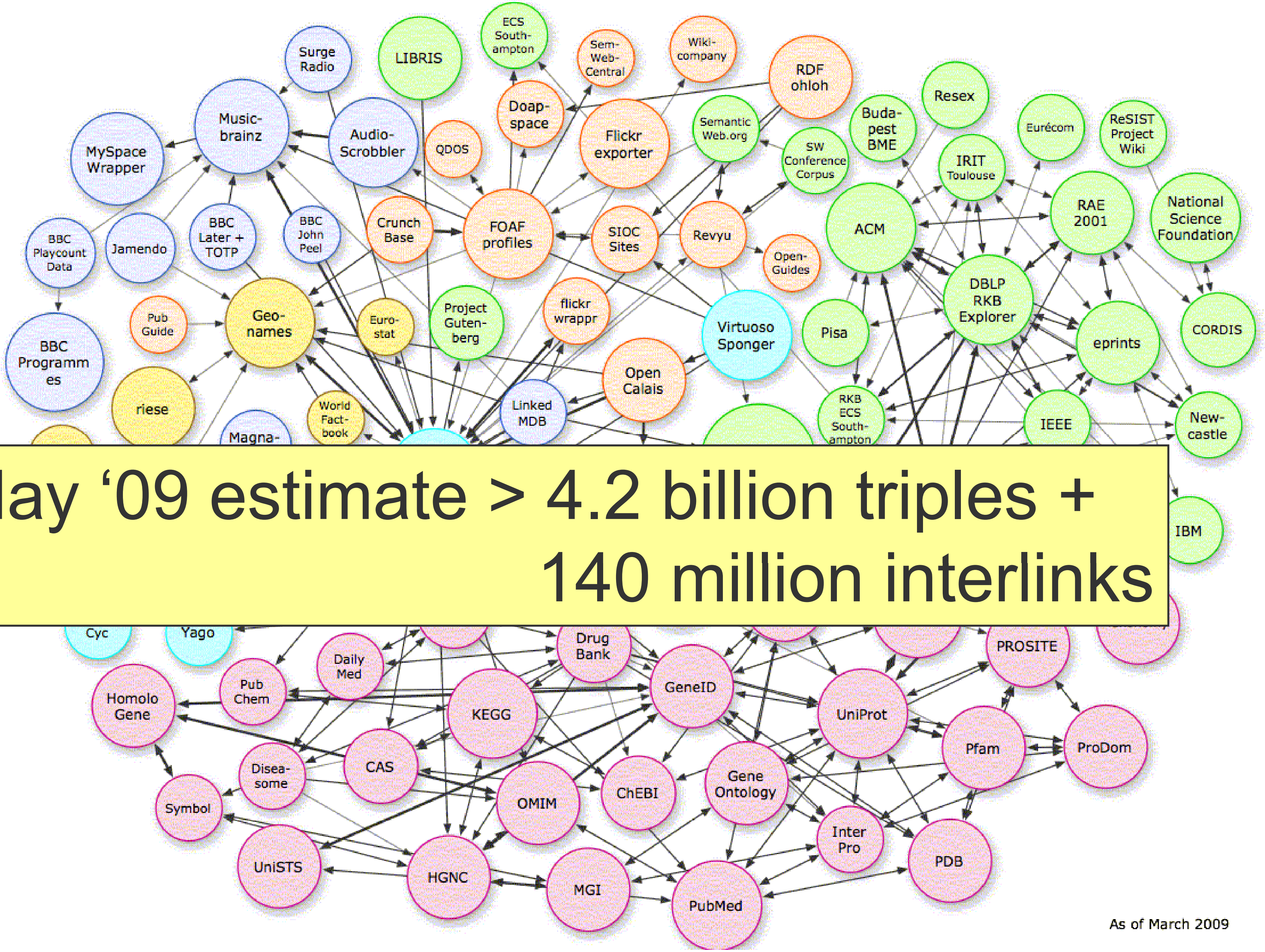
names of artists & art works (10.000's)

Geographic names (millions

Encyclopedia

It gets bigger every month

It gets bigger every month



May '09 estimate > 4.2 billion triples +
140 million interlinks

As of March 2009

All this is “unique in history”

For the first time ever it is now possible to **re-use somebody else's knowledge base without having to talk to them first** (syntax, semantics), and without having to make copies

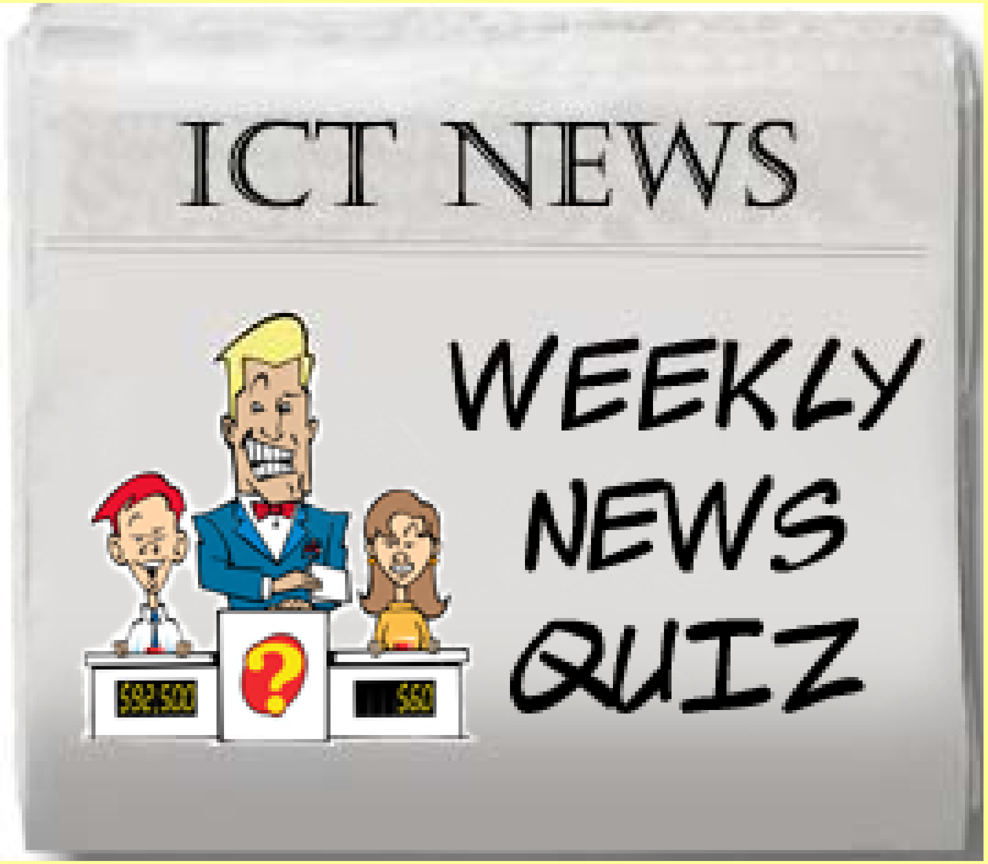
OWL and RDF are now **the most widely used KR languages in history** (by far)

Billion Triple Challenge

2007: “where do we get a billion triples from?”

2008: “**which billion shall we choose**?”

Nice in the lab, but are you getting anywhere in practice?



You choose....

Google
1545 CHA
YAHOO!
ORACLE
BBC
Drupal
Zemanta
amazon
THOMSON
REUTERS
BEST BUY

PALOS VERDES PENINSULA NEWS

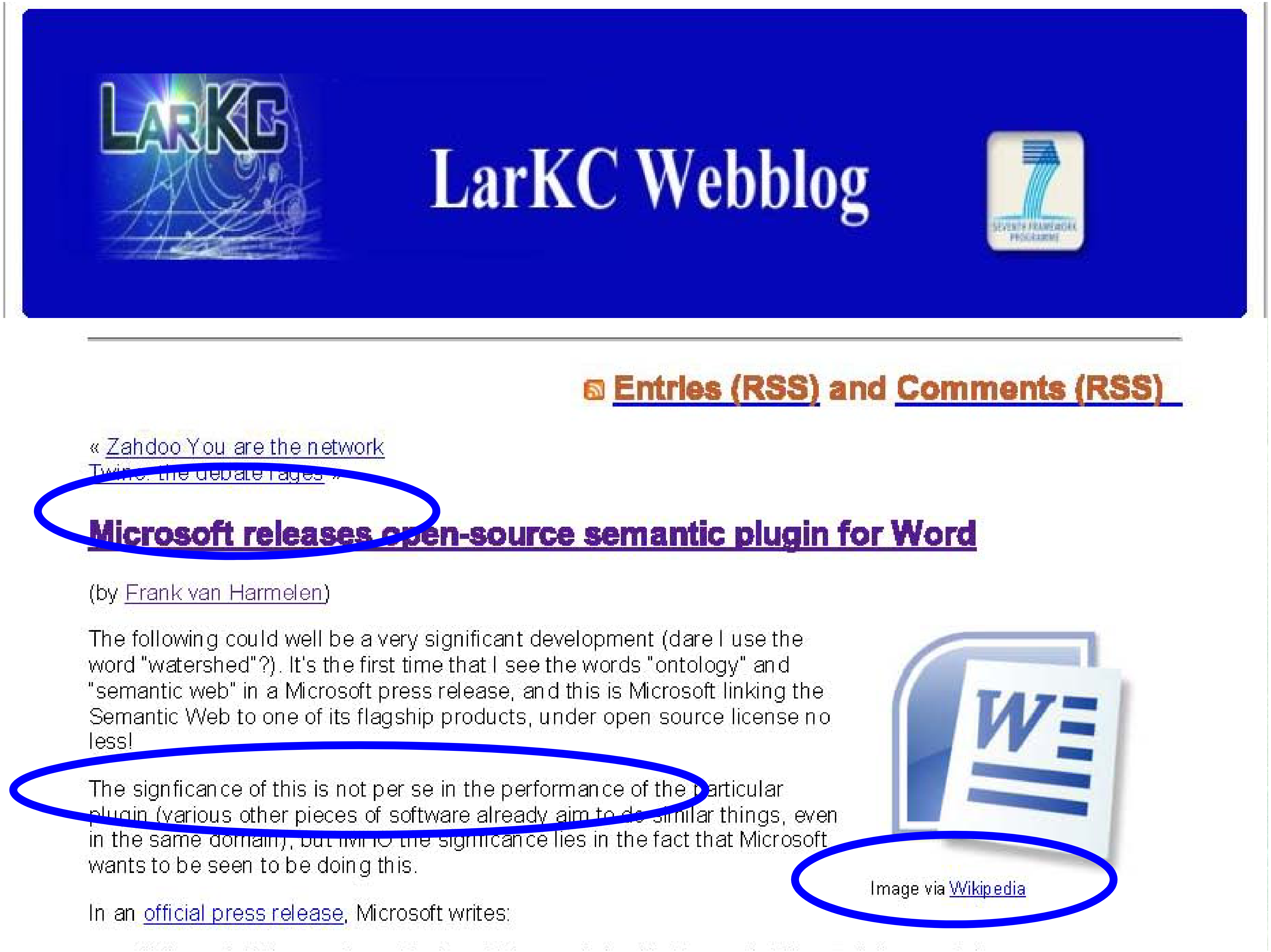
SATURDAY MARCH 15, 2004

GOOGLE TO HARVEST RDF-A

Search giant goes semantic

Late breaking news:
Google Video now also annotated
with RDF-a (using vocabularies from Yahoo and Facebook)

four million documents **per day**



LarKC Webblog

Entries (RSS) and Comments (RSS)

« Zahdoo You are the network
Twine: the debate rages »

Microsoft releases open-source semantic plugin for Word

(by Frank van Harmelen)

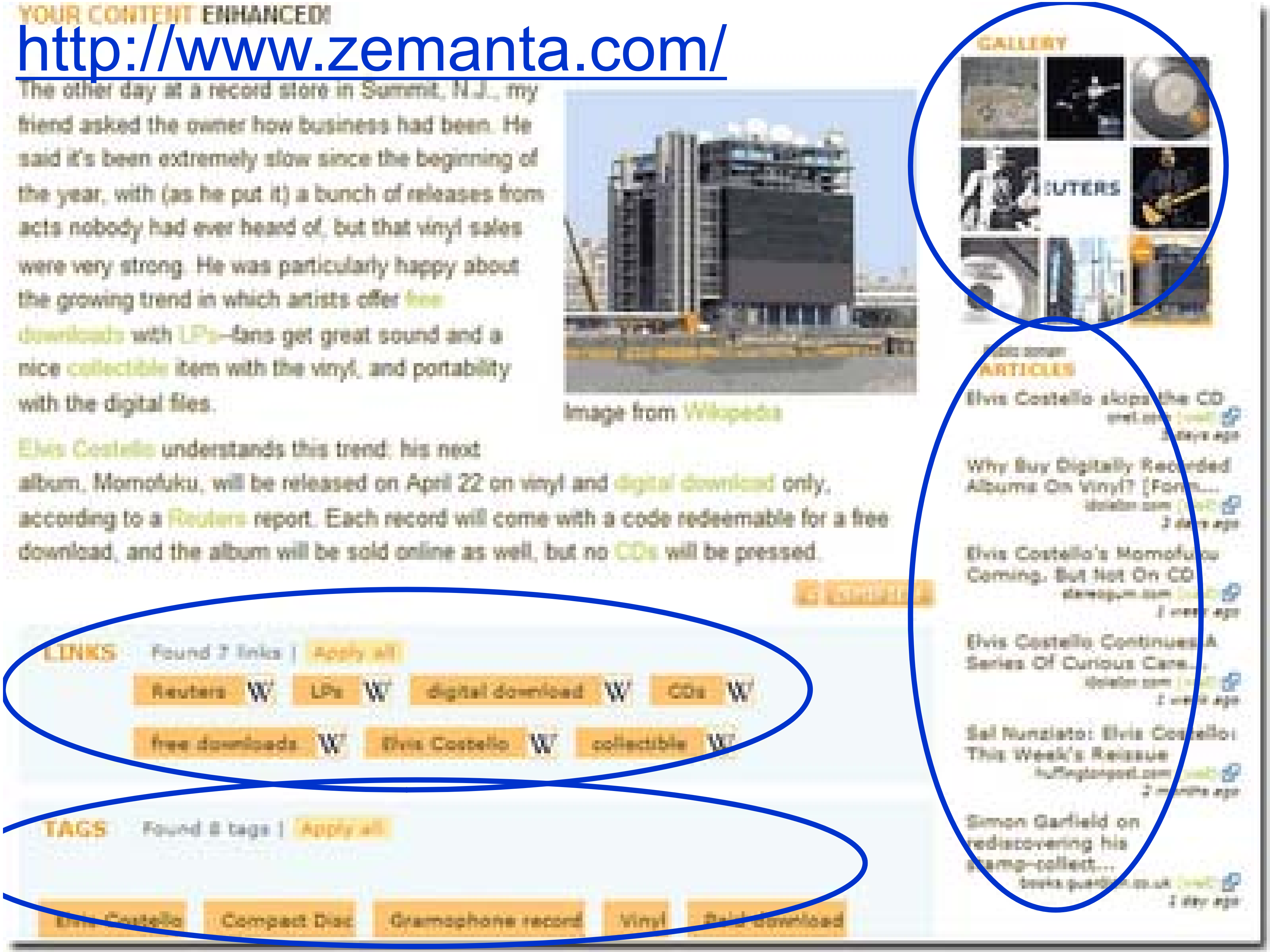
The following could well be a very significant development (dare I use the word "watershed"?). It's the first time that I see the words "ontology" and "semantic web" in a Microsoft press release, and this is Microsoft linking the Semantic Web to one of its flagship products, under open source license no less!

The signficance of this is not per se in the performance of the particular plugin (various other pieces of software already aim to do similar things, even in the same domain), but IMHO the significance lies in the fact that Microsoft wants to be seen to be doing this.

Image via Wikipedia

In an official press release, Microsoft writes:

http://www.zemanta.com/



YOUR CONTENT ENHANCED!

The other day at a record store in Summit, N.J., my friend asked the owner how business had been. He said it's been extremely slow since the beginning of the year, with (as he put it) a bunch of releases from acts nobody had ever heard of, but that vinyl sales were very strong. He was particularly happy about the growing trend in which artists offer free downloads with LPs–fans get great sound and a nice collectible item with the vinyl, and portability with the digital files.

Image from Wikipedia

Elvis Costello understands this trend: his next album, Momofuku, will be released on April 22 on vinyl and digital download only, according to a Reuters report. Each record will come with a code redeemable for a free download, and the album will be sold online as well, but no CDs will be pressed.

LINKS
Found 7 links | Apply all
Reuters W
LPs W
digital download W
CDs W
free downloads W
Elvis Costello W
collectible W
TAGS
Found 8 tags | Apply all
Elvis Costello
Compact Disc
Gramophone record
Vinyl
Paid download
GALLERY
UTERS
ARTICLES
Elvis Costello skips the CD
Why Buy Digitally Recorded Albums On Vinyl? [Form...
Elvis Costello's Momofuku Coming, But Not On CD
Elvis Costello Continues A Series Of Curious Care...
Sal Nunziato: Elvis Costello: This Week's Reissue
huffingtonpost.com
2 months ago
Simon Garfield on rediscovering his stamp-collect...
1 day ago

DISCOVER.
PARTICIPATE.
ENGAGE.

Search Data.gov catalog by category, agency, or both

FEATURED DA

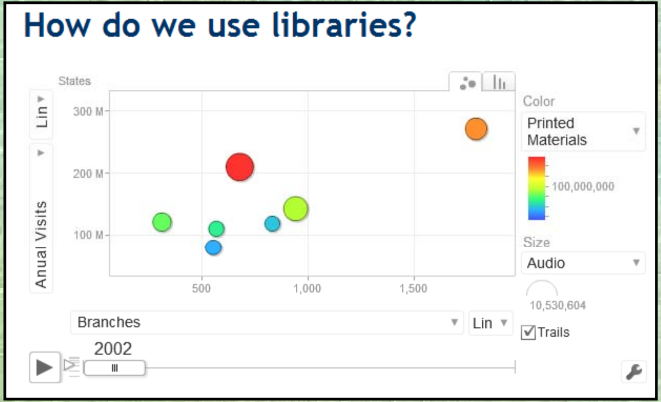
ENERGY INFO

Residential Energy Co

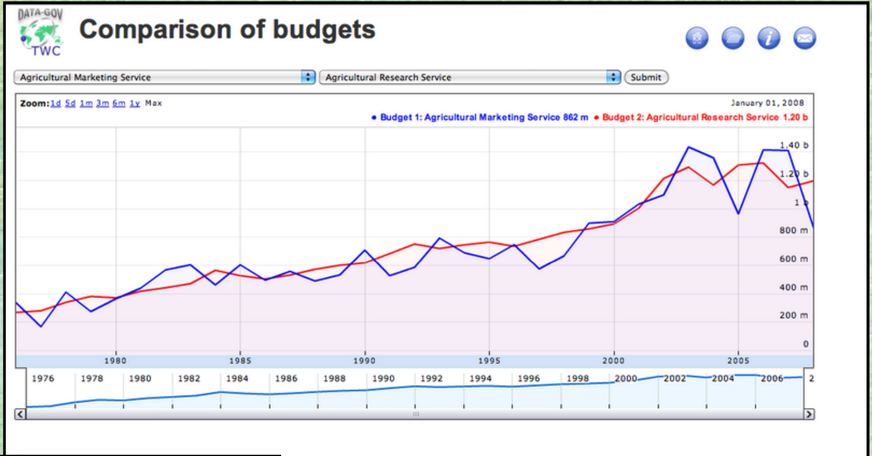
toxic releases
recent earthquakes
crime statistics
assaults on police
social benefits
unemployment rates

consumer expenditure
consumer price index
tornado reports
trade statistics
river elevations
energy consumption

Things to do with data.gov



How do we use libraries?

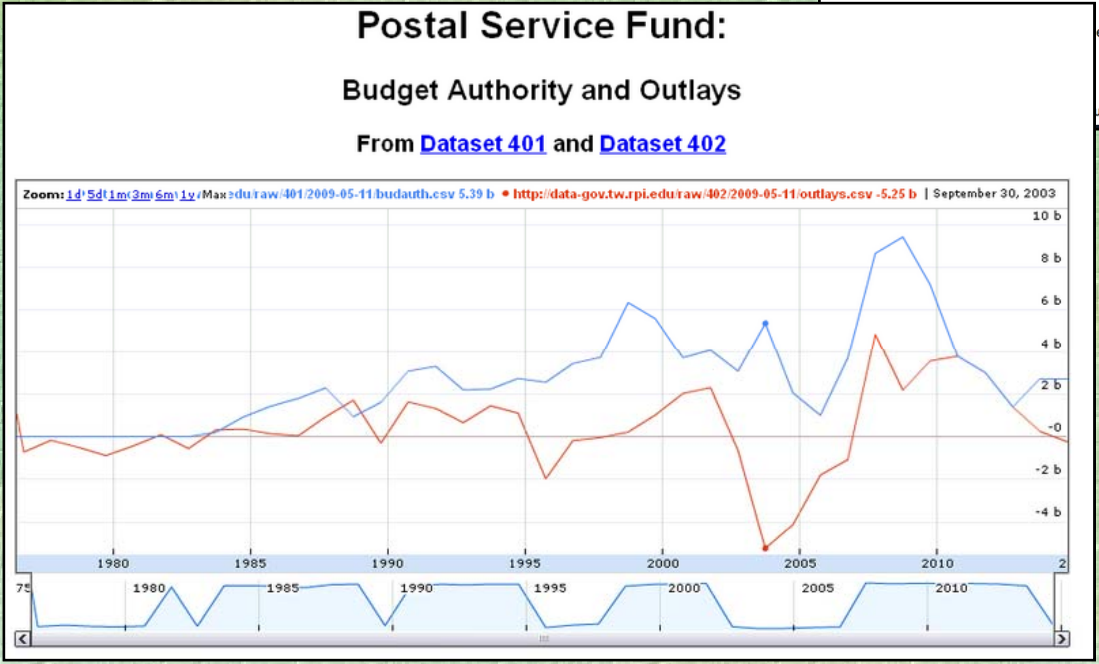


Comparison of budgets

Agricultural Marketing Service | Agricultural Research Service | Submit

etween Agricultural Marketing Service and Agricultural Research

uthority and offsetting receipts 1976-2014".

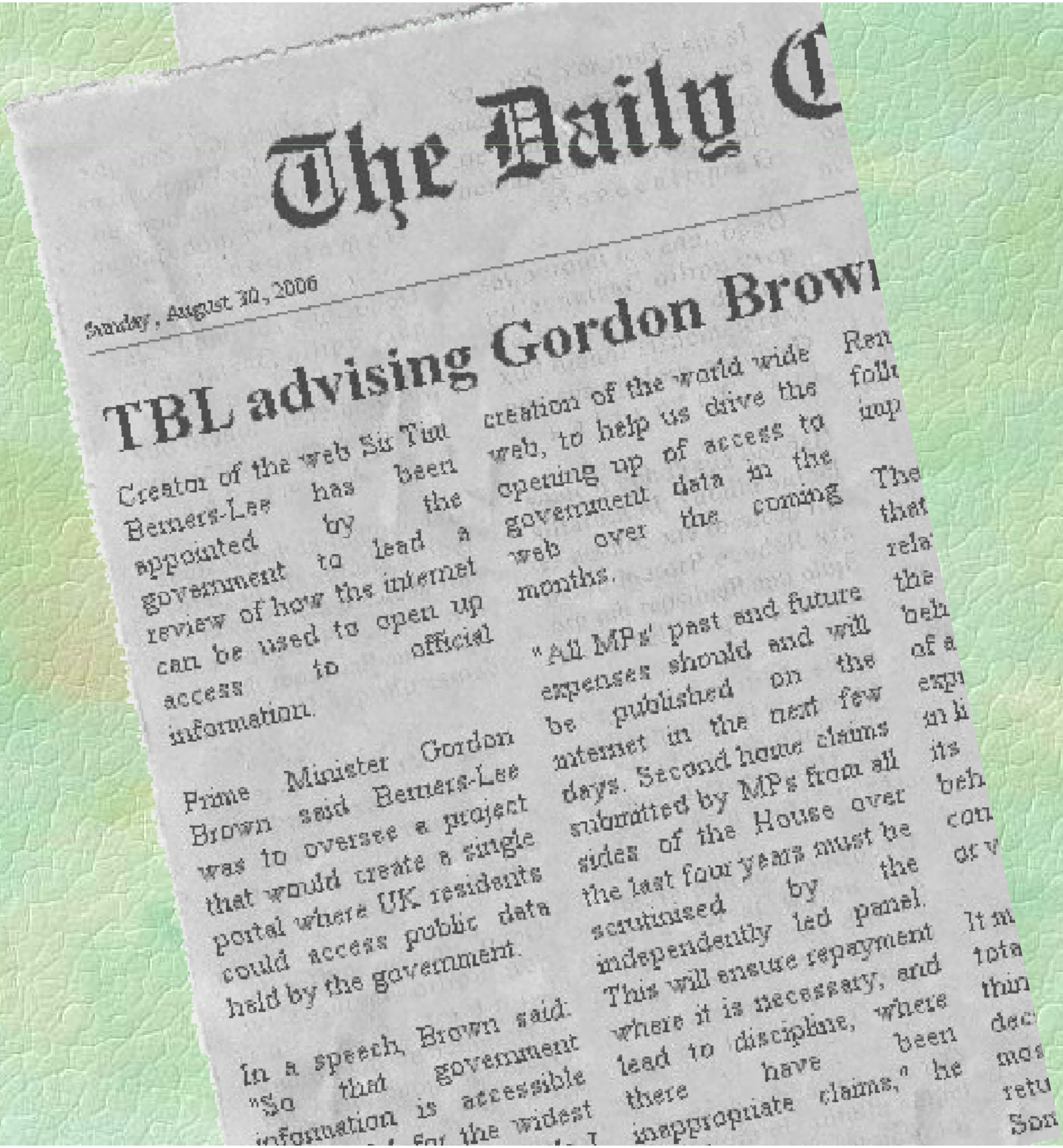


Postal Service Fund:

Budget Authority and Outlays

From Dataset 401 and Dataset 402

Zoom: 1d 5d 1m 3m 6m 1y Max
edu/raw/401/2009-05-11/budauth.csv 5.39 b • http://data-gov.tw.rpi.edu/raw/402/2009-05-11/outlays.csv -5.25 b | September 30, 2003
10 b
8 b
6 b
4 b
2 b
-0
-2 b
-4 b
1980
1985
1990
1995
2000
2005
2010
1980
1985
1990
1995
2000
2005
2010



The Daily C

Sunday, August 30, 2006

TBL advising Gordon Brown

Creator of the web Sir Tim Berners-Lee has been appointed by the government to lead a review of how the internet can be used to open up access to official information.

Prime Minister Gordon Brown said Berners-Lee was to oversee a project that would create a single portal where UK residents could access public data held by the government.

In a speech, Brown said: "So that government information is accessible for the widest

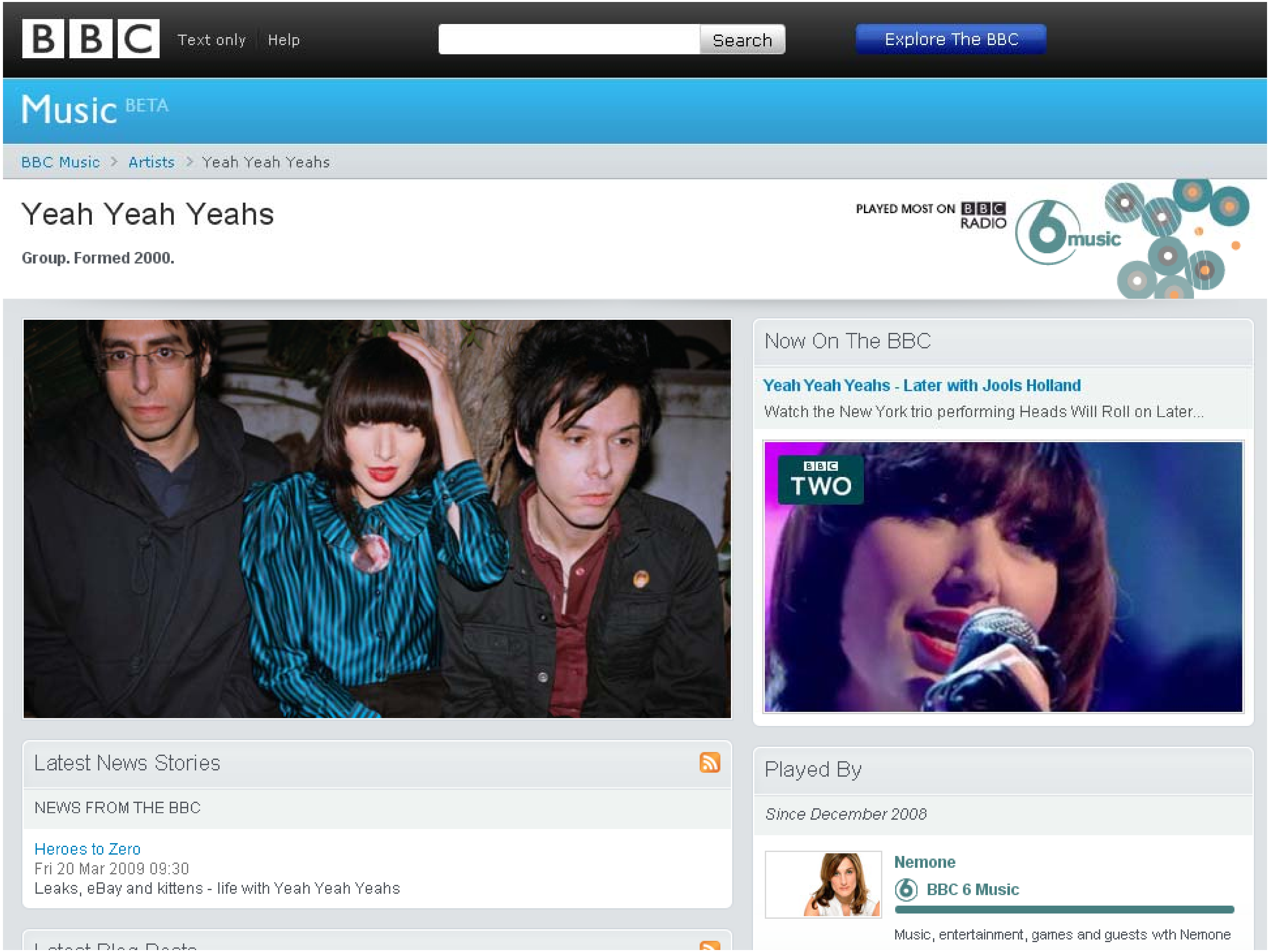
creation of the world wide web, to help us drive the opening up of access to government data in the web over the coming months.

"All MPs' past and future expenses should and will be published on the internet in the next few days. Second home claims submitted by MPs from all sides of the House over the last four years must be scrutinised by the independently led panel. This will ensure repayment where it is necessary, and lead to discipline, where there have been inappropriate claims," he

Rem
foll
imp

The
that
rela
the
beh
of a
exp
in li
its
beh
con
or v

It m
tota
thin
dec
mos
retu
Son


BBC
Text only
Help
Search
Explore The BBC
Music BETA
BBC Music > Artists > Yeah Yeah Yeahs
PLAYED MOST ON BBC RADIO 6 music
BBC TWO

Yeah Yeah Yeahs

Group. Formed 2000.

Now On The BBC

Yeah Yeah Yeahs - Later with Jools Holland

Watch the New York trio performing Heads Will Roll on Later...

Latest News Stories

NEWS FROM THE BBC

Heroes to Zero

Fri 20 Mar 2009 09:30

Leaks, eBay and kittens - life with Yeah Yeah Yeahs

Played By

Since December 2008

Nemone

BBC 6 Music

Music, entertainment, games and guests wth Nemone

Latest Blog Posts

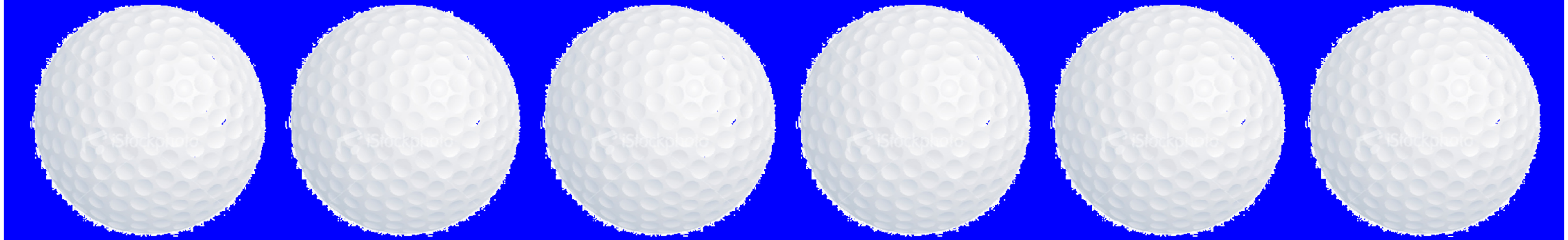
```
<rdf:RDF>
 <rdf:Description rdf:about="/music/artists/584c04d2-4acc-491b-8a0a-e63133f4bfc4.rdf
  <rdfs:label>Description of the artist Yeah Yeah Yeahs</rdfs:label>
  <foaf:primaryTopic rdf:resource="/music/artists/584c04d2-4acc-491b-8a0a-e63133f4bf
 </rdf:Description>
  <mo:MusicArtist rdf:about="/music/artists/584c04d2-4acc-491b-8a0a-e63133f4bfc4#a
  <rdf:type rdf:resource="http://purl.org/ontology/mo/MusicGroup"/>
  <foaf:name>Yeah Yeah Yeahs</foaf:name>
  <ov:sortLabel>Yeah Yeah Yeahs</ov:sortLabel>
  <bio:event>
    <bio:Birth><bio:date rdf:datatype="http://www.w3.org/2001/XMLSchema#dateTime
  </bio:event>
  <owl:sameAs rdf:resource="http://dbpedia.org/resource/Yeah_Yeah_Yeahs"/>
<mo:image rdf:resource="/music/images/artists/7col_in/584c04d2-4acc-491b-8a0a-e63
<foaf:page rdf:resource="/music/artists/584c04d2-4acc-491b-8a0a-e63133f4bfc4.html"/
<mo:musicbrainz rdf:resource="http://musicbrainz.org/artist/584c04d2-4acc-491b-8a0a-
<foaf:homepage rdf:resource="http://www.yeahyeahyeahs.com/"/>
<mo:wikipedia rdf:resource="http://en.wikipedia.org/wiki/Yeah_Yeah_Yeahs"/>
<mo:myspace rdf:resource="http://www.myspace.com/yeahyeahyeahs"/>
<mo:member rdf:resource="/music/artists/a1439b8d-672a-446f-a7ff-6f09d68254b3#art
<mo:member rdf:resource="/music/artists/14d44067-99c2-4f77-b58b-138f0b6911fa#ar
<mo:member rdf:resource="/music/artists/20dc35ec-6cc1-4c66-98a3-4a6116cb3869#ar
...
```

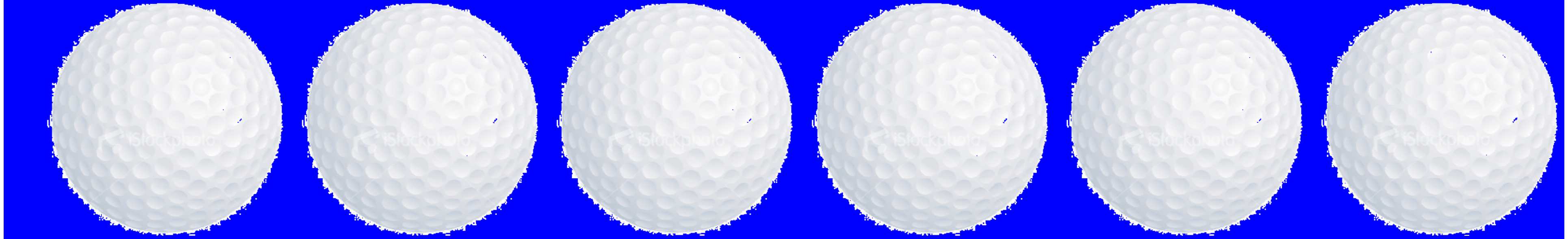
```
<foaf:made>
  <mo:Record>
   <dc:title>It's Blitz!</dc:title>
   <mo:musicbrainz rdf:resource="http://musicbrainz.org/release/9c4177fe-bdce-4f9d-ab
   <rev:hasReview rdf:resource="/music/reviews/hnp2#review"/>
  </mo:Record>
</foaf:made>
.....
<mo:MusicArtist rdf:about="/music/artists/a1439b8d-672a-446f-a7ff-6f09d68254b3#artis
  <foaf:name>Brian Chase</foaf:name>
</mo:MusicArtist>

<mo:MusicArtist rdf:about="/music/artists/14d44067-99c2-4f77-b58b-138f0b6911fa#arti
  <foaf:name>Karen O</foaf:name>
</mo:MusicArtist>

<mo:MusicArtist rdf:about="/music/artists/20dc35ec-6cc1-4c66-98a3-4a6116cb3869#art
   <foaf:name>Nick Zinner</foaf:name>
</mo:MusicArtist>
</rdf:RDF>
```

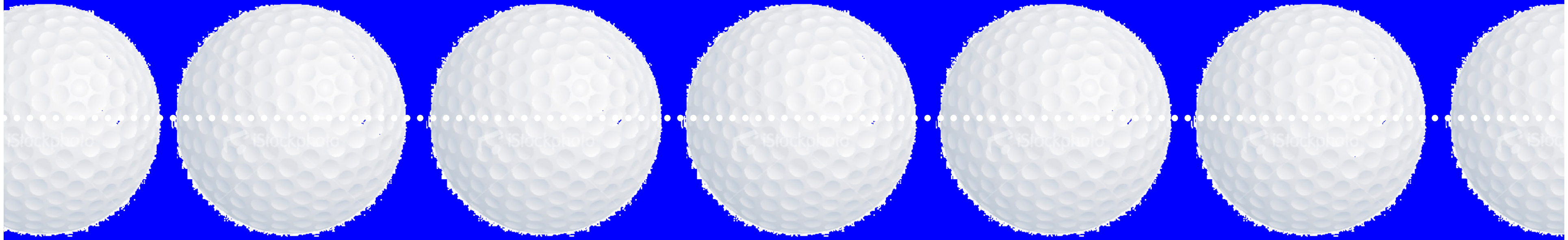
More, much more…

The problem of success

1 triple:

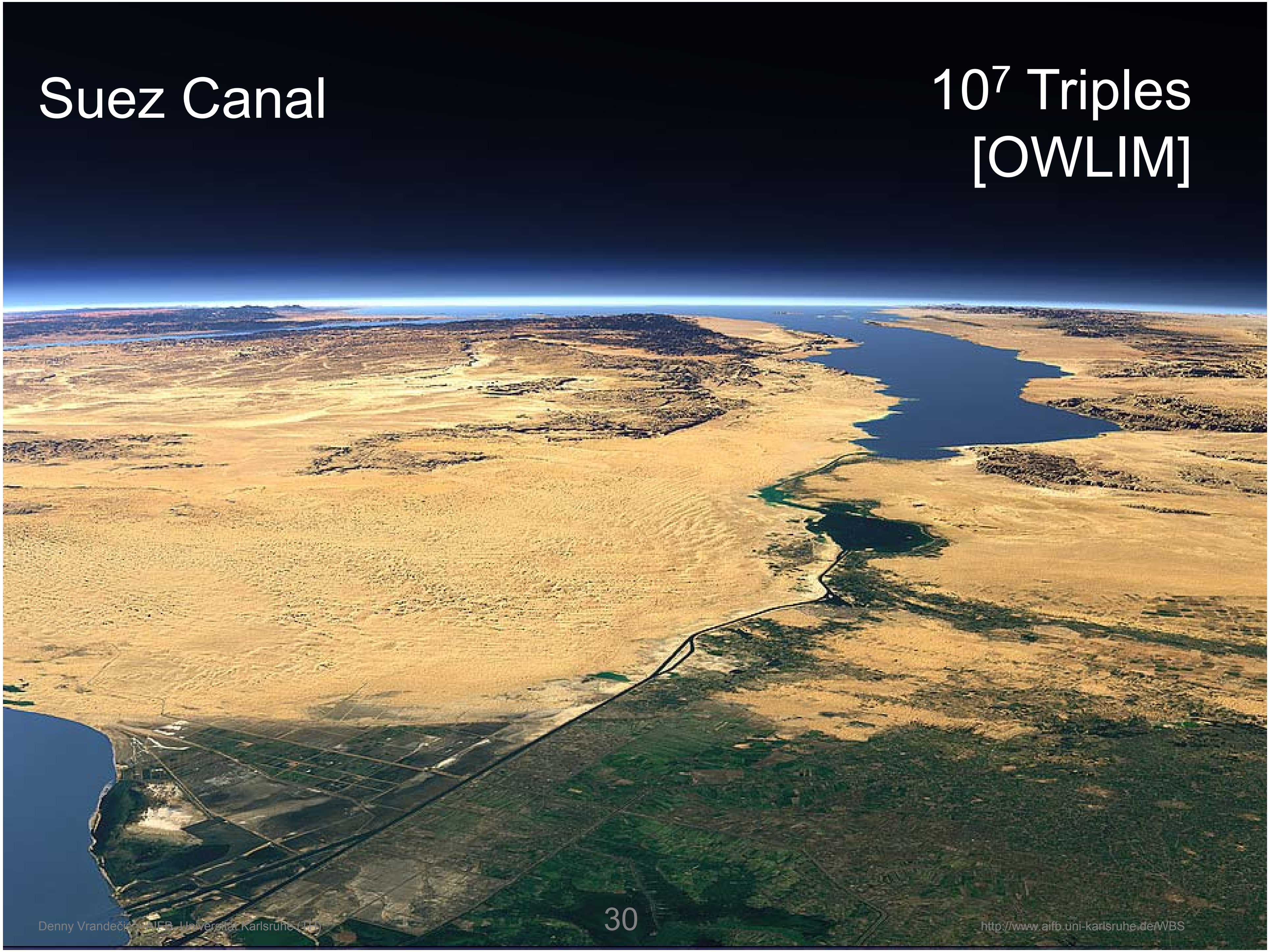






Suez Canal

10^7 Triples
[OWLIM]



Moon

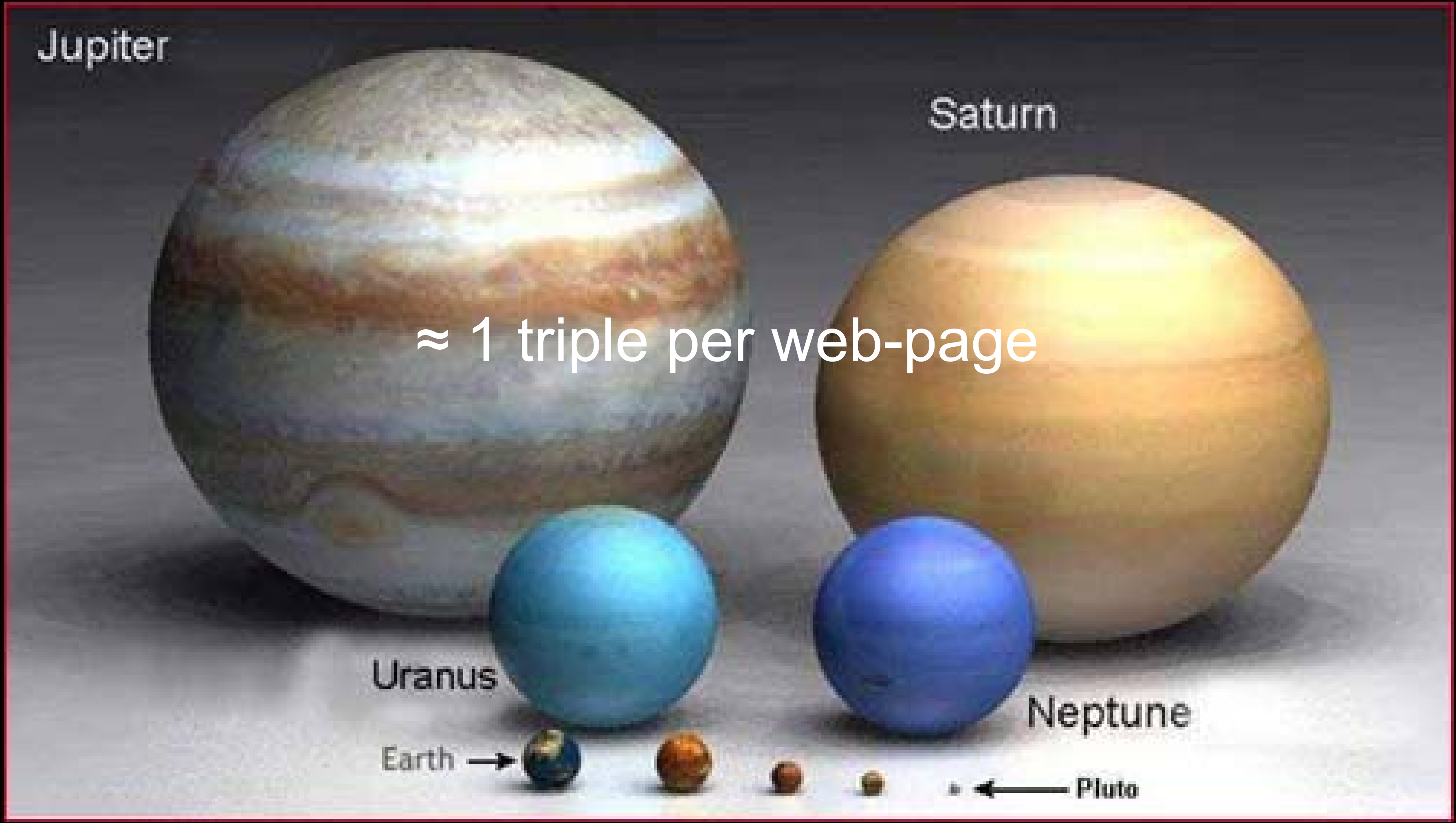
RDF Store subsecond querying
10^8 Triples
[Ingenta]

Earth

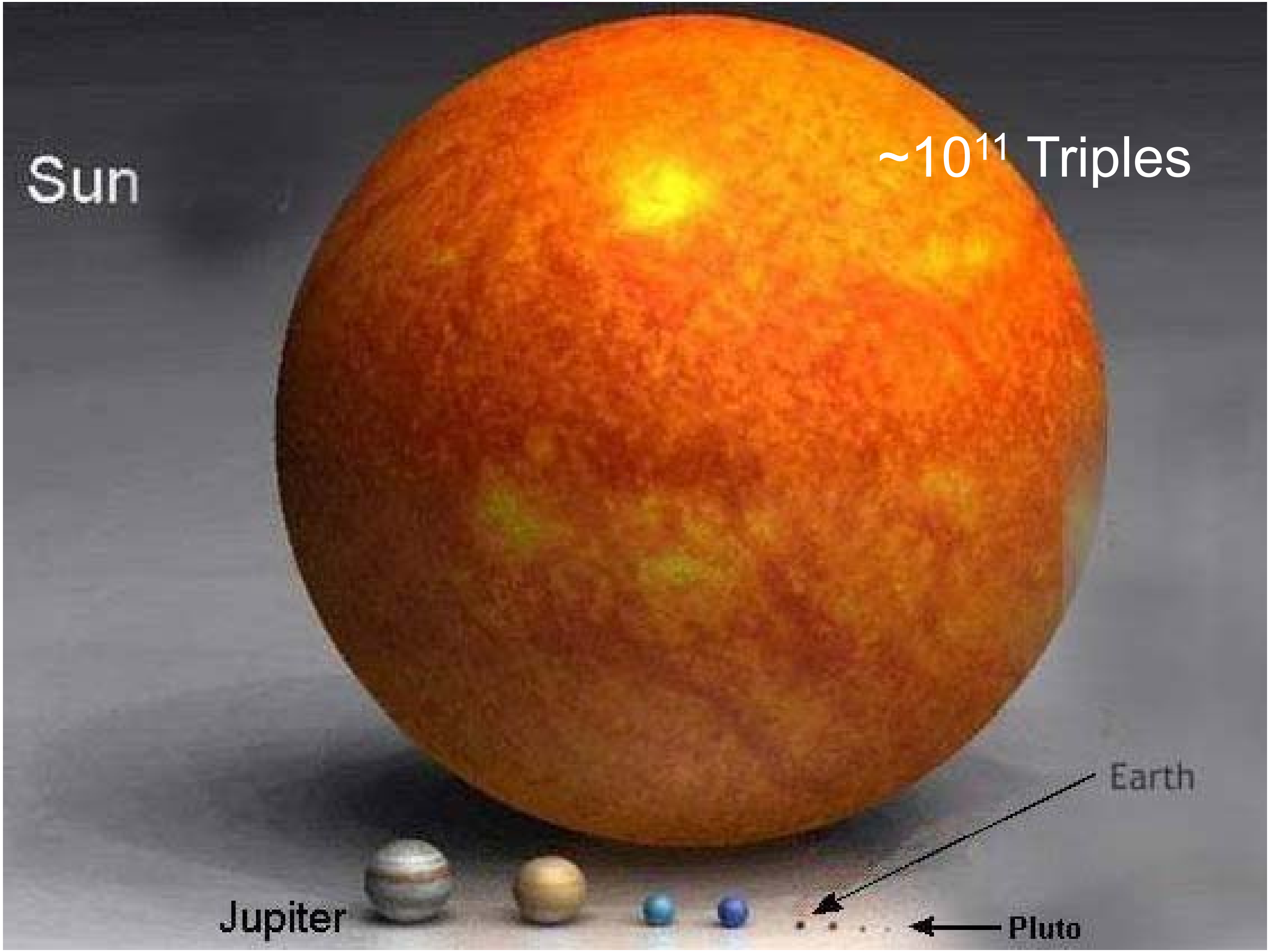
~10^9 Triples

[LarKC proposal]

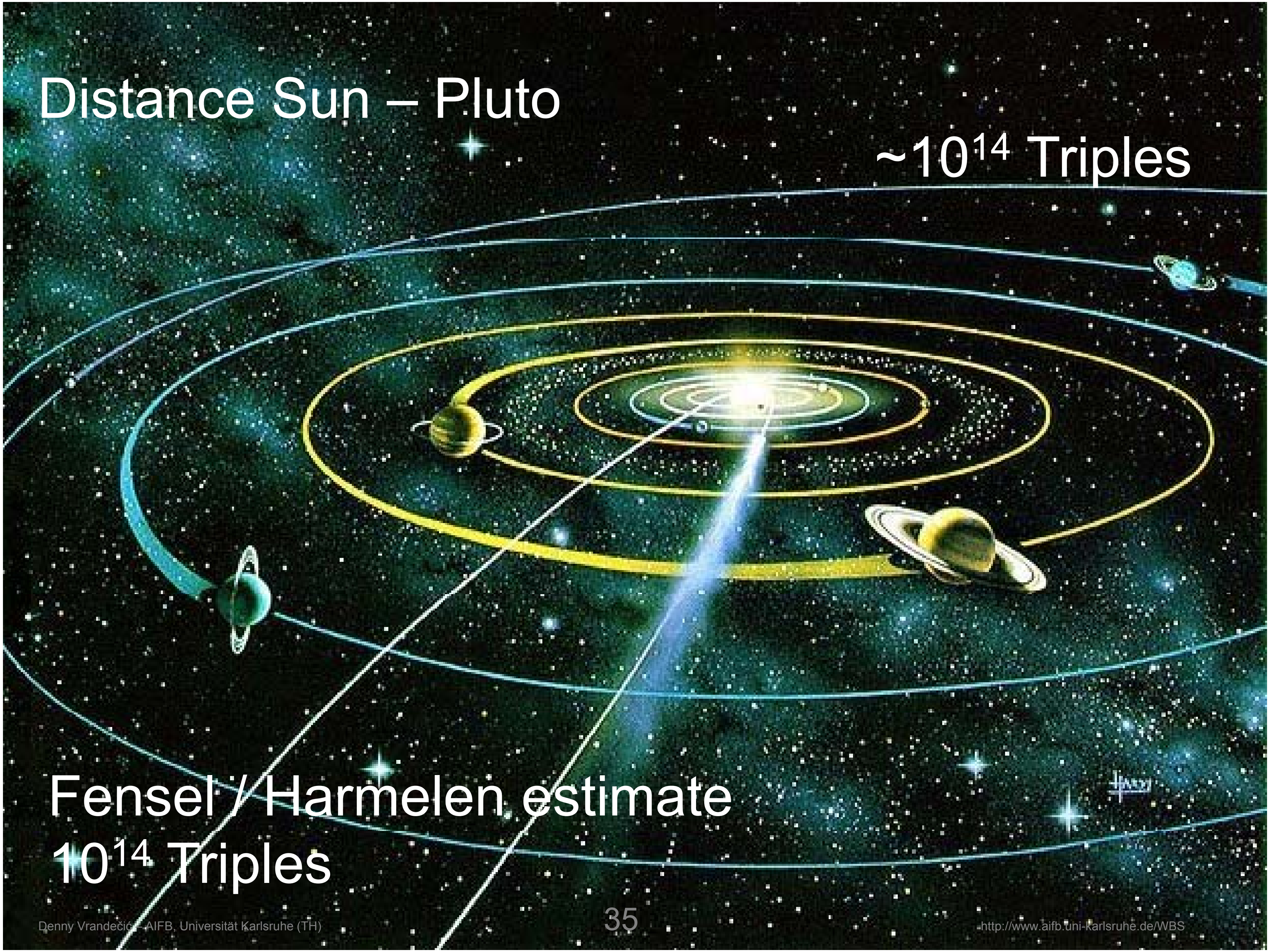
Jupiter ~10^{10} Triples ≈ 1 triple per web-page



≈ 1 triple per web-page



~10^{11} Triples



Distance Sun – Pluto

~10^{14} Triples

Fensel / Harmelen estimate

10^{14} Triples

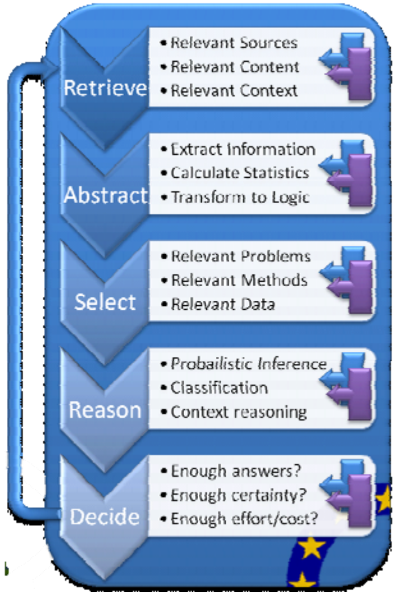
What to about the problem of succes:

LarKC

What to do when success becomes a problem?

The Large Knowledge Collider

a **platform** for infinitely scalable reasoning on the data-web



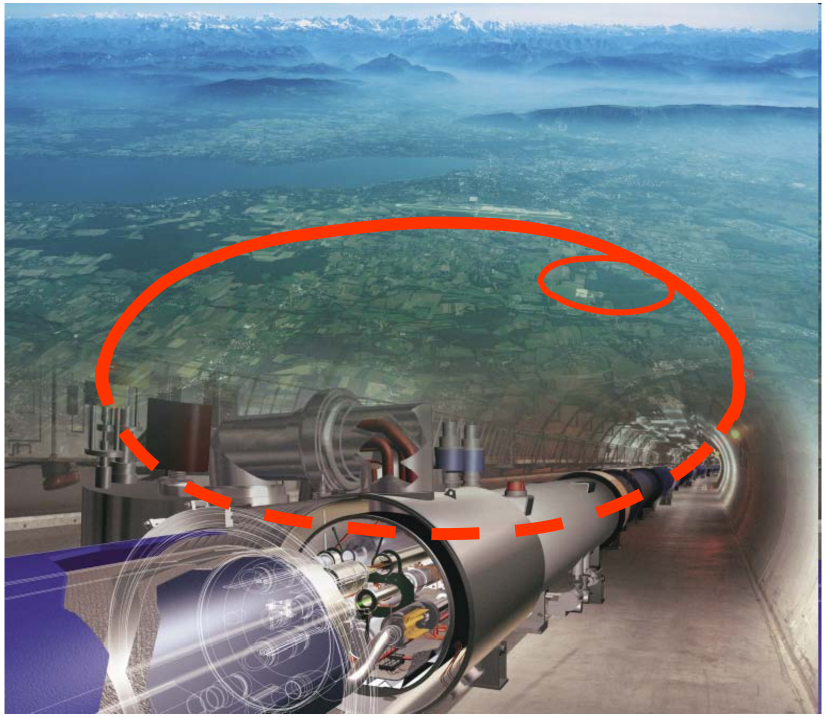
Retrieve
- Relevant Sources
- Relevant Content
- Relevant Context

Abstract
- Extract Information
- Calculate Statistics
- Transform to Logic

Select
- Relevant Problems
- Relevant Methods
- Relevant Data

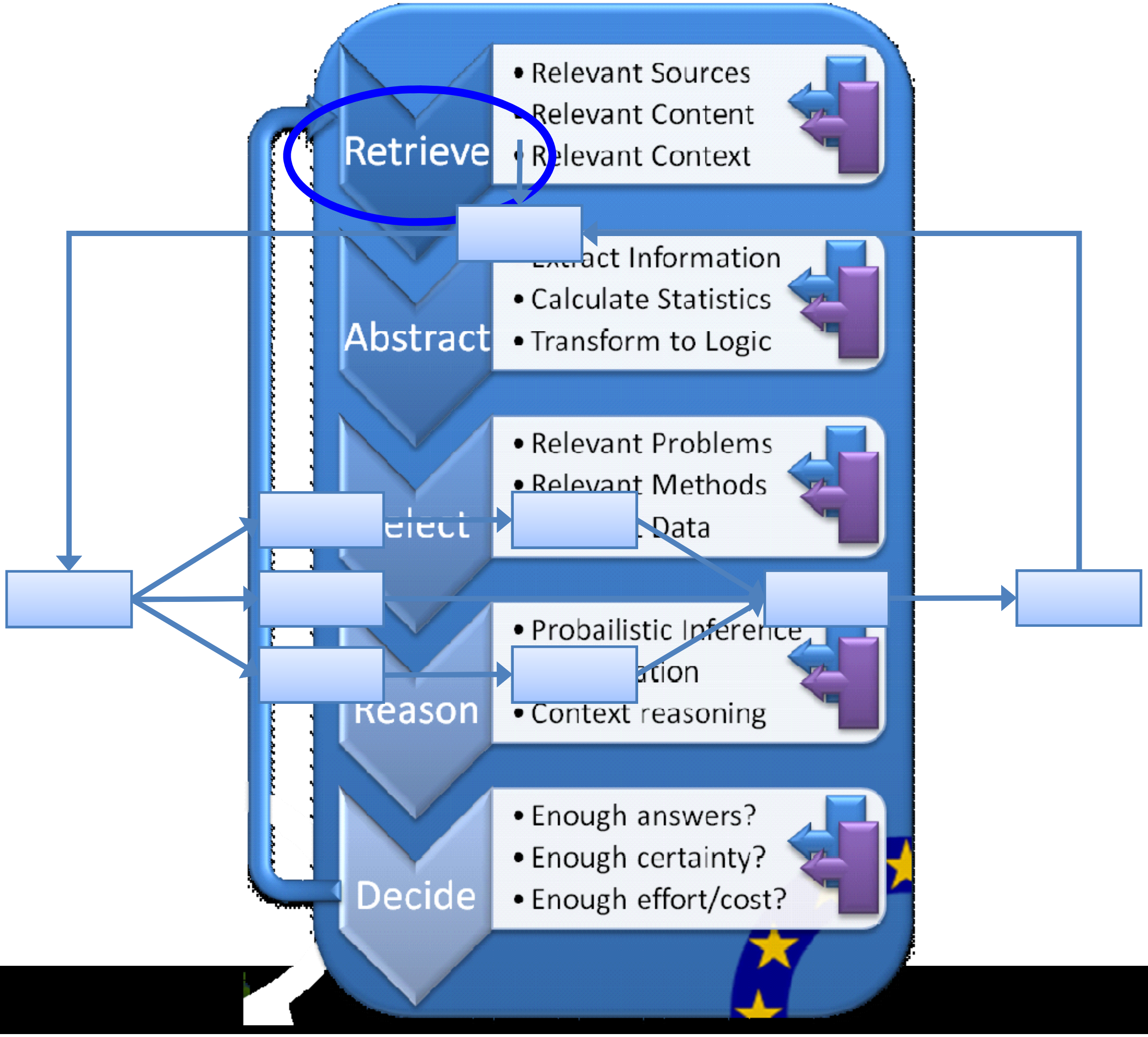
Reason
- Probailistic Inference
- Classification
- Context reasoning

Decide
- Enough answers?
- Enough certainty?
- Enough effort/cost?



“Configurable platform”

“a configurable platform for
infinitely scalable semantic web reasoning”



What to about the problem of succes:

parallelisation

What to do for infinite scalability?(4)

Take parallelisation seriously

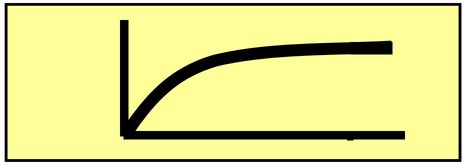
- Different parallel computing models:
  - Map-Reduce
  - Peer-to-peer

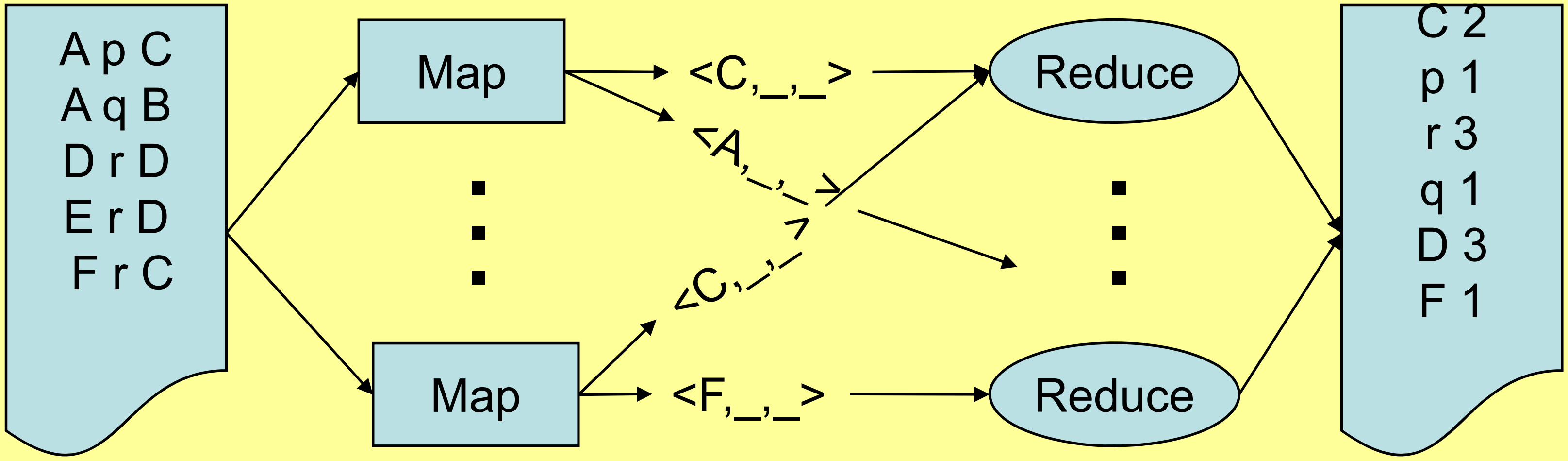
- Very high performance results:
  - Map-Reduce on 64 machines:

Peak inference rates at 8M triples/sec

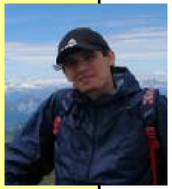
Sustained inference rates at 4M triples/sec

anytime convergence
(more complete over time)



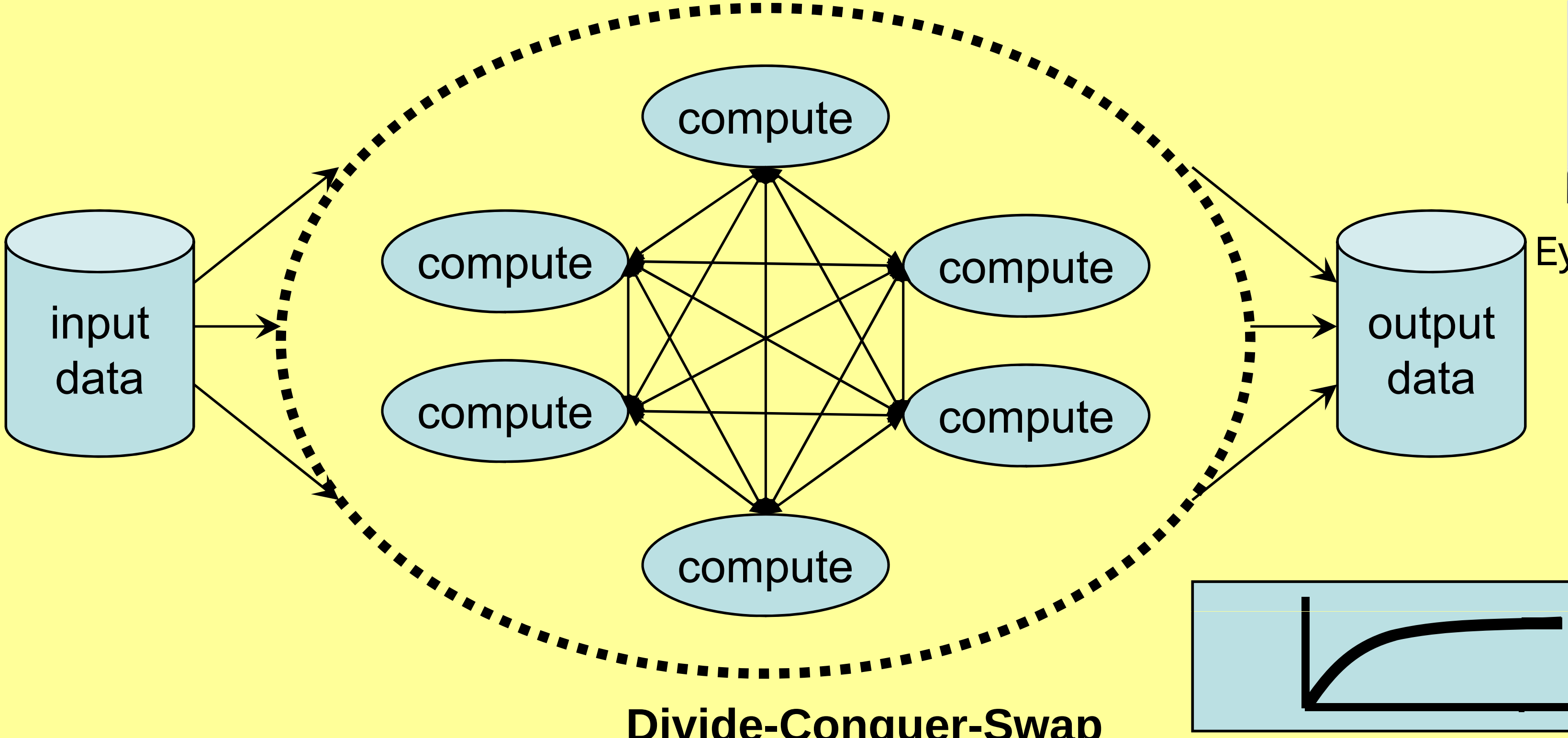


Map-Reduce



Jacopo Urbani

Eyal Oren



Divide-Conquer-Swap

What to about the problem of succes:

cognitive heuristics

What to do for infinite scalability?(6)

- On very large datasets, incompleteness is the rule
- Must stop before we are finished
- When to stop?
- Stopping rules are important
  - determine length of computation (don’t stop too late)
  - quality of result (don’t stop too early)

What to do for infinite scalability?(6)

Take inspiration from economics, biology, psychology

Lael Schooler

Humans have good heuristics for when to stop problem solving:

"Name capital cities in Europe":

London, Paris, Berlin, Rome, Amsterdam, ... Milan, Madrid, ..., ..., Paris, ...,

Time between solutions

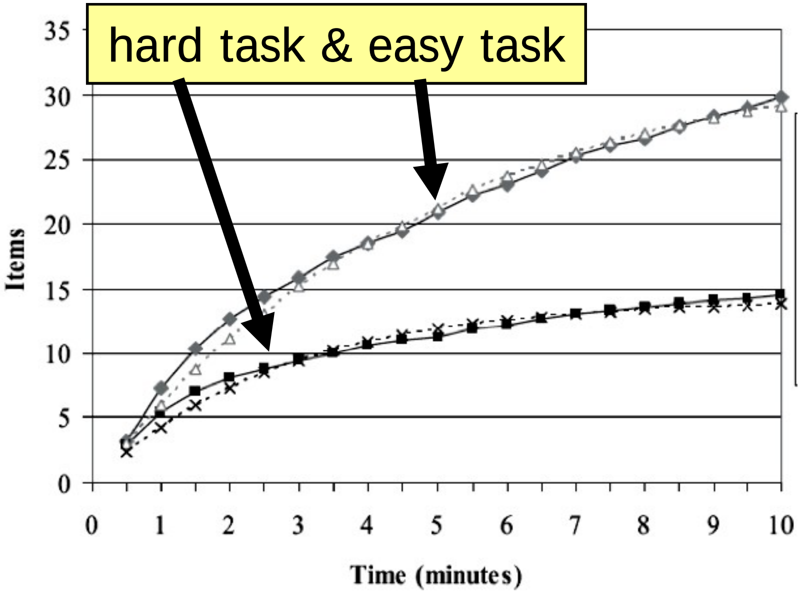
Wrong answers

Repetitions

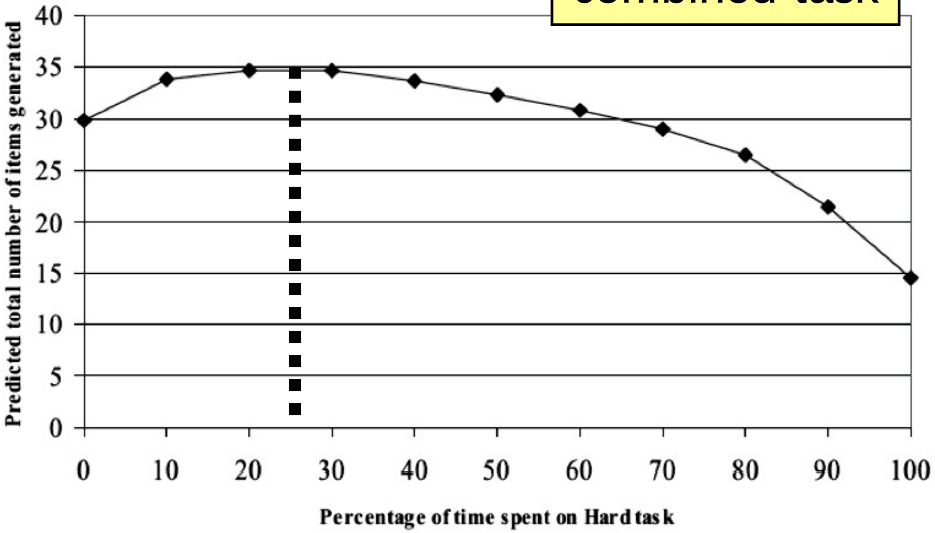
What to do for infinite scalability?(6)

When to switch between tasks?

Lael Schooler



hard task & easy task

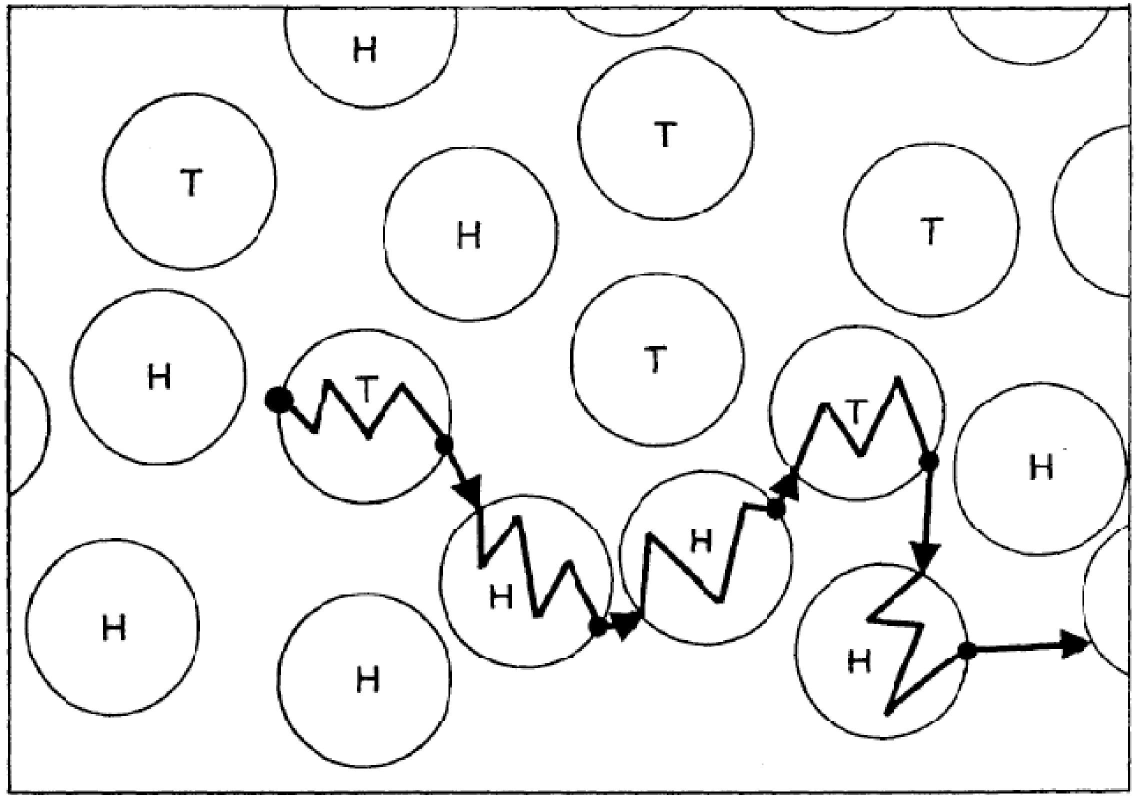


combined task

Humans (& animals) are very good finding this optimum

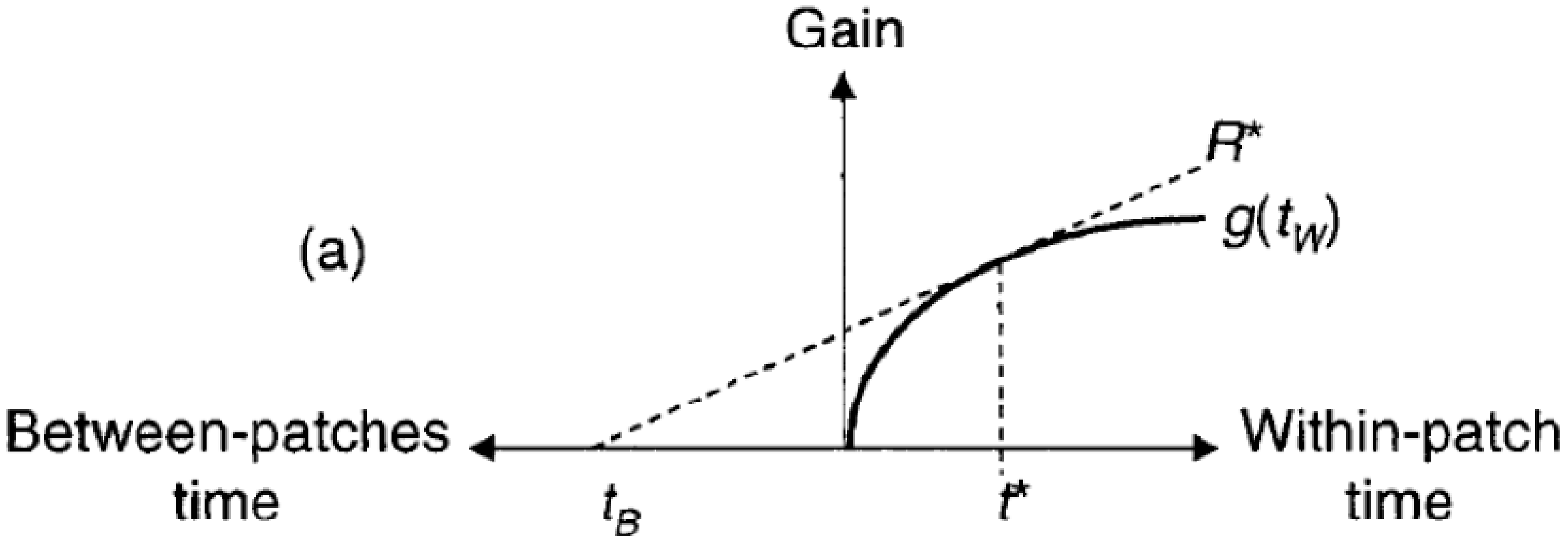
Marginal value thm (Charnov,'76)

- An animal grases various patches
  - Diminishing returns in the patches
  - Switching time between patches



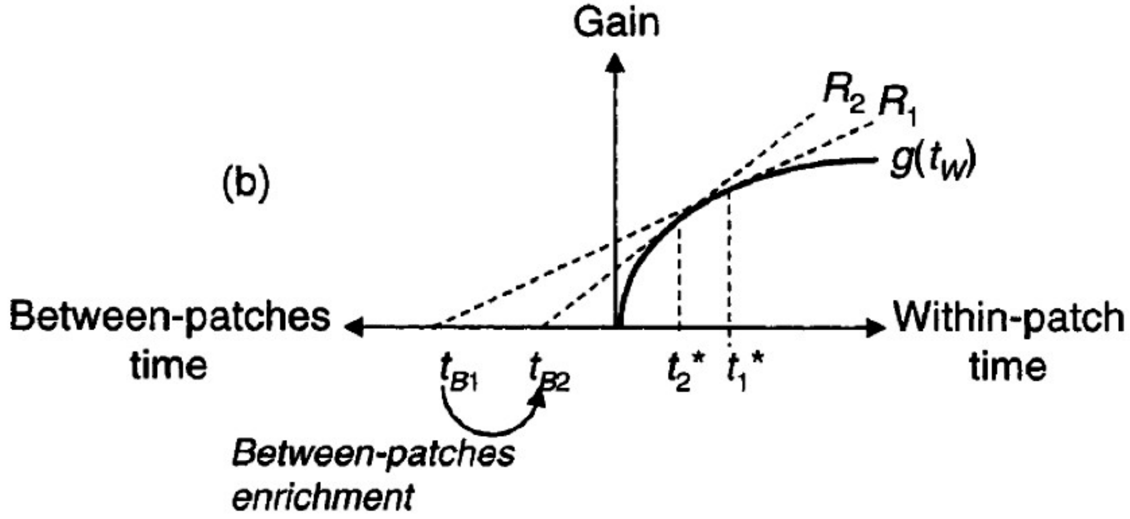
What to do for infinite scalability?(6)

Now take switching costs into account:



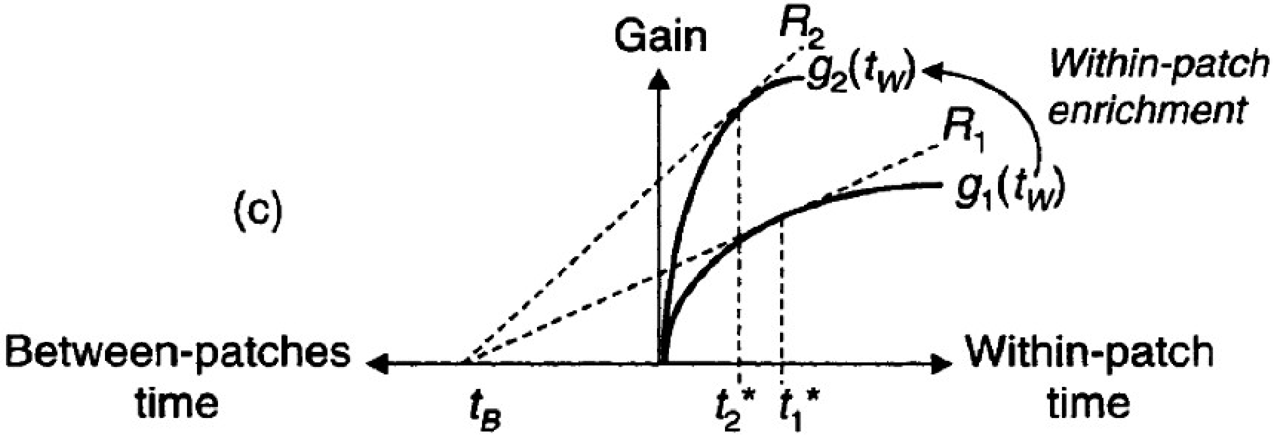
What to do for infinite scalability?(6)

When switching costs decrease, optimal time-per-task decreases



What to do for infinite scalability?(6)

when tasks become easier,
optimal time-on-single-task decreases…



Verified (& exploited) by bumblebees, hummingbirds, woodpeckers, humans and .... machines?

What to about the problem of succes:

data selection

What to do for infinite scalability? (1)

Take data-selection seriously

Zhisheng Huang

- Where do the axioms come from?
- Which subset to use?
- Relevance measures
  - Example: syntactic relevance:
    - $\delta(\alpha,\beta)=1$ if α,β share a concept symbol
    - $\delta(\alpha,\beta)=k$ if $\delta(\alpha,\gamma)=k-1$ and β,γ share a concept symbol
  - very simple measure, very syntactically unstable, but:

Gives a high quality sound approximation (> 90% recall, 100% precision for small k)

What to do for infinite scalability? (2)

Take identifiers seriously

- exploit the grounding of logical symbols in natural language
- Google distance as relevance measure

Zhisheng Huang

$$NGD(x,y) = \frac{\max\{\log f(x), \log f(y)\} - \log f(x,y)}{\log M - \min\{\log f(x), \log f(y)\}}$$

= symmetric conditional probability of co-occurrence

= estimate of semantic distance

Gives almost perfect “forgetting function” for matching class definitions in 2 vocabularies

What to about
the problem of succes:

REAL complexity measures

What to do for infinite scalability?(3)

Give up on worst-case complexity

Dan Weld

- joins are n^2:

  R(X,Y) = married-to(X,Z) & lived-in(Z,Y)
- but almost all people are married to a few people who in turn lived in a few places
- there are a few exceptions (Ramses-II, Liz Taylor), but then it tails of quickly

Approximately Pseudo-Functional relations:

- if almost all cardinalities are bounded by a small k,
- and exceptions are logarithmically rare

APF's scale linearly

Summarising

Summarising

- **The Semantic Web is rapidly becoming real**
- **Scale is becoming a real problem**
- **Different ways of scaling up**:
  - parallelisation
  - exploiting cognitive heuristics
    - theorems from economics & psychology
  - Replace completeness with eventual completeness
  - data-selection

Contact Info

Want to play with LarKC?
Want to contribute plugins?
Want to deploy LarKC?

Frank.van.Harmelen@cs.vu.nl
http://www.larkc.eu

Prof. Ning Zhong
Yi Zeng
@ WICI