INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Indicate Research Pilots
An e-Infrastructure enabled semantic search service

Technical Conference

Catania 20/04/2012

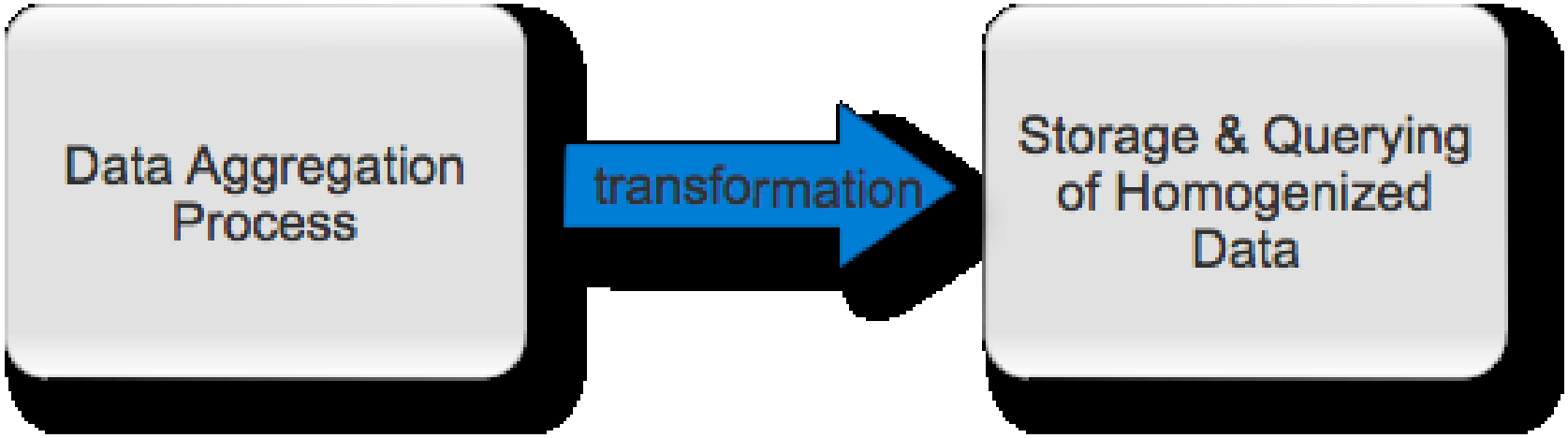
NTUA

Kostas Pardalis

Pilot Objectives

- Establish a search system using MICHAEL data
- Enrich the search system with semantic search capabilities
- Evaluate the feasibility of these requirements using e-infrastructures, presenting the main benefits from this integration

Use Case Scenario



- Adopt a typical but simplified workflow from the digital culture domain consisting of the following steps:
  - Aggregate data
  - Transform data into a common reference schema
  - Data Enrichment
  - Store data into an appropriate semantic repository
  - Semantic search

Implemented Tasks

- Data Manipulation
  - RDFization using a simple data model
  - Semantic Enrichment using DBpedia
- Semantic Repository for data storage
- E-Infrastructures architecture
- Evaluation of the architecture

Data Manipulation - Data Model

- Exploration of data
  - Every xml item represents a collection of digital cultural objects
- Mapping of xml elements to RDF properties for achieving semantic representation of data
  - Language → dcterms:language
  - Digital Format → dcterms:format

Data Manipulation - Enrichment

- Specific values of the examined dataset were discovered as DBpedia resources.
- Additional semantic information is added to the dataset
  - **Countries** : area, capital, density, currency, etc
  - **Languages** : spokenIn, languageFamily, speakers, etc
  - **Famous Persons** : dates of birth death, professions, works, etc

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Enrichment Results

	Total	Found	Percentage
Countries	16429	15987	97.3%
Languages	11090	11032	99.5%
Persons	6442	3632	56.4%

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Semantic Repository for data storage

- Triplestore Evaluation
  - Requirements
    - Distributed
    - Licensing (open source)
    - Sparql language support
    - Web based access
  - Candidates
    - 4store
    - Sesame
    - Bigowlim

Infrastructure Deployment Steps

- Decide about the Cloud platform that is utilized
- Deploy the Semantic Enrichment API
- Deploy the Semantic Repository

E-Infrastructures - Cloud Platform

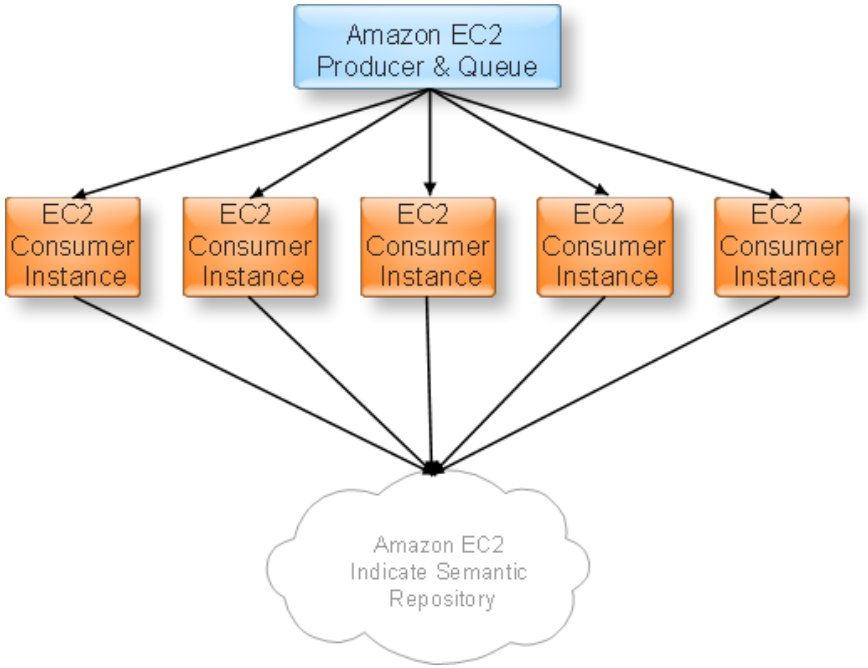
- Amazon EC2 is used as the Cloud environment for deployment.
  - It provides a concrete pricing model for comparisons.
  - It is one of the most technologically mature Cloud environments.

Amazon EC2 Utilized Services

- Amazon Elastic Compute Cloud
  - Large Instance 7.5 GB of memory, 4 EC2 Compute Units (2 virtual cores with 2 EC2 Compute Units each), 850 GB of local instance storage, 64-bit platform, were used to form the Indicate Cluster.
- Elastic IP Addresses
  - Were assigned to each instance to ensure the existence of static IPs
- Amazon Elastic Block Store (EBS)
  - Was used for providing persistence storage to the Indicate Cluster Instances.

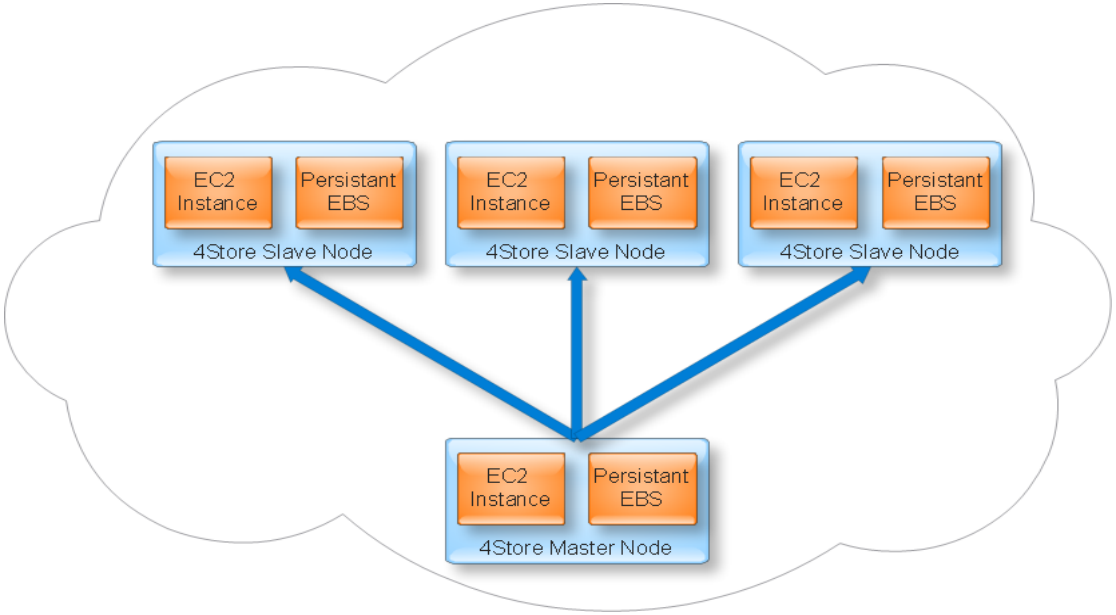
Data Manipulation @ Amazon EC2

- One Amazon EC2 Instance is acting as the producer and hosts the Message Queue (RabbitMQ).
- Five Large Amazon EC2 Instances are hosting the consumers.



Distributed Semantic Repository @ Amazon EC2

- The 4store Distributed Semantic Repository was installed on 4 Large EC2 Instances.
- The number of Nodes attached to the Semantic Repository can be adjusted in order to check scalability and performance.



Evaluation

- Evaluation was performed for rdfization, enrichment and storage (load) tasks using
  - Single thread process on local host
  - Multi-thread process on local cluster (3 nodes)
  - Multi-thread process on Amazon cloud (9 nodes)

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

RabbitMQ Processing

RabbitMQ Management

http://46.137.88.112:55672/#/queues/%2F/publishingQueue

Google

Apple Yahoo! Google Maps YouTube Wikipedia News (234) Popular

RabbitMQ Management | SPARQL httpd server status – size | Tools – Visual Data Web | Amazon EC2 Pricing

RabbitMQ

User: guest

Overview | Connections | Channels | Exchanges | Queues | Users | Virtual Hosts

Queue publishingQueue

Overview

Messages

Ready	Unacknowledged	Total
8331	180	8511

Details

Parameters	durable: true	Consumers	180
Exclusive owner	None	Memory	5.7MB
Status	Idle since 2011-11-28 23:7:7		

Message rates

Incoming

Exchange	publish	confirm
(AMQP default)	0/s	

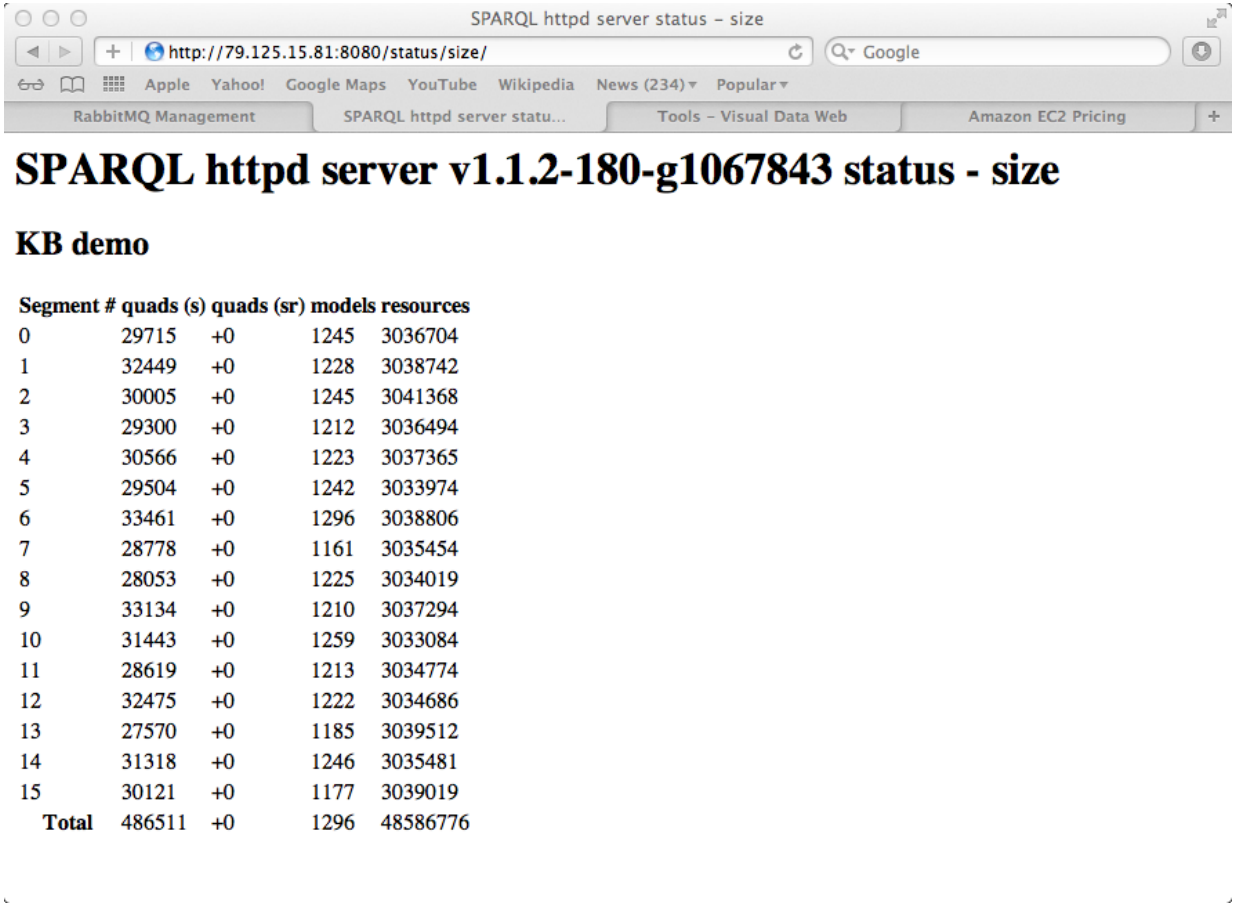
Total: 0/s

Deliveries

Channel	deliver / get	ack
46.137.90.20:49798:1		
46.137.90.20:49806:1		
46.137.90.20:49814:1		

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

4store @ Amazon EC2



SPARQL httpd server v1.1.2-180-g1067843 status - size

KB demo

Segment	# quads (s)	quads (sr)	models	resources
0	29715	+0	1245	3036704
1	32449	+0	1228	3038742
2	30005	+0	1245	3041368
3	29300	+0	1212	3036494
4	30566	+0	1223	3037365
5	29504	+0	1242	3033974
6	33461	+0	1296	3038806
7	28778	+0	1161	3035454
8	28053	+0	1225	3034019
9	33134	+0	1210	3037294
10	31443	+0	1259	3033084
11	28619	+0	1213	3034774
12	32475	+0	1222	3034686
13	27570	+0	1185	3039512
14	31318	+0	1246	3035481
15	30121	+0	1177	3039019
Total	486511	+0	1296	48586776

Results

Method Used	Time in Millisecs
Local Host	22.383.937ms (~6.2hrs)
Local Cluster	5.020.430 (~1.39hrs)
Amazon Cloud	1.422.000 (~23.7 min)

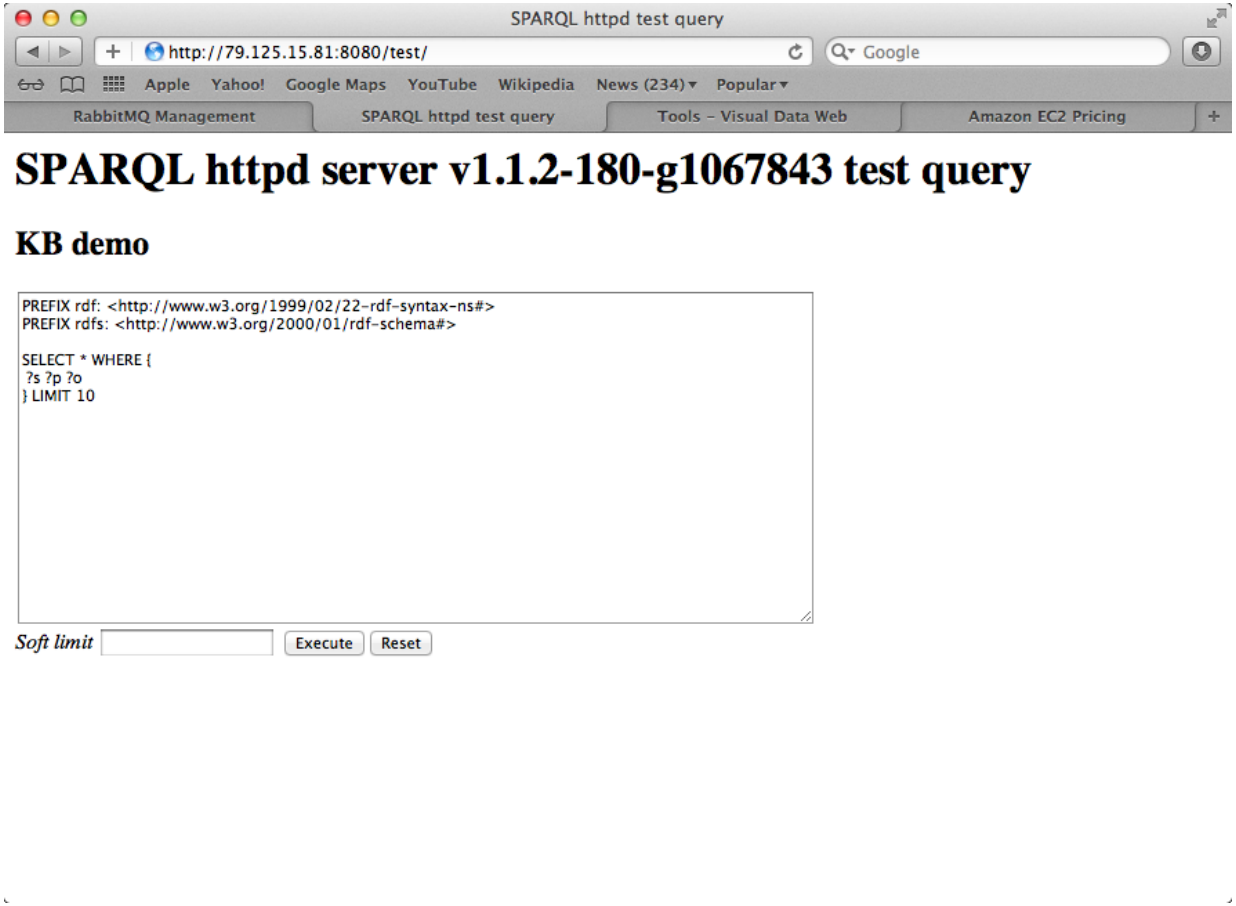
- MICHAEL : 8511 items
- Europeana:~20.000.000 items

Demonstration of Semantic Search

- Querying on data
  - Search for items from a specific country (e.g Greece)
- Semantic Querying using
  - Search for items from a specific country (e.g Greece)
  - Search for items which are hold by Countries of Mediterranean Sea that are about living politicians

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Sparql Endpoint



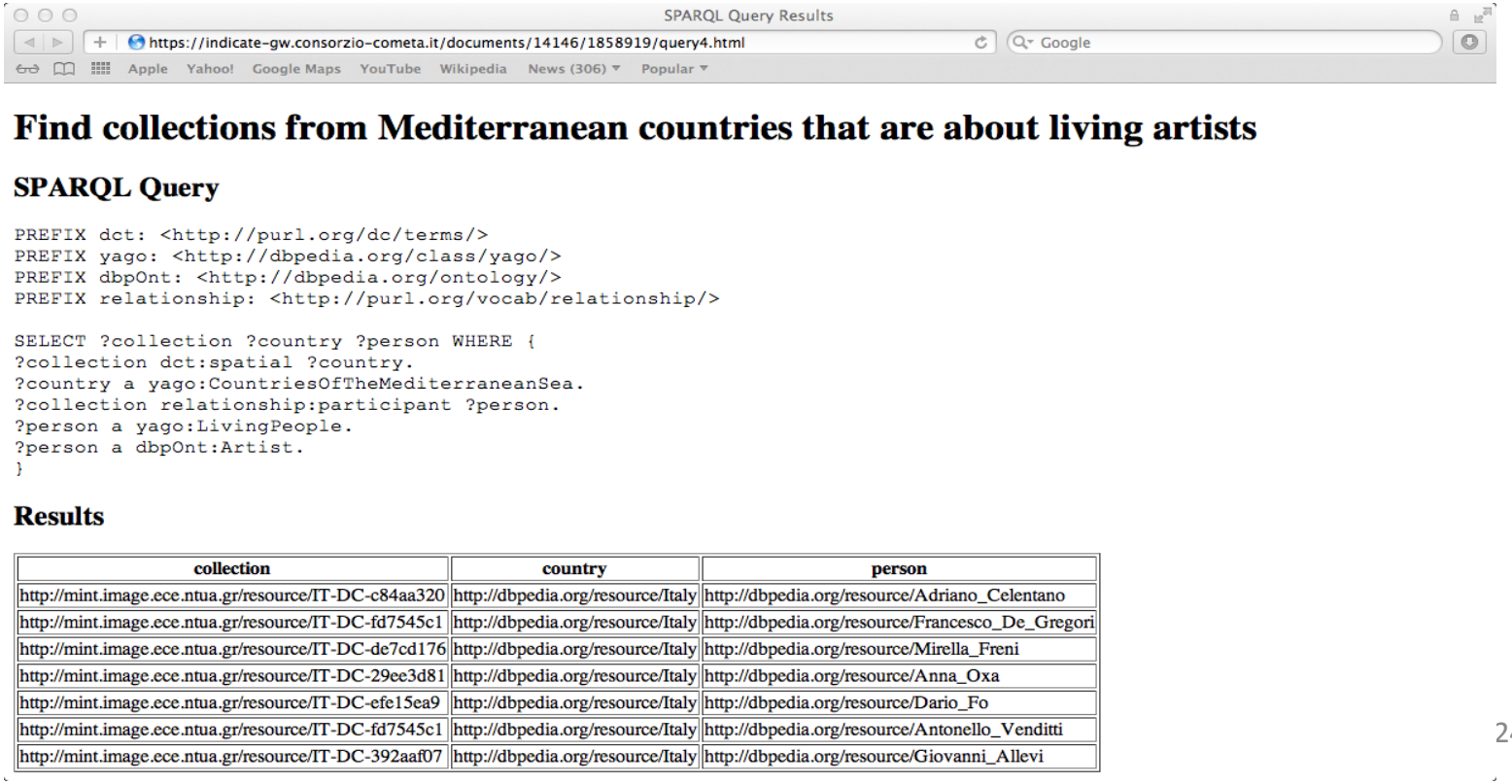
Deployment on ~okeanos

- An IaaS Service.
- Developed by GRNET.
- Aims to deliver production-quality IaaS to the Greek academic and research community.
- Open source (https://docs.dev.grnet.gr/docs/).

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Integration with the Indicate Portal

- http://indicate-gw.consorzio-cometa.it/semantic-search



Find collections from Mediterranean countries that are about living artists

SPARQL Query

```
PREFIX dct: <http://purl.org/dc/terms/>
PREFIX yago: <http://dbpedia.org/class/yago/>
PREFIX dbpOnt: <http://dbpedia.org/ontology/>
PREFIX relationship: <http://purl.org/vocab/relationship/>

SELECT ?collection ?country ?person WHERE {
?collection dct:spatial ?country.
?country a yago:CountriesOfTheMediterraneanSea.
?collection relationship:participant ?person.
?person a yago:LivingPeople.
?person a dbpOnt:Artist.
}
```

Results

collection	country	person
http://mint.image.ece.ntua.gr/resource/IT-DC-c84aa320	http://dbpedia.org/resource/Italy	http://dbpedia.org/resource/Adriano_Celentano
http://mint.image.ece.ntua.gr/resource/IT-DC-fd7545c1	http://dbpedia.org/resource/Italy	http://dbpedia.org/resource/Francesco_De_Gregori
http://mint.image.ece.ntua.gr/resource/IT-DC-de7cd176	http://dbpedia.org/resource/Italy	http://dbpedia.org/resource/Mirella_Freni
http://mint.image.ece.ntua.gr/resource/IT-DC-29ee3d81	http://dbpedia.org/resource/Italy	http://dbpedia.org/resource/Anna_Oxa
http://mint.image.ece.ntua.gr/resource/IT-DC-efe15ea9	http://dbpedia.org/resource/Italy	http://dbpedia.org/resource/Dario_Fo
http://mint.image.ece.ntua.gr/resource/IT-DC-fd7545c1	http://dbpedia.org/resource/Italy	http://dbpedia.org/resource/Antonello_Venditti
http://mint.image.ece.ntua.gr/resource/IT-DC-392aaf07	http://dbpedia.org/resource/Italy	http://dbpedia.org/resource/Giovanni_Allevi

Conclusions

- Semantic Search using e-Infrastructures
  - Provides scalability that is vital for semantic enrichment, since frequent updates are required for remaining consistent.
  - Cost
    - Processing: $ 0.68 per node per hour (~ 1.7 €)
    - Storage: $ 0.11 per Gb per month (~ 4.4 €)

INDICATE
International Network for Digital Cultural Heritage e-Infrastructure

Questions ?