Semantic search

Jérôme Gasperi

WGISS #37
Cocoa Beach, Florida - USA - April 16th, 2014
Semantic search helps users to find the right data.
Semantic search helps users to find the right data
How to add semantics capabilities to EO products search services?
1

Characterize products with relevant information. Think « users », not « experts »
1. Characterize products with relevant information. Think « users », not « experts »

2. Decode users natural language queries
Use footprint to enrich metadata from exogenous data

github.com/jjrom/itag
Tag this footprint with continent, country and Land use

http://goo.gl/WtbcbR
RESTo provides semantic search capabilities. It uses a Query Analyzer to translate query into a set of EO OpenSearch parameters.
Query string analysis algorithm is based on simple recognition of words and patterns.

2. Decode queries

- Split query string into list of unitary words
- Extract «key=value» strings
- Extract Platforms and Instruments
- Remove excluded words and non dictionary words with length < 4 characters
- Extract patterns and dates
- Extract keywords
- Extract location on remaining words

Examples:
- «orbitNumber=4»
- Platforms and instruments list are stored within common dictionary: [https://github.com/jjrom/resto/blob/master/resto/dictionaries/common.php](https://github.com/jjrom/resto/blob/master/resto/dictionaries/common.php)
- «area of Mexico in 2012»
- «acquired in the last 2 days»
- «urban area in France»
- «images acquired in Toulouse»
Recognized patterns

- With "keyword"
- Without "keyword"

- "quantity" <lesser> (than) "numeric" "unit"
- "quantity" <greater> (than) "numeric" "unit"
- "quantity" <equal> (to) "numeric" "unit"
- "lesser" (than) "numeric" "unit" (of) "quantity"
- "greater" (than) "numeric" "unit" (of) "quantity"
- "equal" (to) "numeric" "unit" (of) "quantity"

- "quantity" <between> "numeric" <and> "numeric" ("unit")
- "between" "numeric" <and> "numeric" ("unit")

- <today>
- <yesterday>
- <before> "date"
- <after> "date"
- <between> "date" <and> "date"
- "numeric" 
- "(year|day|month)"
- <ago>
- "(year|day|month)"
- <last>
- "numeric" "(year|day|month)"
- "numeric" <last> "(year|day|month)"
- "(year|day|month)" <last>
- <since> "numeric" "(year|day|month)"
- <since> "month" "year"
- <since> "date"
- <since> "numeric" <last> "(year|day|month)"
- <since> <last> "numeric" "(year|day|month)"
- <since> <last> "(year|day|month)"
- <since> "(year|day|month)" <last>
Words are stored within a dictionary

```php
$dictionary = array(
    'excluded' => array(
        'than',
        'image',
        ...
    ),
    'modifiers' => array(
        'ago' => 'ago',
        'before' => 'before',
        'after' => 'after',
        ...
    ),
    'units' => array(
        'm' => 'm',
        'meter' => 'm',
        'days' => 'days',
        ...
    ),
    'numbers' => array(
        'one' => '1',
        ...
    ),
    'months' => array(
        'january' => '01',
        ...
    ),
    'quantities' => array(
        'resolution' => 'resolution',
        ...
    ),
    'keywords' => array(
        'continent' => array(
            'europe' => 'europe',
            ...
        )
    )
);```
Example

« Images of urban area in the US acquired in the last 10 days with less than 5% of cloud cover »
Decode queries

Example

« Images of urban area in the US acquired in the last 10 days with less than 5% cloud cover »

Example keyword location date acquisition parameter
2. Each search result has a « human readable url » that can be indexed by web crawler (i.e. google robots).

1. Search parameters are derived from Natural Language query.

3. Keywords on resources are links to search requests: they can be indexed by web crawler...and so on.
1. Search parameters are derived from a Natural Language query.

2. Each search result has a "human readable url" that can be indexed by a web crawler (i.e., Google robots).

3. Keywords on resources are links to search requests: they can be indexed by a web crawler and so on.

http://goo.gl/GvMEHj
Issues with keywords approach
Earthquakes in November 2008 in China
Earthquakes in November 2008 in China

Ambiguous since it appears to be a location in New Zealand.
« Linked data is the right way to do Semantic Web »

Tim Berners-Lee
Update REST to JSON model to follow JSON-LD format

```json
{
    "@context": "http://json-ld.org/contexts/person.jsonld",
    "@id": "http://dbpedia.org/resource/John_Lennon",
    "name": "John Lennon",
    "born": "1940-10-09",
    "spouse": "http://dbpedia.org/resource/Cynthia_Lennon"
}
```