Enabling Self-service Data Provisioning through Semantic Enrichment of Data

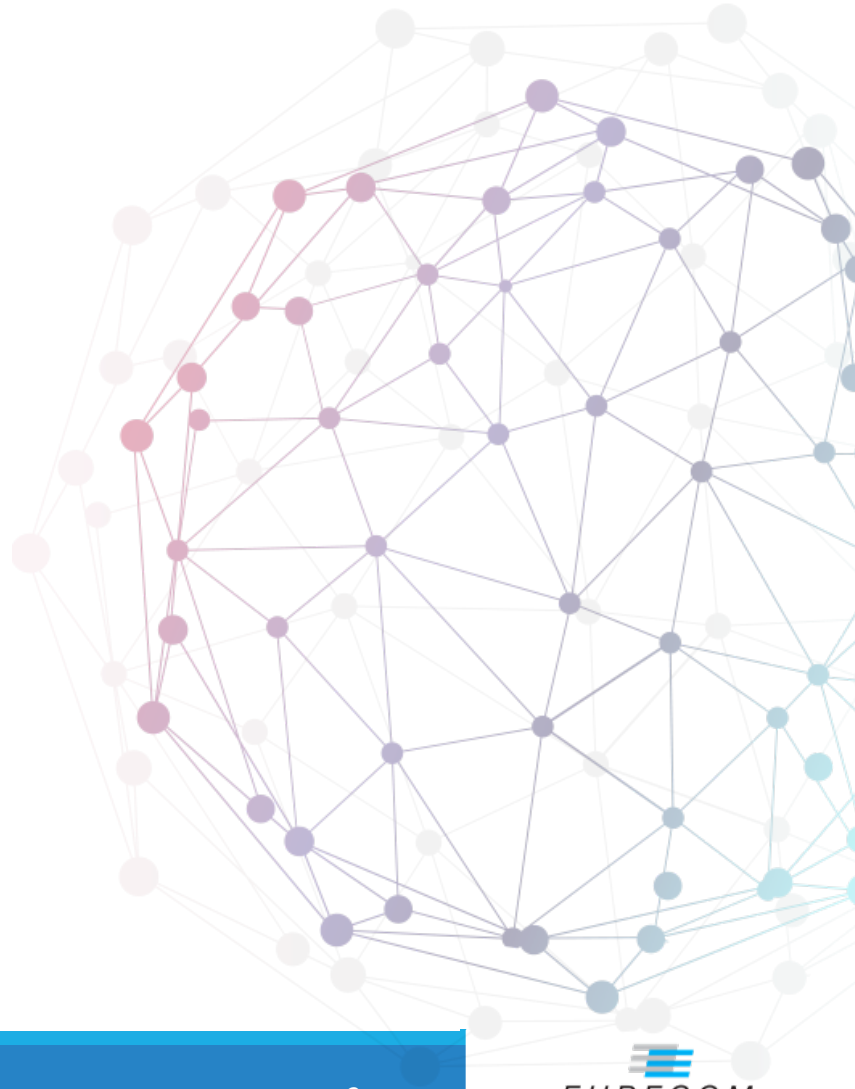
Ahmad Assaf

me@ahmadassaf.com @ahmadaassaf

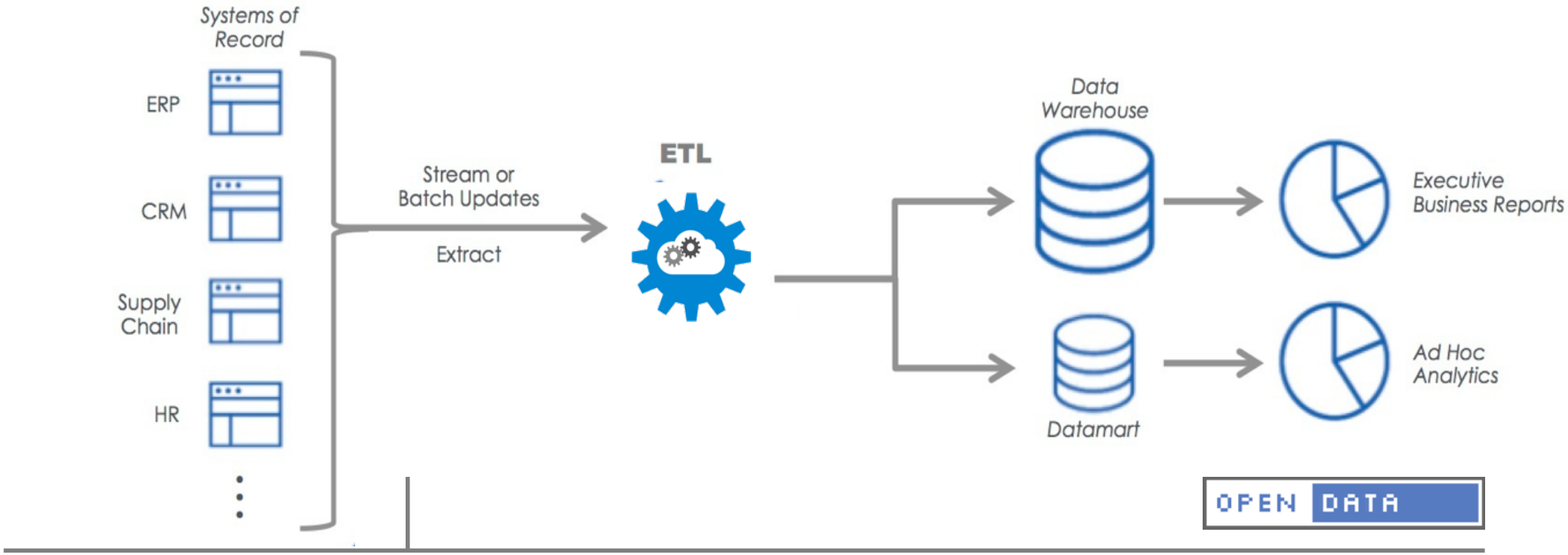
Supervisors: Raphaël Troncy and Aline Senart

Outline

- Introduction
  - Use-case Scenario
  - Research Challenges
- Contributions
  - Towards A Complete Dataset Profile
    - Harmonized Dataset Model (HDL)
    - Automatic generation and validation of dataset profiles (Roomba)
    - Objective quality assessment
  - Towards Enriched Enterprise Data
    - Semantic enrichment in the enterprise
    - Semantic Social News Aggregator (SNARC)
- Conclusion & Future Work
- Publications



Introduction | What is Business Intelligence ?



Introduction | Self-service Data Provisioning

Preparing and gathering data for reporting and visualization is by far the most challenging task in most BI projects large and small

> “Self-service data provisioning aims at reducing the challenge by providing dataset discovery, acquisition and integration techniques intuitively to the end user”

Use-case Scenario - Personas

Dan FERRY

Paris, France

Data Analyst

Ministry of Transport

I collect and acquire #data from multiple data sources, filter and clean data, interpret and analyze results and provide ongoing reports

2 days ago

J'adore #SAP #Lumira #dashboard #visualization #BI
http://go.sap.com/product/analytics/lumira.html

3 days ago

Paul PARKER

London, UK

Data Portal Administrator

data.gov.uk

I monitor the overall health of the portal. Also oversee the creation of users, organizations and datasets

Yesterday

I try to ensure a certain #dataQuality level by continuously checking for spam and manually enhancing dataset descriptions and annotations

1 week ago

Use-case Scenario

- **Dan** receives a memo from his management to create a report comparing the number of car accidents that occurred in **France** for **this year**, to its counterpart in the **United Kingdom (UK)**

- **Dan** is asked to highlight accidents related to **illegal consumption of alcohol** in both countries

Research Challenges [1/2]

Dataset Maintenance & Discovery

- Find useful datasets for specialized domains
- Quickly assess datasets by checking the attached metadata
- Generate and validate datasets descriptions and metadata
- Detect spam and irrelevant datasets

Dataset Quality

- Automatic tools to issue scores or certificates reflecting the objective dataset quality
- Check if the dataset on hand is of a certain degree of quality to be used in reports

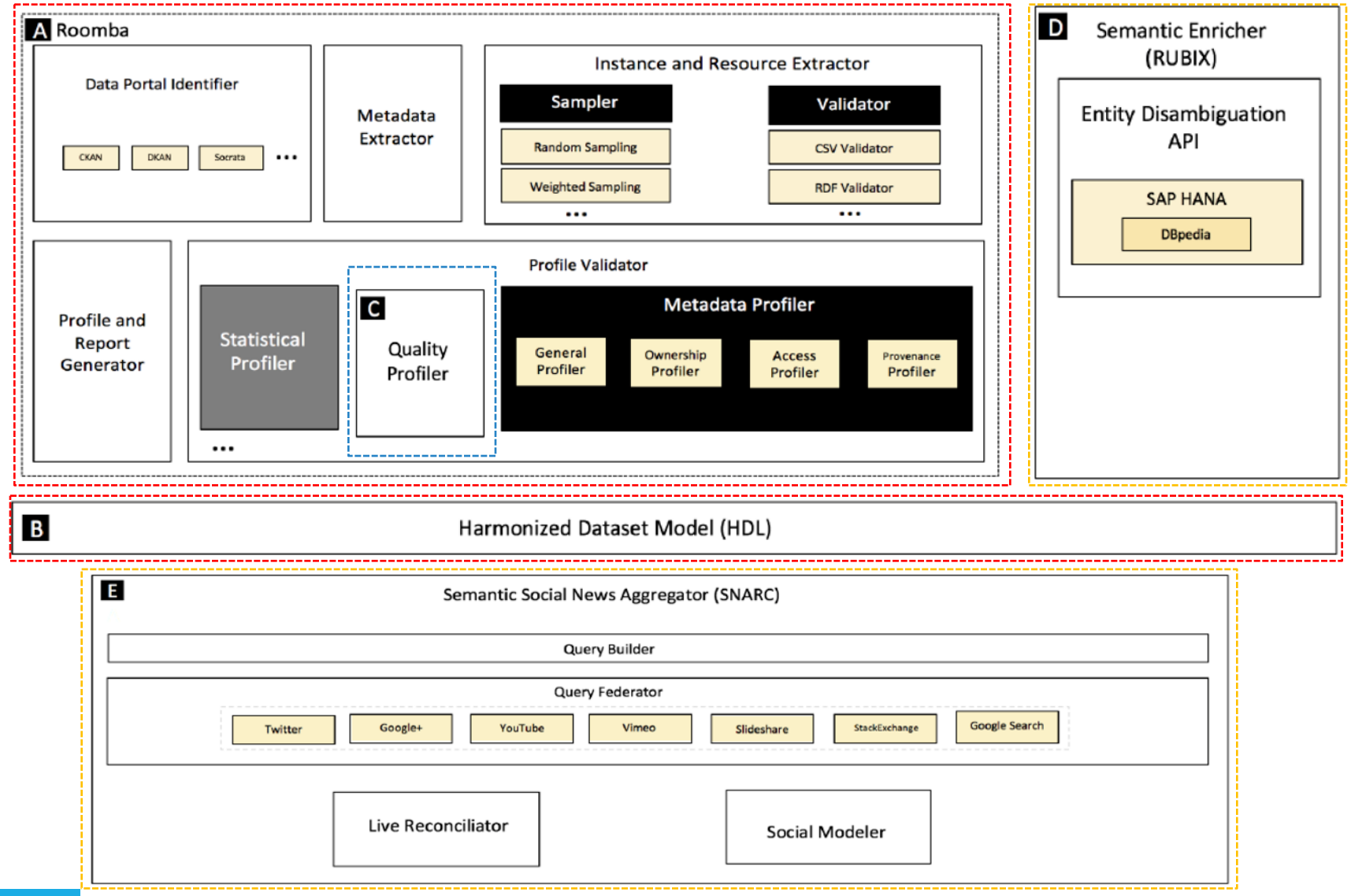
Research Challenges [2/2]

Dataset Integration and Enrichment

- Map and organize the data in order to have a unified view for these heterogeneous and complex data structures 
- Assess which entity's properties are more **important** than others for particular tasks 
- The vast amount of data available in social networks makes it hard to spot what is relevant in a timely manner 

Proposed Framework

- Dataset Maintenance & Discovery
  - HDL
  - Roomba
- Dataset Quality
- Dataset Integration and Enrichment
  - RUBIX
  - SNARC



PART I

Towards A Complete Dataset Profile

Research Challenges

- Dataset Maintenance & Discovery
- Dataset Quality

Data Portals/Data Management Systems

> Metadata is structured information that describes, explains, locates or otherwise makes it easier to retrieve use or manage information resources

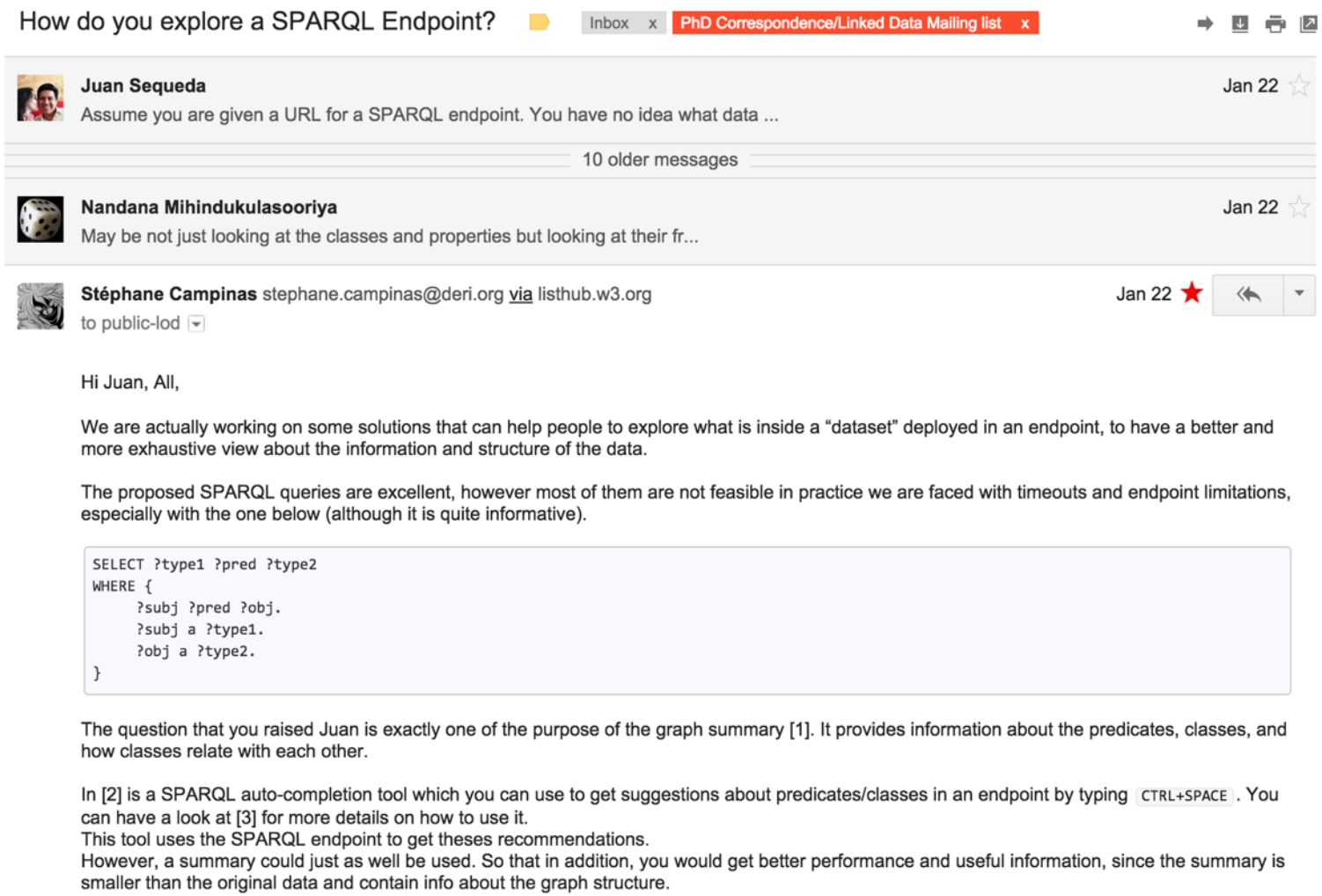
- Entry points to discover published datasets
- Curated collection of datasets metadata providing a set discovery and integration services
- Built on top of Data Management Systems (**DMS**) like CKAN, DKAN and Socrata

Dataset Vocabularies & Models

- Data Catalog Vocabulary (**DCAT**)
  - Three main classes: dcat:Catalog, dcat:Dataset and dcat:Distribution
  - Extensions : **DCAT-AP,** The Asset Description Metadata Schema (**ADMS)**

- Vocabulary of Interlinked Datasets (**VoID**)
  - Three main classes: void:Dataset, void:Linkset and void:subset
  - + links between datasets

Why a Harmonized Model ?



How do you explore a SPARQL Endpoint? Inbox x PhD Correspondence/Linked Data Mailing list x

Juan Sequeda Jan 22
Assume you are given a URL for a SPARQL endpoint. You have no idea what data ...

10 older messages

Nandana Mihindukulasooriya Jan 22
May be not just looking at the classes and properties but looking at their fr...

Stéphane Campinas stephane.campinas@deri.org via listhub.w3.org Jan 22
to public-lod

Hi Juan, All,

We are actually working on some solutions that can help people to explore what is inside a "dataset" deployed in an endpoint, to have a better and more exhaustive view about the information and structure of the data.

The proposed SPARQL queries are excellent, however most of them are not feasible in practice we are faced with timeouts and endpoint limitations, especially with the one below (although it is quite informative).

```
SELECT ?type1 ?pred ?type2
WHERE {
     ?subj ?pred ?obj.
     ?subj a ?type1.
     ?obj a ?type2.
}
```

The question that you raised Juan is exactly one of the purpose of the graph summary [1]. It provides information about the predicates, classes, and how classes relate with each other.

In [2] is a SPARQL auto-completion tool which you can use to get suggestions about predicates/classes in an endpoint by typing `CTRL+SPACE`. You can have a look at [3] for more details on how to use it.
This tool uses the SPARQL endpoint to get theses recommendations.
However, a summary could just as well be used. So that in addition, you would get better performance and useful information, since the summary is smaller than the original data and contain info about the graph structure.

- Exploring/discovering datasets for (re)use
- Defining a "minimal" set of information needed to build a "profile"
- Building tools that will automatically generate/validate metadata models

Metadata Model Classification | Information groups

A dataset can contain four main Information groups:

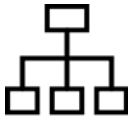
Resource

Actual raw data e.g. JSON, CSV, SPARQL endpoint

Tag

Descriptive knowledge: simple textual tags, semantically rich controlled terms

Causalities, roads, roadsafety, accidents



Group

Organizational units that share common semantics, a cluster or curation based on shared themes/categories

Health, Travel and Transport

Organization

Clustering or curation solely based on associations with specific administration parties

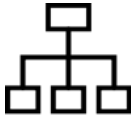
Department for Transport, Ministère de l'Écologie

Metadata Model Classification | Information types

A dataset can contain the following information types:

Resource

Tag



Group

Organization

General information

title, description, id

Statistical information

max_value, uniques, average

Access information

URL, license_title, license_id

Quality information

rating, availability, freshness

Ownership information

author, maintainer_email

Geospatial information

bbox, layers

Provenance information

version, creation_date, update_date

Temporal information

coverage_from, coverage_to

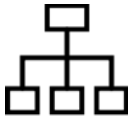
EURECOM

Mapping Datasets Models | Information groups

1 Map the information groups [resource, tag, group, organization]

Resource

Tag



Group

Organization

CKAN	DKAN	POD	DCAT	VoID	Schema.org	Socrata
resources	resources	distribution	Dcat:Distribution	Void:Dataset →void:dataDump	Dataset:distribution	attachments
tags	tags	keyword	Dcat:Dataset → :keyword	Void:Dataset → :keyword	creativeWork:keywords	tags

Mapping Datasets Models | Information types

2 Map the information types [general, ownership, provenance, etc.]

CKAN maintainer_email

DKAN maintainer_email

POD ContactPoint -> hasEmail

DCAT dcat:Dataset -> dct:creator -> foaf:Person:mbox

VoID void:Dataset -> dct:creator -> foaf:Person:mbox

Schema.org CreativeWork:producer -> Person:email


```
"license_title": "Creative Commons Attribution Share-Alike",
"maintainer": "DBpedia Team - http://wiki.dbpedia.org/Imprint",
"relationships_as_object": [],
"private": false,
"maintainer_email": "dbpedia-discussion@lists.sourceforge.net",
"revision_timestamp": "2013-10-10T20:29:25.819614",
"id": "dcc6715c-bf94-4a89-bbf3-35933da795a5",
"metadata_created": "2007-07-04T17:01:04",
"owner_org": "a27ebdd0-6059-46e5-b59c-5aaee0168568",
"metadata_modified": "2013-10-10T20:29:25.819614",
"author": "DBpedia Team - http://wiki.dbpedia.org/Imprint",
"author_email": "dbpedia-discussion@lists.sourceforge.net",
"state": "active",
"version": "2010-09-02 (3.7)",
"license_id": "cc-by-sa",
"type": "dataset",
"resources": [
  {
    "mimetype": "",
    "cache_url": "",
    "mimetype_inner": "",
    "hash": "",
    "description": "Download page (N-Triples, bz2-compressed)",
    "format": "",
    "url": "http://wiki.dbpedia.org/Downloads",
    "cache_last_updated": null,
    "tracking_summary": {
      "total": 173,
      "recent": 5
    },
    "revision_timestamp": "2012-08-02T09:16:44.343222",
    "webstore_url": "",
    "resource_group_id": "fd2b6dd2-d32d-bc1c-616f-6aff1472cf7d",
    "state": "active",
    "last_modified": null,
    "webstore_last_updated": null,
    "position": 0,
    "revision_id": "d0b5f9c2-f394-4c40-b9fa-c43341589586",
    "size": null,
    "id": "07e05667-36d3-4e7f-8479-21a14d952213",
    "resource_type": "file",
    "name": ""
```

Harmonized Dataset Model (HDL)

- There is an abundance of extensions and application profiles that try to fill in those gaps, but they are usually domain specific addressing specific issues like geographic or temporal information

HDL

5 Classes

61 properties

+ 3 Controlled field values

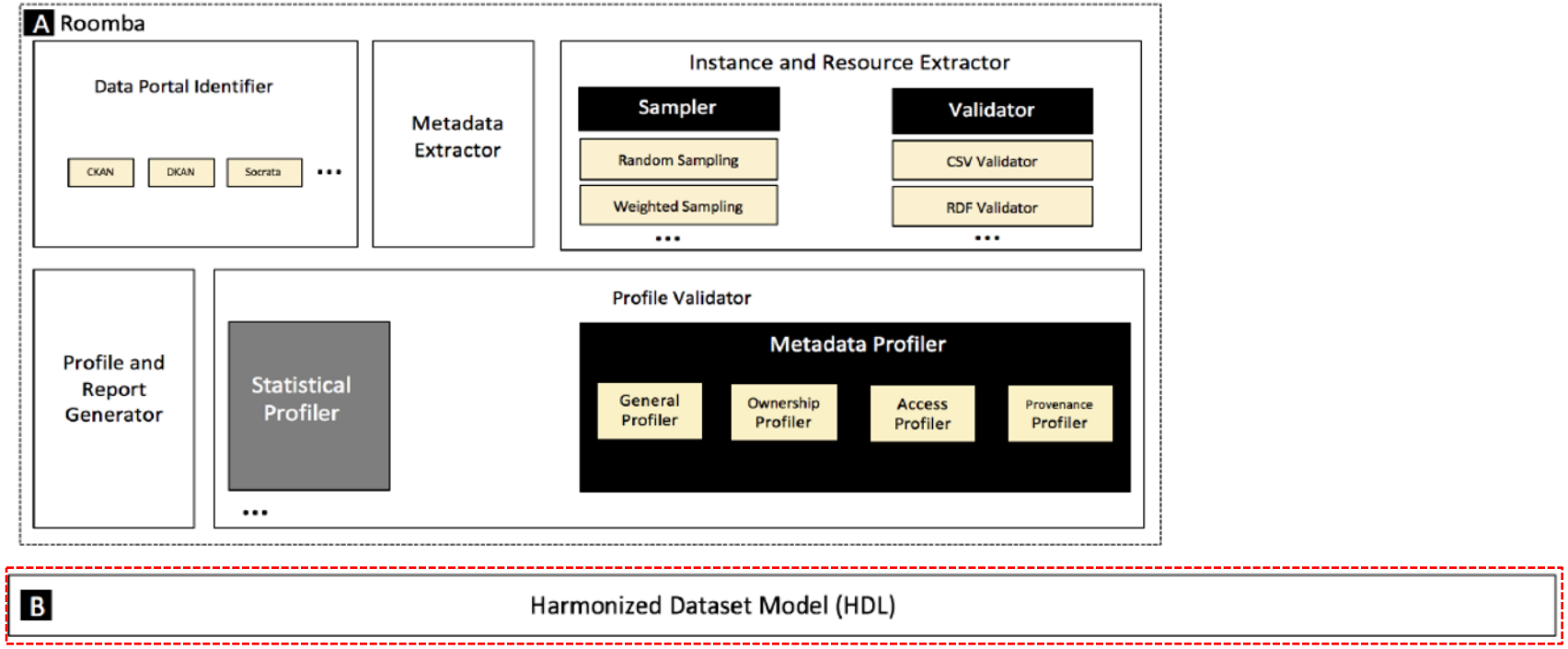
`resource_type: direct, accessible_bitstream, file.upload, api, visualization, code, documentation`

Flashback | Revisiting Scenario

- **Paul** knows what are the major dataset models out there, and what kind of metadata data owners need to fully represent their dataset 
- **Paul** will be able to use HDL as a basis to extend and present the datasets he controls 
- **Paul** can use HDL and the proposed mappings as a basis to extend Roomba to support various dataset models like DKAN or Socrata 
- **Dan** will be able to make fast decisions whether the dataset examined is suitable or not by examining the rich datasets metadata presented in HDL 

Proposed Framework

- Dataset Maintenance & Discovery
  - HDL



Dataset Profiling

- Data profiling is the process of creating descriptive information and collect statistics about that data. It is a cardinal activity when facing an unfamiliar dataset
- Profiles reflect the importance of datasets without the need for detailed inspection of the raw data

Statistical Profiling
[Abedjan 2014 , Makela 2014]

Metadata Profiling

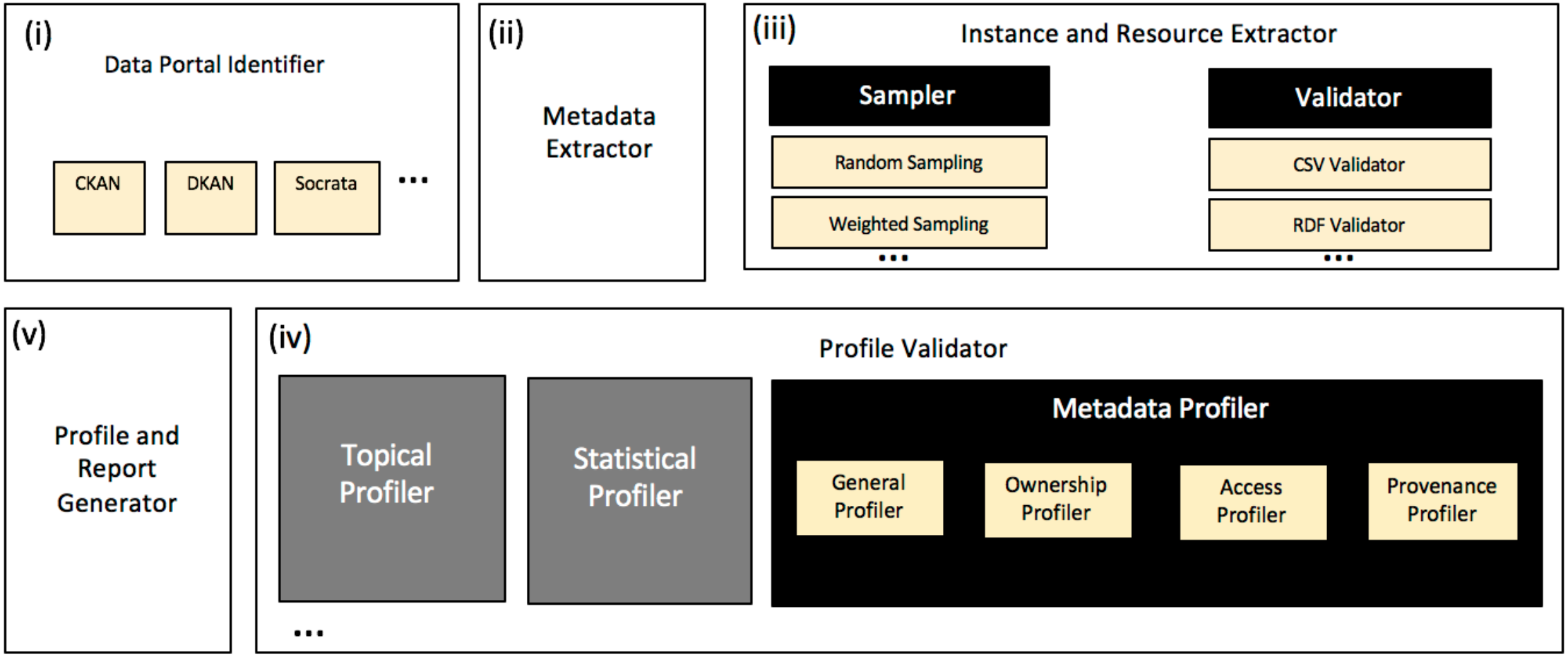
Topical Profiling
[Bohm 2012, Fetahu 2014]

Metadata Profiling | Related Work

- Flemming's data quality assessment tool provides metadata assessment based on manual user input [Flemming 2010]
- ODI certificate* provides descriptions of the published data quality in plain English based on an extensive survey filled by the publisher
- Project Open Data Dashboard** tracks and measures how US governments implement Open Data principles
- The Datahub Validator*** gives an overview of data sources cataloged on The Datahub

* https://certificates.theodi.org/en/
** https://project-open-data.cio.gov/
*** http://validator.lod-cloud.net/

Roomba | An Extensible Pipeline to Generate and Validate Profiles

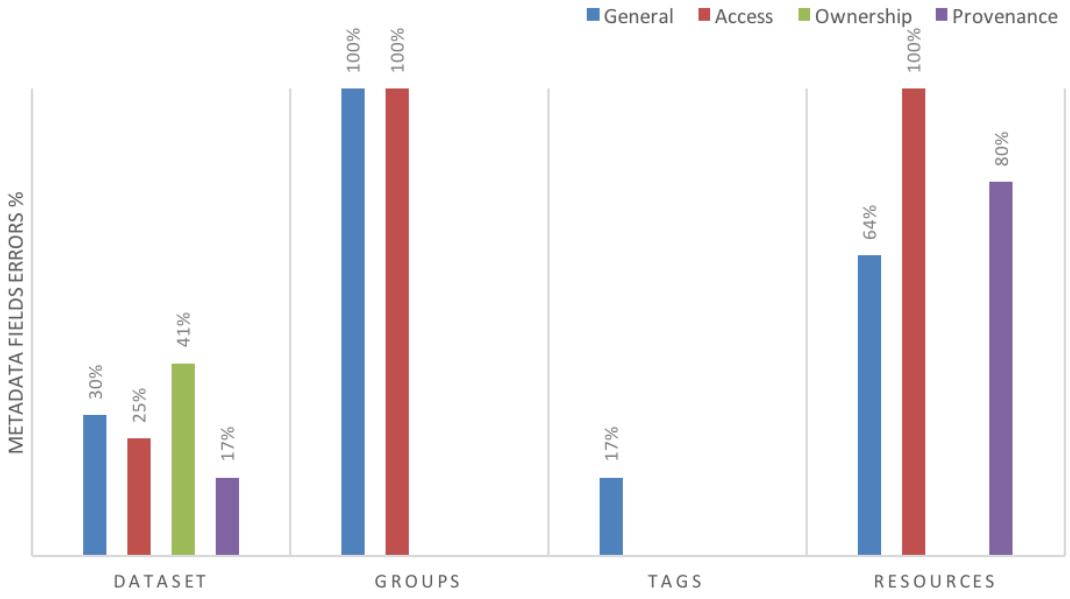


https://github.com/ahmadassaf/opendata-checker/


12:17:03
...projects/Semantic Web, IR and Data Analysis/Dataset Crawler
master

$

LOD Cloud State | Top Metadata Errors



Metadata Field		Error %	Section	Error Type	Auto Fix
General	group	100%	Dataset	Missing	-
	vocabulary_id	100%	Tag	Undefined	-
	url-type	96.82%	Resource	Missing	-
	mimetype_inner	95.88%	Resource	Undefined	Yes
	hash	95.51%	Resource	Undefined	Yes
	size	81.55%	Resource	Undefined	Yes
Access	cache_url	96.9%	Resource	Undefined	-
	webstore_url	91.29%	Resource	Undefined	-
	license_url	54.44%	Dataset	Missing	Yes
	url	30.89%	Resource	Unreachable	-
	license_title	16.6%	Dataset	Undefined	Yes
Provenance	cache_last_updated	96.91%	Resource	Undefined	Yes
	webstore_last_updated	95.88%	Resource	Undefined	Yes
	created	86.8%	Resource	Missing	Yes
	last_modified	79.87%	Resource	Undefined	Yes
	version	60.23%	Dataset	Undefined	-
Ownership	maintainer_email	55.21%	Dataset	Undefined	-
	maintainer	51.35%	Dataset	Undefined	-
	author_email	15.06%	Dataset	Undefined	-
	organization_image_url	10.81%	Dataset	Undefined	-
	author	2.32%	Dataset	Undefined	-

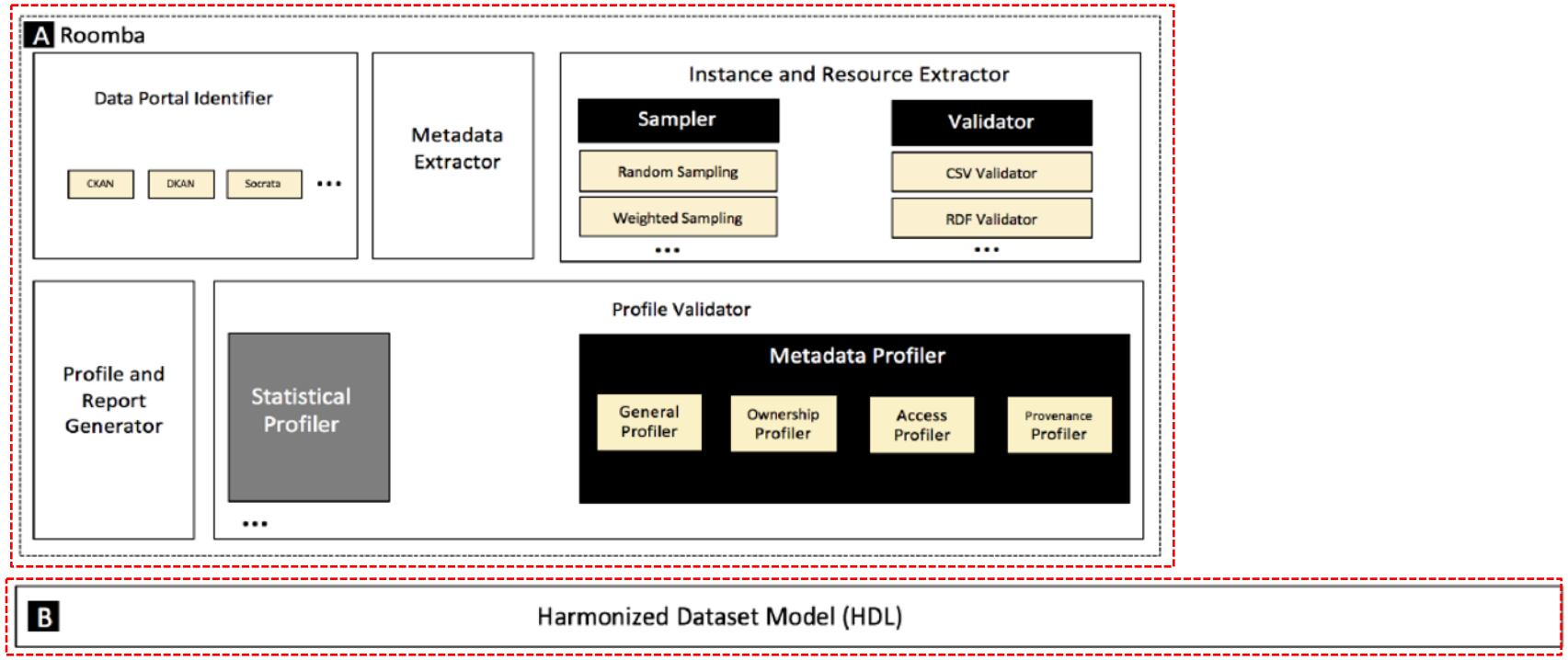
Table 1: Top metadata fields error % by information type

Flashback | Revisiting Scenario

- **Paul** can use Roomba to automatically fix datasets metadata issues, and notify the datasets owners of the other issues to be manually fixed 
- **Paul** will be able now to identify spam datasets. In addition to that, data available in his portal will now have rich semantic information attached to it 
- **Dan** will be able to have direct access to rich and high-quality dataset descriptions generated by Roomba 

Proposed Framework

- Dataset Maintenance & Discovery
  - HDL
  - Roomba



Objective Data Quality Assessment

- Data quality assessment is the process of evaluating if a piece of data meets the consumers need in a specific use case
- In [Zaveri 2012] lots of the defined quality dimensions are not objective e.g., like low latency, high throughput or scalability
  - In addition, there were some missing objective indicators vital to the quality of LOD e.g., indication of the openness of the dataset

We propose an objective framework to evaluate the quality of Linked Data sources.

Objective Data Quality Assessment

10 Quality Measures

Availability

Licensing

Freshness

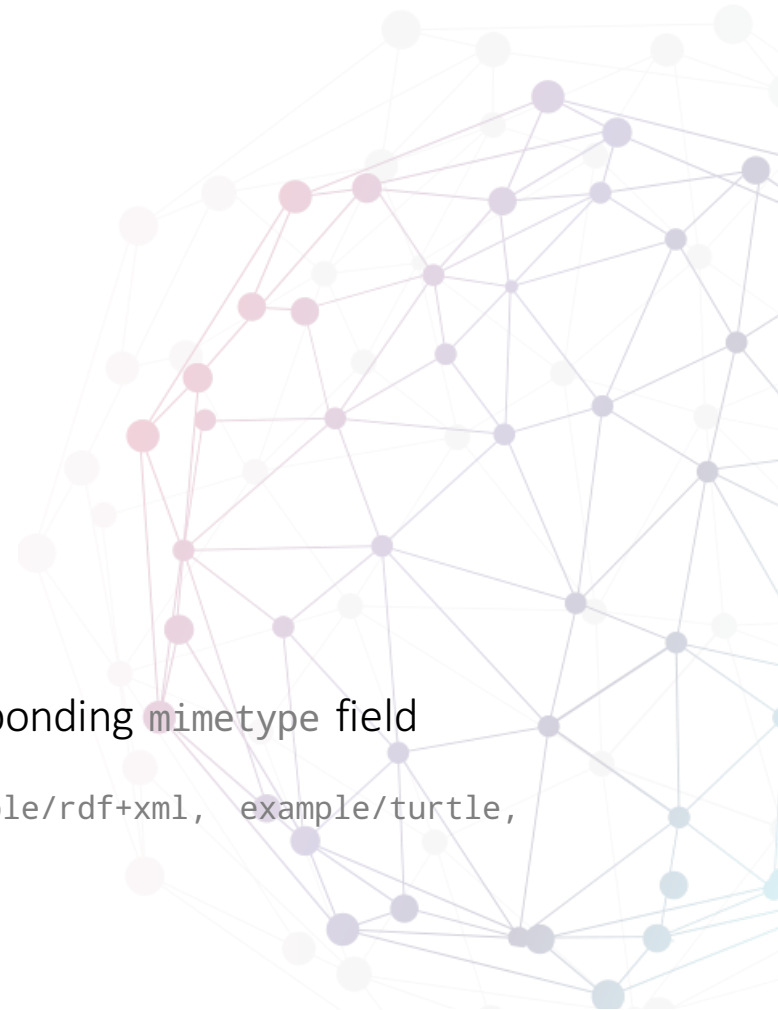
64 Quality Indicators

Check the resources format field for meta/void value

Check the format and `mimetype` fields for resources

Check if the content-type extracted from the a valid `HTTP` request is equal to the corresponding `mimetype` field

Check if there is at least one resource with a format value corresponding to one of `example/rdf+xml`, `example/turtle`, `example/ntriples`, `example/x-quads`, `example/rdfa`, `example/x-trig`



Objective Data Quality Assessment Tools

- Large amount of those indicators can be examined automatically from attached datasets metadata found in data portals Extensive survey on Linked Data quality tools
  - Models and ontologies quality covered in the literature [Kontokostas 2014, Ruckhaus 2014, Cherix 2014]
  - Lack in automatic tools to check the dataset quality especially in its completeness, licensing and provenance measures

Quality Score Calculation

- The quality indicator score is an error ratio based on a ratio between the number of violations **V** and the total number of instances where the rule applies **T** multiplied by the specified weight for that indicator

$$Q = (V/T) * weight$$

- A quality measure score should reflect the alignment of the dataset with respect to the quality indicators
- The quality measure score **M** is calculated by dividing the weighted quality indicator scores sum by the total number of instances in its context

$$M = 1 - ((\sum_{i=1}^{n} Q_i)/ \mid Q_i \mid)$$

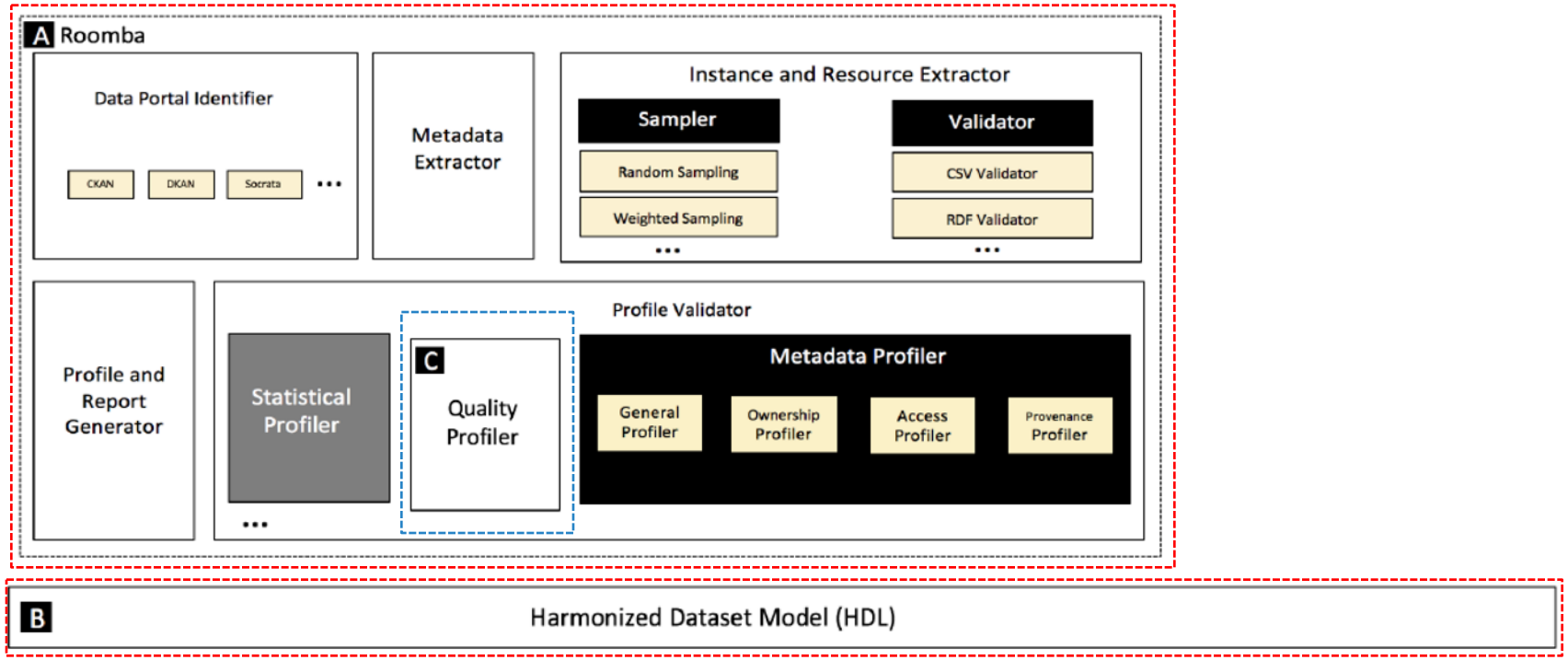
Roomba Experiments & Evaluation

- Profiling **correctness** and **completeness**
- To analyze the completeness, we manually constructed a synthetic set of profiles
  - Incorrect resources mimtype or size
  - Invalid number of tags or resources
  - Correct normalization of license information: license_id or license_title
  - Syntactically invalid emails and urls : author_email or maintainer_email

Tool\Indicator	1	2	3	4	5	6	7	8	9	18	19	20	21	22	23	24	25	26	27	28	29	37	38	39	40	44	45	46	63	64
LOV	●		●	●	◗		◗		●	●		●	◗									●		◗		●		●		
Data.gov	●				◗	●			●			●				◗	●							●		◗				
Roomba	●	●	●	●	●	●	●		●	●	●	●	●	●		●	●	●	●	●		●		●	●	●		●		

Flashback | Revisiting Scenario, Proposed Framework

- Dataset Maintenance & Discovery
  - HDL
  - Roomba
- Dataset Quality



- **Paul** will be able now to identify low quality datasets 
- **Paul** will be able to generate a quality score for his datasets 
- **Dan** will be able to have access to cleaner, richer set of datasets 

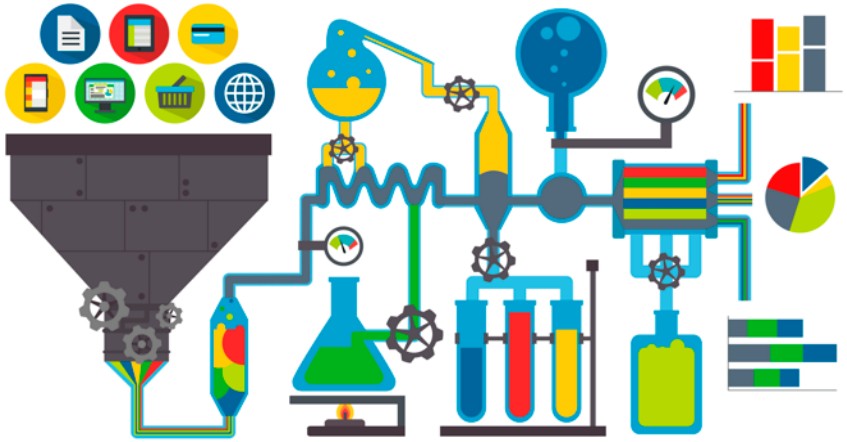
PART II

Towards Enriched Enterprise Data

Research Challenges

- Dataset Integration and Enrichment

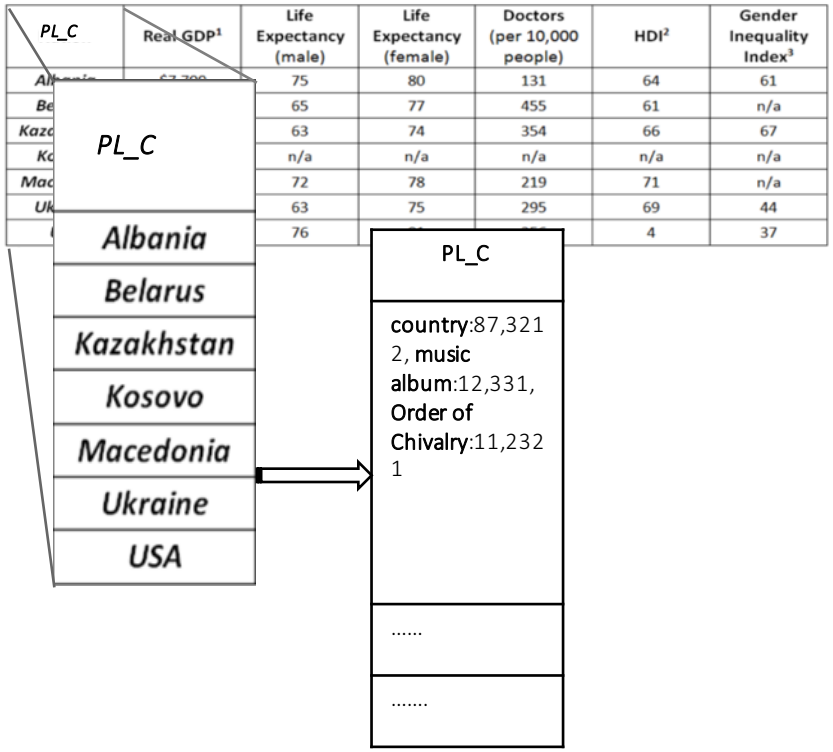
Enterprise Knowledge Base Services



- Search service (ranked, contextual)
- Entities and properties recommendation
- Enhancing schema matching

Data Enrichment

- Annotates datasets based on the semantics of the data at the instance level
- We represent each instance at the cell level with a set of types retrieved from our service
- The column is represented now as a vector of rich types and their corresponding confidence
- Select the most common type to annotate and label the column



Data Enrichment | Properties Ranking

dbo:wikiPageID
dbo:wikiPageRevisionID
dbp:constituencyWestminster
dbp:country
dbp:dialCode
dbp:hasPhotoCollection
dbp:latitude
dbp:longitude
dbp:metropolitanBorough
dbp:metropolitanCounty
dbp:officialName
dbp:osGridReference
dbp:population
dbp:postTown
dbp:postcodeArea
dbp:postcodeDistrict
dbp:region
dbp:staticImageName
dbp:wordnet_type
dct:subject

georss:point
rdf:type


```
"label": {
  "uri": "http://dbpedia.org/property/label",
  "count": 100
},
"description": {
  "uri": "http://dbpedia.org/property/description",
  "count": 100
},
"type": {
  "uri": "http://dbpedia.org/property/type",
  "count": 100
},
"area": {
  "uri": "http://dbpedia.org/property/area",
  "count": 95
},
"population": {
  "uri": "http://dbpedia.org/property/population",
  "count": 93
},
"province": {
  "uri": "http://dbpedia.org/property/state/province",
  "count": 1
},
"founded": {
  "uri": "http://dbpedia.org/property/founded",
  "count": 2
}
},
"infoboxless": [],
"Unmapped_Properties": {
  "weather": 1,
  "localTime": 1,
  "university": 1,
  "seeResultsAbout": 1,
  "collegesAndUniversities": 1,
  "pointsOfInterest": 1,
  "unemploymentRate": 1
}
}
```


Bootle

Town in England

Bootle is a town within the Metropolitan Borough of Sefton in Merseyside. The town was formerly known as Bootle-cum-Linacre and has a total resident population of 77,640. Wikipedia

Weather: 12°C, Wind SE at 21 km/h, 90% Humidity

Getting there: 4 h 30 min flight, around €205. View flights

Metropolitan county: Merseyside

Local time: Tuesday 9:29 PM

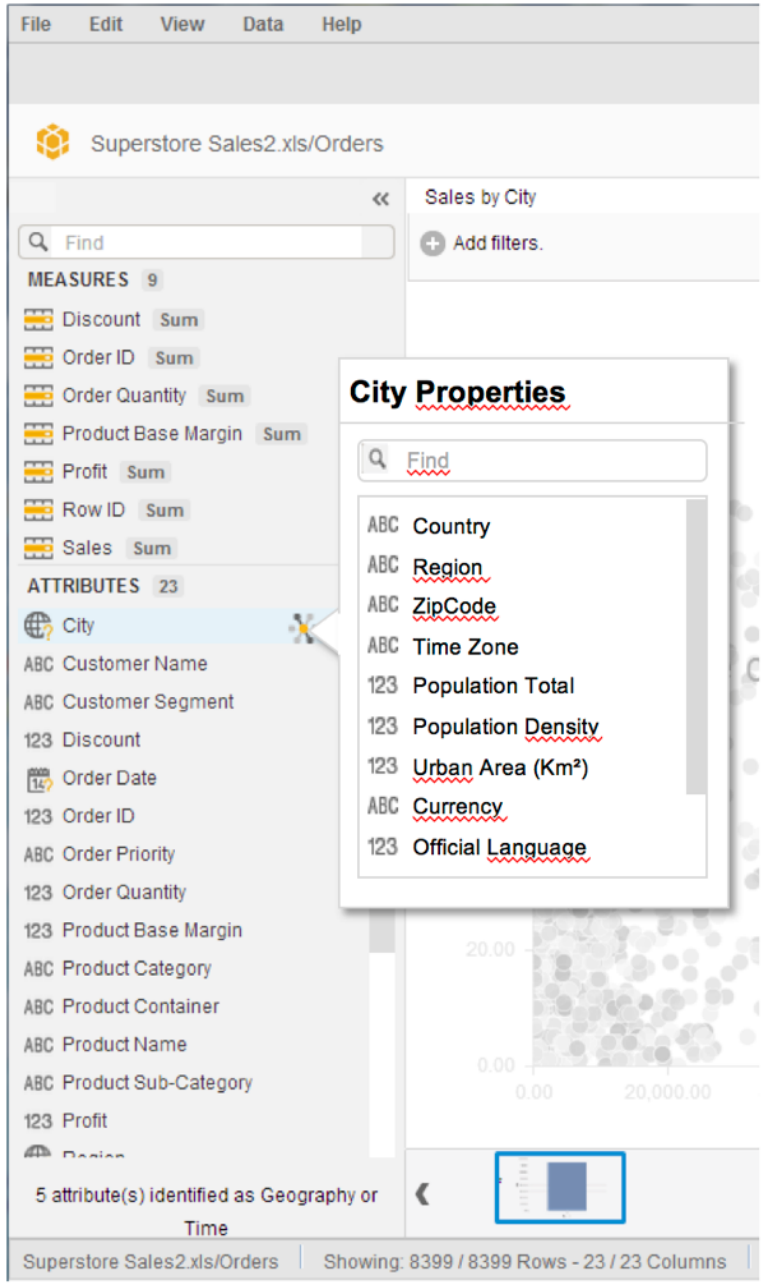
Dialling codes: 0151

University: Hugh Baird College

Data Enrichment | Properties Ranking

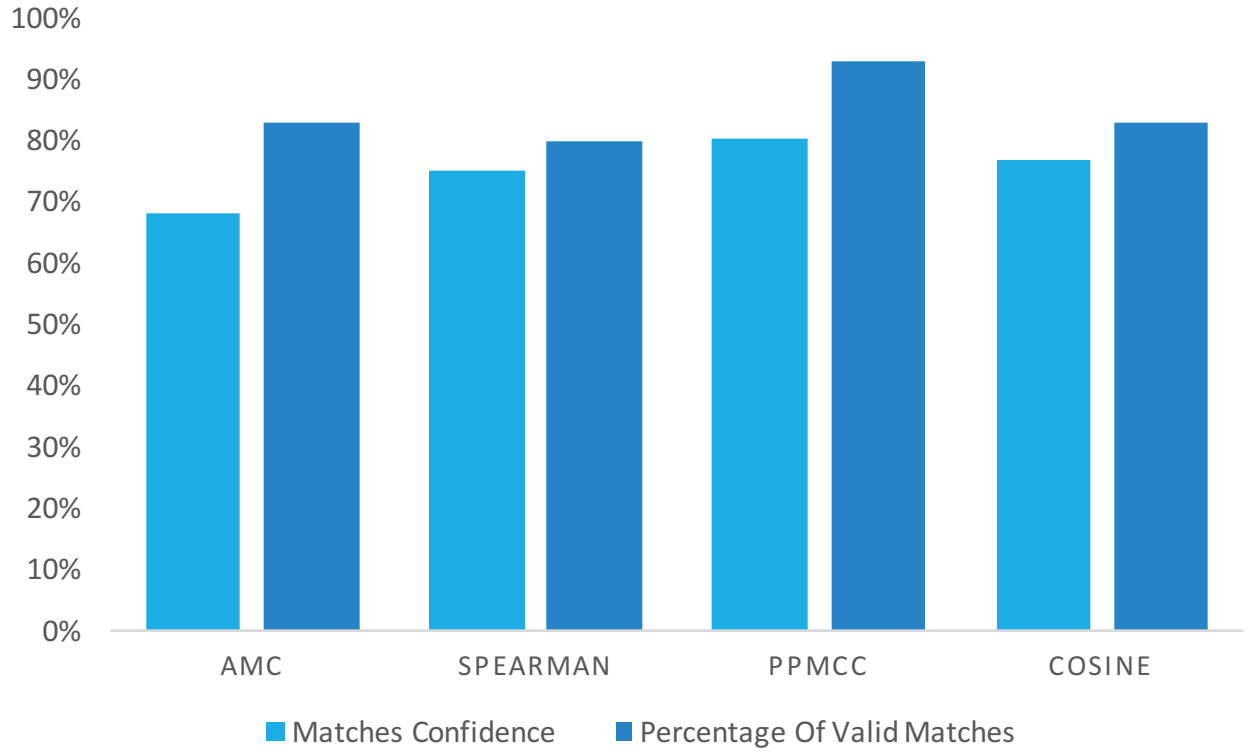
Algorithm 1 Google Knowledge Panel reverse engineering algorithm

```
INITIALIZE equivalentClasses(DBpedia, Freebase) AS vectorClasses
Upload vectorClasses for querying processing
Set n AS number-of-instances-to-query
for each conceptType ∈ vectorClasses do
    SELECT n instances
    listInstances ← SELECT-SPARQL(conceptType, n)
    for each instance ∈ listInstances do
        CALL http://www.google.com/search?q=instance
        if knowledgePanel exists then
            SCRAP GOOGLE KNOWLEDGE PANEL
        else
            CALL http://www.google.com/search?q=instance + conceptType
            SCRAP GOOGLE KNOWLEDGE PANEL
        end if
        gkpProperties ← GetData(DOM, EXIST(GKP))
    end for
    COMPUTE occurrences for each prop ∈ gkpProperties
end for
gkpProperties
```



Improving Schema Matching

- Added the semantic matching to the Auto Mapping Core (AMC)
- Evaluated against data coming from two SAP systems: The Event Tracker and Travel Expense Manager
- Increased the overall confidence score with an average of 11% and the number of valid matches found with an average of 10%



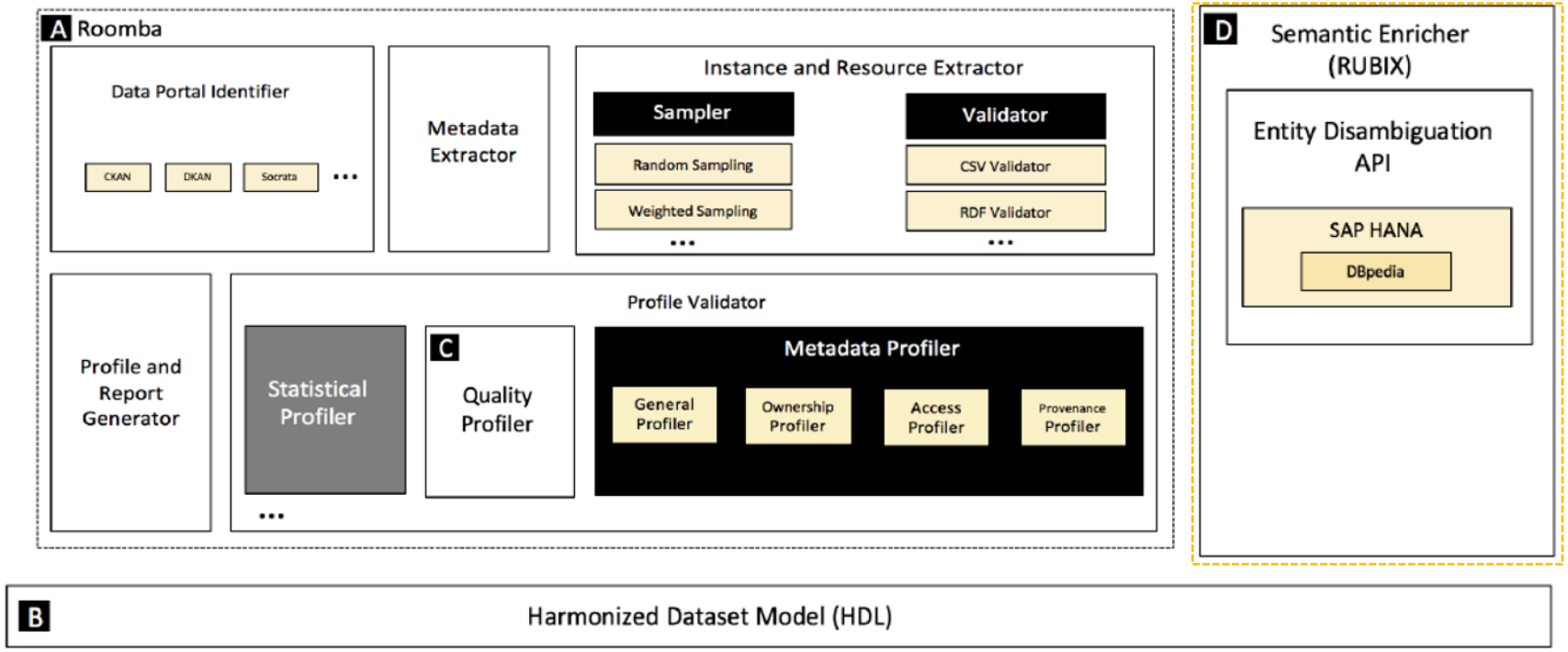
Flashback | Revisiting Scenario

Dan now has access to various datasets that he found matching his query to the portal administered by **Paul**

- Having imported those dataset into Lumira, **Dan** will be also able to use the internal knowledge base to apply various semantic enrichments on this data 
- Dan will be also able to use the schema matching services to find and merge those datasets in his reports 
- Dan will be able to annotate and enrich his reports with external data 

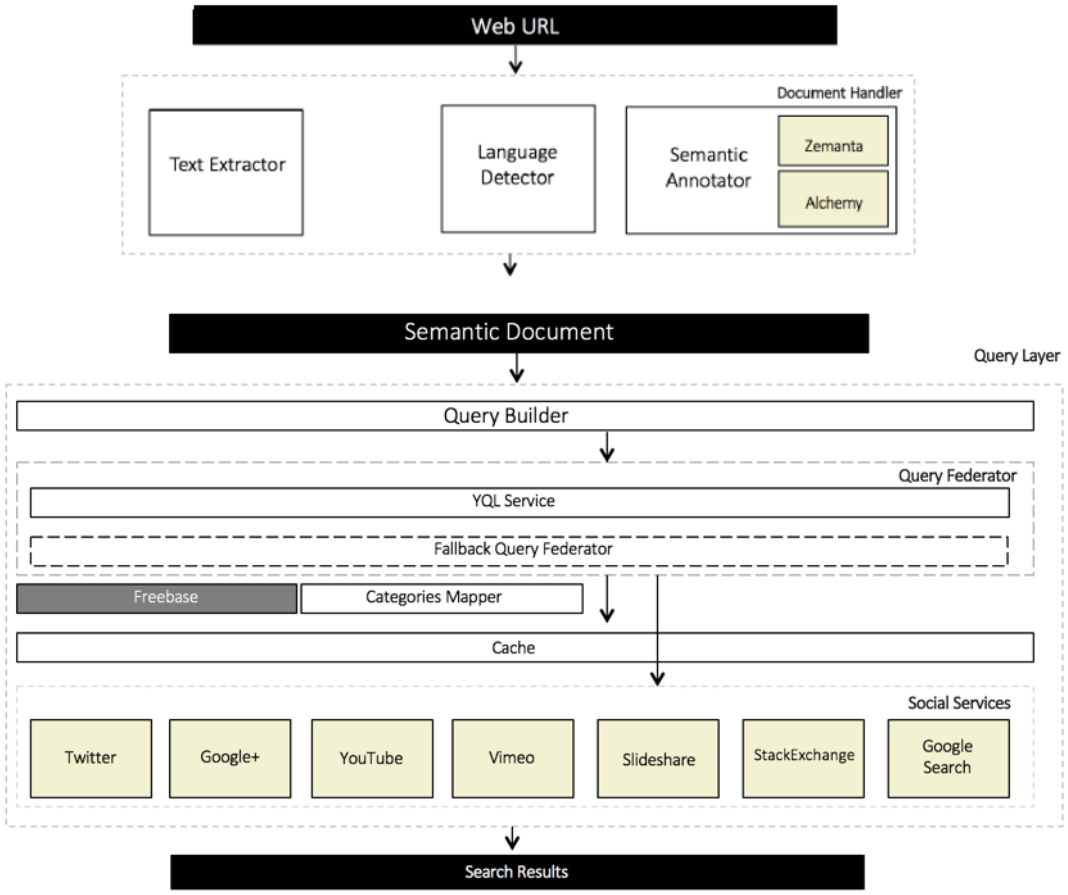
Proposed Framework

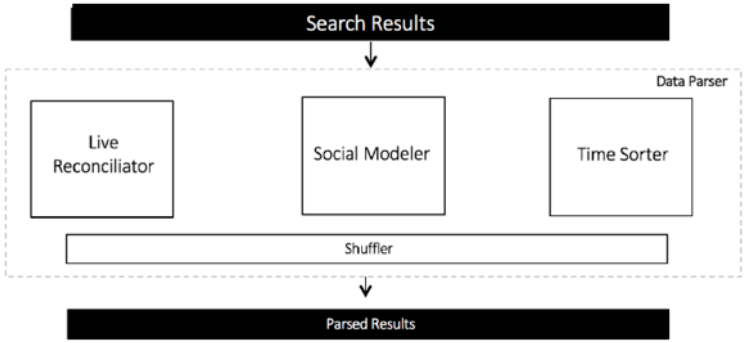
- Dataset Integration and Enrichment
  - RUBIX



Semantic Social News Aggregation | SNARC

- SNARC is a service that extracts the semantic context of documents in order to recommend related content from the web and social media





Semantic Social News Aggregation | SNARC

Drink Driving Laws In The UK

LANGUAGE: ENGLISH LAW_CRIME

Extracted Keywords

DRIVER AND 100%

TRAFFIC COLLISION 100%

LABOUR PARTY 100%

UNITED KINGDOM 100%

BREATH TEST 100%

ALLSTATE 100%

ROAD SIDE 80.0%

ALCOHOL 74.8%

PRESCRIBED LEGAL 66.3%

PRELIMINARY ROAD 62.2%

Extracted Entities

BREATH TEST FIELDTERMINOLOGY

DRIVER AND VEHICLE LICENSING AGENCY GOVERNMENT GOVERNMENT_AGENCY

TRAFFIC COLLISION

MAGISTRATES' COURT (ENGLAND AND WALES)

ENGLAND AND WALES

BREATHALYZER

UNITED KINGDOM LOCATION COUNTRY

SHOW ALL CLOSE

@Wiseman Lawyers High Range 0.162 DUI Drink Driving With Collision

4 hours ago

@JonesChevy1 Available Rear Cross Traffic Alert in the Chevy Equinox helps the driver avoid a collision with an approaching vehi… https://t.co/8rTTKwGTnY

1 day ago

via Intermark Social Media

DVLAgov

5 years ago

@sandiegojournal Suspected DUI driver arrested after injuring motorcyclist: A head-on collision between a truck and a motorcycle… https://t.co/IrdwvWdNgd

17 hours ago

via dlvr.it

My Arrest record: 2015

2015 - My Arrest record - 2015 (all at Exeter Crown Court & Magistrates court , Exeter, UK)(public record) 27/12/14 - Arrested under suspicion of driving under the influence of drink/drugs. 12/1/15 - Arrested and interview for dangerous driving/ 3 bad tyres.s 3/3/15 - ...

5 hours ago

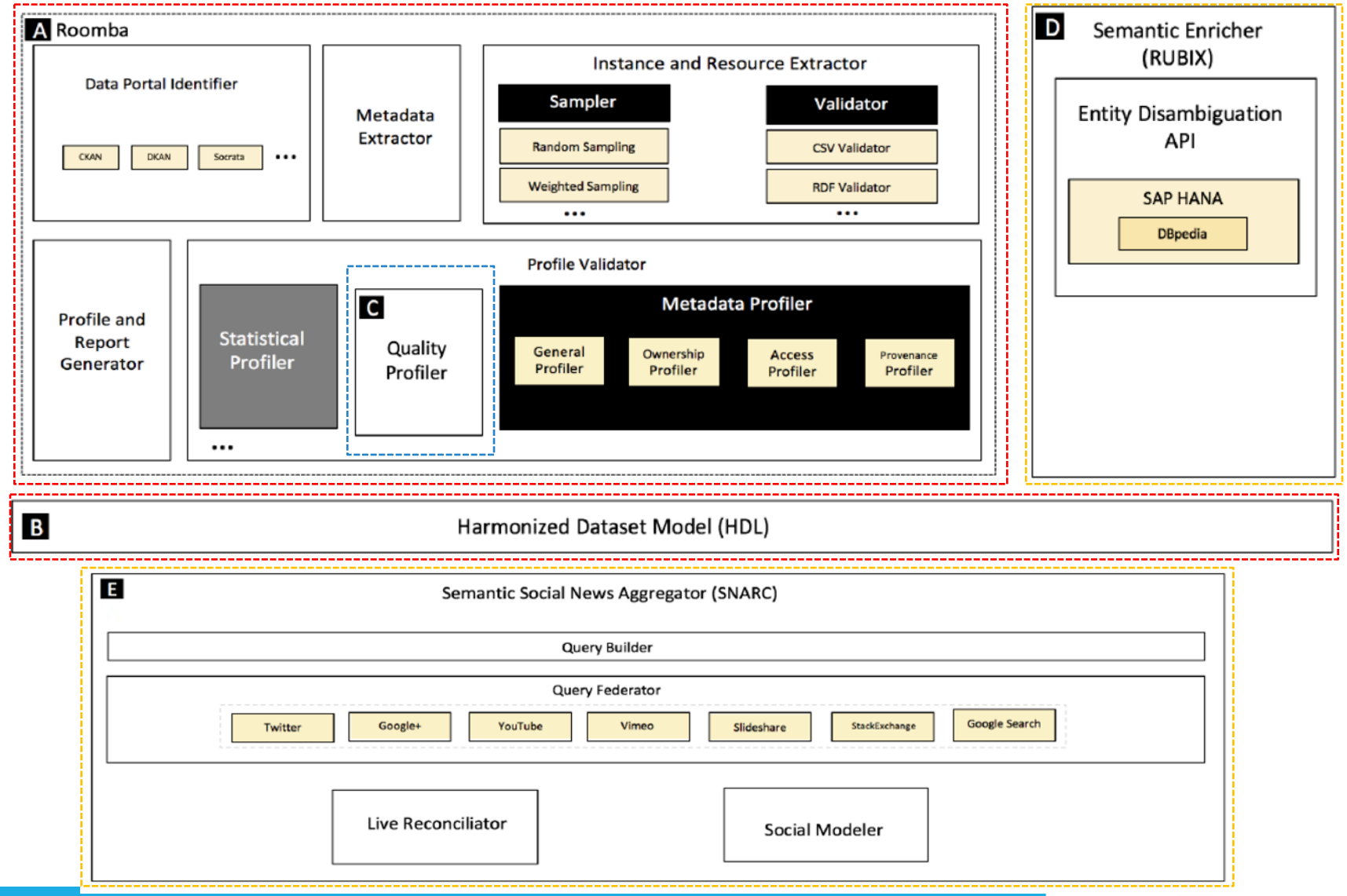
@Andrew Wiseman High Range 0.162 DUI Drink Driving With Collision

4 hours ago

EURECOM

Proposed Framework

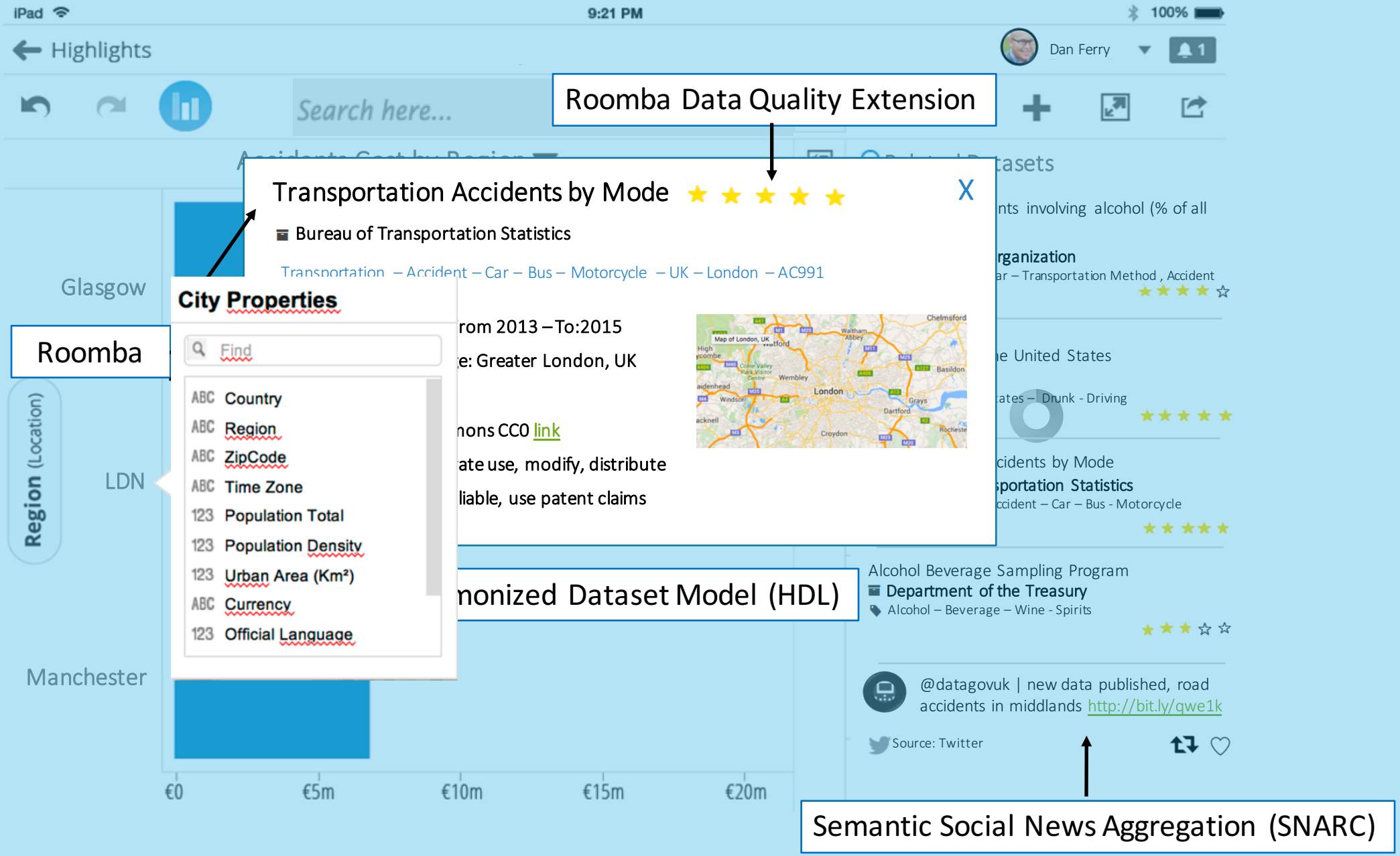
- Dataset Maintenance & Discovery
  - HDL
  - Roomba
- Dataset Quality
- Dataset Integration and Enrichment
  - RUBIX
  - SNARC



Revisit: Use-case Scenario

- **Dan** receives a memo from his management to create a report comparing the number of car accidents that occurred in **France** for **this year**, to its counterpart in the **United Kingdom (UK)**.

- **Dan** asked to highlight accidents related to **illegal consumption of alcohol** in both countries.


iPad
9:21 PM
100%
Highlights
Dan Ferry
Search here...
Roomba Data Quality Extension
Transportation Accidents by Mode
Bureau of Transportation Statistics
Transportation – Accident – Car – Bus – Motorcycle – UK – London – AC991
X
City Properties
Find
ABC Country
ABC Region
ABC ZipCode
ABC Time Zone
123 Population Total
123 Population Density
123 Urban Area (Km²)
ABC Currency
123 Official Language
Roomba
Region (Location)
Glasgow
LDN
Manchester
rom 2013 –To:2015
e: Greater London, UK
nons CC0 link
ate use, modify, distribute
liable, use patent claims
monized Dataset Model (HDL)
Alcohol Beverage Sampling Program
Department of the Treasury
Alcohol – Beverage – Wine - Spirits
@datagovuk | new data published, road accidents in middlands http://bit.ly/qwe1k
Source: Twitter
€0
€5m
€10m
€15m
€20m
Semantic Social News Aggregation (SNARC)

Future Work | Data Profile Representation

- [S] HDL as an ontology
- [S] Extend HDL with a set of enumerations as values to ensure a unified fine-grained representation of a dataset
- [M] Evaluate the usage of HDL in data portals and their effect on data portal traffic, number of downloads, number of datasets published, social popularity, etc.

Future Work | Automatic Dataset Profiling

- [S] Roomba support to other data portal types like DKAN or Socrata
- [M] Integrating statistical and topical profilers to allow generation of full comprehensive profiles
- [M] Add the ability to correct the rest of the metadata either automatically or through intuitive manually-driven interfaces
- [M] Investigate various resource sampling techniques for profiling purposes (random, weighted, centrality)

Future Work | Objective Linked Data Quality

- [M] Integrate tools assessing models quality in addition to syntactic checkers
- [M] Suggest weights to those indicators which will result in a more objective quality calculation process
- [L] Monitor the evolution of datasets quality (indicators and measures)

Future Work | Enterprise Data Integration

- [S] Integrating additional linked open data sources of semantic types such as YAGO
- [S] Reverse engineer other knowledge graphs like Bing and compare the various results
- [M] Evaluate our matching results against instance-based ontology alignment benchmarks
- [L] Better evaluation of SNARC as a I) dataset search tool over social media II) social ranking tool for datasets popularity

Publications

Journals

- Ahmad Assaf, Raphaël Troncy and Aline Senart: Towards An Objective Assessment Framework for Linked Data Quality. International Journal On Semantic Web and Information Systems, Minor revision, 2015

Conference Demo, Poster and Challenges

- Ahmad Assaf, Raphaël Troncy and Aline Senart: Automatic Validation, Correction and Generation of Dataset Metadata - Enhancing Dataset Search and Spam Detection. In 24th International World Wide Web Conference (WWW 2015), Demo Track, May 2015, Florence, Italy
- Ahmad Assaf, Ghislain Atemezing, Raphaël Troncy and Elena Cabrio: What are the important properties of an entity? Comparing users and knowledge graph point of view. In 11th Extended Semantic Web Conference (ESWC 2014), Demo Track, May 2014, Heraklion, Crete
- Ahmad Assaf, Aline Senart and Raphaël Troncy: SNARC - An Approach for Aggregating and Recommending Contextualized Social Content. In 10th Extended Semantic Web Conference (ESWC 2013), Sattelite Events, May 2013, Montpellier, France.

1st Prize Winner of the AI Mashup Challenge

Publications

Workshops

- Ahmad Assaf, Raphaël Troncy and Aline Senart: What's up LOD Cloud - Observing The State of Linked Open Data Cloud Metadata. In 2nd Workshop on Linked Data Quality (LDQ), May 2015, Portoroz, Slovenia

   Best paper award

- Ahmad Assaf, Raphaël Troncy and Aline Senart: HDL - Towards A Harmonized Dataset Model for Open Data Portals. In 2nd International Workshop on Dataset PROFIling & fEderated Search for Linked Data (PROFILES), May 2015, Portoroz, Slovenia
- Ahmad Assaf, Raphaël Troncy and Aline Senart: An Extensible Framework to Validate and Build Dataset Profiles. In 2nd International Workshop on Dataset PROFIling & fEderated Search for Linked Data (PROFILES), May 2015, Portoroz, Slovenia.

   Best paper award

- Ahmad Assaf, Aline Senart and Raphaël Troncy: Data Quality Principles in the Semantic Web. In International Workshop on Data Quality Management and Semantic Technologies (DQMST), July 2012, Palermo, Italy
- Ahmad Assaf, Eldad Louw, Aline Senart, Corentin Follenfant Raphaël Troncy and David Trastour: RUBIX: a Framework for Improving Data Integration with Linked Data. In 1st International Workshop on Open Data (WOD), June 2012, Nantes, France.

References

1. Annika Flemming. **Quality Characteristics of Linked Data Publishing Datasources**. Master's thesis, Humboldt-Universitt zu Berlin, 2010
2. Amrapali Zaveri, Anisa Rula, Andrea Maurino, Ricardo Pietrobon, Jens Lehmann, and Soren Auer. **Quality Assessment Methodologies for Linked Open Data**. Semantic Web Journal, 2012
3. Tim Berners-Lee. **Linked Data - Design Issues**. W3C Personal Notes, 2006 http://www.w3.org/DesignIssues/LinkedData
4. Dimitris Kontokostas, Patrick Westphal, Soren Auer, Sebastian Hellmann, Jens Lehmann, Roland Cornelissen, and Amrapali Zaveri. **Test-driven Evaluation of Linked Data Quality**. In 23rd International Conference on World Wide Web (WWW'14), 2014
5. Edna Ruckhaus, Oriana Baldizan, and Maria-Esther Vidal. **Analyzing Linked Data Quality with LiQuate**. In 11th European Semantic Web Conference (ESWC), 2014
6. Didier Cherix, Ricardo Usbeck, Andreas Both, and Jens Lehmann. **CROCUS: Cluster-based ontology data cleansing**. In 2nd International Workshop on Semantic Web Enterprise Adoption and Best Practice, 2014
7. Ziawasch Abedjan, Tony Gruetze, Anja Jentzsch, and Felix Naumann. **Profiling and mining RDF data with ProLOD++.** In 30th IEEE International Conference on Data Engineering (ICDE)
8. Eetu Makela. **Aether - Generating and Viewing Extended VoID Statistical Descriptions of RDF Datasets**. In 11th European Semantic Web Conference (ESWC), Demo Track, Heraklion, Greece, 2014
9. Christoph Bohm, Gjergji Kasneci, and Felix Naumann. **Latent Topics in Graph structured Data**. In 21st ACM International Conference on Information and Knowledge Management (CIKM)
10. Besnik Fetahu, Stefan Dietze, Bernardo Pereira Nunes, Marco Antonio Casanova, Davide Taibi, and Wolfgang Nejdl. **A Scalable Approach for Efficiently Generating Structured Dataset Topic Profiles**. In 11th European Semantic Web Conference (ESWC), 2014

THANK YOU

Ahmad Assaf

http://ahmadassaf.com/

@ahmadaassaf

http://github.com/ahmadassaf