From Linked Data to Networked Knowledge

Prof. Stefan Decker
A Network of Knowledge

- Interconnected
- Universal
- All encompassing

- assists humans, organisations and systems with problem solving
- enabling innovation and increased productivity

Enabling networked knowledge
A Network of Knowledge

- Interconnected
- Universal
- All encompassing
- assists humans, organisations and systems with problem solving
- enabling innovation and increased productivity

Linked Data

- Search
- Collaboration
- Information Mining
- Middleware

Research Domains

- Application
- Commercialization

Enabling networked knowledge
1. **Scalability**: No growth scalability problem (e.g., no back links from HTML pages)

2. **No censorship**: no lengthy permission or review process

3. **Positive feedback loop**: exploit Metcalf’s Law
Metcalfe's law:
The value of a network is proportional to the square of the number of connected members.

The Systemic Value of Compatibly Communicating Devices Grows as the Square of Their Number:

\[ \text{Value} = N^2 \]

\[ \text{Cost} \approx N \]

Critical Mass Crossover
Metcalfes’s Law 1: Links

Enabling networked knowledge
Metcalfé’s Law 2
Requirements for a Data Web

1. **Scalability.** No centralized infrastructure (e.g., a central object repository) required.

2. **No censorship.** It must be possible to publish data without having to ask for prior permission.

3. **Positive feedback loop.** Similar to the World Wide Web it must be possible to exploit Metcalfe’s Law.
Enabling Metcalfe’s Law

1. **Global Object Identity.**

2. **Composability:** The value of data can be increased if it can be combined with other data. Composability has a number of consequences:

   1. Data format must be *schema-less.* (Combined data originating from difference sources unlikely to conform to a schema)

   2. Data format must be *self-describing*

   3. Data format must be “*object centric*”. In order to integrate information about different entities data must be related to these entities.

   4. Data format at least to be *graph-based.* The composition of multiple object-centric data sources results in a graph in the general case.
Two Key Ingredients

1. **RDF – Resource Description Framework**
   - Graph based Data – nodes and arcs
     - Identifies objects (URIs)
     - Interlink information (Relationships)

2. **Vocabularies (Ontologies)**
   - Provide shared understanding of a domain
   - Organise knowledge in a machine-comprehensible way
   - Give an exploitable meaning to the data

Enabling *networked* knowledge
Why Graphs and Ontologies?

**Wikipedia.org**

Cities:Dublin

Geo:hasCapital

Geo:hasLargestCity

84421km²

Geo:locatedOn

Geo:IslandOfIreland

**Gov.ie**

Person:EndaKenny

Gov:hasTaoiseach

Gov:hasDepartment

EU:RepublicOfIreland

IE:DepartmentOfFinance

Enabling **networked** knowledge
Linked Open Data cloud

http://lod-cloud.net/

Over 200 open data sets with more than 25 billion facts, interlinked by 400 million typed links, doubling every 10 month!
From Linked Data to Networked Knowledge

1. Data awareness
2. Modeling
3. Publishing
4. Discovery
5. Integration
6. Use cases

Enabling *networked* knowledge
To raise awareness regarding government data (in Ireland) and establish a community around open data access and usage.

1. Data awareness
   - opendata.ie
   - LOD cloud

2. Modeling
   - Neologism
   - DataCube
   - prefix.cc

3. Publishing
   - Google Refine plug-in
   - RDB2RDF/D2R

4. Discovery
   - VoID/DCAT
   - Sindice
   - CKAN

5. Integration
   - LATC
   - Sig.ma

6. Use cases
   - school explorer
   - DERI in-house

Enabling networked knowledge
Welcome to CKAN Ireland

CKAN Ireland is an open registry of data and content packages. Harnessing the CKAN software, this site makes it easy to find, share and reuse content and data, especially in ways that are machine automatatable.

264 registered data packages available.

E.g. 'geo', 'shakespeare', 'science'

Filters:  open license  downloadable

Top Tags

disability  qnhs  income  certificate  electricity  saps  household  exports  s  fuel  employment  sapmap  education  transport  balance  agriculture  agricultural  loans  injury  construction  health  gas  farming  insurance  water  deaths  tr  energy  capital  work  hospital  consumption  census  nationality  accounts  housing  sap  illness  births  map  family  goods  social  financial  population  labour
Open data is...

- Information about our country and our government.
- Published on the web, under an open license.
- In a machine-readable format.

Read more »

New Irish Data

World Bank data
List of all schools
Oireachtas Bills

Planning Applications to Galway City Council

Posts from Open Data Ireland

Blog roll
- KildareStreet news
- The Story
- Jolitics Blog
- Irish Election
- Gavin Sheridan
- Open Knowledge Foundation
- EchoLibre
- LiDRC blog

Help free Ireland

Advocate for Open Data
You can get involved.

Add pointers
Add links to Web sites that contain important information.

Get data into the Irish datastore
Modeling

1. Data awareness
   - opendata.ie
   - LOD cloud

2. Modeling
   - Neologism
   - DataCube
   - prefix.cc

3. Publishing
   - Google Refine plug-in
   - RDB2RDF/D2R

4. Discovery
   - VoID/DCAT
   - Sindice
   - CKAN

5. Integration
   - LATC
   - Sig.ma

6. Use cases
   - school explorer
   - DERI in-house

Creation and adoption of Web of Data vocabularies
http://linked-statistics.org/datacube/
1. Data awareness
  - opendata.ie
  - LOD cloud

2. Modeling
  - Neologism
  - DataCube
  - prefix.cc

3. Publishing
  - Google Refine plug-in
  - RDB2RDF/D2R

4. Discovery
  - VoID/DCAT
  - Sindice
  - CKAN

5. Integration
  - LATC
  - Sig.ma

6. Use cases
  - school explorer
  - DERI in-house

Making legacy data (RDB, spreadsheets, etc.) usable
Publishing: Clean up


1. Machine-readable catalogues
2. Data clean up
3. Reconciliation
4. Conversion to RDF & interlinking
5. Sharing
Enable discovery of datasets

1. Data awareness
   - opendata.ie
   - LOD cloud

2. Modeling
   - Neologism
   - DataCube
   - prefix.cc

3. Publishing
   - Google Refine plug-in
   - RDB2RDF/D2R

4. Discovery
   - VoID/DCAT
   - Sindice
   - CKAN

5. Integration
   - LATC
   - Sig.ma

6. Use cases
   - school explorer
   - DERI in-house

Enabling networked knowledge
Discovery

- Model for dataset description (VoID vocabulary)
- Users in industry and governments
- Published at W3C [http://www.w3.org/TR/void](http://www.w3.org/TR/void)
- Significant uptake in research
Integration

1. Data awareness
   - opendata.ie
   - LOD cloud

2. Modeling
   - Neologism
   - DataCube
   - prefix.cc

3. Publishing
   - Google Refine plug-in
   - RDB2RDF/D2R

4. Discovery
   - voiD/DCAT
   - Sindice
   - CKAN

5. Integration
   - LATC
   - Sig.ma

6. Use cases
   - school explorer
   - DERI in-house

Large-scale entity interlinking

Enabling networked knowledge
Challenges

- **Open Questions**
  - How to find sources that describe topics and entities of interest?
  - How to cleanse and transform said data to fit one’s needs?
  - How to visualize, verify, and consume said data for creating new competitive advantages for my business?

- **Building a “Dataspace” infrastructure to**
  - Make data access easier
    - Search and Exploration
  - Make data exploitation easier
    - Integration and Reuse
Dataspace Example: Irish Tourism

Sindice Dataspace Infrastructure

Search and Querying

Visualisation

Vertical Applications
E.g., e-advertissement

Hotels

Restaurants
Enabling networked knowledge
Semantic Information Retrieval Engine

Goal:
Efficient data structures for semantic IR, effective ranking algorithms

Semantic Search Platform

Web Sites & Services

Messaging System

Desktop

DMS

Archives & Storage

Relational Database

Social Media

Web Database
Progress to Date – SIREn

Data Structure / Compression / Query Processing:

- Data structure for semi-structured inverted indexes
- High-performance compression technique

**Query Rate**

<table>
<thead>
<tr>
<th>Number of attribute-value pairs</th>
<th>Queries per second</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>2</td>
<td>50</td>
</tr>
<tr>
<td>4</td>
<td>100</td>
</tr>
<tr>
<td>8</td>
<td>150</td>
</tr>
<tr>
<td>16</td>
<td>200</td>
</tr>
</tbody>
</table>

- Small (1.2B)
- Medium (2.5B)
- Large (4.6B)

**Commit time (ms)**

- Constant time

Enabling networked knowledge
Cloud Scalable Dataspace Infrastructure

- Map Reduce (Hadoop)
- NOSQL technologies
- Semantic Information Retrieval
Sindice.com: Portal for Sindice Services

Sindice - Data Web Services

Billion pieces of reusable information can already be found across hundreds of millions of web pages which embed RDF and Microformats. Start consuming this data today with Sindice Data Web services.

LATEST DATA
11:03:13 (sirgun) 11 triples http://semanticoeurope.com...00946g505p525 rdf
11:03:13 (sirgun) 57 triples http://www.freebase.com...9_11_01_archive.html
11:03:12 (site_manager) 2 triples http://www.freebase.com...sic/album/genre
11:03:12 (site_manager) 2 triples http://www.freebase.com...episode/writer

HIGHLIGHTS
Anything to Triples

Introduction

any23

SINIDCE TWEET
Tue Sep 07 09:49:30 2010 Sindice participates in the new EU LDDZ project, 4 years long, just started. http://lod2.eu Truly great competences in the consortium.

SINIDCE BLOG
Sindice now supports Efficient Data discovery and Sync.
So far semantic web search engines and semantic aggregation services have been inserting datasets by hand or have been based on "raw... (More —>)

Sindice planned downtime this weekend
HiDue to an expansion of one of our datacentres (and the electrical work that this implies), Sindice and related services such as sig.ma w... (More —>)

Enabling networked knowledge
Building on Sindice: Sig.ma

Giovanni Tummarello

given name: Giovanni
family name: Tummarello
is creator of:
- A Node Indexing Scheme for Web Entity Retrieval
- Hierarchical Link Analysis for Ranking Web Data
- ESWC 2006 Demo: DBin - enabling SW P2P communities
- Rapid Prototyping of Semantic Mash-Ups through Semantic Web Pipes
- Context Dependent Reasoning for Semantic Documents in Sindice
- An Entity Name System for Linking Semantic Web Data
- Semantic Sitemaps: Efficient and Flexible Access to Datasets on the Semantic Web
- Sindice.com: Weaving the Open Linked Data
- RDFS\textit{sync}: efficient remote synchronization of RDF models
- Exposing Large Datasets with Semantic Sitemaps
- Enabling Semantic Web communities with DBin: an overview

show 100 more values
Data Reuse and Repurposing

“Got an I Like It button(*)? You are 1 click away to leverage the entire Web of Data (for your needs specific visitor needs).”

Sindice Site Services - for website owners

Sindice Widgets

Get Web Data Powered widgets to deliver highly valuable extra information and power smart recommendations across your pages.
Sindice.com/YOURSITE/movies/search

**DEMO.SINDICE.NET - Data Site Search**

**Tiles • Table**

**get_more_info_about_movie_exhibit**

sorted by: labels; then by... • ✓ grouped as sorted

1. **Title:** 10,000 BC
   **Image:**
   **Starring:** Steven Strait
   **Writer:** Harald Kloser
   **Director:** Roland Emmerich
   **Budget:** 105000000
   **Runtime:** 109 min.

2. **Title:** The Day After Tomorrow
   **Image:**
   **Starring:** Jake Gyllenhaal
   **Writer:** Roland Emmerich
   **Director:** Roland Emmerich
   **Budget:** $125,000,000 (estimated)
   **Runtime:** 124 mins.

3. **Title:** Alien 3
   **Image:**
   **Starring:** Sigourney Weaver
   **Writer:** Dan O'Bannon
   **Director:** David Fincher
   **Budget:** 50000000
   **Runtime:** Theatrical Cut: 114 min. Special Edition: 144 min.

4. **Title:** Alien Resurrection
   **Image:**

5. **Title:** The Thirteenth Floor
   **Image:**

6. **Title:** Aliens
   **Image:**

**Title**

1. 10,000 BC
2. Alien 3
3. Alien Resurrection
4. Aliens
5. The Day After Tomorrow
6. The Thirteenth Floor

**Starring**

1. Armin Mueller-Stahl
2. Jake Gyllenhaal
3. Sigourney Weaver
4. Steven Strait

**Writer**

1. Dan O'Bannon
2. Daniel F. Galouye
3. Harald Kloser
4. James Cameron
5. Roland Emmerich
Use Cases

Demonstrate how to benefit from Linked Open Data

1. Data awareness
   - opendata.ie
   - LOD cloud

2. Modeling
   - Neologism
   - DataCube
   - prefix.cc

3. Publishing
   - Google Refine plug-in
   - RDB2RDF/D2R

4. Discovery
   - VoID/DCAT
   - Sindice
   - CKAN

5. Integration
   - LATC
   - Sig.ma

6. Use cases
   - school explorer
   - DERI in-house

Enabling networked knowledge
Human burden to hold everything together

Organising a project showcase at a conference: involved Documents, contacts, time-schedule

E-Emails

Meeting Presentations

Travel Website

Time Schedule

Conference Website

Enabling networked knowledge
Challenges....

What are the sweet spots in this continuum?

• Scientific Communication
• Enterprise Data
• Collaboration
Science Communication

The process of scientific communication

- (reading)
- Scientist
- (authorship)
- Manuscript
- (searching)
- Database
- (indexing)
- Journal
- (publication)
- Paper
- Record
- Item

Enabling networked knowledge
■ Lightweight semantics and Linked Data on top of existing well-know applications

☐ User-driven semantics

☐ Combined with advanced browsing and querying interface

Enabling networked knowledge
Exploiting and enhancing Linked Data: Semantic Collaboration
Vision: Networked Collaborative Knowledge

- Free the Silos - Bridge the Islands
- Interconnected Knowledge
- Ensure local and global Collaboration

Enabling networked knowledge
Personal Information Model

E-Emails

Model PIMO

Time Schedule

Meeting Presentations

Travel Website

Conference Website

Interconnect disperse Information

http://www.semanticdesktop.org/ontologies/
Personal Semantic Web: a semantically enlarged intimate supplement to memory

Social protocols and distributed search

Social semantic peers

Enabling networked knowledge
Saffron - Expert Finder

Ian Horrocks
Oxford University Computing Laboratory
Homepage

Topics
1. OBO
2. enhanced traversal algorithm
3. OBO ontologies
4. subsumption tests
5. Description Logics
6. OBO parsers using automated tools
7. OBO language
8. Semantic Web
9. life sciences
10. OWL DL

Publications
1. OBO and OWL: Leveraging Semantic Web Technologies for the Life Sciences
2. Exploiting Partial Information in Taxonomy Construction
3. Efficient Query Answering for OWL 2
4. The OBO to OWL mapping, GO to OWL 1.1!
5. Extracting Modules from Ontologies: A Logic-based Approach
6. Adding Integrity Constraints to OWL
Enabling networked knowledge
Changing the World..

Enabling networked knowledge