Building a High Performance Environment for RDF Publishing

Pascal Christoph

SWIB12
Semantic Web in Libraries

hbz

These slides and all the graphics made by the author and those taken from https://openclipart.org/ are dedicated to the public domain : https://creativecommons.org/about/cc0 .

All marks mentioned may be trademarks or registered trademarks of their respective owners.

Read about the license of „The scream“ of Edward Munch at https://en.wikipedia.org/wiki/File:The_Scream.jpg

Linking Open Data cloud diagram, by Richard Cyganiak and Anja Jentzsch. http://lod-cloud.net/

hbz

Overview

Publishing is for Consuming
- Mandatory
- Nice to have

Story so far - experiences with lobid.org
- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch
- Benefits
- Some more details
- Caveats

Future prospects

Overview

Publishing is for Consuming
- Mandatory
- Nice to have

Story so far - experiences with lobid.org
- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch
- Benefits
- Some more details
- Caveats

Future prospects

Publishing is for Consuming

Publishing is for Consuming

Publishing is for Consuming

Mandatory

A resource:

Publishing is for Consuming

Mandatory

A resource:

gets a dereferenceable URI:

http://lobid.org/resource/HT002948556

Publishing is for Consuming

Mandatory

A resource: 

gets a dereferenceable URI: http://lobid.org/resource/HT002948556

which provides RDF:

<http://lobid.org/resource/HT002948556> <http://purl.org/dc/terms/title> "With reference to reference" .
<http://lobid.org/resource/HT002948556> <http://purl.org/dc/terms/issued> "1983" .
<http://lobid.org/resource/HT002948556> <http://purl.org/ontology/bibo/isbn13> "9780915145539" .
<http://lobid.org/resource/HT002948556><http://purl.org/dc/elements/1.1/creator><http://d-nb.info/gnd/135539897> .

Publishing is for Consuming

Mandatory

=> basic LOD publishing is very simple:

you just need a Webserver

Publishing is for Consuming

Nice to have

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (best: RDFa in HTML)
- Data searchable
- Timely updates
- High Availability
- Versioning
- Web developers want simple APIs providing JSON
- ...

Publishing is for Consuming

SPARQL Endpoint

- (Dumps)
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (best: RDFa in HTML)
- Data searchable
- Timely updates
- High Availability
- Versioning
- Web developers want simple APIs providing JSON
- ...

Publishing is for Consuming

SPARQL Endpoint

- (Dumps): but may be painfully slow when having lots of data
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (best: RDFa in HTML)
- (Data searchable) : maybe painfully slow
- Timely updates
- High Availability
- Versioning
- Web developers want simple APIs providing JSON
  - most triple stores provides JSON/RDF
  - ~~Simple~~ powerful API : too powerful/complex ?
-

Publishing is for Consuming

Nice to have

In principle, web developers already got simple APIs :

LOD *is* the API !

Publishing is for Consuming

Nice to have

In principle, web developers already got simple APIs :

Remember:

Publishing is for Consuming

Mandatory

A resource: 

gets a dereferenceable URI: http://lobid.org/resource/HT002948556

which provides the data (in RDF):

<http://lobid.org/resource/HT002948556> <http://purl.org/dc/terms/title> "With reference to reference" .
<http://lobid.org/resource/HT002948556> <http://purl.org/dc/terms/issued> "1983" .
<http://lobid.org/resource/HT002948556> <http://purl.org/ontology/bibo/isbn13> "9780915145539" .
<http://lobid.org/resource/HT002948556><http://purl.org/dc/elements/1.1/creator><http://d-nb.info/gnd/135539897> .

Publishing is for Consuming

Nice to have

In principle, web developers already got powerful APIs :

RESTful SPARQL

RESTful SPARQL example

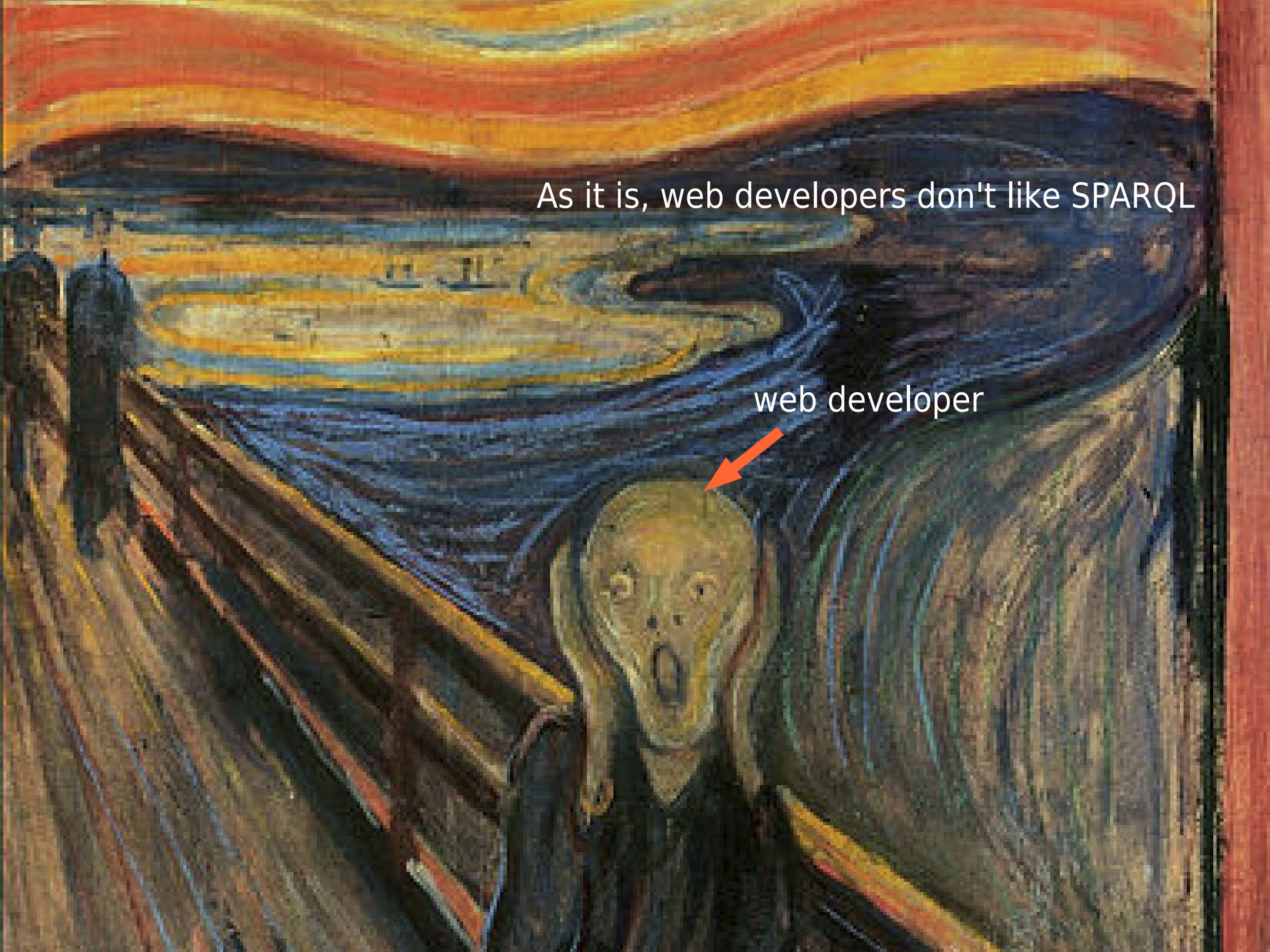
getting all data of all resources having a particular ISBN:

```
curl   -H "Accept: application/json"  --data-urlencode 'query=
prefix bibo: <http://purl.org/ontology/bibo/>
SELECT * WHERE {
 ?s bibo:isbn13 "9780851706238" ;
    ?p ?o .
 } LIMIT 100
' http://lobid.org/sparql/
```

RESTful SPARQL example

… and the JSON/RDF result:


```
{
   "head": {
      "vars": [ "s", "p","o"]
   },
   "results": {
      "bindings": [ {
            "o": {
               "type": "uri",
               "value": "http://openlibrary.org/works/OL2109573W"
      },
      "p": {
         "type": "uri",
         "value": "http://rdvocab.info/RDARelationshipsWEMI/workManifested"
      },
      "s": {
         "type": "uri",
         "value": "http://lobid.org/resource/HT007824357"
      }
   },
   {
      "o": { ...
```


As it is, web developers don't like SPARQL
web developer

Publishing is for Consuming

Nice to have

Web developers want APIs like:

http://lobid.org/resources/api/isbn/$isbn


Happy web developer

Publishing is for Consuming
- Mandatory
- Nice to have

Story so far - experiences with lobid.org
- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch
- Benefits
- Some more details
- Caveats

Future prospects

lobid.org

What is lobid.org ?

- lobid := **l**inking **o**pen **b**ibliographic **d**ata
- LOD services of the hbz
  - lobid-resources :
    - exposes 85% of the hbz cooperative catalogue
    - entries coming from > 200 scientific German libraries
    - ~ 16 M records with 700 M triples
    - with links to ~ 5 M other resources
    - with links to ~ 32 M items (consisting of 300 M triples)
  - lobid-organisations :
    - exposes German Sigelverzeichnis and MARC-Isil directory
    - ~ 40 k descriptions of institutions

What's missing?

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable

- Timely updates
- High Availability
- Versioning
- Web developers want simple APIs providing JSON
- ...

Overview

Publishing is for Consuming

- Mandatory
- Nice to have

Story so far - experiences with lobid.org

- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch

- Benefits
- Some more details
- Caveats

Future prospects

2010 - 2011, lobid-organisation

Filesystem :

+ easy to maintain
+ reliable
+ fast
- no search
- no SPARQL
- ...

lobid today

Triple Store (4store) : 4store

+ power of SPARQL

+/- depending on the query: fast to horribly slow

+/- search (but string searches often slow and limited)

- sometimes gets stuck !

lobid today

Search engine (elasticsearch):

+ fast search
+ stemming, linguistics …
+ wildcard searching
+ facets
+ geo search
+ JSON
+ schema-less
+ simple RESTful API
+ many plugins
+ ...
+ easy to achieve High Availability
+ scales nicely

lobid today

storing/getting the data

Transformation via hbz-tools:
Aleph XML to RDF (Dublin Core, Bibo ...)

Transformation via hbz-tools:
Aleph XML to JSON, ES Index Profile

Aleph DB

4store

SPARQL Endpoint

external RDF data

Apache HTTP Server

The Apache Software Foundation
http://www.apache.org/

Webserver

elasticsearch.

Searchmachine

PHRESNEL
Web Frontend of RDF data

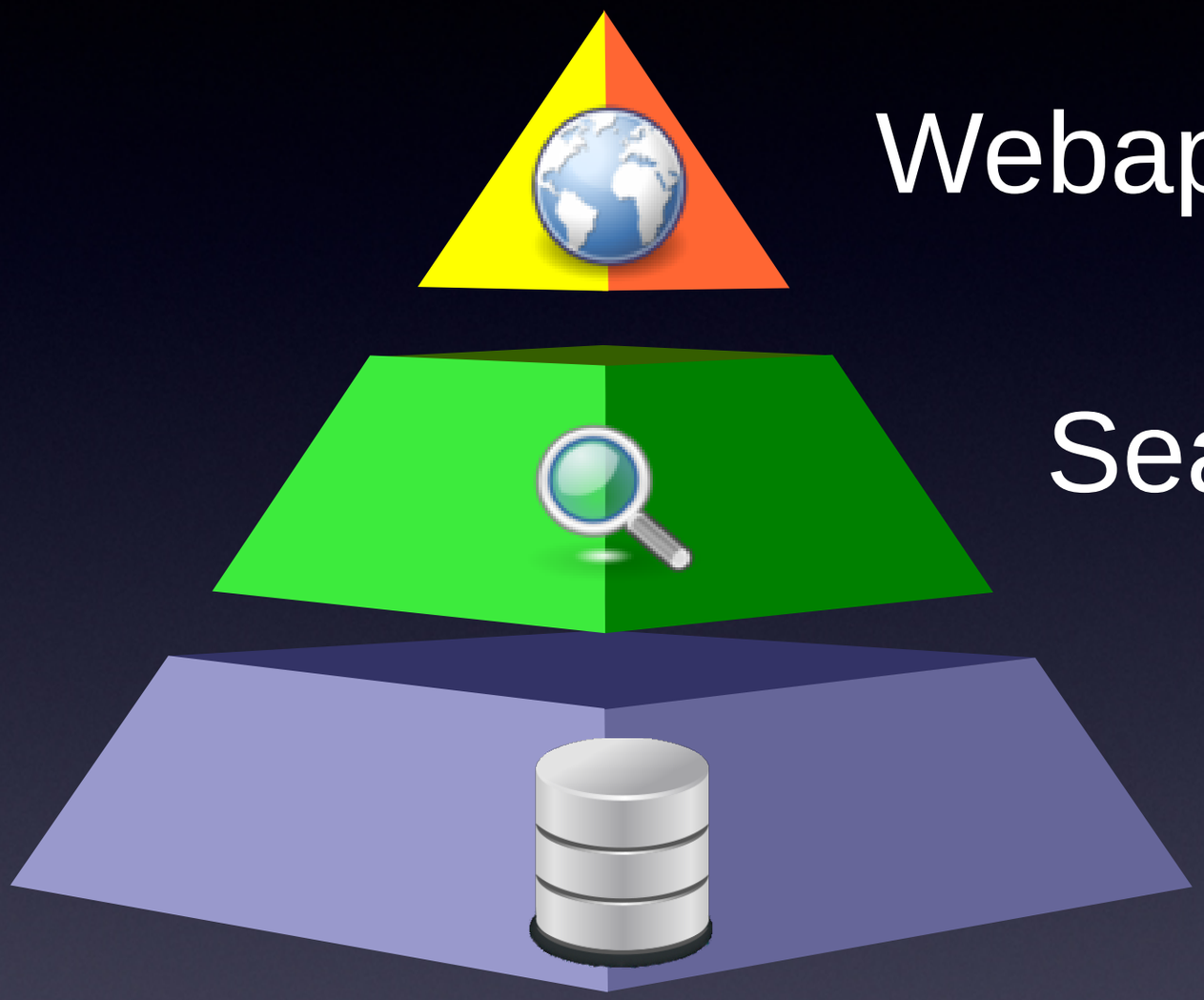
Publishing is for Consuming
- Mandatory
- Nice to have

Story so far - experiences with lobid.org
- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch
- Benefits
- Some more details
- Caveats

Future prospects

lobid : technology/dependency stack

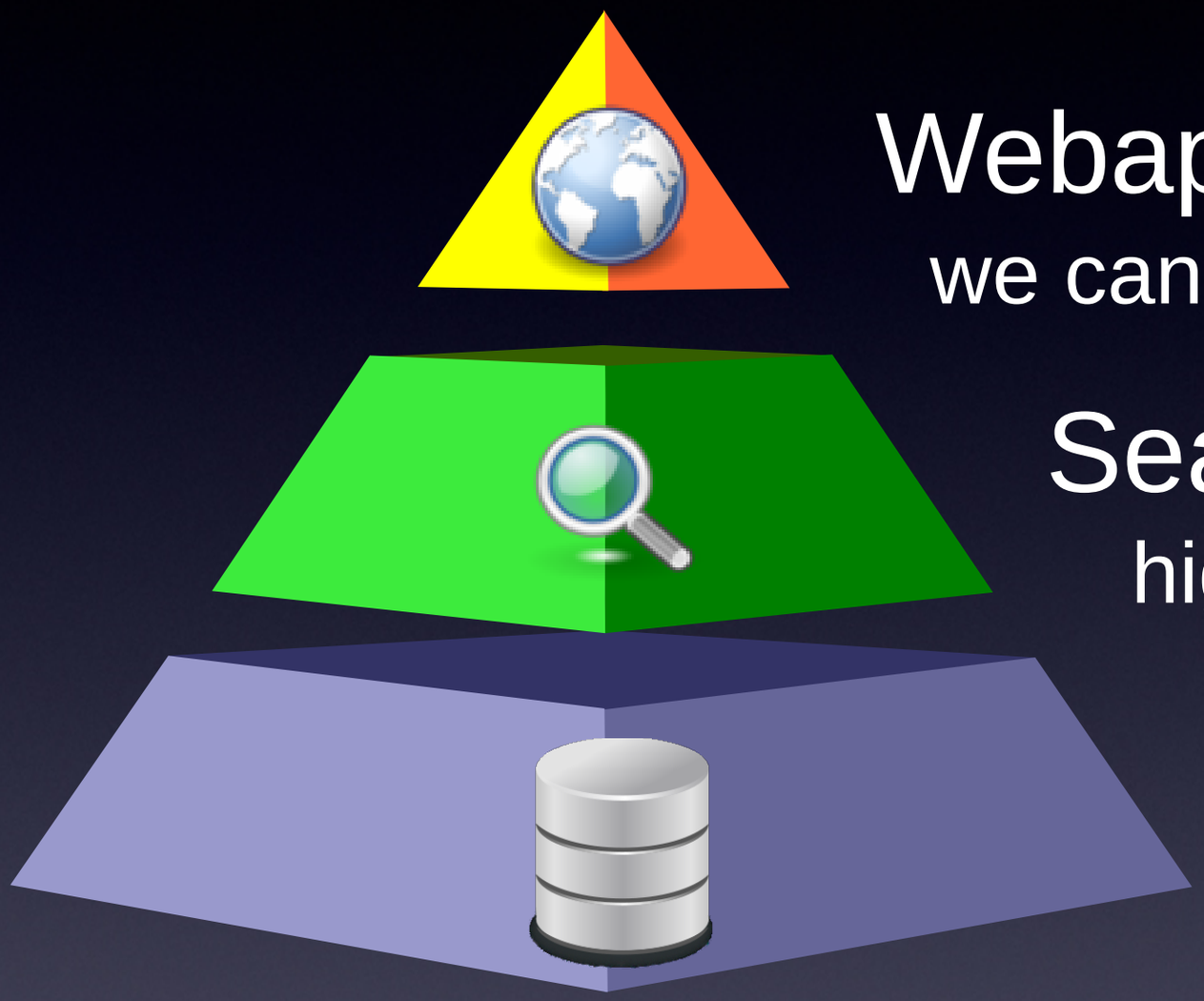


Webapp

Search Engine

Triple Store

lobid : technology/dependency stack



Webapp

we can do that

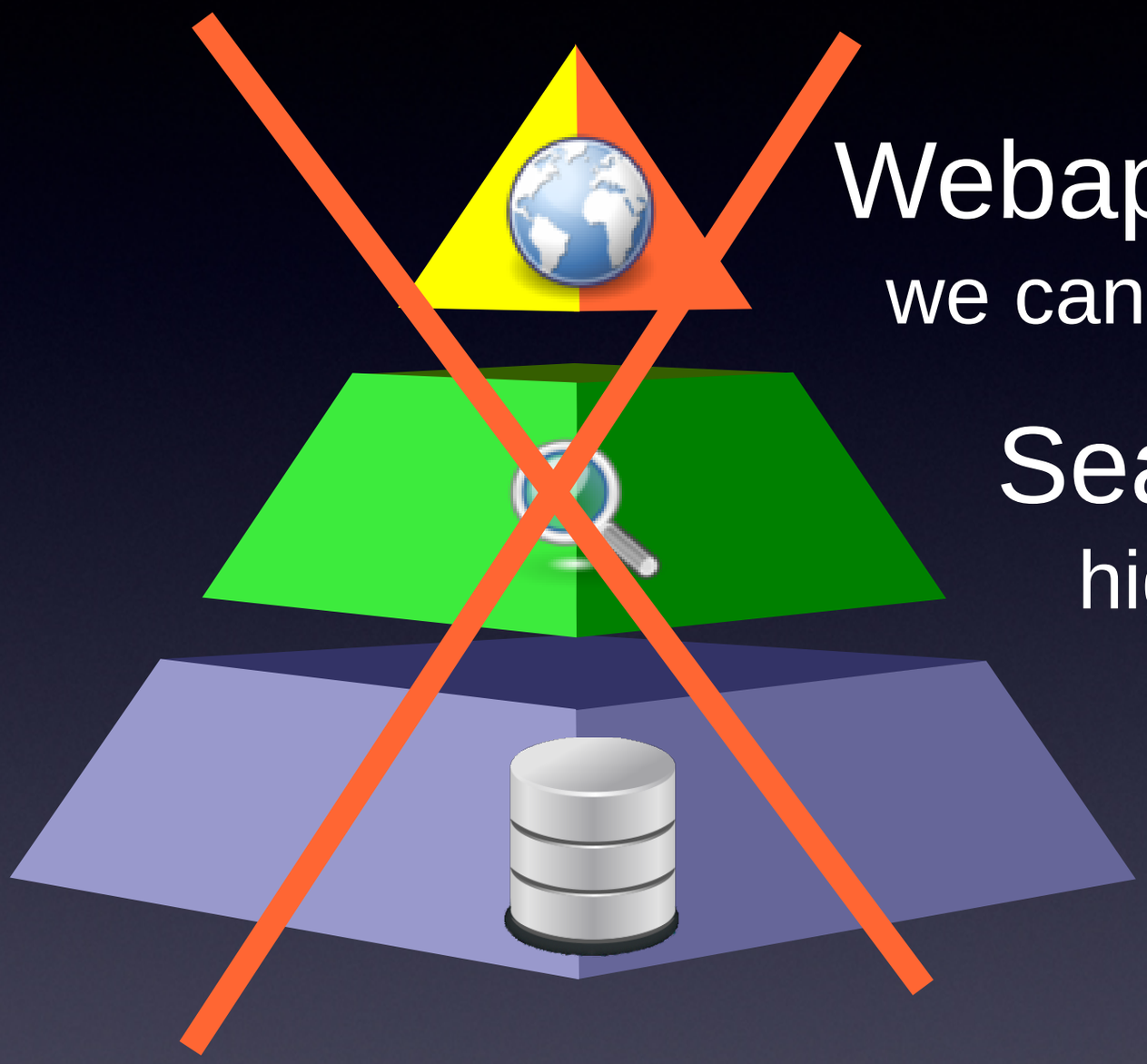
Search Engine

highly available !

Triple Store

sometimes gets stuck!

lobid : technology/dependency stack



Webapp

we can do that

Search Engine

highly available !

Triple Store

sometimes gets stuck!

lobid : technology/dependency stack

Webapp

we can do that

Search Engine

highly available !

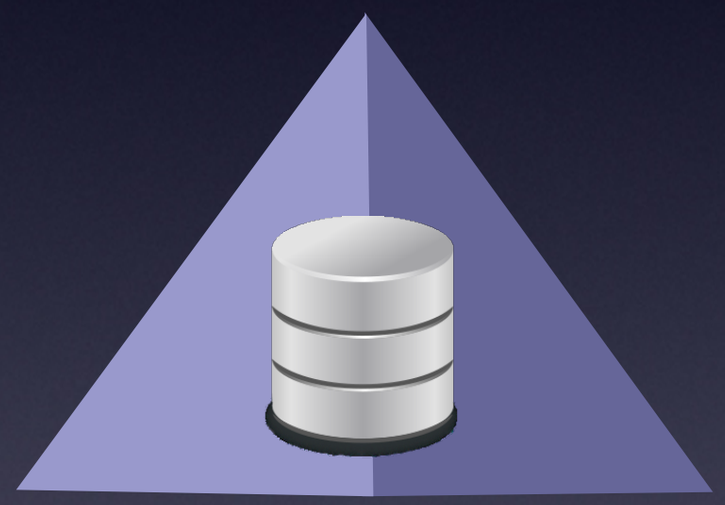
Triple Store

sometimes gets stuck!

Variant 1 : technology/dependency stack

Triple Store

Closed, internal. Will be safe from malign queries.

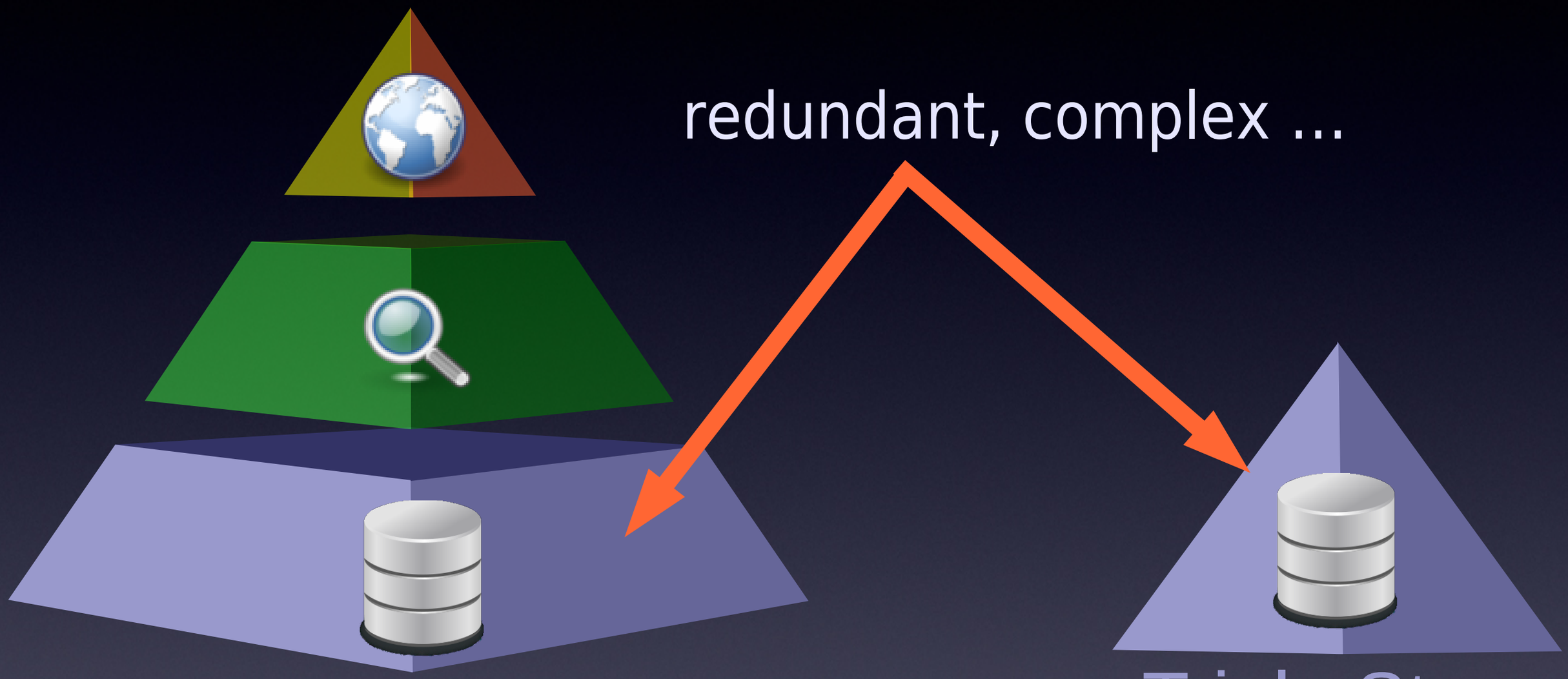


Triple Store

For external access.

Sometimes gets stuck!

Variant 1 : technology/dependency stack



Triple Store

Closed, internal. Will be safe from malign queries.

Triple Store

For external access.
Sometimes gets stuck!

Overview

Publishing is for Consuming

- Mandatory
- Nice to have

Story so far - experiences with lobid.org

- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch

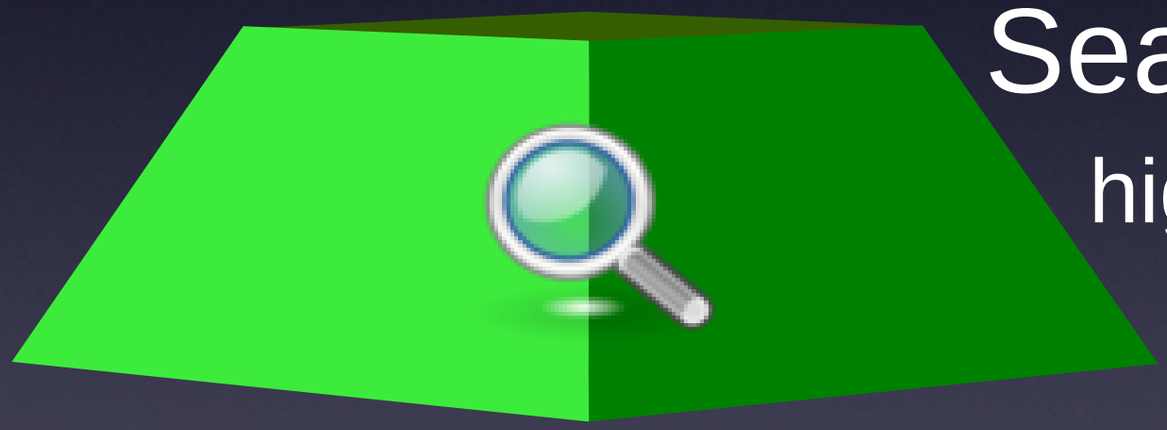
- Benefits
- Some more details
- Caveats

Future prospects

Variant 2: technology/dependency stack

Webapp

we can do that



Search Engine

highly available !

Triple Store

For external access and some fancy nice-to-have stuff.

Sometimes gets stuck!

LOD basis functionality (and some other APIs) are highly available

Benefits

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable
- Near Real Time updates
- High Availability
- (Versioning)
- Web developers want simple APIs returning JSON
- ...

Benefits

fast, scalable search engine

elasticsearch.

4store

Publishing LOD with elasticsearch

performance test

Data: 10 M records <=> 300 M triple
Case-insensitive query: „beach“

```
SELECT  ?s
WHERE
  { ?s <http://purl.org/dc/terms/title> ?o
    FILTER regex(str(?o), "beach", "i")
  }
#### => SPARQL execution time for Q8316: 108.7s, returned 2815 rows.

http://$ip:9200/_search?q=beach&from=0&size=2800
# => Elasticsearch needed 0.4s
```

=> Elasticsearch is 250 times faster

elasticsearch.

4store

performance test

(there is a support for text indexing in 4store, have not tested that.)

Publishing LOD with elasticsearch

performance test

Elasticsearch: 18 M records , 6 GB RAM: 5 hour
4store: 1 B triples, having 72 GB RAM: 7 hours

CPU: Quad Core mit 2.4 GhZ und Hyperthreading => 8 CPUs
HD: 6 x 2.5" 10k U/min a 146GB

(Don't take benchmarks too seriously – they just give a clue !)

Benefits

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable 
- Near Real Time updates
- High Availability
- (Versioning)
  - Web developers want simple APIs providing JSON
- ...

Benefits

build to be easily made highly available !

Benefits

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable 
- Near Real Time updates
- High Availability 
- (Versioning)
  - Web developers want simple APIs providing JSON
- ...

Benefits

Versioning with elasticsearch:

Not out-of-the-box, but comes at least e.g. with

* concurrency control
* documents have a version number

=> implementing versioning is not hard

Publishing LOD with elasticsearch

Benefits, relying on elasticsearch as basic LOD storage

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable
- Near Real Time updates
- High Availability
- Versionizing
- Web developers want:
  - JSON (LD)
  - Simple APIs
- ...

Benefits

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable

- Near Real Time updates
- High Availability

- (Versioning)
  - Web developers want simple APIs providing JSON

- ...

Why JSON-LD?

JSON is :

- stored natively by many tools (e.g. elasticsearch)
- loved by consumers (web developers)

JSON-LD is :

- supported by RDF libraries (e.g. transforming to NTriples)

Benefits

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable

- Near Real Time updates
- High Availability

- (Versioning)
  - Web developers want simple APIs providing JSON

- ...

Benefits

RESTful elasticsearch API, e. g. :

```
http://lobid.org/resources/_search?q=isbn:$isbn
```

Benefits

- … and many other nice things come with elasticsearch
  - geo-search : „Query only libraries/items residing up to 10 km from me.“
  - …

Publishing LOD with elasticsearch

Benefits

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable
- Near Real Time updates
- High Availability
- (Versioning)
- Web developers want simple APIs providing JSON
- ...

Mission accomplished !

Publishing LOD with elasticsearch

(… ok, something is left to be done !)

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML)
- Data searchable
- Near Real Time updates
- High Availability
- Versionizing
- Web developers want simple APIs providing JSON
- ...

Overview

Publishing is for Consuming
- Mandatory
- Nice to have

Story so far - experiences with lobid.org
- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch
- Benefits
- Caveats
- Auto suggest demo

Conclusion

Caveats

- Dumps
- Content Negotiation (different RDF serializations)
- SPARQL
- Human readable representation (RDFa in HTML) ☹
- Data searchable 😐
- Near Real Time updates
- High Availability
- Versionizing 😐
- Web developers want simple APIs providing JSON
- ...

Caveats

How to integrate semantic search into a document storage ?

dct:contributor --------> dct:creator -------> dc:creator
\--------> dc:contributor
\--------> bibo:translator
...

There is no inferencing as comes with SPARQL !

Caveats

Our data flow :

from records to RDF triples to records

Caveats

Our data flow :

from records to RDF triples to records

Caveats

from records to RDF triples to records

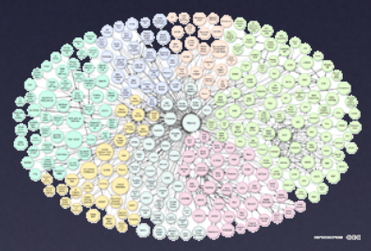
Caveats

From records to RDF triples

|-----> graph-database

'------> computing ---> record-database

MARC/MAB/PICA...



JSON-LD

Caveats

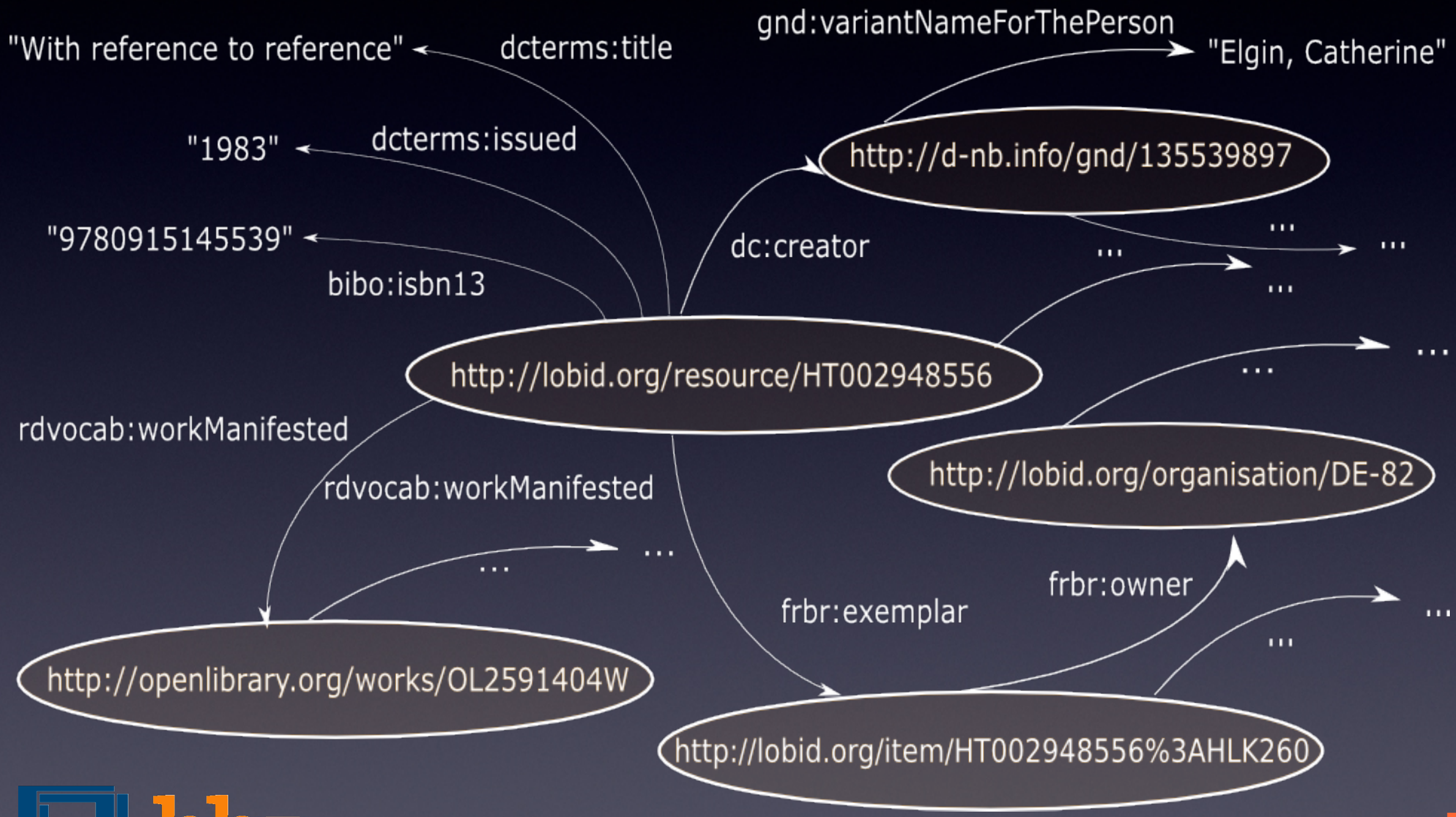
tree-based vs graph-based:

Pre-render the whole document?

What *is* the document ?

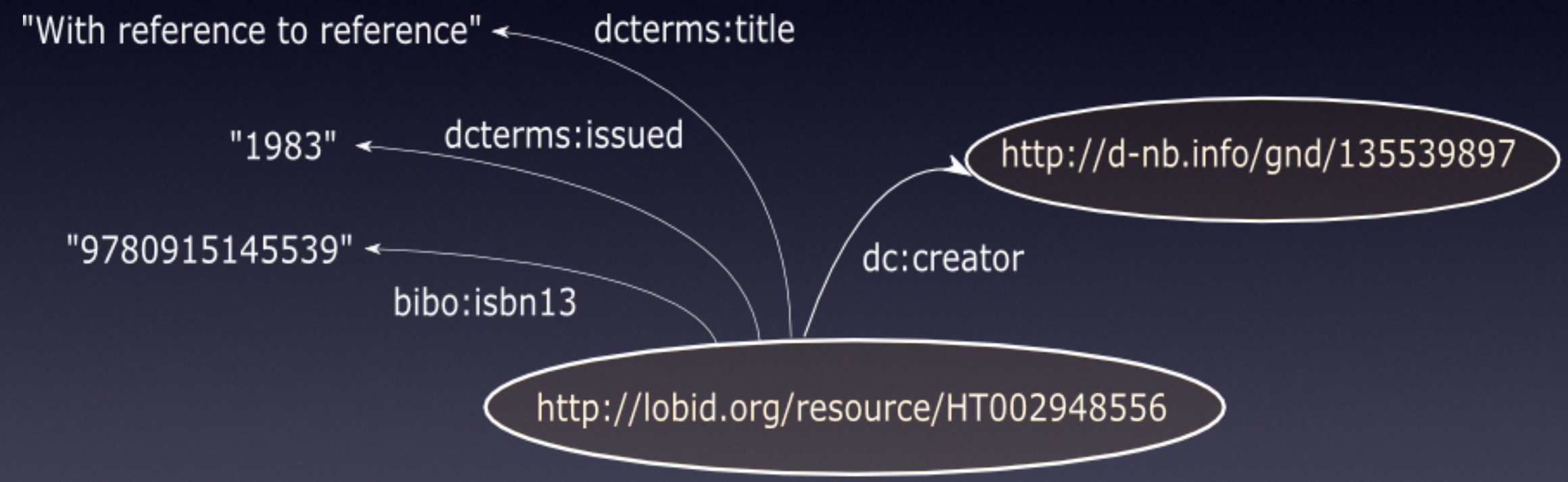
Publishing LOD with elasticsearch

Caveats



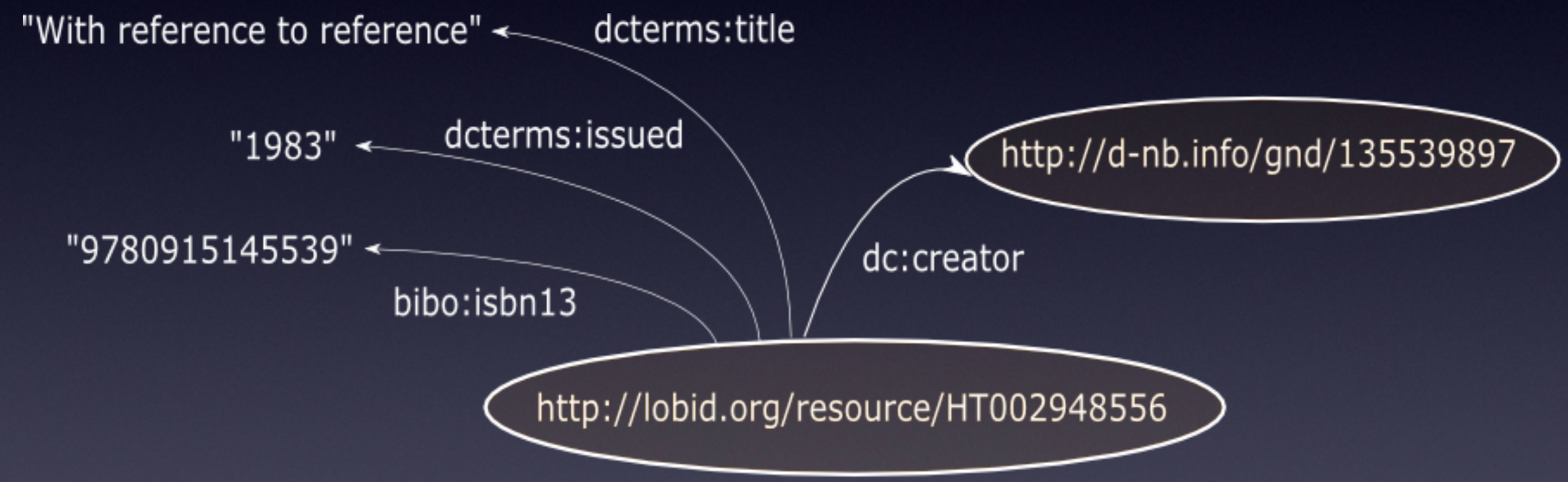
Caveats

What *is* the document ? Only the top-level node ?



Caveats

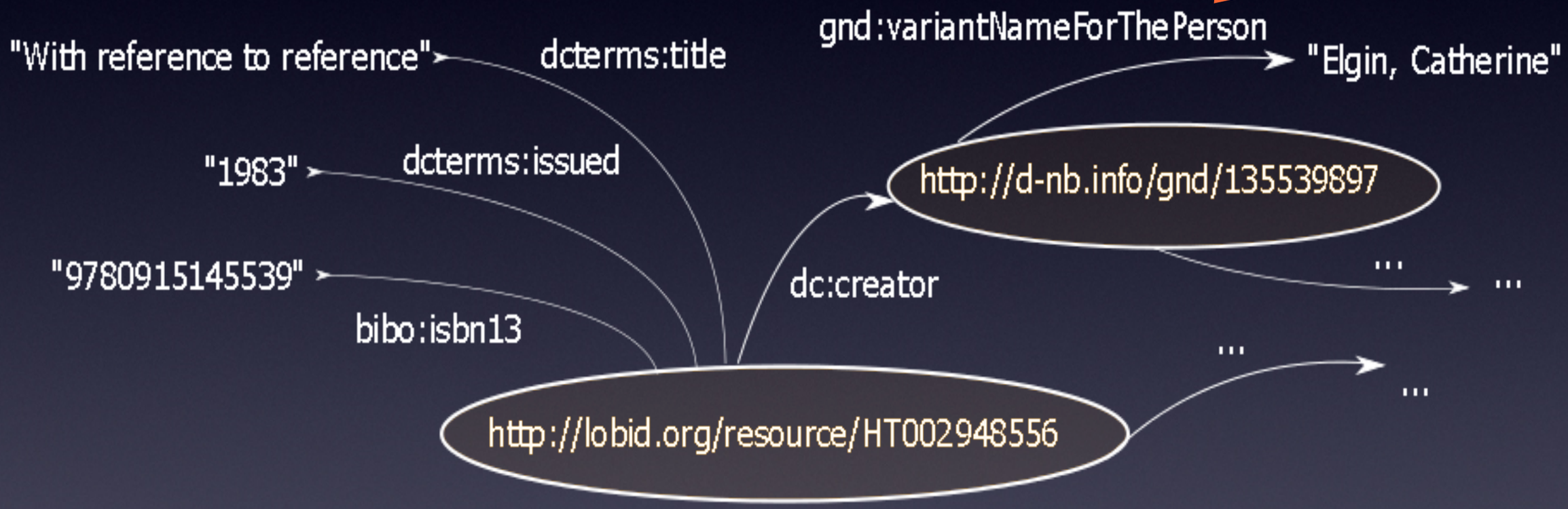
What *is* the document ? Only the top-level node ?



... but then you couldn't even search the authors name !

Caveats

searching needs integration of some fields from subgraphs into the document



Overview

Publishing is for Consuming
- Mandatory
- Nice to have

Story so far - experiences with lobid.org
- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch
- Benefits
- Caveats
- Auto suggest demo

Conclusion

auto suggest

authority IDs must be easily found

auto suggest

authority IDs must be easily found

=> in need of auto suggest

auto suggest

auto suggests needs fast searching

Demo

auto suggest

gnd-index

Karl Schmid

Schmidt, Karl (1935-)
Schmid, Karl
Schmidt, L. F. Karl
Schmidt, Karl A.
Schmidt, Karl (1867-)
Schmidt, F. L. Karl
Schmidt, Karl (1913-)
Schmidt, Karl (1902-1945)
Schmid, Karl (1910-)

Publishing LOD with elasticsearch

nt for 'Schmidt, Karl (1902-1945)'

fo/gnd/142017272

{"@context":{"rdfs":"http://www.w3.org/2000/01/rdf-schema#","xsd":"http://www.w3.org/2001/XMLSchema#","rdf":"http://www.w
ns#"},"http://d-nb.info/standards/elementset/gnd#oldAuthorityNumber":"(DE-588a)142017272","http://d-nb.info/standards/elem
/gnd#preferredNameForThePerson":"Schmidt, Karl","http://d-nb.info/standards/elementset/gnd#professionOrOccupation":{"@
/gnd/4177294-5"},"@type":"http://d-nb.info/standards/elementset/gnd#DifferentiatedPerson","http://d-nb.info/standards/eleme
/gnd#biographicalOrHistoricalInformation":"Als politischer Gefangener auf einem Schiff interniert , bei der Versenkung ums Le
nb.info/standards/elementset/gnd#placeOfBirth":{"@id":"http://d-nb.info/gnd/4036934-1"},"http://d-nb.info/standards/elementse
//d-nb.info/gnd/2016052-5"},"http://d-nb.info/standards/elementset/gnd#preferredNameEntityForThePerson":{"@id":"_:bn0"},"h
/elementset/gnd#gndIdentifier":"142017272","http://d-nb.info/standards/elementset/gnd#geographicAreaCode":{"@id":"http://
/gnd/geographic-area-code#XA-DE"},"http://d-nb.info/standards/elementset/gnd#dateOfDeath":"1945","@id":"http://d-nb.info/g
//d-nb.info/standards/elementset/gnd#gender":{"@id":"http://d-nb.info/standards/vocab/gnd/Gender#notKnown"},"http://d-nb.in
/gnd#dateOfBirth":"1902"}

NT

<http://d-nb.info/gnd/142017272> <http://d-nb.info/standards/elementset/gnd#affiliation> <http://d-nb.info/gnd/2016052-5> . <
/gnd/142017272> <http://d-nb.info/standards/elementset/gnd#gender> <http://d-nb.info/standards/vocab/gnd/Gender#notKn
/gnd/142017272> <http://d-nb.info/standards/elementset/gnd#preferredNameEntityForThePerson> <_:c14n0> . <http://d-nb.i
nb.info/standards/elementset/gnd#professionOrOccupation> <http://d-nb.info/gnd/4177294-5> . <http://d-nb.info/gnd/142017

Publishing LOD with elasticsearch

auto suggest

RESTful APIs:

http://demo.lobid.org/search?format=short&index=gnd-index&author=Schmidt%2C+Karl

http://demo.lobid.org/search?format=page&index=gnd-index&author=Schmidt%2C+Karl

http://demo.lobid.org/search?format=full&index=gnd-index&author=Schmidt%2C+Karl

...

API usage:
GET /search?format=<page|full|short>&index=<lobid-index|gnd-index>&author=<query>

easy to enhance with the play framework and the elasticsearch API

auto suggest

RESTful APIs:

http://demo.lobid.org/search
?format=short&index=gnd-index&author=Schmidt%2C+Karl


```
[
    "Schmidt, Karl (1894-1945)",
    "Schmidt, Karl",
    "Schmidt, Karl (1910-)",
    "Schmidt, Karl (1846-1928)",
    "Schmidt, Karl (1913-)",
    "Schmidt, Karl (1899-)",
    "Schmidt, Karl (1924-)",
    "Schmidt, Karl (1836-1888)",
    "Schmidt, L. F. Karl",
    "Schmidt, Karl (1902-1945)",
    "Schmidt, Karl J.",
    "Schmidt, Karl (1848-1905)",
    "Schmidt, Karl (1817-1882)",
    "Schmidt, Karl R.",
    "Schmidt, Karl (1954-)",
    "Schmidt, Karl (1888-)",
    "Schmidt, Karl (1867-)",
  ...
]
```

auto suggest

GND authority file

in

lobid-resources

Publishing LOD with elasticsearch

lobid-index | Schmidt, Karl (1859-) | Search

1 Document for 'Schmidt, Karl (1859-)'

http://lobid.org/resource/HT000800543

JLD

{"@context":{"rdfs":"http://www.w3.org/2000/01/rdf-schema#","xsd":"http://www.w3.org/2001/XMLSchema#","rdf
syntax-ns#"},"http://purl.org/dc/elements/1.1/creator#dateOfBirth":"1859","http://purl.org/dc/terms/language":{"@
/iso639-2/deu"},"http://purl.org/dc/terms/source":["Bericht / Königl. Kaiser-Wilhelms-Realgymnasium zu Berlin","
["http://purl.org/vocab/frbr/core#Manifestation","http://purl.org/ontology/bibo/Article","http://purl.org/dc/terms/Bibli
/dc/terms/issued":"1894","http://purl.org/dc/elements/1.1/creator#preferredNameForThePerson":"Schmidt, Karl",
Gründe des Bedeutungswandels : ein semasiologischer Vergleich","http://xmlns.com/foaf/0.1/isPrimaryTopicOf"
/F/?func=find-c&ccl_term=IDN%3DHT000800543"},"http://iflastandards.info/ns/isbd/elements/P1016":"Berlin","h
/elements/P1004":"Die Gründe des Bedeutungswandels","http://purl.org/dc/terms/isPartOf":{"@id":"http://lobid.o
//iflastandards.info/ns/isbd/elements/P1006":"ein semasiologischer Vergleich","@id":"http://lobid.org/resource/HT
/2007/05/powder-s#describedby":{"@id":"http://lobid.org/resource/HT000800543/about"},"http://purl.org/dc/eleme
/gnd/142004529"}}

NT

http://lobid.org/resource/HT000800543> <http://www.w3.org/1999/02/22-rdf-syntax-ns#type> <http://purl.org/vo
<http://lobid.org/resource/HT000800543> <http://iflastandards.info/ns/isbd/elements/P1004> "Die Gr\u00FCnde d

lobid.org

*l*inking *o*pen *bi*bliographic *d*ata

DIENSTE

SPARQL Endpoint

SPARQL Webform

RESSOURCEN

Suche

RESTful HTML-Beispiel

Open-Data-Dumps

ORGANISATIONEN

Suche

RESTful HTML-Beispiel

META

LOD-Wiki des hbz

Impressum

Kontakt

http://lobid.org/resource/HT000800543

Titel	Die Gründe des Bedeutungswandels
Titelzusatz	ein semasiologischer Vergleich
Autor	http://d-nb.info/gnd/142004529
Erscheinungsjahr	1894
Typ	<http://purl.org/dc/terms/BibliographicResource>
Typ	<http://purl.org/ontology/bibo/Article>
Typ	<http://purl.org/vocab/frbr/core#Manifestation>
Sprache	http://id.loc.gov/vocabulary/iso639-2/deu
Erscheinungsort	Berlin
Übergeordnetes Werk	http://lobid.org/resource/HT012537724
<http://purl.org/dc/terms/source>	1894, S. [3] - 44

Overview

Publishing is for Consuming

- Mandatory
- Nice to have

Story so far - experiences with lobid.org

- What is lobid.org ?
- Storing the data
- Getting the data

Publishing RDF through elasticsearch

- Benefits
- Caveats
- Auto suggest demo

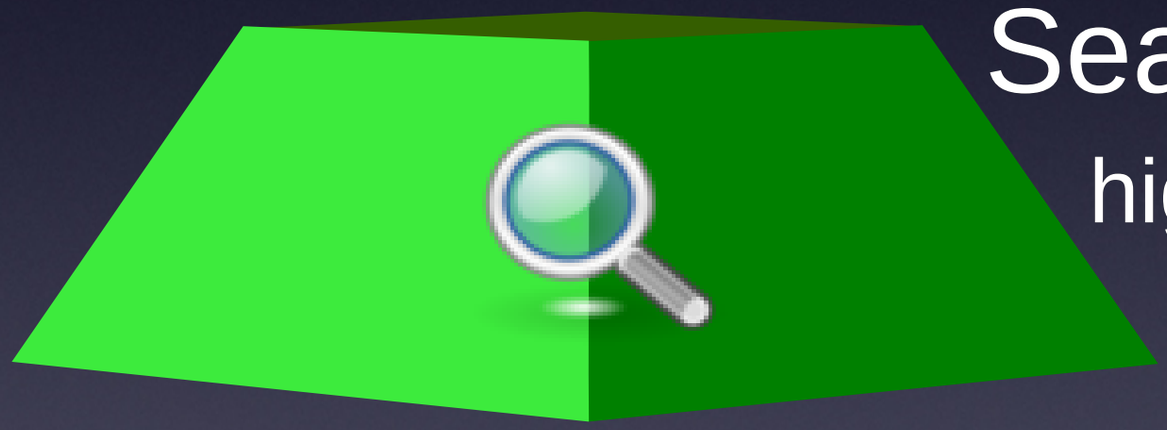
Conclusion

Conclusion

a highly customizable/reliable/feature-rich LOD service

Webapp

we can do that



Search Engine

highly available !

LOD basis functionality (and some other APIs) are highly available

Triple Store

For external access and some fancy nice-to-have stuff.

Sometimes gets stuck!

Publishing LOD with elasticsearch

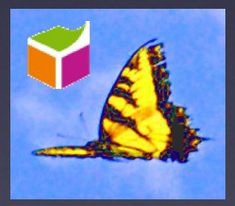
the software is Open Source:

elasticsearch. http://elasticsearch.org/

https://hadoop.apache.org/

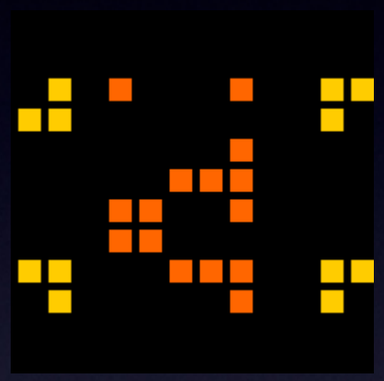
http://www.playframework.org/

http://4store.org/



https://github.com/lobid/

Any Questions ?



Pascal Christoph

hbz christoph@hbz-nrw.de

semweb@hbz-nrw.de

Using a dark background, this presentation saves maybe 70% of energy