JSONpedia

Facilitating consumption of MediaWiki content

Michele Mostarda <mostarda@fbk.eu>, Twitter: @micmos

v1.2

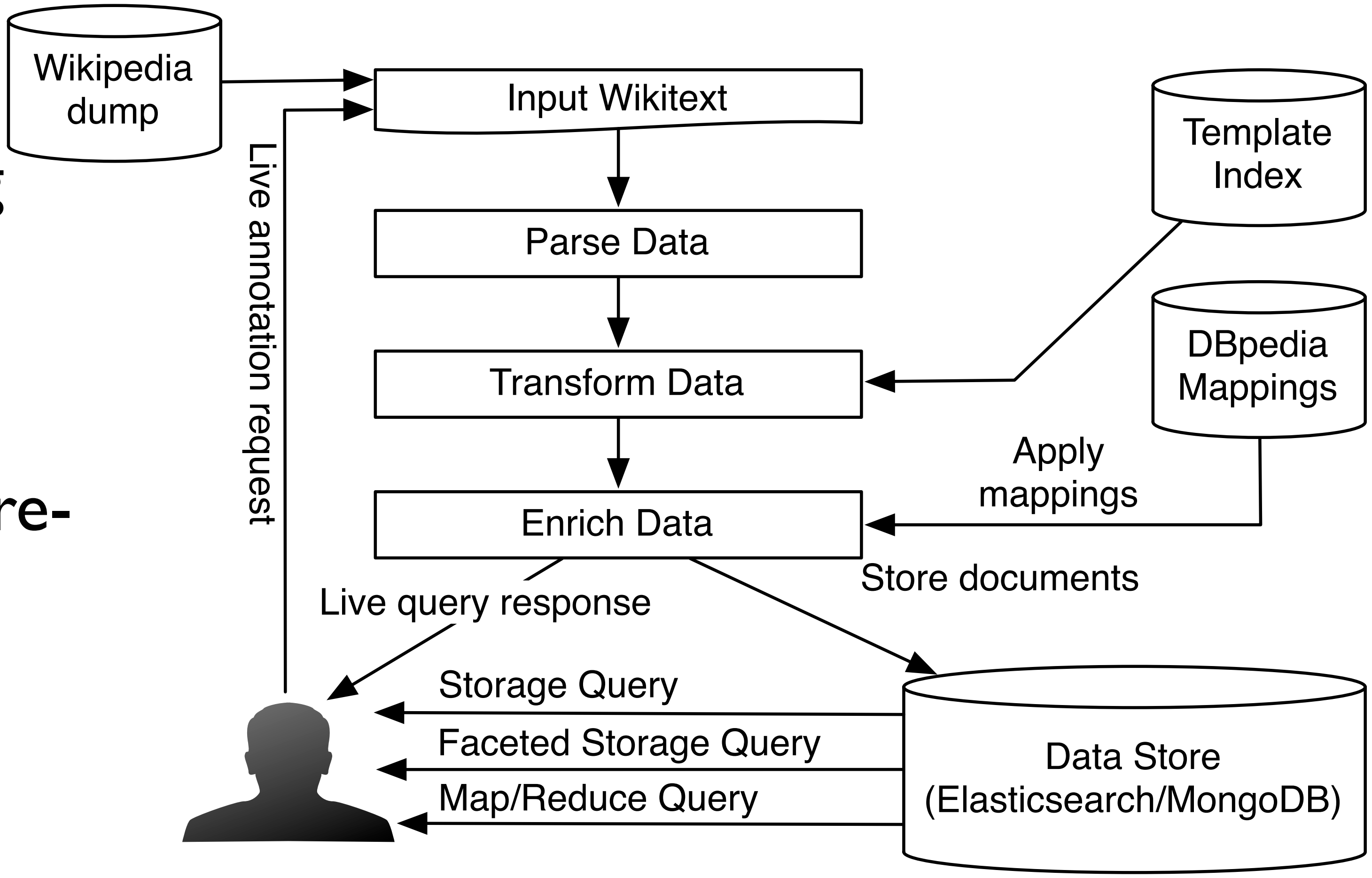
Outline

- What is JSONpedia
- How does it work
- Main features
- Online demo
- Web App
- REST API
- jQuery plugin
- Code snippets
- Internals
- GSoC 2014
- History & previous work
- Forthcoming features
- Next release
- Online resources
- Support up
- Acknowledgements

What is JSONpedia

JSONpedia is a Java library and a REST service meant to read MediaWiki pages as JSON.

How does it work



- A user can perform a **live annotation** requests providing **Wikitext** or a reference to a Wikipedia page.
- A user can perform a **storage query** over the data storage pre-populated with the Wikipedia dump.
- A user can perform a **faceted storage query** over the data storage pre-populated with the Wikipedia dump.

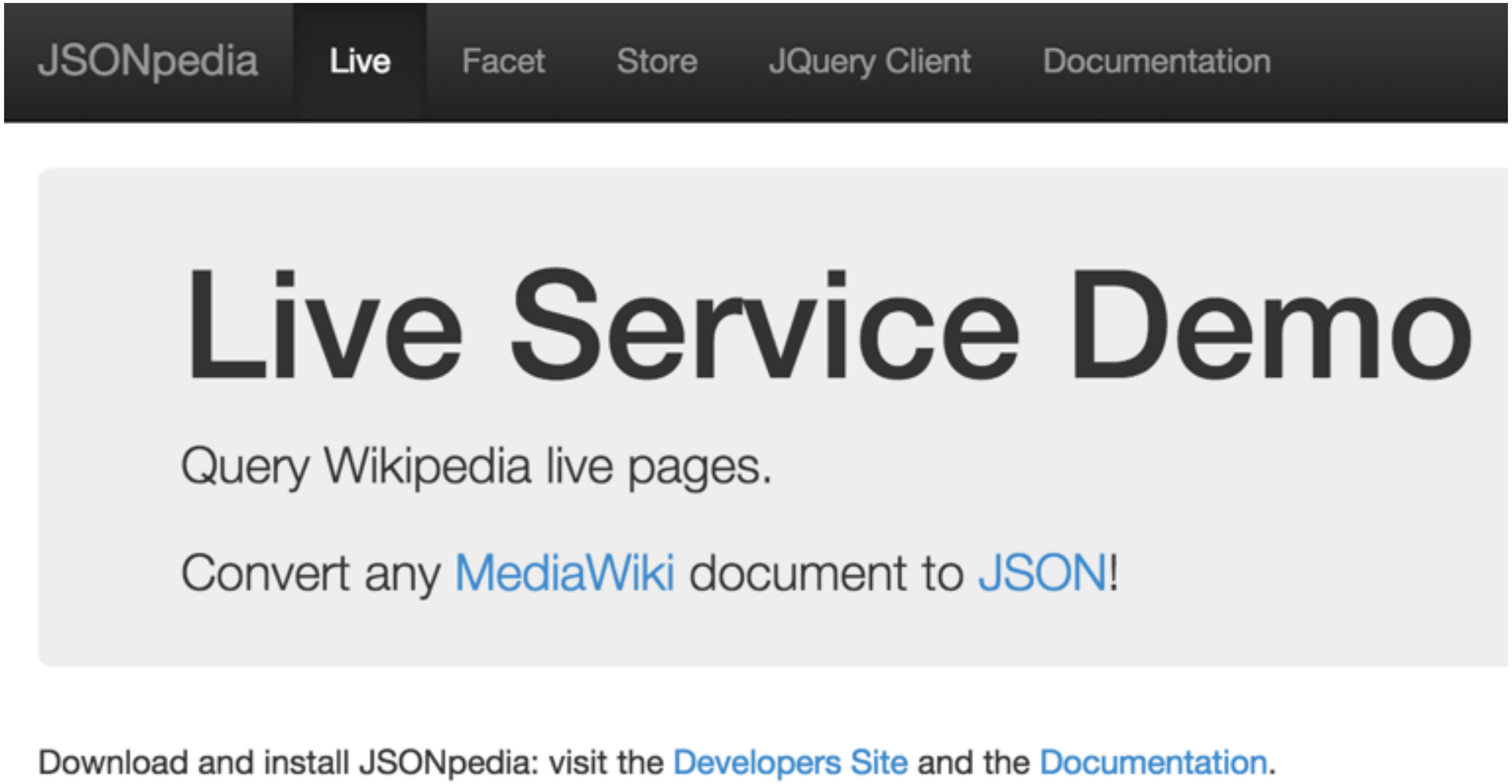
- A user can perform a **faceted storage query** over the data storage pre-populated with the Wikipedia dump.

- A user can perform a **map/reduce storage query** over the data storage pre-populated with the Wikipedia dump.
- Any provided Wikitext is **parsed** (Parse Data), **templates are expanded** and new **metadata** is generated (Transform Data), **external data sources** are linked (Enrich Data), the final model is **converted in JSON** and stored into the Data Store.

Main features

- WikiText event-based parser
- Configurable page processing pipeline
- Wikimedia template processing support
- DBpedia mapping integration
- RESTful interface
- MongoDB storage and map/reduce support
- Elasticsearch query support
- Elasticsearch faceting support
- Web frontend
- HTML data rendering
- CLI interface

Online Demo



Live Service Demo

Query Wikipedia live pages.

Convert any MediaWiki document to JSON!

Download and install JSONpedia: visit the Developers Site and the Documentation.

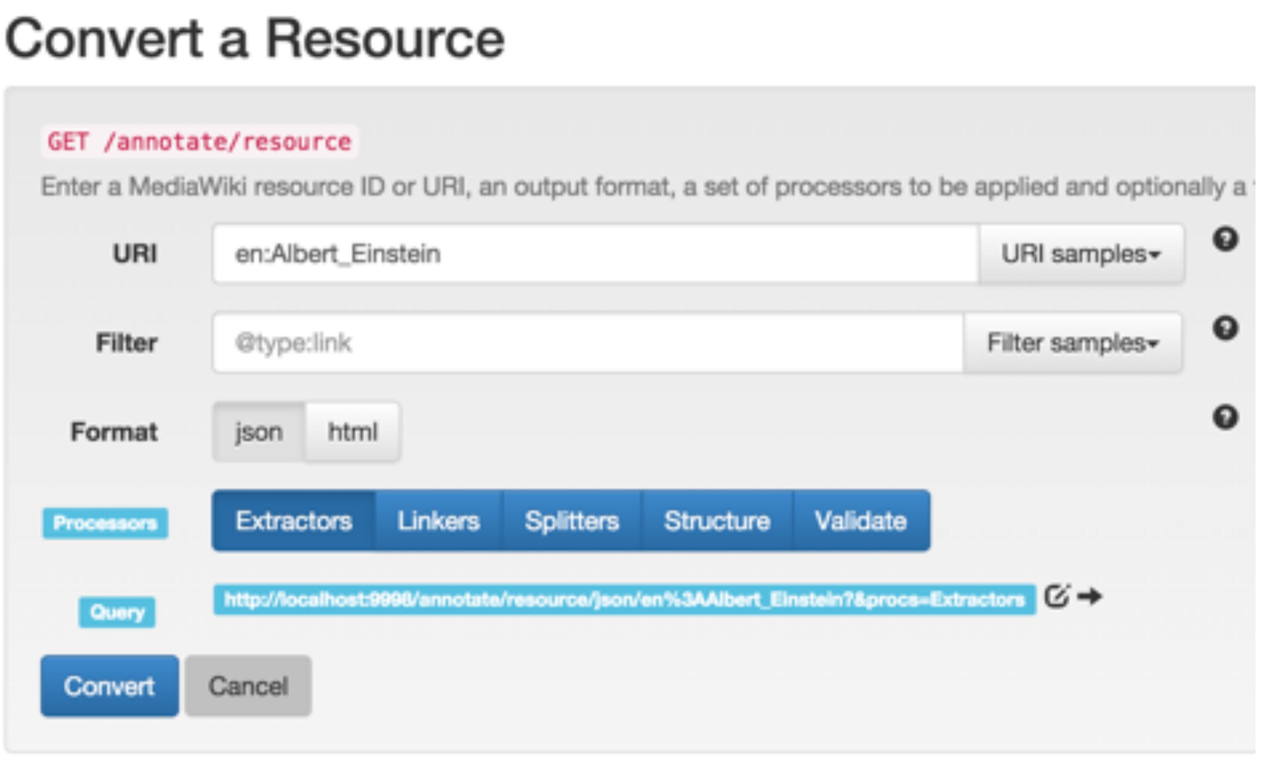
The official JSONpedia online demo is available at http://jsonpedia.org

Web App

The JSONpedia web app allows to experiment with the REST service through a comfortable UX

Live Panel

Analyze any MediaWiki page live or
directly copy/paste WikiText



Convert a Resource

`GET /annotate/resource`

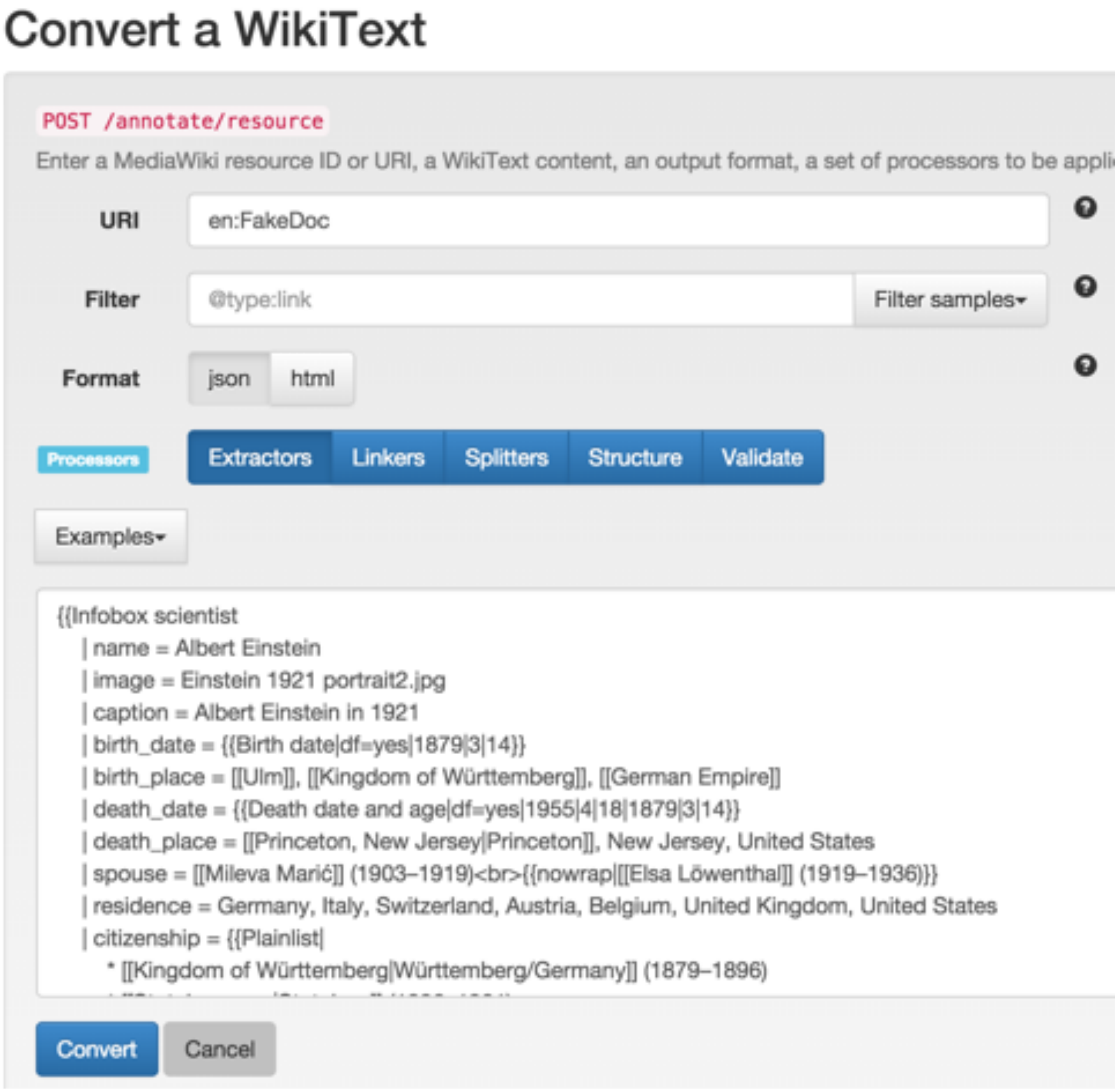
Enter a MediaWiki resource ID or URI, an output format, a set of processors to be applied and optionally a

URI: en:Albert_Einstein

Filter: @type:link

Format: json html

Query: http://localhost:9998/annotate/resource/json/en%3AAlbert_Einstein?&procs=Extractors



Convert a WikiText

`POST /annotate/resource`

Enter a MediaWiki resource ID or URI, a WikiText content, an output format, a set of processors to be appli

URI: en:FakeDoc

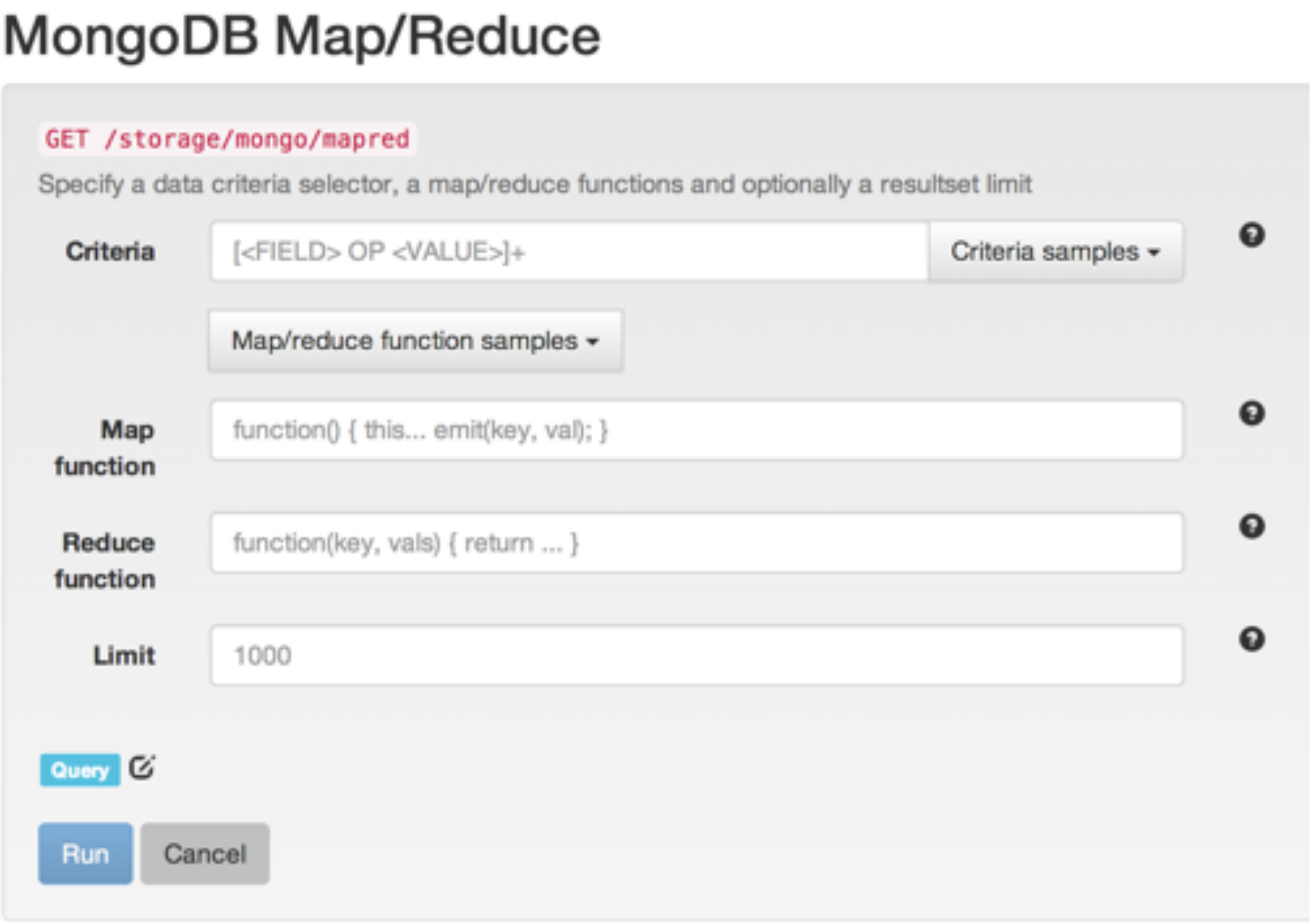
Filter: @type:link

Format: json html

```
{{Infobox scientist
   | name = Albert Einstein
   | image = Einstein 1921 portrait2.jpg
   | caption = Albert Einstein in 1921
   | birth_date = {{Birth date|df=yes|1879|3|14}}
   | birth_place = [[Ulm]], [[Kingdom of Württemberg]], [[German Empire]]
   | death_date = {{Death date and age|df=yes|1955|4|18|1879|3|14}}
   | death_place = [[Princeton, New Jersey|Princeton]], New Jersey, United States
   | spouse = [[Mileva Marić]] (1903–1919)<br>{{nowrap|[[Elsa Löwenthal]] (1919–1936)}}
   | residence = Germany, Italy, Switzerland, Austria, Belgium, United Kingdom, United States
   | citizenship = {{Plainlist|
     * [[Kingdom of Württemberg|Württemberg/Germany]] (1879–1896)
```

Query panel: MongoDB

Query MediaWiki pages stored in MongoDB



MongoDB Map/Reduce

`GET /storage/mongo/mapred`

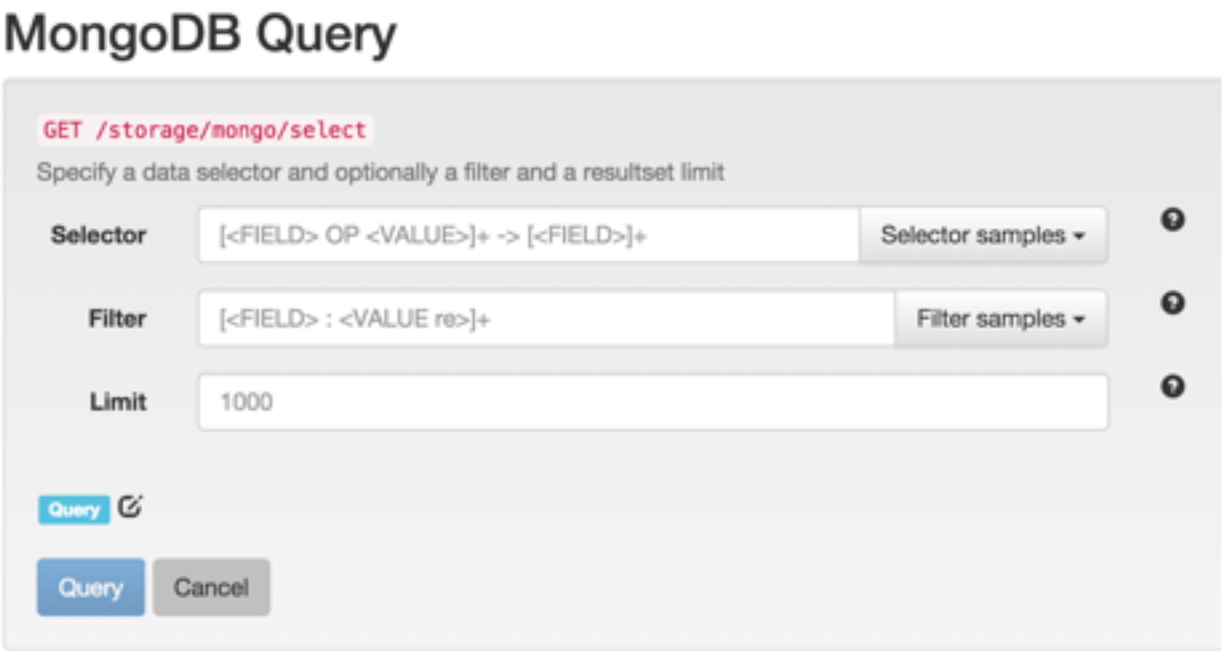
Specify a data criteria selector, a map/reduce functions and optionally a resultset limit

Criteria [<FIELD> OP <VALUE>]+

Map function function() { this... emit(key, val); }

Reduce function function(key, vals) { return ... }

Limit 1000



MongoDB Query

`GET /storage/mongo/select`

Specify a data selector and optionally a filter and a resultset limit

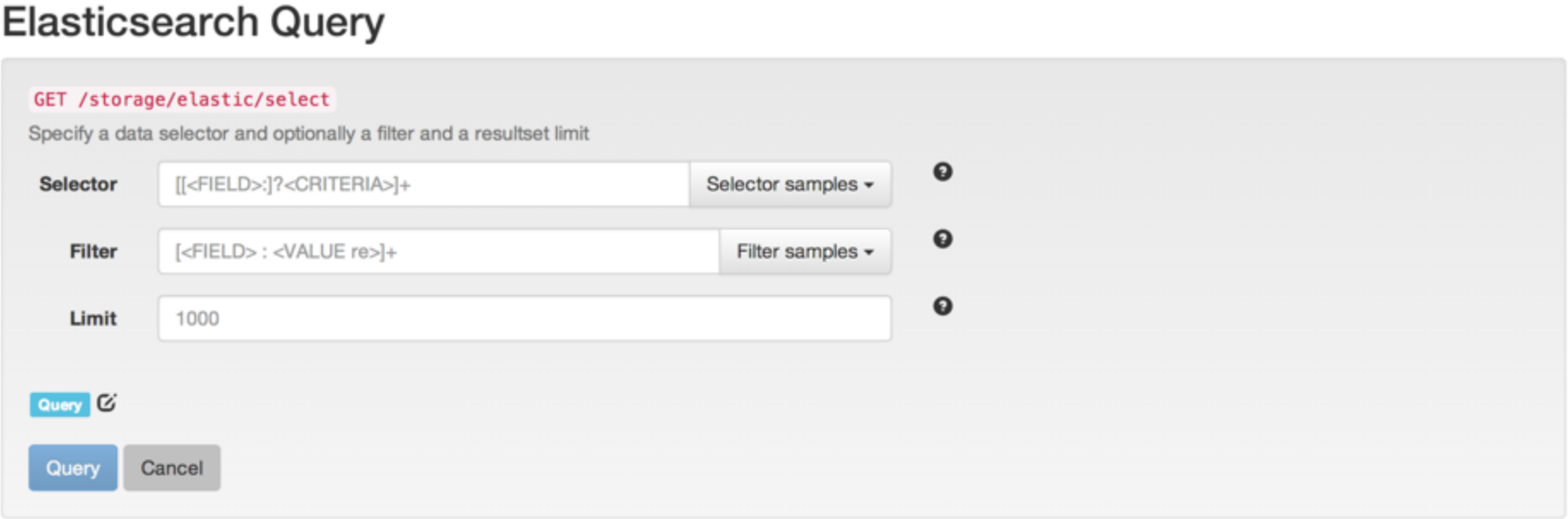
Selector [<FIELD> OP <VALUE>]+ -> [<FIELD>]+

Filter [<FIELD> : <VALUE re>]+

Limit 1000

Query panel: Elasticsearch

Query the latest Wikipedia dump with **Elasticsearch**



Elasticsearch Query

`GET /storage/elastic/select`

Specify a data selector and optionally a filter and a resultset limit

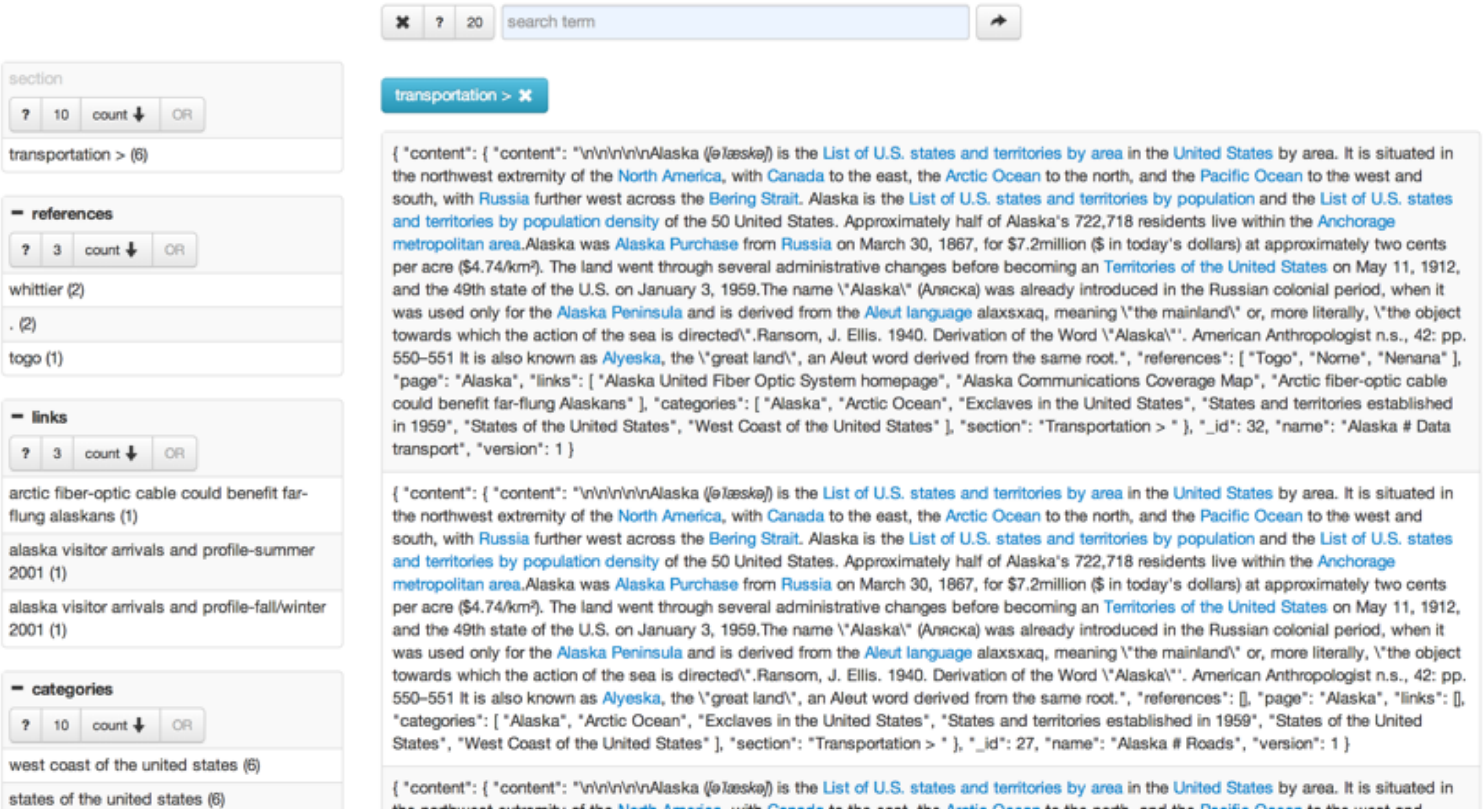
Selector [[<FIELD>:]?<CRITERIA>]+

Filter [<FIELD> : <VALUE re>]+

Limit 1000

Query panel: Elasticsearch

Explore the latest Wikipedia dump with
Elasticsearch *FacetView*



transportation > (6)

whittier (2)

. (2)

togo (1)

arctic fiber-optic cable could benefit far-flung alaskans (1)

alaska visitor arrivals and profile-summer 2001 (1)

alaska visitor arrivals and profile-fall/winter 2001 (1)

west coast of the united states (6)

states of the united states (6)

{ "content": { "content": "\n\n\n\n\nAlaska (/əˈlæskə/) is the List of U.S. states and territories by area in the United States by area. It is situated in the northwest extremity of the North America, with Canada to the east, the Arctic Ocean to the north, and the Pacific Ocean to the west and south, with Russia further west across the Bering Strait. Alaska is the List of U.S. states and territories by population and the List of U.S. states and territories by population density of the 50 United States. Approximately half of Alaska's 722,718 residents live within the Anchorage metropolitan area.Alaska was Alaska Purchase from Russia on March 30, 1867, for $7.2million ($ in today's dollars) at approximately two cents per acre ($4.74/km²). The land went through several administrative changes before becoming an Territories of the United States on May 11, 1912, and the 49th state of the U.S. on January 3, 1959.The name \"Alaska\" (Аляска) was already introduced in the Russian colonial period, when it was used only for the Alaska Peninsula and is derived from the Aleut language alaxsxaq, meaning \"the mainland\" or, more literally, \"the object towards which the action of the sea is directed\".Ransom, J. Ellis. 1940. Derivation of the Word \"Alaska\"'. American Anthropologist n.s., 42: pp. 550–551 It is also known as Alyeska, the \"great land\", an Aleut word derived from the same root.", "references": ["Togo", "Nome", "Nenana"], "page": "Alaska", "links": ["Alaska United Fiber Optic System homepage", "Alaska Communications Coverage Map", "Arctic fiber-optic cable could benefit far-flung Alaskans"], "categories": ["Alaska", "Arctic Ocean", "Exclaves in the United States", "States and territories established in 1959", "States of the United States", "West Coast of the United States"], "section": "Transportation > " }, "_id": 32, "name": "Alaska # Data transport", "version": 1 }

{ "content": { "content": "\n\n\n\n\nAlaska (/əˈlæskə/) is the List of U.S. states and territories by area in the United States by area. It is situated in the northwest extremity of the North America, with Canada to the east, the Arctic Ocean to the north, and the Pacific Ocean to the west and south, with Russia further west across the Bering Strait. Alaska is the List of U.S. states and territories by population and the List of U.S. states and territories by population density of the 50 United States. Approximately half of Alaska's 722,718 residents live within the Anchorage metropolitan area.Alaska was Alaska Purchase from Russia on March 30, 1867, for $7.2million ($ in today's dollars) at approximately two cents per acre ($4.74/km²). The land went through several administrative changes before becoming an Territories of the United States on May 11, 1912, and the 49th state of the U.S. on January 3, 1959.The name \"Alaska\" (Аляска) was already introduced in the Russian colonial period, when it was used only for the Alaska Peninsula and is derived from the Aleut language alaxsxaq, meaning \"the mainland\" or, more literally, \"the object towards which the action of the sea is directed\".Ransom, J. Ellis. 1940. Derivation of the Word \"Alaska\"'. American Anthropologist n.s., 42: pp. 550–551 It is also known as Alyeska, the \"great land\", an Aleut word derived from the same root.", "references": [], "page": "Alaska", "links": [], "categories": ["Alaska", "Arctic Ocean", "Exclaves in the United States", "States and territories established in 1959", "States of the United States", "West Coast of the United States"], "section": "Transportation > " }, "_id": 27, "name": "Alaska # Roads", "version": 1 }

{ "content": { "content": "\n\n\n\n\nAlaska (/əˈlæskə/) is the List of U.S. states and territories by area in the United States by area. It is situated in

REST API

```
GET /annotate/resource/{json|html}/{res-id|res-url}
```

Process a live WikiMedia resource

```
POST /annotate/resource
```

(wikitext, format, processors, filter)
Process arbitrary WikiText markup

```
GET /storage/mongo/select
?q=<query>&filter=<filter>&limit=<limit>
```

Query the Wikipedia dump with MongoDB

```
GET /storage/mongo/mapred
?map=<map-func>&red=<red-func>&criteria=<criteria-exp>&limit=<limit>
```

Query the Wikipedia dump with MongoDB Map / Reduce

```
GET /storage/elastic/select
?q=<query>&filter=<filter>&limit=<limit>
```

Query the Wikipedia dump with Elasticsearch

jQuery Plugin

JSONpedia comes with a jQuery 1.8 plugin providing facilitated access to the REST service.

http://jsonpedia.org/frontend/js/jsonpedia.js

Code Snippets

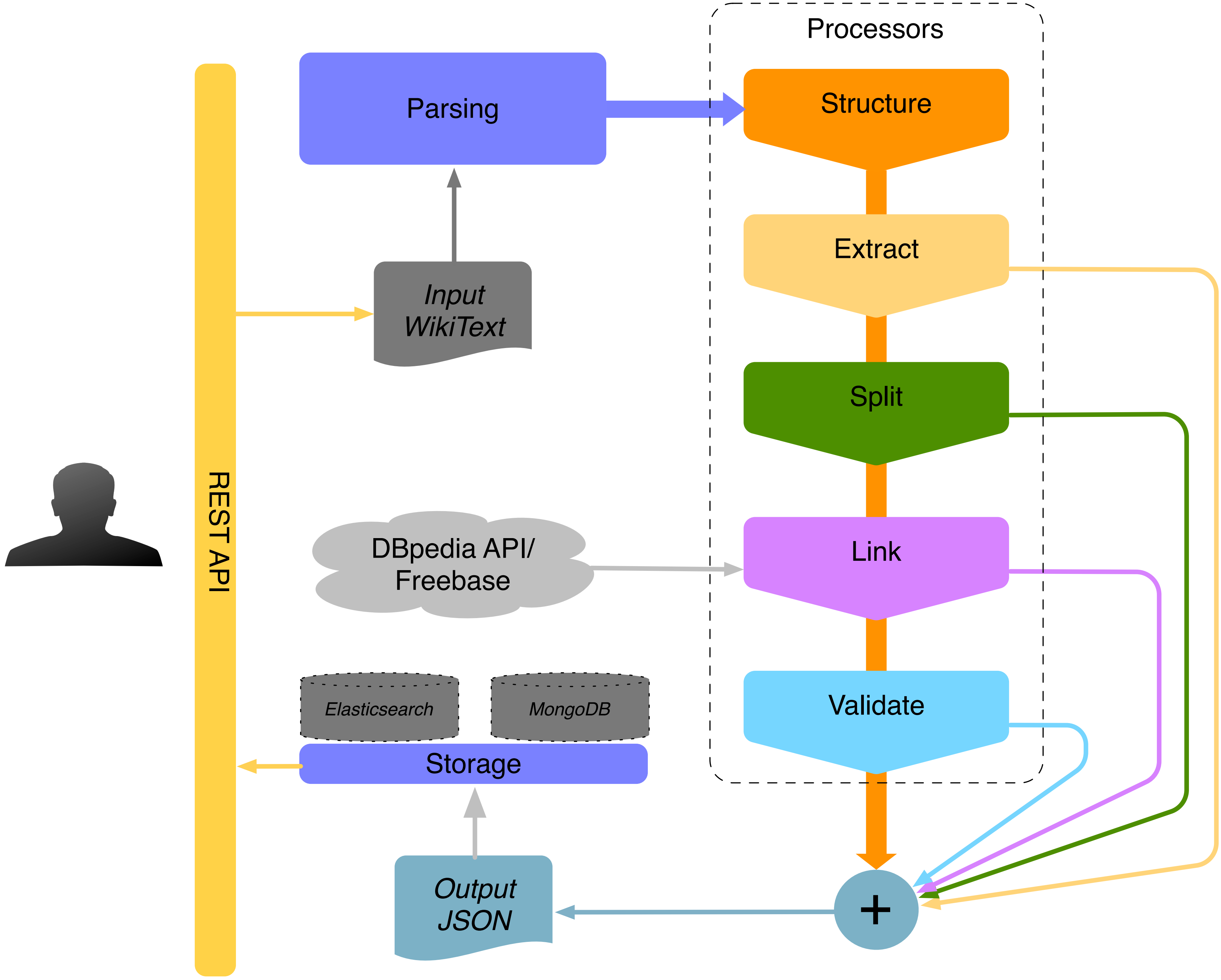
Example:

retrieve content of page London from English Wikipedia, extract the DOM structure, filter nodes of type "section", get first of them and render as HTML.

```
import com.machinelinking.main.JSONpedia;
import org.codehaus.jackson.JsonNode;

JSONpedia jsonpedia = JSONpedia.instance();
JsonNode root = jsonpedia.process("en:London").flags("Structure").json();

JsonNode[] sections = jsonpedia.applyFilter("@type:section", root);
String firstSectionHTML = jsonpedia.render("en:London", sections[0]);
```

Internals

Processing Pipeline



This picture shows the processing pipeline implemented in JSONpedia

Types of Processor

A Processor receives a stream of events generated by parser and perform data enrichment and transformation.

- Structure
- Extractors
- Linkers
- Splitters
- Validator

Structure

The *Structure* Processor receives a stream of WikiText parsing events and builds a 1-1 JSON representation of the document DOM.

Extract

Extractors are specific Processors that collect a certain type of data from the event stream.

For example the SectionsExtractor collects a list of all sections declared in the document stream

Split

A *Splitter* is a Processor cutting sub-trees of the JSON document built by the Structure processor.

An example of Splitter is the TableSplitter which collects the JSON nodes representing all tables found in document.

Link

A *Linker* is a Processor which links the detected document entities to other information acquired from external sources.

An example of Linker is the FreebaseLinker which connects an entity to the same representation in Freebase if any.

Validate

A *Validator* is a Processor performing the check of data structures parsed from a document.

WikiText event based parser messages

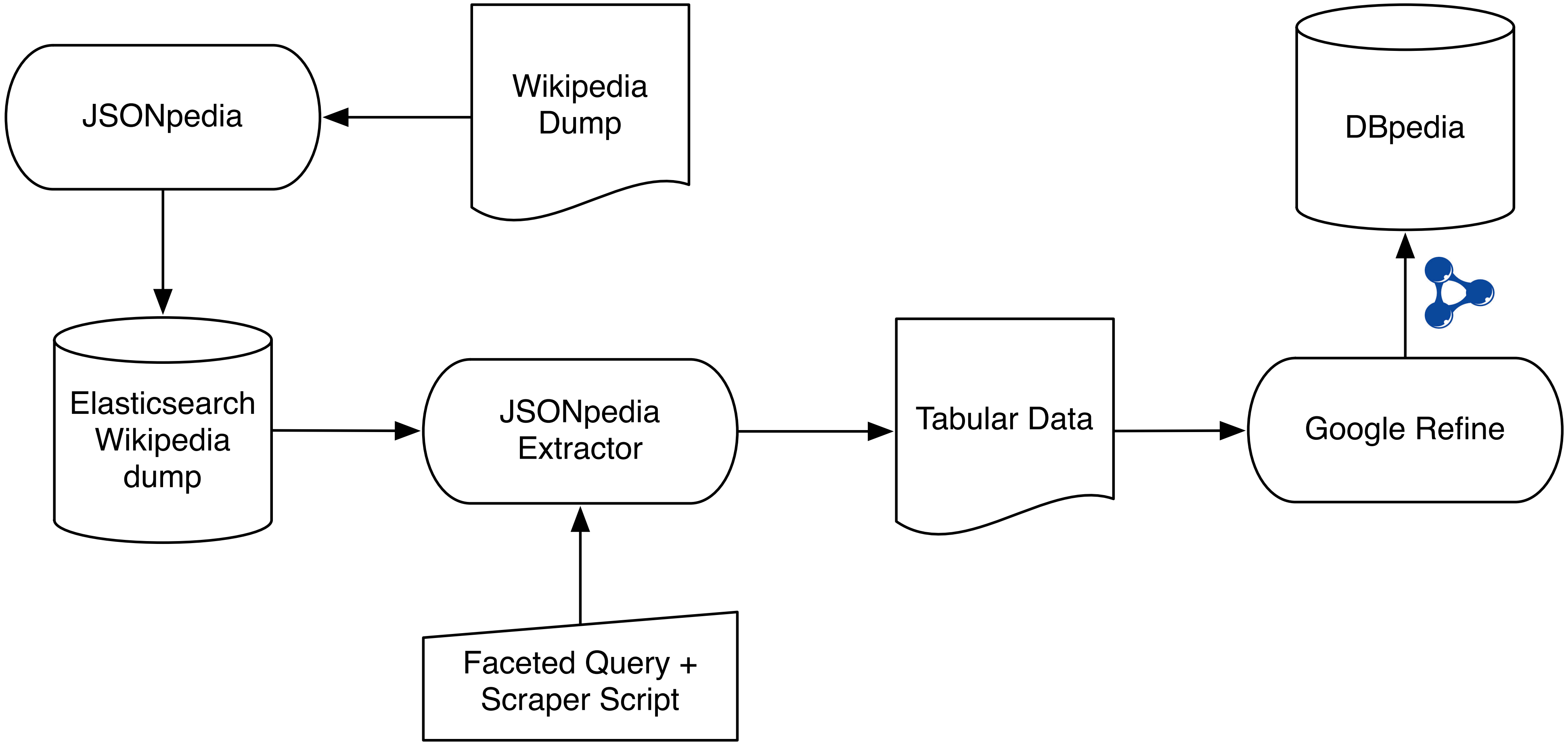
```
// Document bounding.
void beginDocument(URL document);
void endDocument();

// Error handling.
void parseWarning(String msg,
ParserLocation location);
void parseError(Exception e,
ParserLocation location);

// Tag handling.
void beginTag(String node, Attribute[]
attributes);
void endTag(String node);
void inlineTag(String node,
Attribute[] attributes);
void commentTag(String comment);

// Sections
void section(String title, int level);

// References
void beginReference(String label);
void endReference(String label);
```

```
// Links
void beginLink(String url);
void endLink(String url);

// lists
void beginList();
void listItem();
void endList();

// Templates
void beginTemplate(String name);
void endTemplate(String name);

// Tables
void beginTable();
void headCell(int row, int col);
void bodyCell(int row, int col);
void endTable();

// Generic parameter
void parameter(String param);
// Plain text
void text(String content);
```

JSONpedia @Google Summer of Code 2014

Project:
JSONpedia Extractor
Organization:
DBpedia & DBpedia Spotlight
Student:
Roberto Bampi
Mentor:
Michele Mostarda
Description:
Create a general infrastructure to create DBpedia extractors based on JSONpedia.
Public Repo:
https://github.com/dbpedia/jsonpedia-extractor/

The JSONpedia extractor for DBpedia relies on a Wikipedia dump processed with JSONpedia and stored in Elasticsearch, and allows to build scriptable data scrapers based on faceted queries.



Extraction Samples

Discography
Extract artist, album, year and reference for all discographies defined in Wikipedia

Painter works
Extract painter, work, year and link for any paining defined in Wikipedia

Public Gardens
Extract city, garden, description for any public garden defined in Wikipedia

Forthcoming Features

- JSONpedia dumps will be available for download.
- RDF output.
- Online data model Exporter Tool (CSV).

Follow the updates here: https://bitbucket.org/hardest/jsonpedia/issues

History & Previous Work

History

- Initially conceived as a tool to generate machine learning datasets.
- The REST service,inspired by Sweeble Crystalball,produces JSON and a browsable HTML data.
- Written over a context-dependent event based parser to be more performant than a regex matcher (like the WikiParser) or a DOM based parser (like Sweeble).

Differences with DBpedia

- JSONpedia produces JSON, DBpedia RDF.
- JSONpedia includes all the structural elements of a page: links, references, lists, sections, template, tables, XML markup.
- JSONpedia produces low-refined data which requires further processing to be consumed, DBpedia produces ready to use high quality data.
- JSONpedia is a not competitor of DBpedia but rather a complement.

Differences with Sweeble

- Lightweight Event based parser vs DOM parser.
- More tolerant to frequent syntax errors present within WikiText pages.
- Serializes to JSON output which is easier to consume!

Next Release

End of March 2015 v1.2

Online resources

live demo:

http://jsonpedia.org/

source code:

https://bitbucket.org/hardest/jsonpedia

Acknowledgements

DBpedia Association for supporting JSONpedia in GSoC 2014 and 2015.

SpazioDati for hosting the JSONpedia online demo.

Fondazione Bruno Kessler for sponsoring part of the development effort.

Marco Fossati - FBK WeD, PhD student, DBpedia community member.

@hjfocs

Roberto Bampi - SpazioDati, Backend Developer, JSONpedia contributor and student in GSoC 2014.

@BampiRoberto

Thanks for reading!

Michele Mostarda <mostarda@fbk.eu>, Twitter: @micmos