Semantic Tags for Open Data Portals: Metadata Enhancements for Searchable Open Data

Alan Freihof Tygel

D.Sc. Thesis Defence

PPGI/UFRJ

Supervisor: Maria Luiza Machado Campos

Co-supervisor: Sören Auer (University of Bonn/Germany)

Rio de Janeiro, 21/07/2016

Agenda

Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

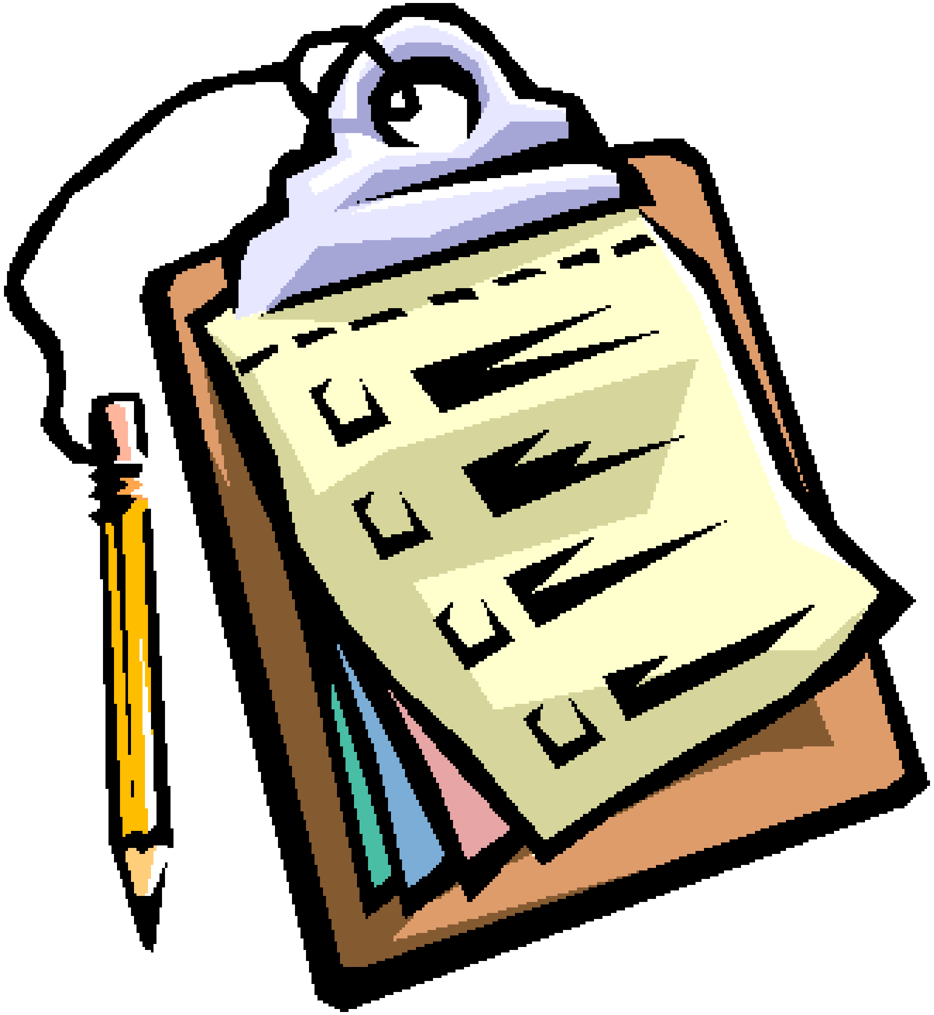
Specific Literature Review

Analysis of Situation

Solution Approach

Evaluation

Conclusions



Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

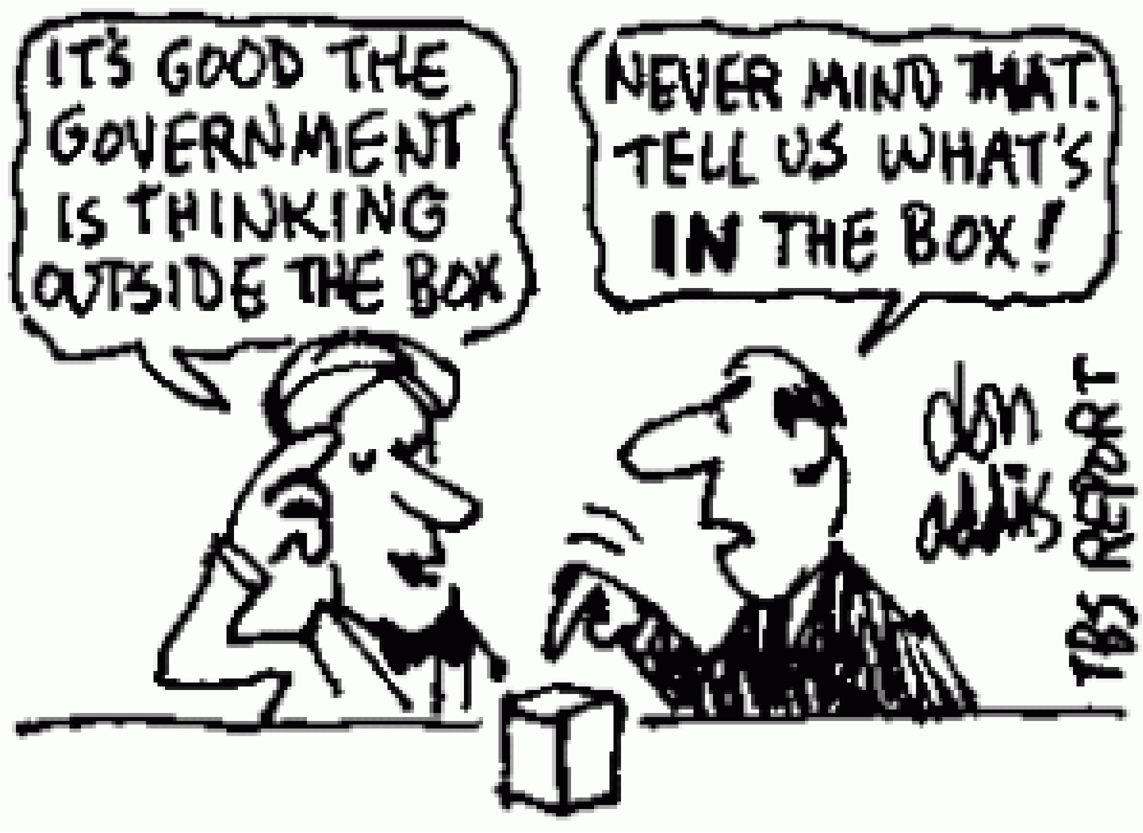
Specific Literature Review

Analysis of Situation

Solution Approach

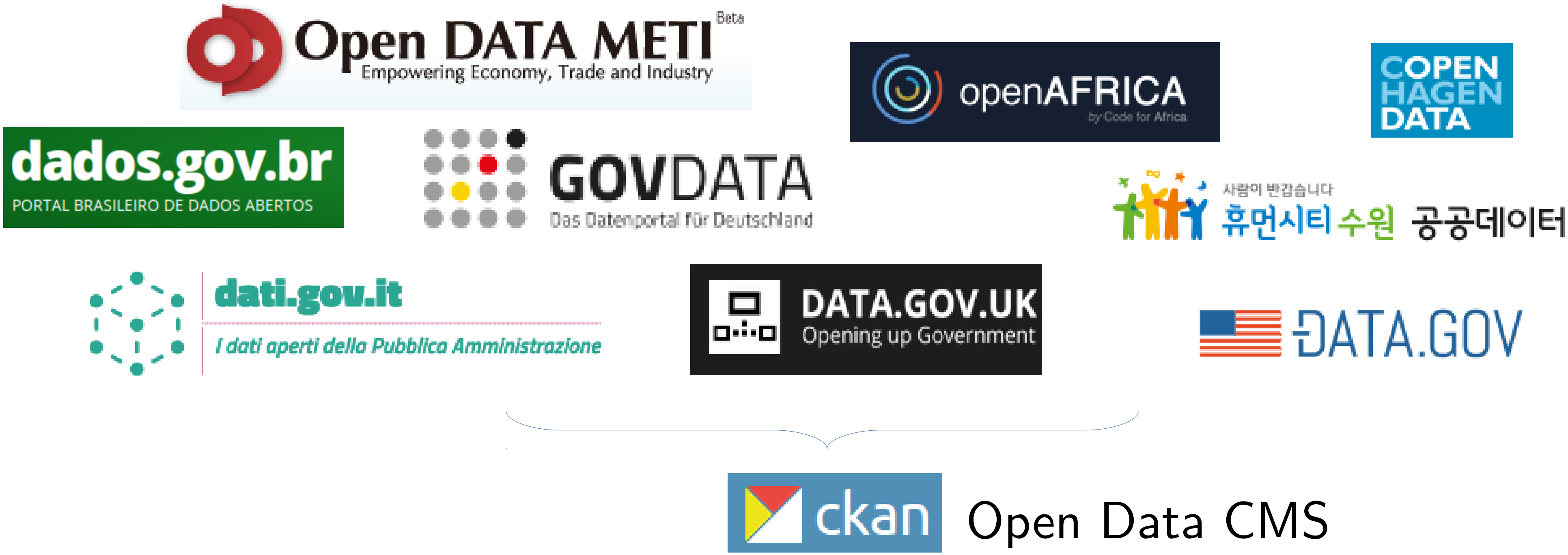
Evaluation

Conclusions


IT'S GOOD THE GOVERNMENT IS THINKING OUTSIDE THE BOX
NEVER MIND THAT. TELL US WHAT'S IN THE BOX!
don addis
TBS REPORT

Motivation - Open Data

A worldwide movement!



Open Data CMS

> Open Source Software + Freedom of Information

Motivation - Open Data

What is open data?

Published on the Web

Machine Readable

Open License

Why open data?

Transparency | Participation | Value creation

Motivation - Open Data Challenges

Some authors dedicated to **open data critique**:

ZUIDERWIJK et al., 2012; ZUIDERWIJK; JANSSEN, 2014a; GURSTEIN, 2011; BATES, 2014; ROSEIRA, 2016; PARYCEK; SCHÖLLHAMMER; SCHOSSBÖCK, 2016; DAVIES; BAWA, 2012

Challenges can be divided into:

>> **Problems**: caused implementation difficulties

>> **Perils**: risks caused by the correct implementation

Motivation - Open Data Challenges

Problems: availability and access, find ability, usability, understand ability, quality, linking and combining data, comparability and compatibility and metadata (ZUIDERWIJK et al., 2012)

Perils: creating/enlarging a "data divide", setting limits between public and private data, open data versus political interests, ... (GURSTEIN, 2011; ZUIDERWIJK; JANSSEN, 2014a;)

Motivation - Open Data Challenges

Among these critiques, **access to data** is in the root of several challenges:

Data that is **not adequately described** can hardly be found and used

+

Inequalities in data skills results that only specific groups can take advantage of accessing data.

Objective and Hypothesis

Hypothesis:

Cleaning up, reconciling and enriching metadata leads to a *higher searchability* of open datasets.

Objective:

To develop an approach to *enhance the description* of open datasets, with the perspective of *facilitating access to open data*, and consequently improving the *realisation of its benefits* in democratic way.

Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

Specific Literature Review

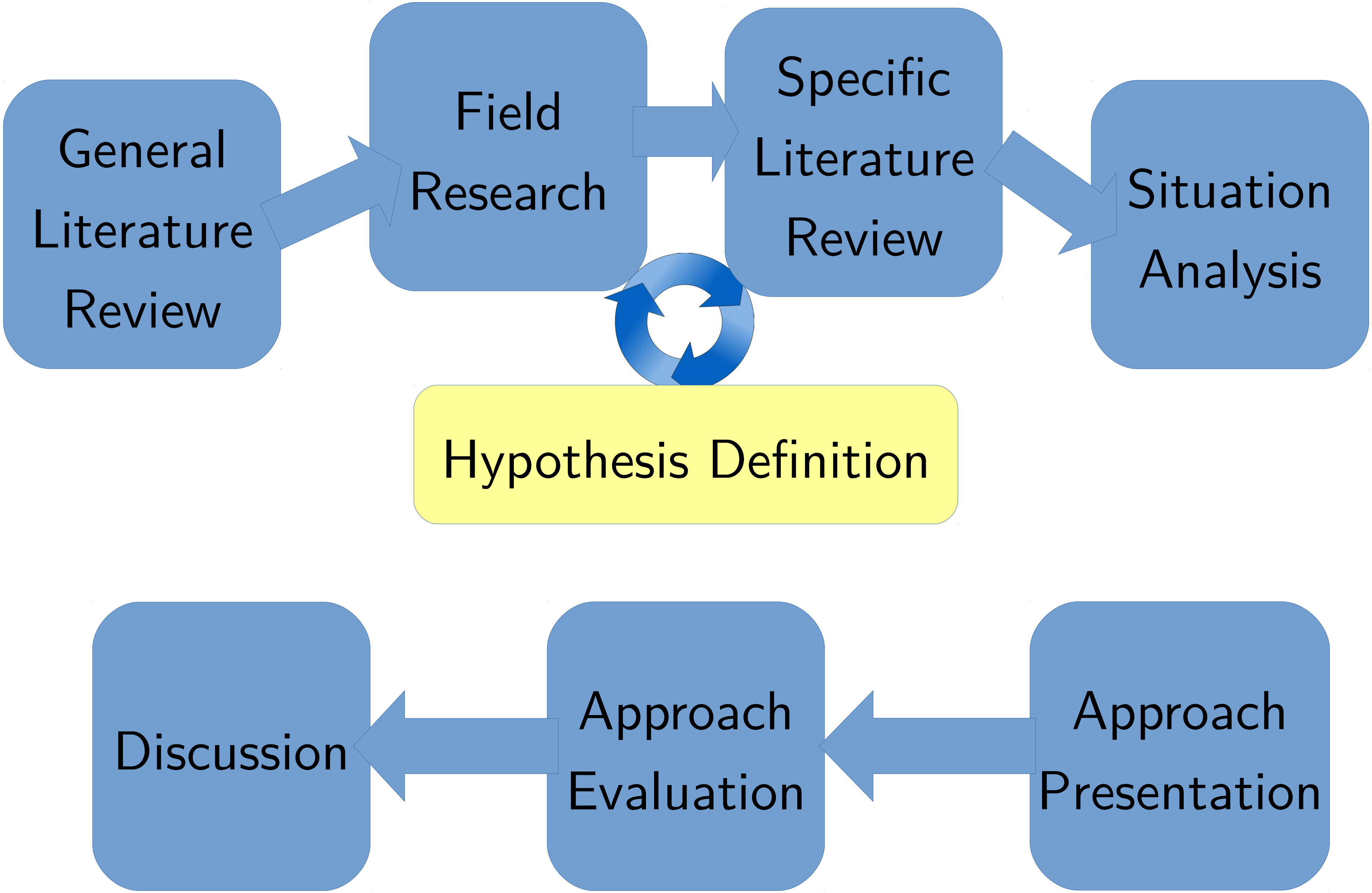
Analysis of Situation

Solution Approach

Evaluation

Conclusions

Methodology



General Literature Review

Field Research

Specific Literature Review

Situation Analysis

Hypothesis Definition

Approach Presentation

Approach Evaluation

Discussion

Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

Specific Literature Review

Analysis of Situation

Solution Approach

Evaluation

Conclusions

Open Data Critique - Literature

- Evidences from the literature that open data description / access to open data is a problem
- Zuiderwijk et al. (2012): “absence of commonly agreed metadata”, “insufficiency of metadata”, “lack of interoperability” and “difficulty in searching and browsing data”
- Roseira (2016):
    - Most datasets have incomplete or non-existent metadata.
    - Generates a higher workload on cleaning and harmonizing data.
    - Advances on datasets standardization in order to boost open data economic value creation at national and international levels.

Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

Specific Literature Review

Analysis of Situation

Solution Approach

Evaluation

Conclusions

Field Research

Main objective: to find out the motivations and impediments for open data use

Specific group: social movement activists

>> Personal experience + Interest in data for activism/advocacy

>> In general, low computer/internet skills

Methodology: data literacy course

>> Participatory research: not only collecting data, but also offering open data training

Field Research – Data Literacy

- A methodology for a Data Literacy Course was developed and applied to five classes with a total of 52 participants
- Courses were evaluated through observation and a questionnaire filled by students
- As a result, impediments, motivations and desired improvements were systematised: “Data organisation is confusing”, “Finding data in the web is hard”, “Government agencies do not follow common data standards”

Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

Specific Literature Review

Analysis of Situation

Solution Approach

Evaluation

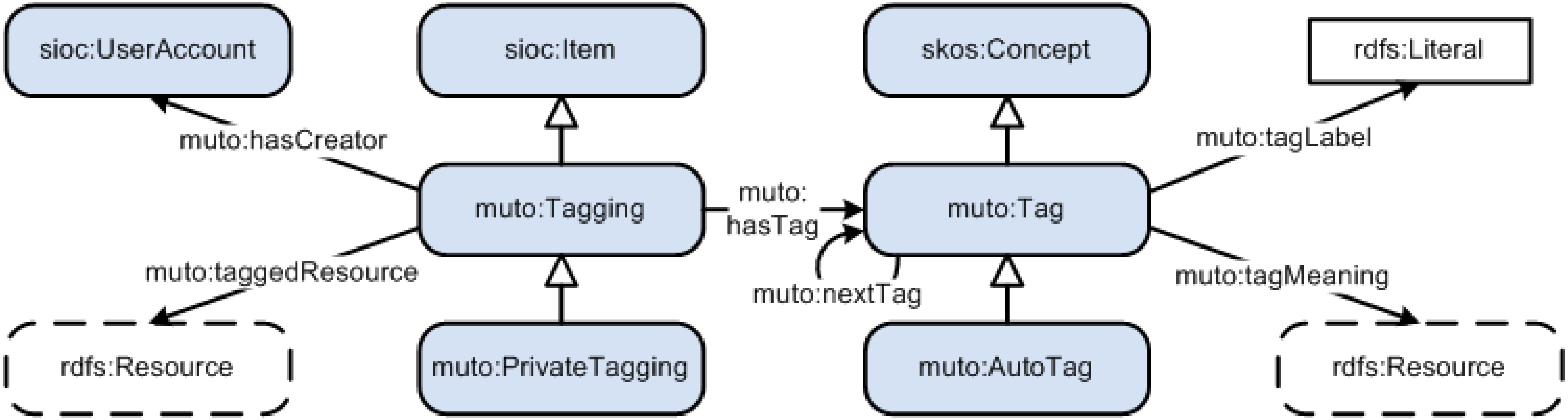
Conclusions

Metadata meets semantic web

- Folksonomies versus Ontologies?
  - Conceptualization of the act of tagging (GRUBBER 2007) $>$ *T (user, resource, tag, context)*
  - Folksonomy is a “lightweight, dynamic and limited in sharing scope” ontology. (MIKA 2007)
- Problems of Metadata without semantics:
  - Polysemy, synonyms, miss-spelling, no relations …

Metadata meets semantic web

Semantic tags (MUTO):



Enhancing dataset description

- Clean-up
  - Determining possible lexical representations for each tag (plural/singular, verb tenses, synonyms etc.)
- Reconciliation
  - Searching for equivalence between tags and semantic resources
- Structure emergence
  - Establishing relationships between dataset descriptors

Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

Specific Literature Review

Analysis of Situation

Solution Approach

Evaluation

Conclusions

ODP Metadata

In order to understand the actual situation of metadata in ODPs, 87 portals were analysed

Local Metrics:

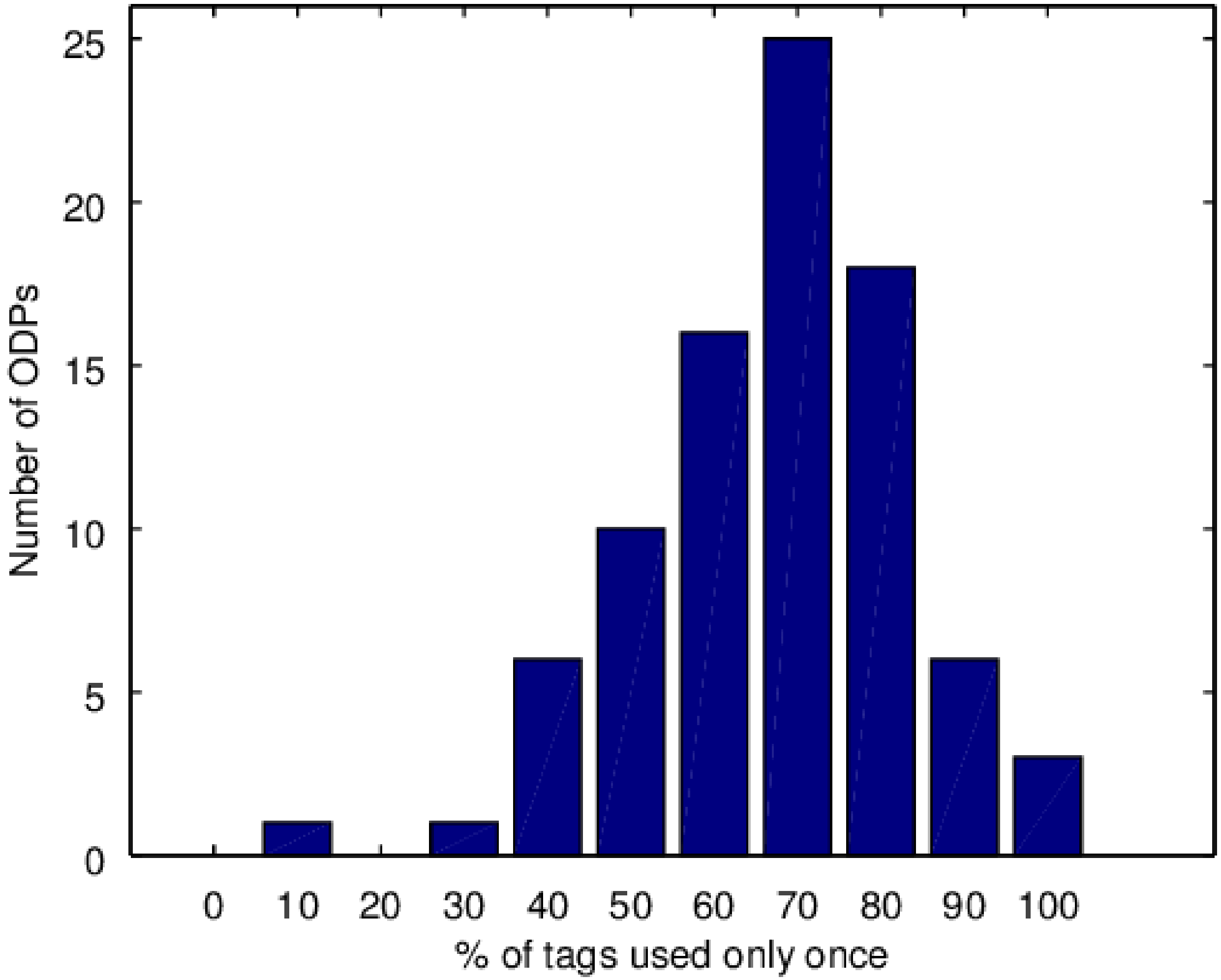
Tag reuse
Tags per dataset
Tag similarity

Global Metrics:

Coincident tags between portals
Tag expressiveness

ODP Metadata

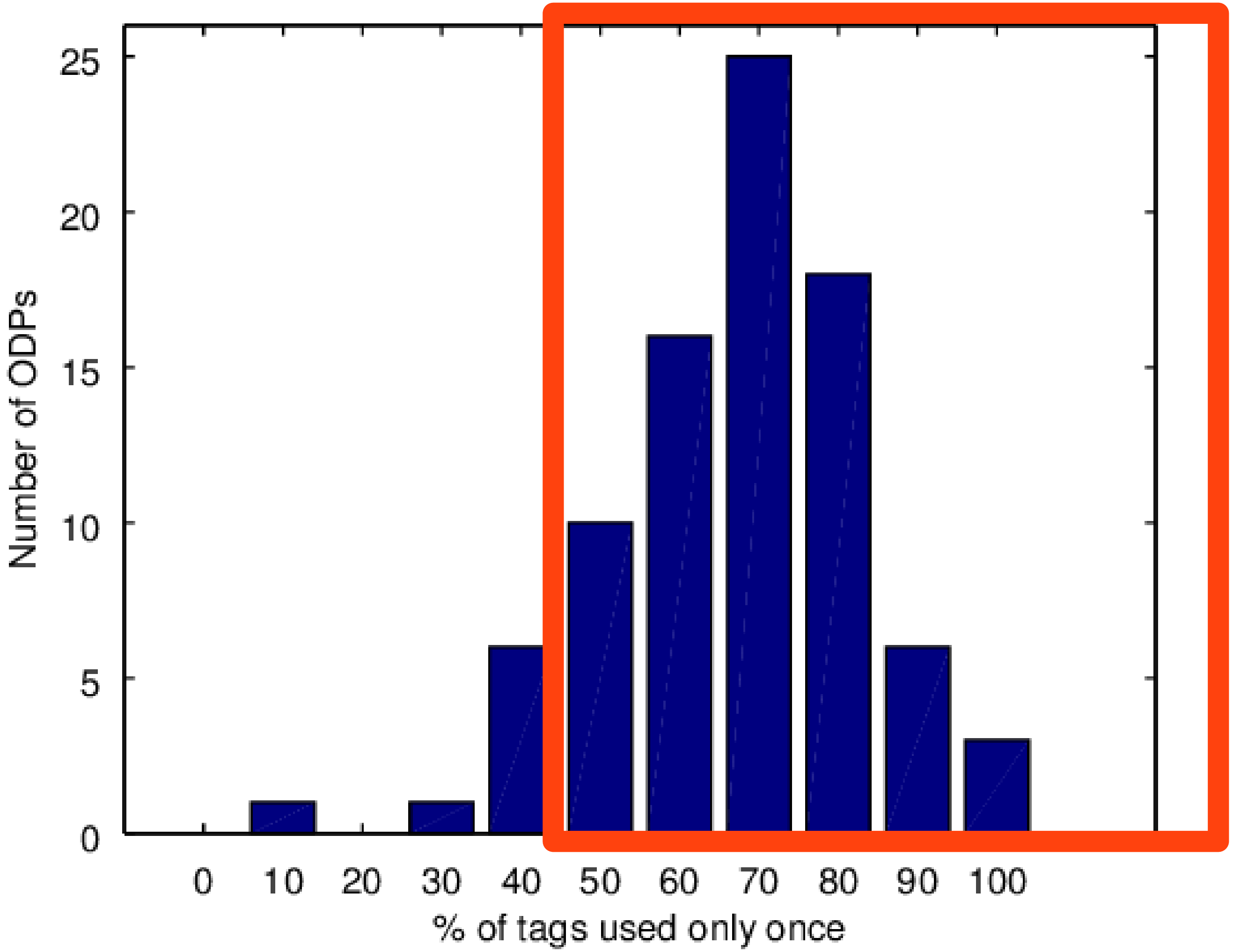
Tag reuse:



From the 87 portals, 75 use more than 50% of the tags only once.

ODP Metadata

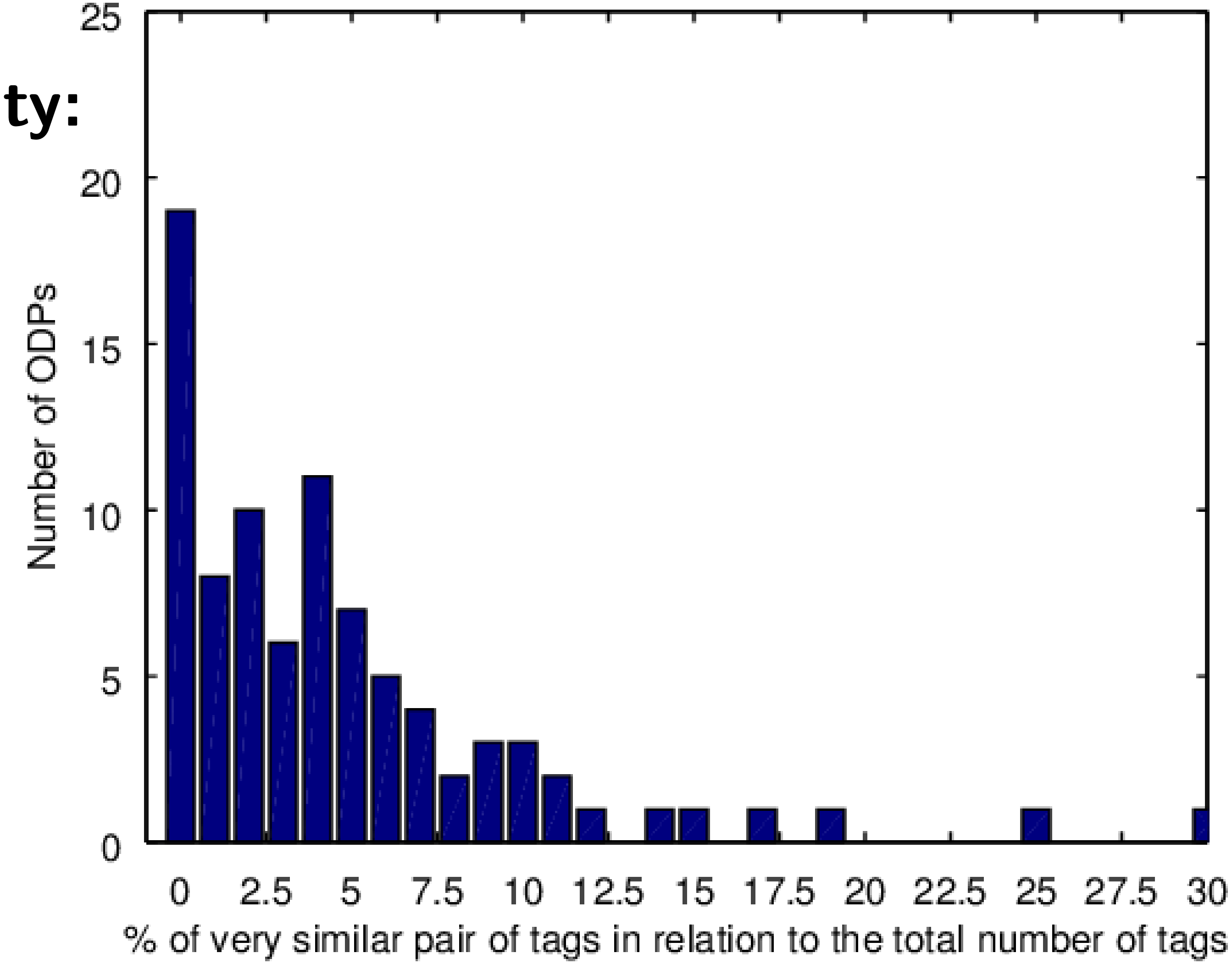
Tag reuse:



From the 87 portals, 75 use more than 50% of the tags only once.

ODP Metadata

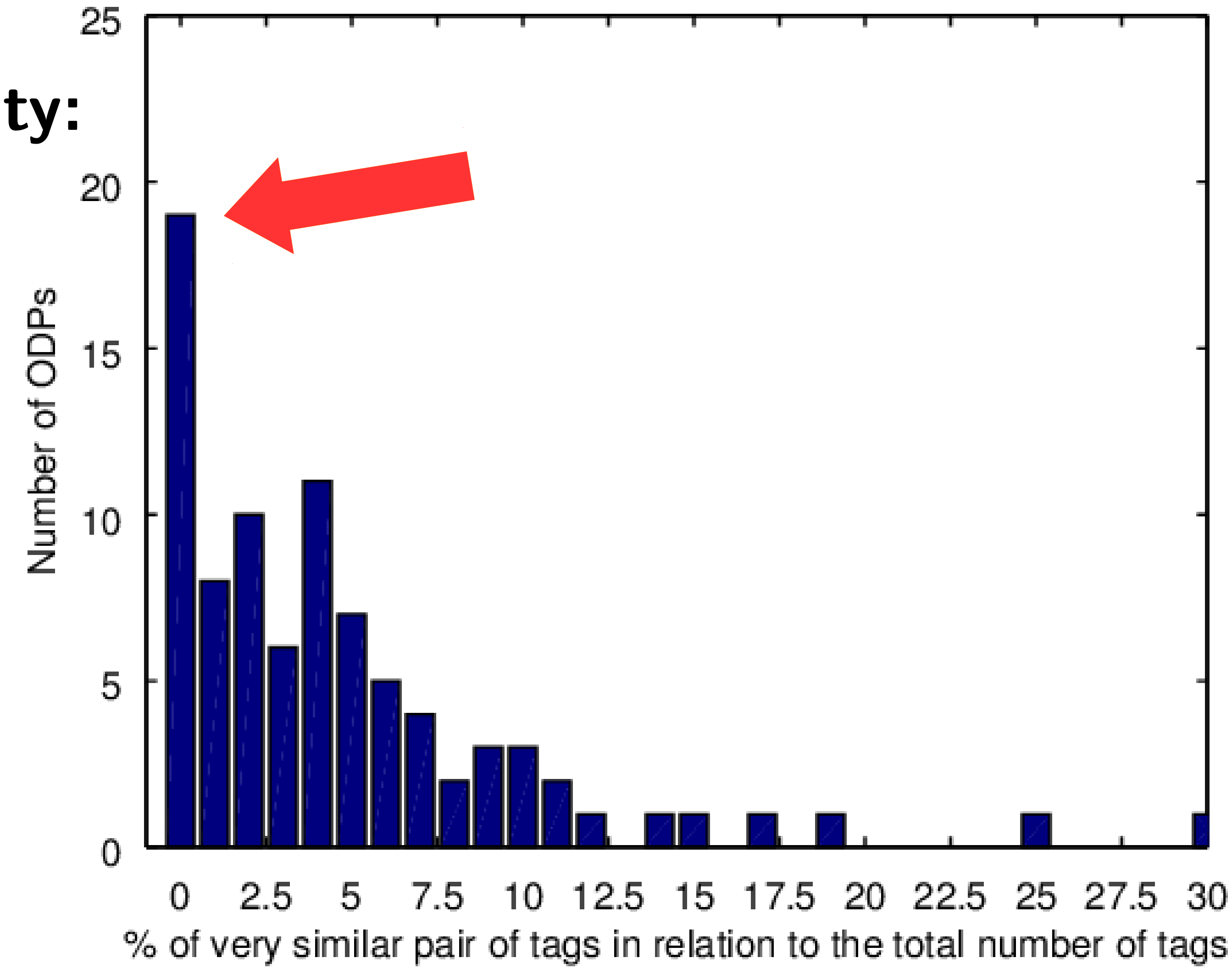
Tag similarity:



Only 20 portals, out of 87, revealed no similar tags at all

ODP Metadata

Tag similarity:



Only 19 portals, out of 87, revealed no similar tags at all

ODP Metadata

- Most ODPs apply between 1 and 7 tags to each dataset
- 28% tags appeared in more than one ODP, which represents 79,882 tags
- The majority of tags (73.65%) did not correspond to any semantic resource. For 26.35% of the tags, at least one meaning was found

TYGEL, A. F. et al. Towards Cleaning-up Open Data Portals: A Metadata Reconciliation Approach. In: Proc. of the 10th International Conference on Semantic Computing. Laguna Hills, California, 2016. p. 8.

Motivation

Hypothesis and Objective

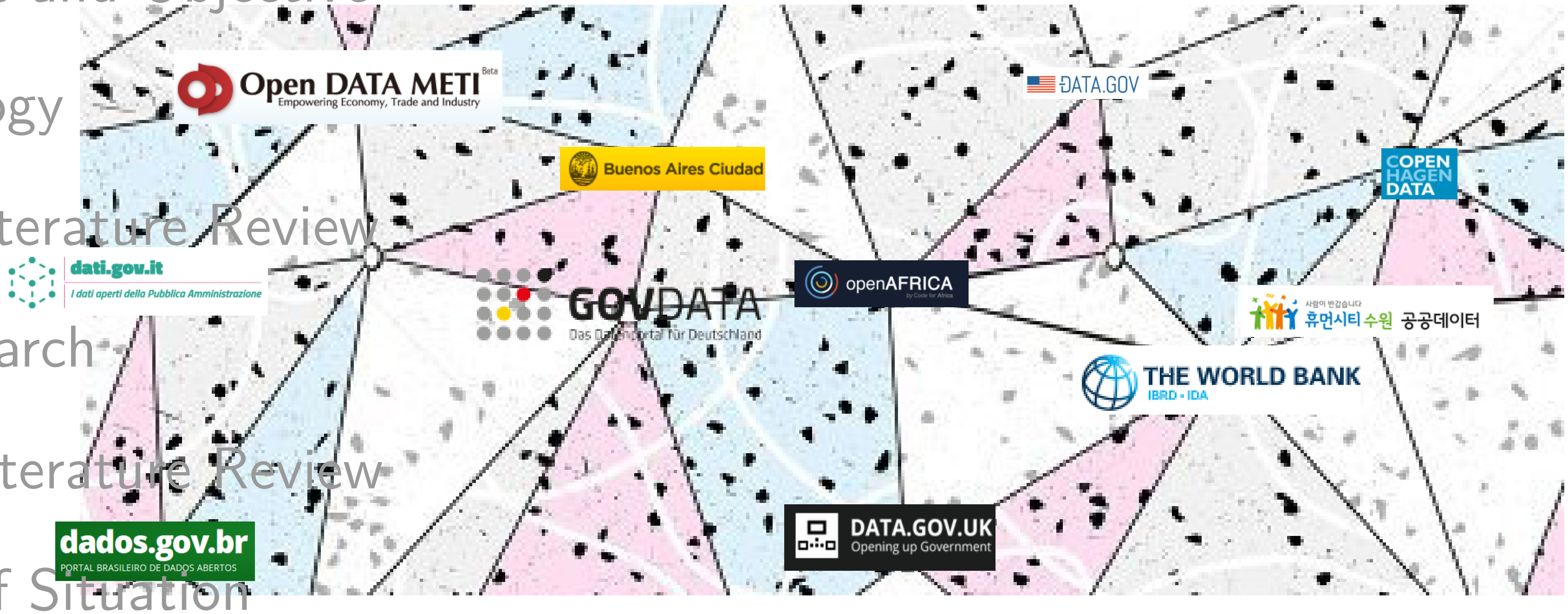
Methodology

General Literature Review

Field Research

Specific Literature Review

Analysis of Situation



Solution Approach

Evaluation

Conclusions

STODaP Approach

Objective

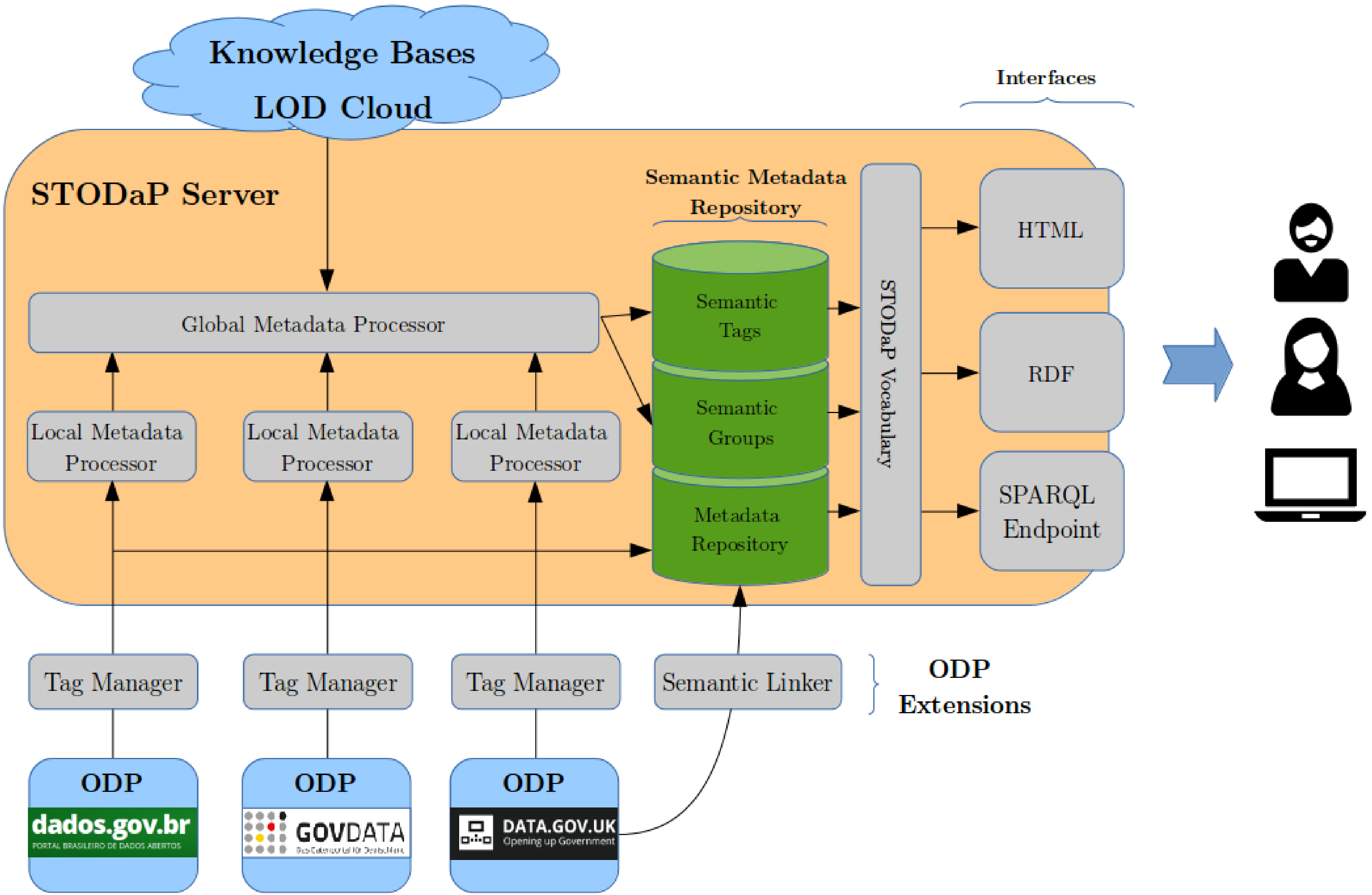
cleaning up and reconciling metadata in Open Data Portals, providing semantic connections between open datasets

Composed by

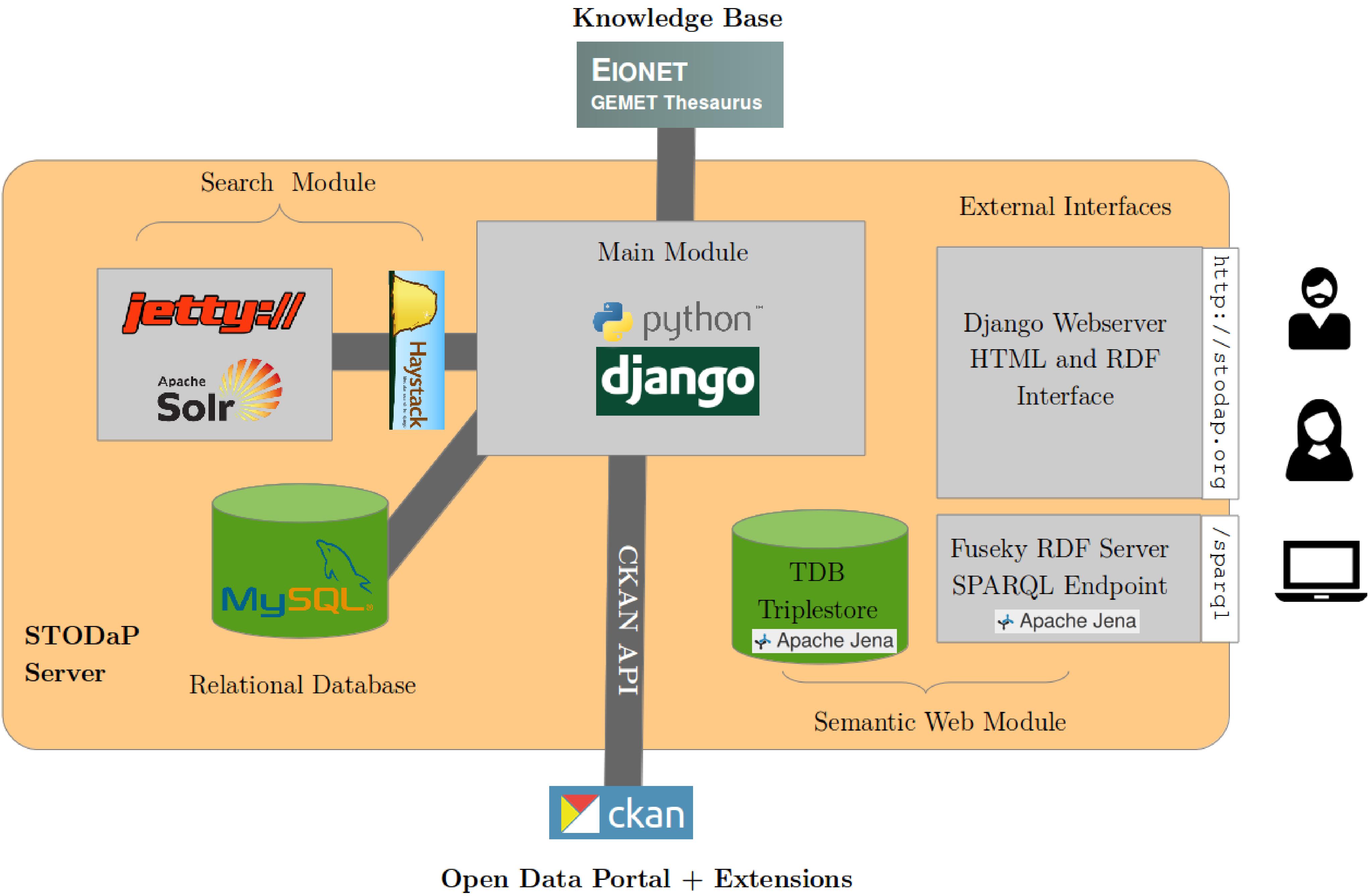
>> Global Part: Semantic Metadata Server

>> Local Part: Tag Manager and Semantic Linker

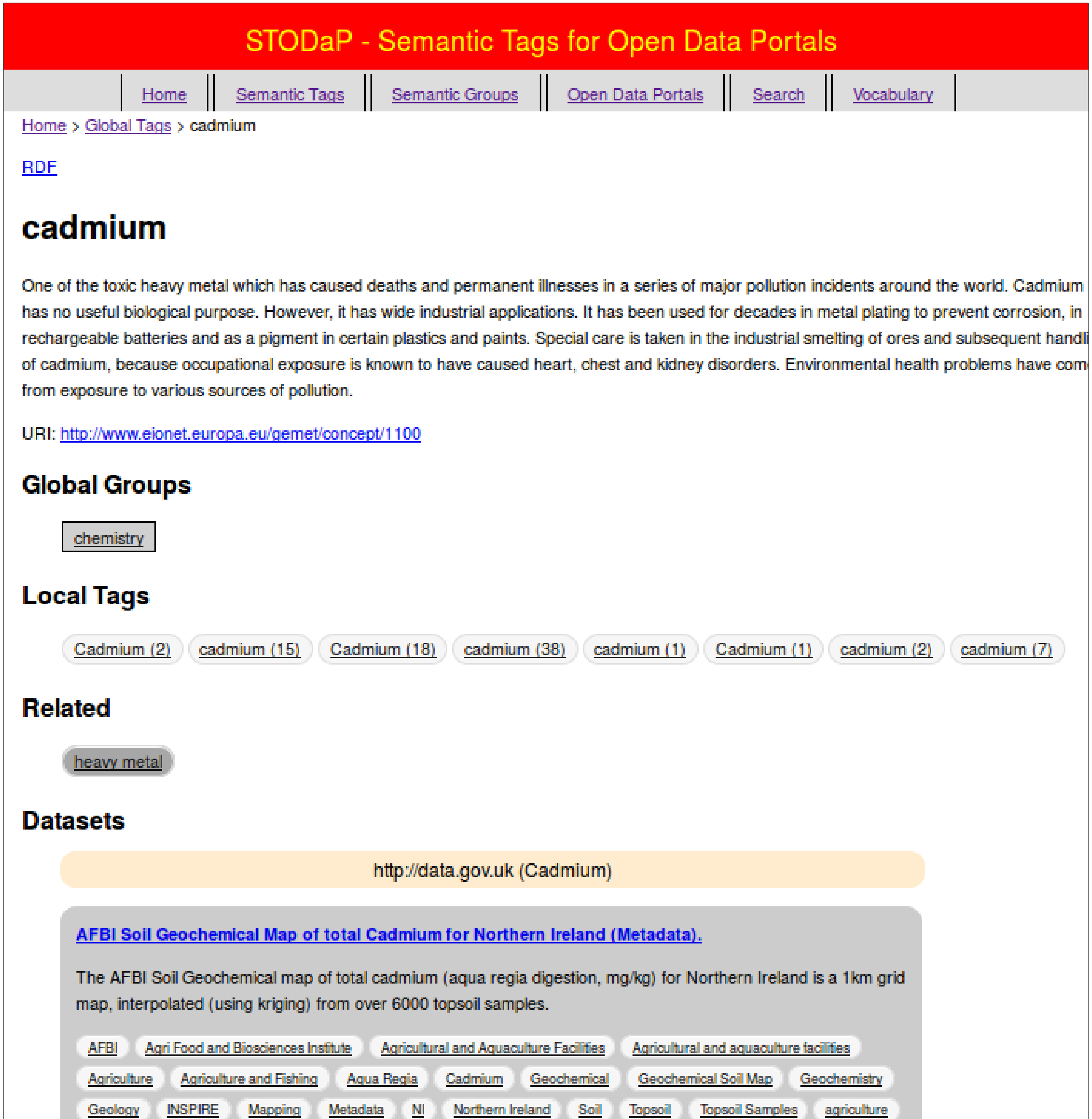
STODaP Approach – Architecture



STODaP Implementation



STODaP Navigation Interface



Home > Global Tags > cadmium

RDF

cadmium

One of the toxic heavy metal which has caused deaths and permanent illnesses in a series of major pollution incidents around the world. Cadmium has no useful biological purpose. However, it has wide industrial applications. It has been used for decades in metal plating to prevent corrosion, in rechargeable batteries and as a pigment in certain plastics and paints. Special care is taken in the industrial smelting of ores and subsequent handli of cadmium, because occupational exposure is known to have caused heart, chest and kidney disorders. Environmental health problems have com from exposure to various sources of pollution.

URI: http://www.eionet.europa.eu/gemet/concept/1100

Global Groups

chemistry

Local Tags

Cadmium (2) cadmium (15) Cadmium (18) cadmium (38) cadmium (1) Cadmium (1) cadmium (2) cadmium (7)

Related

heavy metal

Datasets

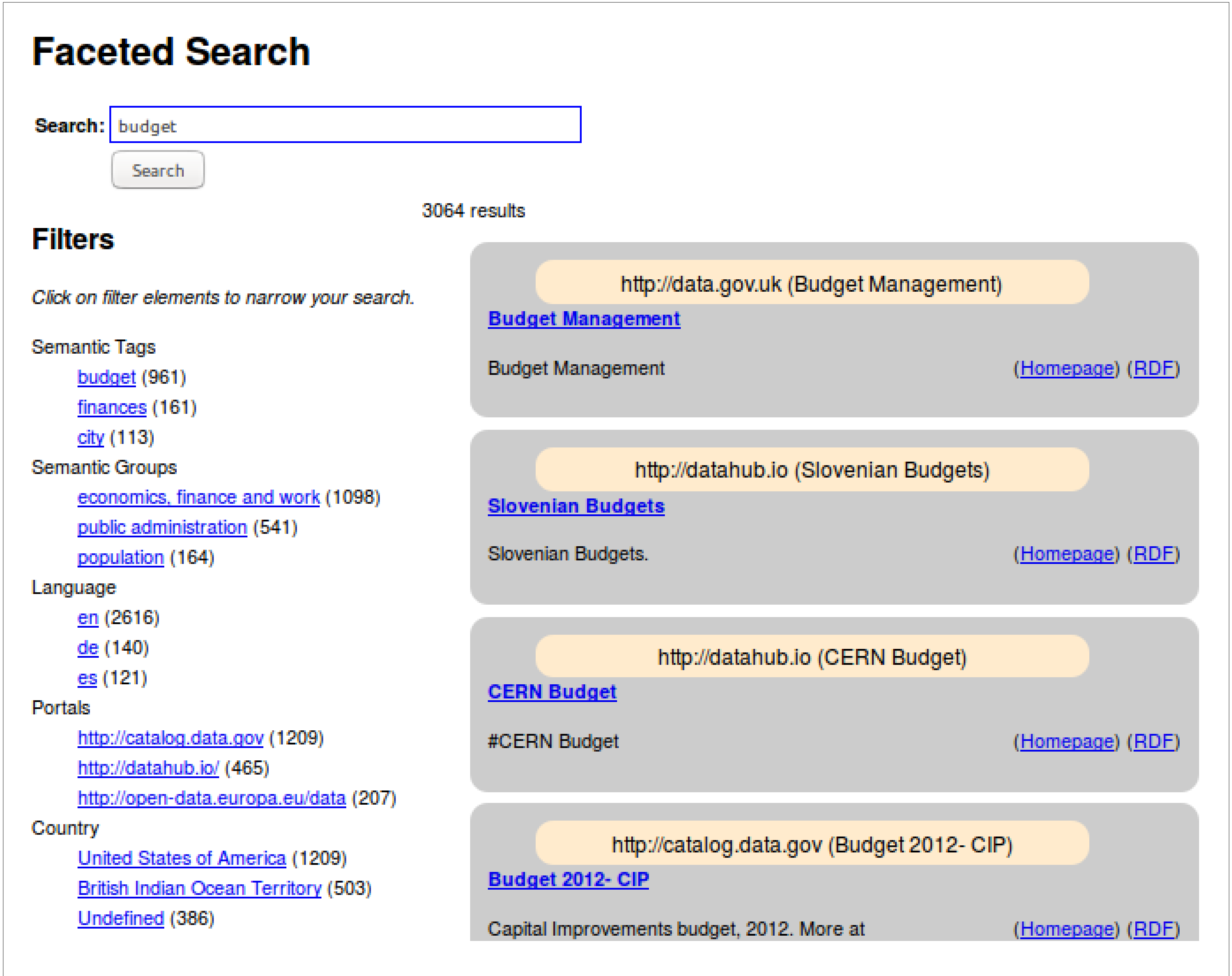
http://data.gov.uk (Cadmium)

AFBI Soil Geochemical Map of total Cadmium for Northern Ireland (Metadata).

The AFBI Soil Geochemical map of total cadmium (aqua regia digestion, mg/kg) for Northern Ireland is a 1km grid map, interpolated (using kriging) from over 6000 topsoil samples.

AFBI Agri Food and Biosciences Institute Agricultural and Aquaculture Facilities Agricultural and aquaculture facilities Agriculture Agriculture and Fishing Aqua Regia Cadmium Geochemical Geochemical Soil Map Geochemistry Geology INSPIRE Mapping Metadata NI Northern Ireland Soil Topsoil Topsoil Samples agriculture

STODaP Search Interface



Faceted Search

Search: budget

3064 results

Filters

Click on filter elements to narrow your search.

Semantic Tags
- budget (961)
- finances (161)
- city (113)

Semantic Groups
- economics, finance and work (1098)
- public administration (541)
- population (164)

Language
- en (2616)
- de (140)
- es (121)

Portals
- http://catalog.data.gov (1209)
- http://datahub.io/ (465)
- http://open-data.europa.eu/data (207)

Country
- United States of America (1209)
- British Indian Ocean Territory (503)
- Undefined (386)

http://data.gov.uk (Budget Management)

Budget Management

Budget Management (Homepage) (RDF)

http://datahub.io (Slovenian Budgets)

Slovenian Budgets

Slovenian Budgets. (Homepage) (RDF)

http://datahub.io (CERN Budget)

CERN Budget

#CERN Budget (Homepage) (RDF)

http://catalog.data.gov (Budget 2012- CIP)

Budget 2012- CIP

Capital Improvements budget, 2012. More at (Homepage) (RDF)

Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

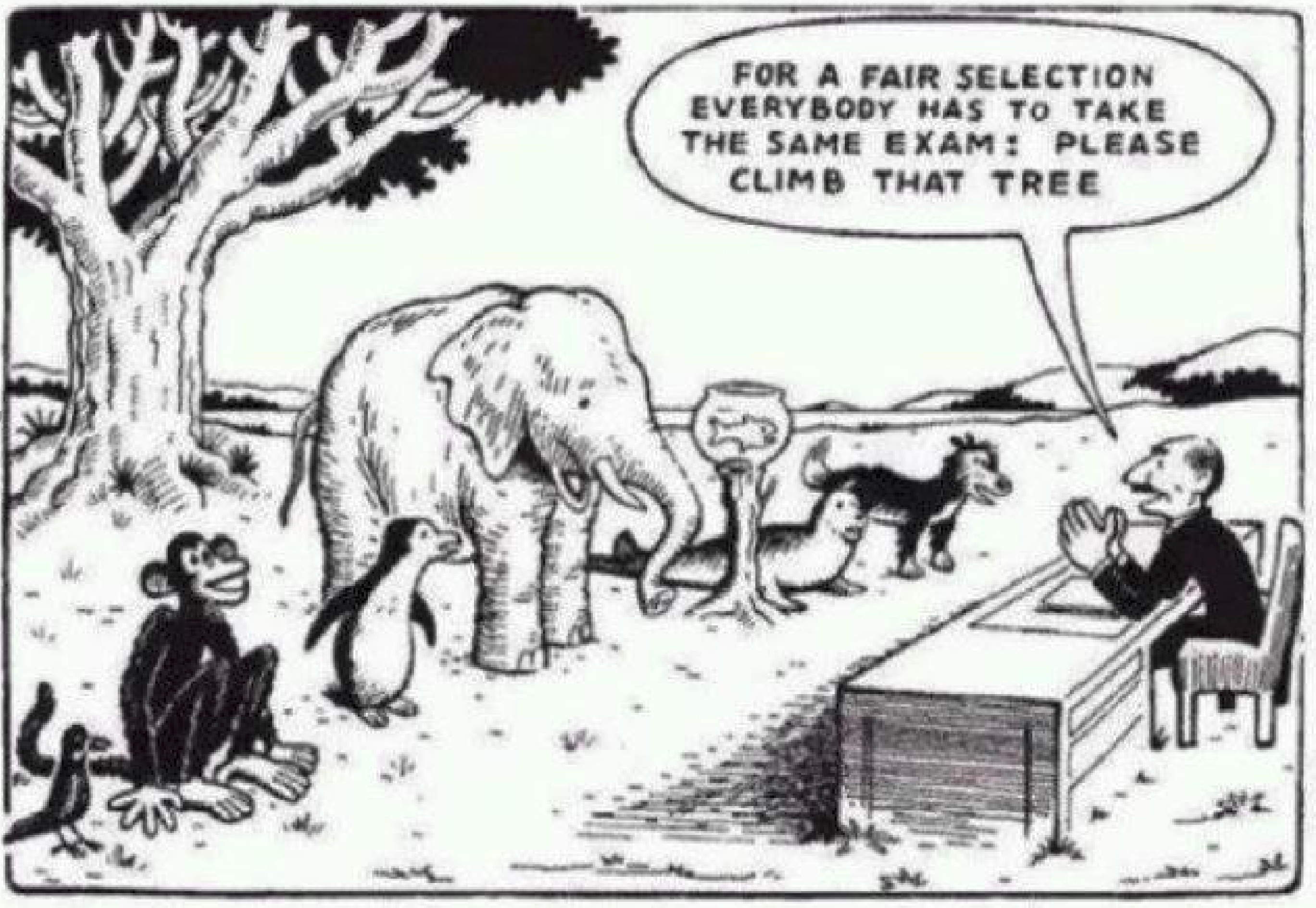
Specific Literature Review

Analysis of Situation

Solution Approach

Evaluation

Conclusions


FOR A FAIR SELECTION EVERYBODY HAS TO TAKE THE SAME EXAM: PLEASE CLIMB THAT TREE

STODaP Evaluation

Background: Dataset Search Engine Evaluation

Goals:

G1: When searching for open datasets, how does the STODaP server compares to other data-specific and general search engines?

Metrics:

>> Task Completion Time, Precision

G2: Is the STODaP server an useful tool for searching open datasets?

Metrics:

>> subjective evaluation

STODaP Evaluation

Questions

> Q1: Find open datasets about **water quality** on **7 different rivers outside Europe**.

> Q2: Find open datasets containing **2015 budget** data from locations in **5 different countries**.

> Q3: Find open datasets containing **procurement information** in **3 different languages**.

Search methods

> **Exversion**: Data specific search engine

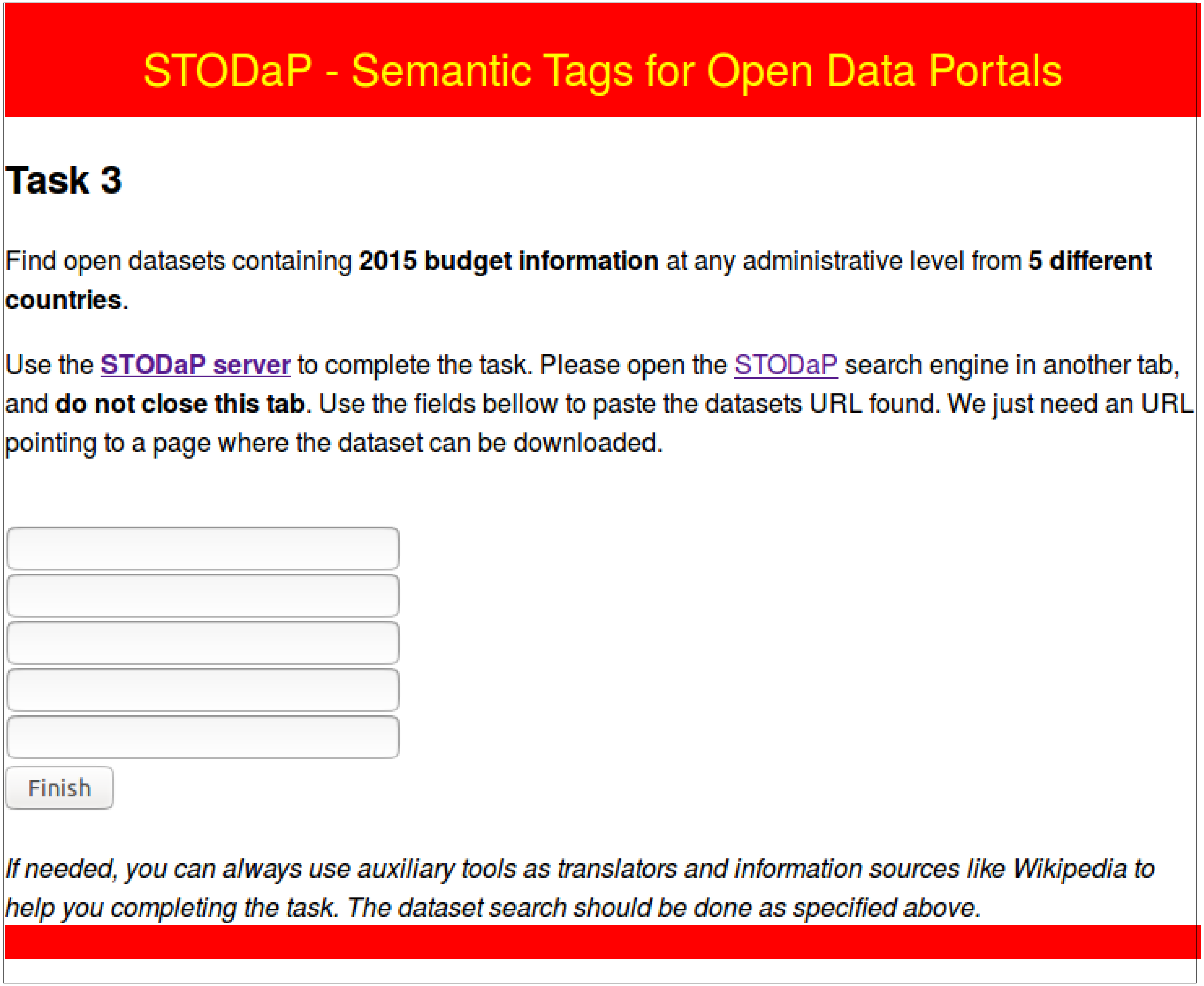
> **Free**: Generic Web Search Engines, freely chosen by users

STODaP Evaluation - Task

Question

Search Method

Answers



STODaP - Semantic Tags for Open Data Portals

Task 3

Find open datasets containing **2015 budget information** at any administrative level from **5 different countries**.

Use the **STODaP server** to complete the task. Please open the STODaP search engine in another tab, and **do not close this tab**. Use the fields bellow to paste the datasets URL found. We just need an URL pointing to a page where the dataset can be downloaded.

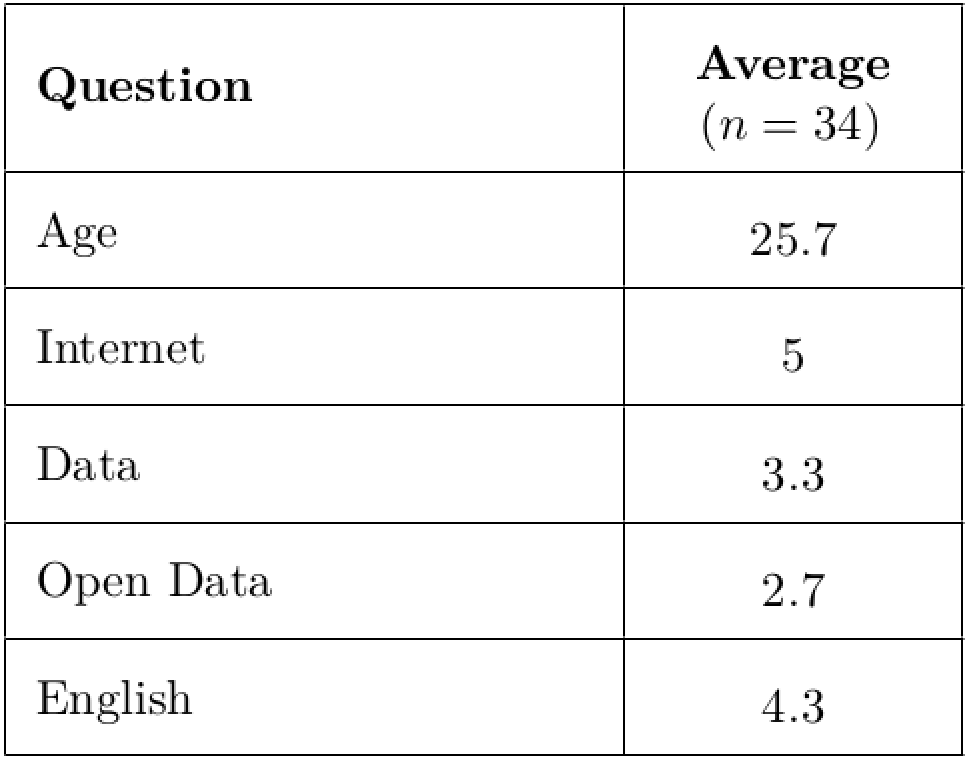
If needed, you can always use auxiliary tools as translators and information sources like Wikipedia to help you completing the task. The dataset search should be done as specified above.

TCT: T/5

P: Ac/5

STODaP Evaluation

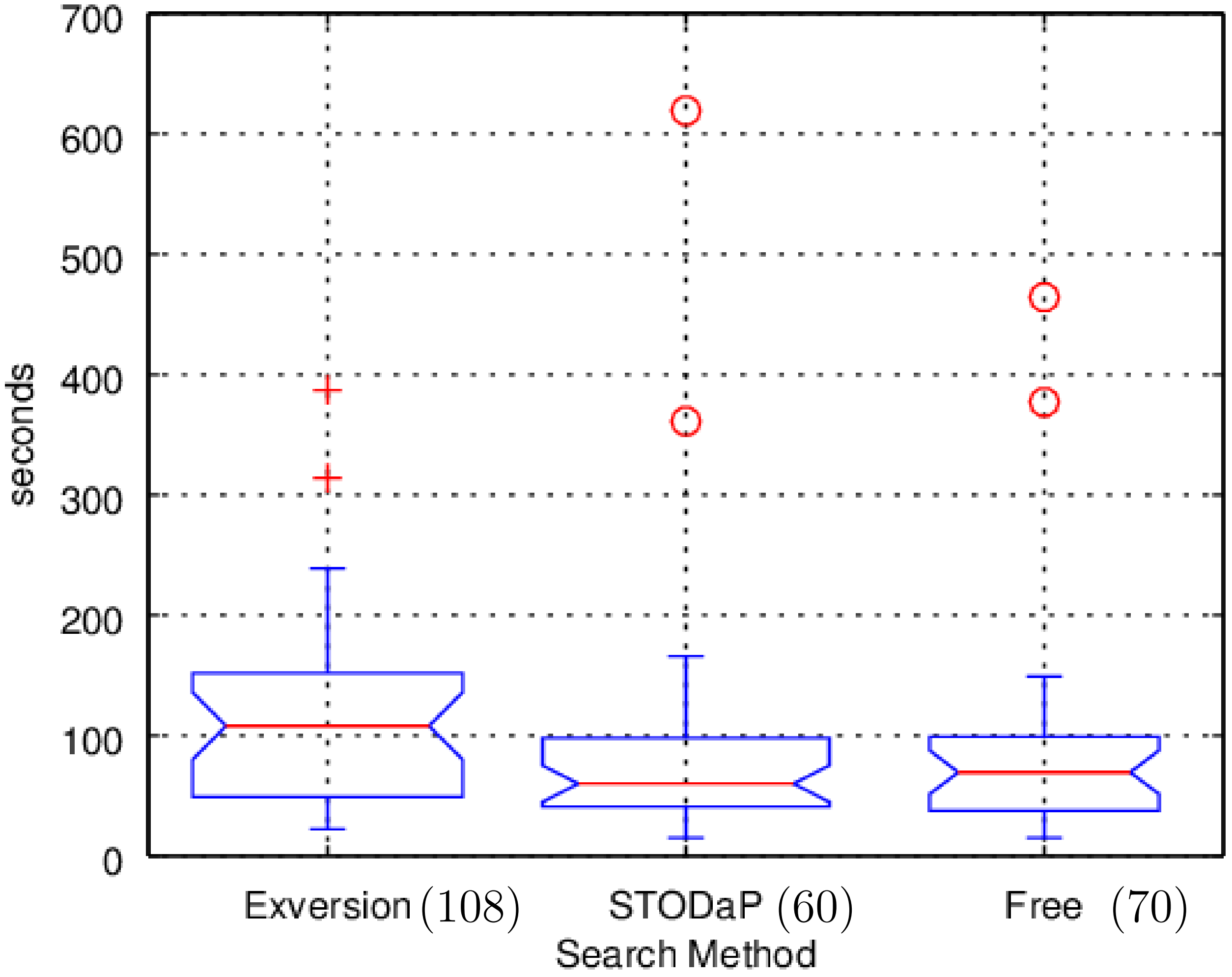
Participants Profile



Question	**Average** $(n = 34)$
Age	25.7
Internet	5
Data	3.3
Open Data	2.7
English	4.3

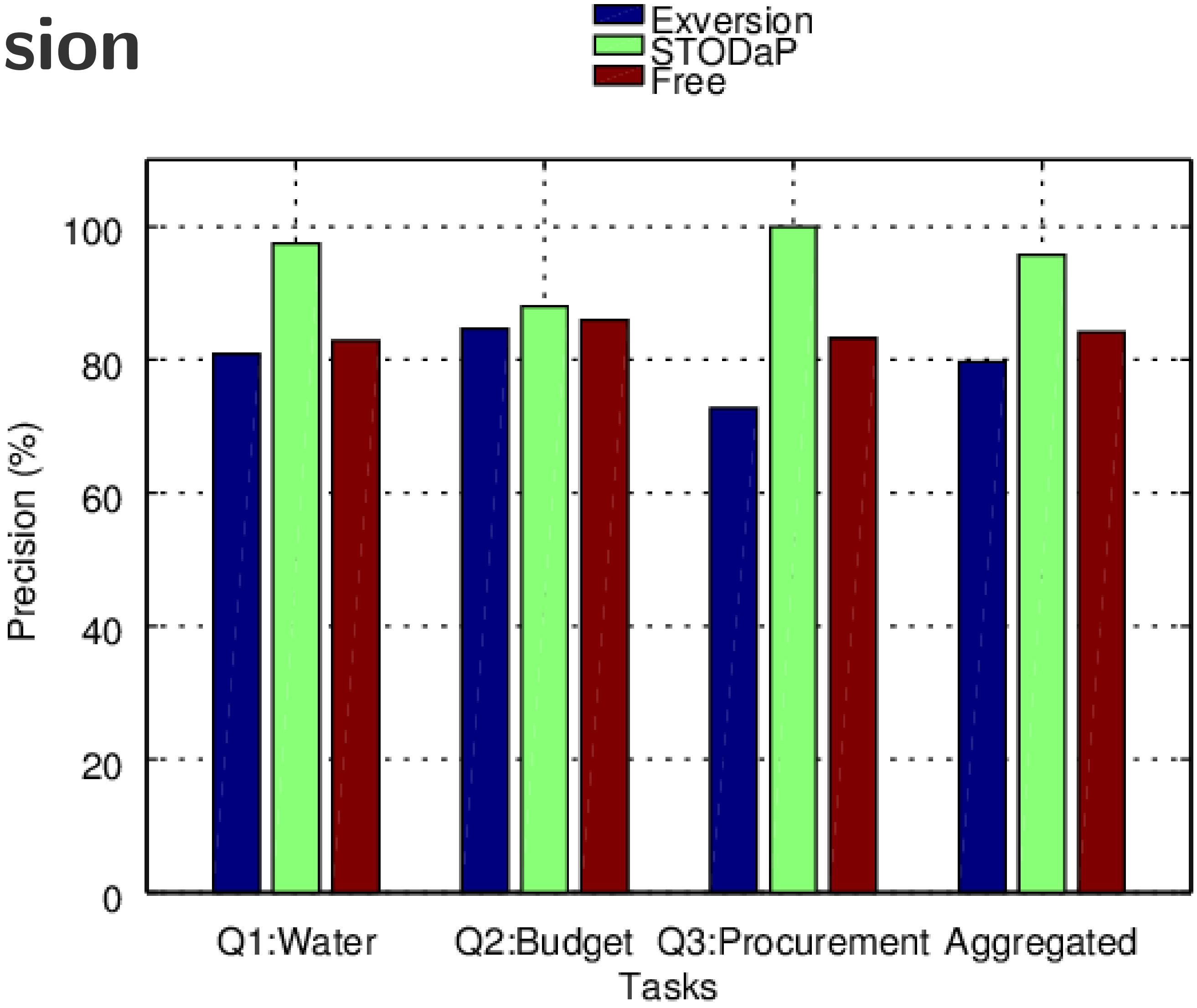
STODaP Evaluation - Results

TCT



STODaP Evaluation – Results

Precision



STODaP Evaluation - Results

Subjective Evaluation

Do you think STODaP is a useful tool for finding data on the web?

How easy it was to get the data you need using STODaP in comparison with other methods?

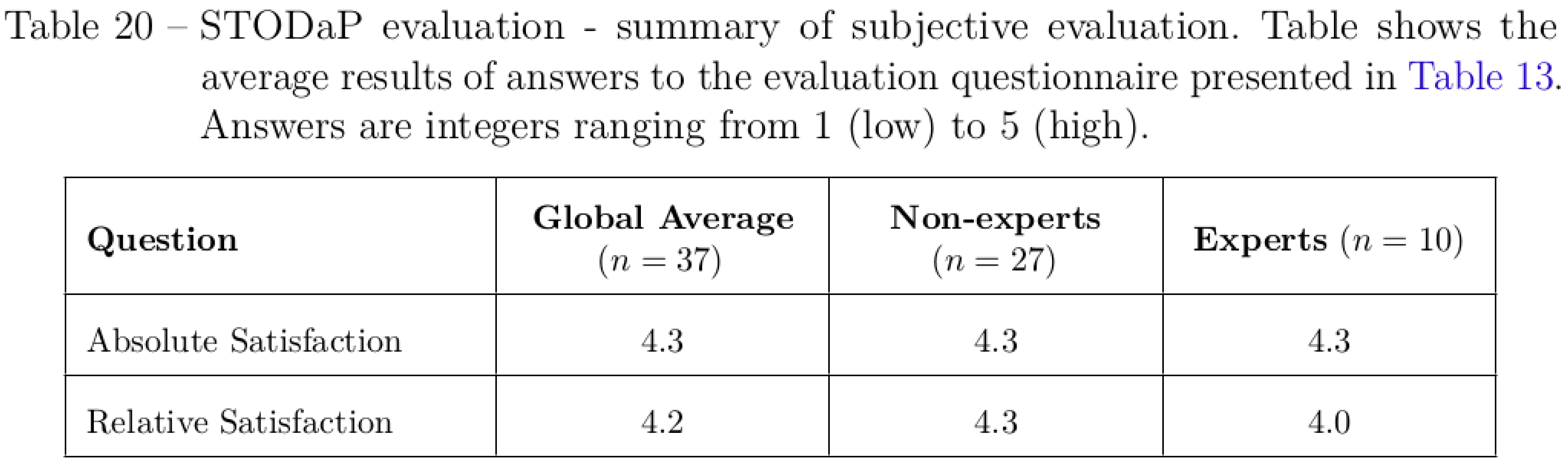


Table 20 – STODaP evaluation - summary of subjective evaluation. Table shows the average results of answers to the evaluation questionnaire presented in Table 13. Answers are integers ranging from 1 (low) to 5 (high).

Question	**Global Average** $(n = 37)$	**Non-experts** $(n = 27)$	**Experts** $(n = 10)$
Absolute Satisfaction	4.3	4.3	4.3
Relative Satisfaction	4.2	4.3	4.0

STODaP Evaluation Results Analysis

- In general, participants searching datasets using STODaP were able to retrieve open datasets **faster and more precisely**
- However, regarding Q1, free search achieved an **equivalent TCT**, and for Q3, a **faster TCT**
- For Q2, **precision was equivalent** among all methods
- Negative correlation between **open data ability** and **relative satisfaction** (low confidence) > Higher satisfaction for non-experts

Motivation

Hypothesis and Objective

Methodology

General Literature Review

Field Research

Specific Literature Review

Analysis of Situation

Solution Approach

Evaluation

Conclusions

Contribution

Main Contribution: STODaP approach

- *For users with at least an intermediate level of English, daily internet use, and average data experience, STODaP open data search engine delivers open datasets with a higher precision in less time than other search methods when searching for relevant open data topics.*

Limitations

- **Limitations** of the evaluation:
  - Participants profile: distinct from field research
  - Topics: different results for different questions
  - Non-assessed components: extensions, navigation, semantic relations
- Balance between **generic** and **specific** tasks

Contribution

Contributions on Data Literacy

Theoretical contribution regarding Data Literacy and Popular Education

Methodological contribution regarding Data Literacy Course

A practical contribution, regarding the systematisation of impediments, benefits and improvements of open data according to social movement activists.

Limitation: evaluation

Future Work

Enhancements on STODaP implementation:

- Layout:
- Semantic Lifting Quality $>$ como?
- Increase number of ODP

Enlarge the evaluation scope

- Connect data literacy and STODaP evaluations
- Enlarge topics coverage

Conclusion

- Open Data potential for consolidating a “Data Revolution” is high
- However, access to data must be enhanced:
  - Breaking the silos of data
  - Boosting open data skills on the society
- Transdisciplinary approaches are fundamental to understand the problem and propose solutions

Publications related to the Thesis

TYGEL, A. F.; AUER, S.; DEBATTISTA, J., ORLANDI, F.; CAMPOS, M. L. M. .Towards Cleaning-up Open Data Portals: A Metadata Reconciliation Approach. 10th International Conference on Semantic Computing, Laguna Hills, California. February 3-5 2016.

TYGEL, A. F.; ATTARD, J.; ORLANDI, F.; CAMPOS, M. L. M. ; AUER, S. . "How much?" Is Not Enough - An Analysis of Open Budget Initiatives. ICEGOV 2016, Montevideo, March 1-3 2016.

TYGEL, A. F. ; KIRSCH, R. . Contributions of Paulo Freire for a Critical Data Literacy: a Popular Education Approach. , to appear in Journal of Community Informatics.

TYGEL, A. F. ; CAMPOS, M. L. M. ; ALVEAR, C. A. S. . Teaching Open Data for Social Movements: a Research Strategy. Journal of Community Informatics, v. 11, p. 1, 2015.