Cultural Linked Open Data

2014-02-06

Lars Marius Garshol, larsga@bouvet.no, http://twitter.com/larsga
The importance of data

• Most web sites are data-driven
  – if you have the data, you can add functionality
  – if you don’t have the data, you’re stuck

• Example: Google Maps
  – imagine you have the application, the server farm, the scaling and monitoring, etc
  – but you don’t have the actual map data
  – not only are you stuck, but creating the data is *much* harder than making the service
Customers Who Bought This Item Also Bought

- Concerts: Bregenz Munich
  - Keith Jarrett
  - Audio CD
  - $21.35

- Six Sonatas for Violin and Piano
  - Keith Jarrett
  - Audio CD
  - $20.79

- Setting Standards
  - Keith Jarrett
  - Audio CD
  - $26.44

- Free Flying
  - Fred Hersch and Julian Gearhardt
  - Audio CD
  - $13.82

Product Details

- Audio CD (May 28, 2013)
- Original Release Date: 2013
- Number of Discs: 1
- Label: ECM Records
- ASIN: B005Y1FBUQ
- Also Available in: Audio CD | MP3 Music
- Average Customer Review: 💫💫💫💫💫 (9) (33 customer reviews)
- Amazon Best Sellers Rank: #1,862 in Music (See Top 100 in Music)
  #16 in Music > Jazz > Traditional Jazz & Ragtime

Did we miss any relevant features for this product? Tell us what we missed.
Would you like to update product info, give feedback on images, or tell us about a lower price?

Editorial Reviews

Now in its 30th year, the Keith Jarrett Trio is widely considered, as the NY Times recently remarked, to have set the go
to jazz groups, and this sparkling concert recording from 2009 is issued to mark a milestone anniversary.

The Somewhere in which the Standards trio find themselves is Lucerne, Switzerland with a performance both explorato
the-tradition. The Neue Zürcher Zeitung headlined its review of the show Kontrollierte Ekstase controlled ecstasy an ap
for a set that begins in improvisational Deep Space modulates into Miles Davis Solar, soars through the standards Stars
Alabama and Between The Devil And The Deep Blue Sea and climaxes with an extended romp through West Side Story,
Bernstein's Somewhere and Tonight are bridged by the freely associative Jarrett original Everywhere.

Customer Reviews

- 5 star (33)
- 4 star
- 3 star
- 2 star
- 1 star

- They played standards in splendid ways and gave us all beautiful moments we needed from the trio like this CD. ”
  - Mikio Miyaki | 7 reviewers made a similar statement

- My Highest Recommendation”
  - RSPRODS | 4 reviewers made a similar statement

- The Jarrett Trio is probably the best jazz trio of all jazz history.”
  - Manuel Gross & Galvan | 1 reviewer made a similar statement
Environmental friendly behaviour (Use Case 2)

- **Problem statement**: Faced with different transportation options for a short trip, which are the most environmental-friendly options given constraints like time, weather, traffic and private preferences.

- **Typically different options**
  - Public transportation (bus/tram/metro/train)
  - Private car (electric/gas/diesel) car, taxi
  - Cycling, walking

- **Constraints**: Time, avoid bad weather, polluted zones, traffic, private preferences

- **Environmental parameters**: CO2 emissions, energy efficiency

- **Added value proposition**: Enable smarter/faster environmental friendly decision making for local trips when options are available
  - Assist the user's decision making wrt travelling from his current position to the position of the next event in the user's calendar
Must be at meeting at 1345. Three transport alternatives.

Research project by SINTEF and Computas
Data sources

Research project by SINTEF and Computas
Data is raw material for building services!
Possible users of cultural data

• Any kind of web store
  – publishers
  – streaming services
  – ...

• Travel businesses
  – public sector, hotels, tour organizers, event organizers, ...

• Media
  – newspapers, broadcasting, ...

• Lots of public sector uses
  – education, ...

• Many things none of us can’t imagine now
Where is the data?!?
Only linked data is usable

NRK/Skole

Cappelen Damm
Linked Open Data

• Movement to publish open data online
  – in machine-readable form
  – linked to other data sets

• Based on some key technologies
  – URLs for identifiers
  – RDF for data

• Gaining a lot of traction in the cultural sector
  – BBC
  – Europeana
  – Smithsonian Institution
  – ...
The technology

• Provides simple data representation
  – graph model (RDF)
  – has ready-made formats (XML, text, JSON, ...)
  – standard query language (SPARQL)
  – lots of RDF databases available

• Allows anyone to refer to anything
  – a museum can say explicitly that one object in their collection has a specific relation to an object in another collection
  – liberation from the ID scheme confusion

• Can reuse terminology from other authorities
  – can also easily extend that terminology
About: **Knut Faldbakken**

An Entity of Type: **agent**, from Named Graph: [http://dbpedia.org](http://dbpedia.org), within Data Space: [dbpedia.org](http://dbpedia.org)

Knut Faldbakken (born Hamar, 31 August 1941) is a Norwegian novelist.

<table>
<thead>
<tr>
<th>Property</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>dbpedia-owl:birthDate</td>
<td>1941-08-31 (xsd:date)</td>
</tr>
<tr>
<td>dbpedia-owl:wikiPageID</td>
<td>13524052 (xsd:integer)</td>
</tr>
<tr>
<td>dbpedia-owl:wikiPageRevisionID</td>
<td>547830169 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:dateOfBirth</td>
<td>31 (xsd:integer)</td>
</tr>
<tr>
<td>dbpprop:hasPhotoCollection</td>
<td><a href="http://wifo5-03.informatik.uni-mannheim.de/flickrwrapp/photos/Knut_Faldbakken">http://wifo5-03.informatik.uni-mannheim.de/flickrwrapp/photos/Knut_Faldbakken</a></td>
</tr>
<tr>
<td>dbpprop:name</td>
<td>Faldbakken, Knut</td>
</tr>
<tr>
<td>dcterms:subject</td>
<td>category:People_from_Hamar</td>
</tr>
<tr>
<td></td>
<td>category:Hamar_Katedralskole_alumni</td>
</tr>
<tr>
<td></td>
<td>category:1941_births</td>
</tr>
<tr>
<td></td>
<td>category:Living_people</td>
</tr>
<tr>
<td></td>
<td>category:Norwegian_writers</td>
</tr>
<tr>
<td>rdf:type</td>
<td>owl:Thing</td>
</tr>
<tr>
<td></td>
<td>dbpedia-owl:Agent</td>
</tr>
</tbody>
</table>
http://dbpedia.org/resource/BouvetASA

- **Globally unique**
  - across all systems and organizations

- **Distributed**
  - if you have a domain, you can make URIs

- **Self-documenting**
  - just follow the link to find documentation

- **Can be used anywhere**
  - anyone can point at anything
Today

- Flat, unlinked data
- No navigation
- No connections
- Poor characterization – doesn’t say *what* it is
As linked data

- **edm:ProvidedCHO**
  - **nv:Photograph**
    - rdf:type: Bergliot Ibsen
    - dc:title: 1903
    - dc:date: Bergliot Ibsen
    - dc:subject: Bergliot Ibsen
    - 1953-02-02
    - dbp:died: 1869-06-10
    - dbp:born: Bergliot Ibsen
    - foaf:Person
  - http://dbpedia.org/resource/Bergliot_Ibsen

- **nv:provider**
  - http://dbpedia.org/resource/Aulestad
  - rdfs:label: Aulestad
  - grs:point: 61.2173 10.265952
Weaknesses in traditional formats

• Cannot be imported directly
  – must always be translated/interpreted somehow
• Cannot be automatically merged
  – no concept of identity
• Vocabularies cannot be linked
  – types and properties not connected to other types and properties
• Linked Data solves all of these problems
  – Linked Data ≈ RDF
The power of owl:sameAs

• Different IDs in these two systems
  – using “sameAs” we know which are the same
  – can do queries across the systems
  – can also translate IDs from one system to the other

• Triple stores can do reasoning
  – basically map any number of sameAs connections automatically
Choice of tools

Modelling
- TopBraid Composer™
- protégé
- ENTERPRISE ARCHITECT

Reasoners
- RacerPro 2.0
- KAON2
- clarkparsia
- pellet

APIs
- Yena
- dotNetRDF

Triple stores
- ORACLE
- Stardog
- VIRTUOSO
- IBM DB2
Great, but how can we actually link the data?
data.deichman.no

data.deichman.no is a gateway to raw RDF data from Oslo Public Library and a platform for developers to access these data and build applications that use them.

The main part of data.deichman.no is an enriched RDF version of the library catalogue and its authority files. This includes data about books, music, movies, persons, organizations, topics and genres. The dataset consists of an RDF version of the main MARC catalogue, enriched with book cover URLs harvested from external datasets, as well as a FRBR-like work-manifestation-structure that identifies the unique works described in the catalogue, and clusters the different editions and translations of each work. The dataset is updated daily to account for additions, deletions and changes made in the catalogue.

data.deichman.no also contains a collection of book reviews produced by librarians at Oslo public library, or through the «Ønskebok» project. Book reviews are linked to works in the main catalogue dataset.

As of January 2014, data.deichman.no contains descriptions of:

~ 415 000 documents (books, CDs, DVDs, periodicals etc.)
~ 200 000 persons
~ 75 000 organizations
~ 65 000 topics
~ 2000 genres
~ 7500 book reviews

To look up data in the web browser, simply enter a resource URI in the address bar:

Example: http://data.deichman.no/resource/tnr_749919

As Turtle:
http://data.deichman.no/resource/tnr_749919.rdf

As JSON:
http://data.deichman.no/resource/tnr_749919.json
“Do they have Knut Faldbakken in here?”
Yes, but not connected to anything ...  
...can we do anything about that?
Record linkage to the rescue

• Active research field  
  – dating back to the 1940s
• Can connect data  
  without common IDs  
  – measure similarity instead
• Tools exist, with  
  – value cleaning  
  – statistical analysis  
  – sophisticated comparators  
  – fast search backends
• One example is Duke  
  – http://code.google.com/p/duke/  
  – Java and open source
Connect to DBpedia

http://data.deichman.no/...dbakken_Knut_1941-

NAME: Faldbakken, Knut
LIFESPAN: 1941-
NATIONALITY: n

http://dbpedia.org/resource/Knut_Faldbakken

NAME: Knut Faldbakken
BIRTHDATE: 1941-08-31


---NAME
'kнут faldbakken' ~ 'knut faldbakken': 1.0 (prob 0.8)
'kнут faldbakken' ~ 'knut faldbakken': 1.0 (prob 0.8)
Result: 0.5 -> 0.8

---YEAROFBIRTH
'1941' ~ '1941': 1.0 (prob 0.52)
Result: 0.8 -> 0.8125

---YEAROFDEATH
---NATIONALITY
Overall: 0.8125

Complete recipe here
Training with genetic algorithm

http://www.garshol.priv.no/blog/262.html
Conclusion

• Linked Open Data has tremendous potential
  – vastly easier reuse of data
  – hugely empowering for consumers
• Growing use in cultural sector
  – both internationally and in Norway
• To learn more
  – http://www.slideshare.net/larsga/linked-open-data-14964163
  – http://data.norge.no/veiledning
  – http://linkeddatabook.com/editions/1.0/
Hafslund SESAM

DIREKTE/AUTOMATISK PÅSLAG AV METADATA