DAGOBAH

An End-to-End Context-Free Tabular Data Semantic Annotation System

Yoan Chabot
Orange Labs
@yoan_chabot

Thomas Labbé
Orange Labs
@tau_labbe

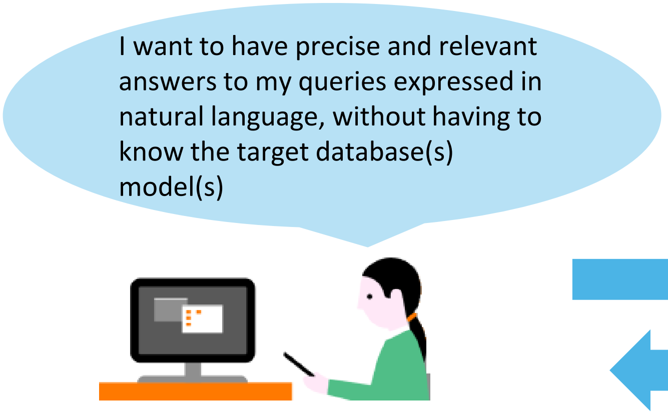
Jixiong Liu
Orange Labs

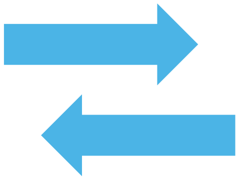
Raphaël Troncy
EURECOM
@rtroncy

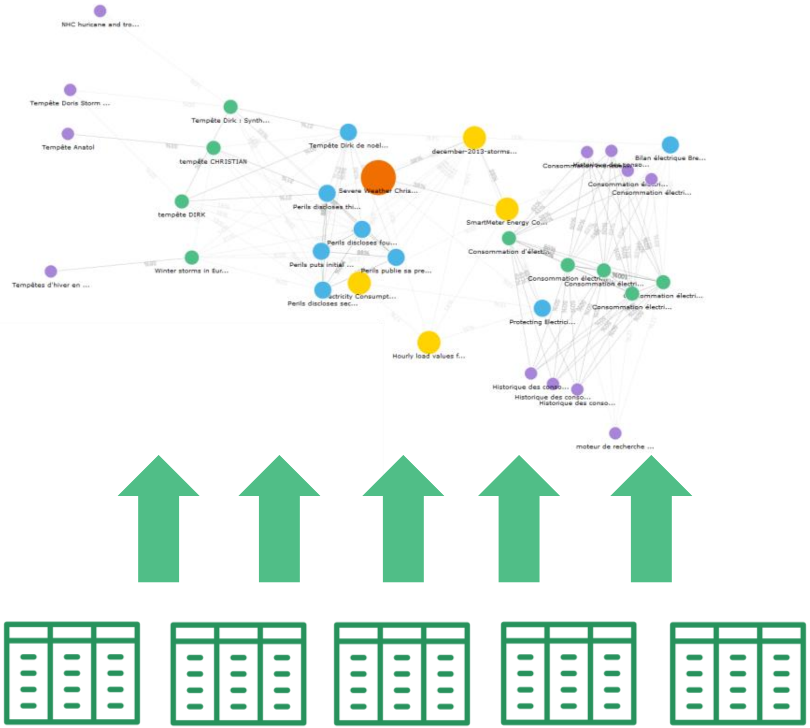
Auckland 2019 ISWC

Context & Goals

- Design a semantic engine able to query (semi-)structured data





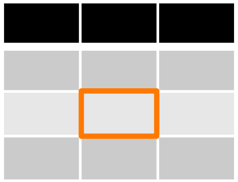


- We focus on tabular data: annotate the content and structure of tabular data for searching and recommending datasets

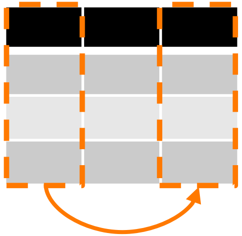
Tabular Data to Knowledge Graph Matching

- Goals

CTA
Column-Type Annotation



CEA
Cell-Entity Annotation

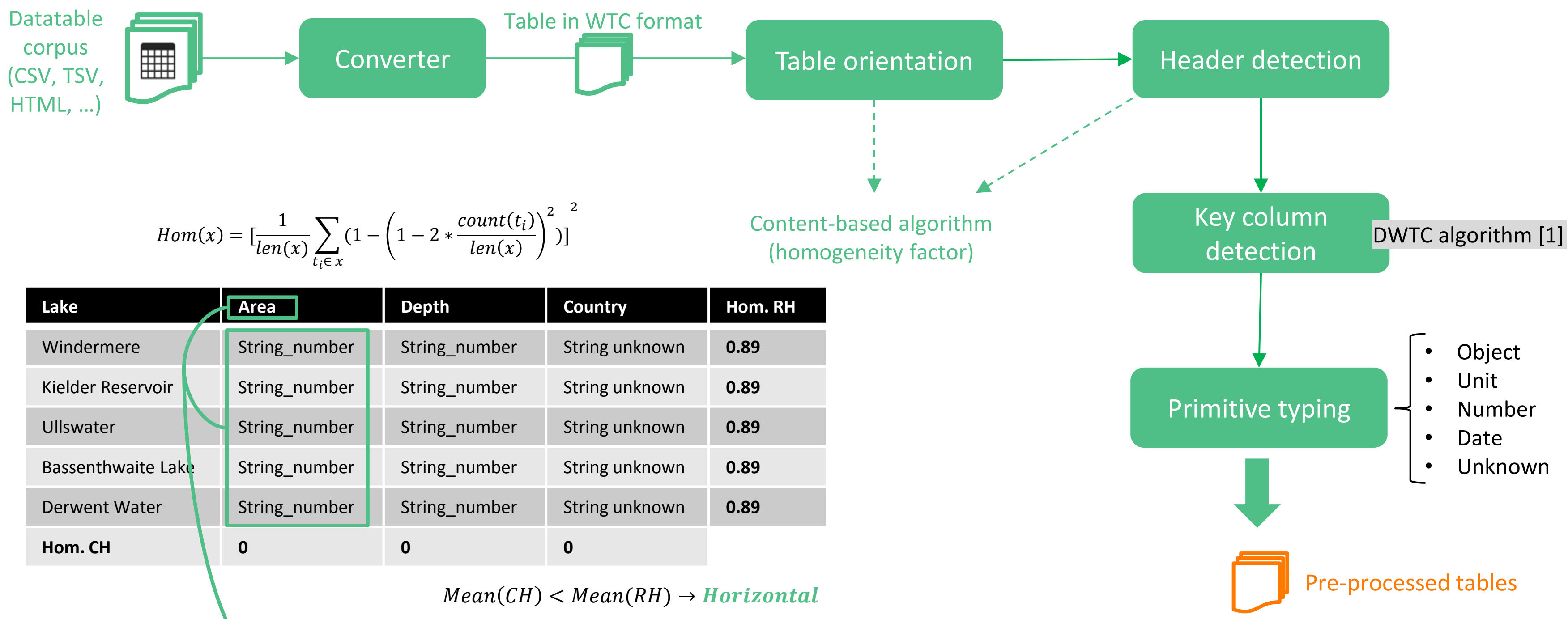


CPA
Columns-Property Annotation

- 1st step: preprocessing to identify tables characteristics (orientation, key-column…)
- 2nd step: annotations workflows
  - Method 1: Baseline lookups
  - Method 2: Embedding approach
- We focus on the CTA and CEA tasks

  CPA processing: list of properties associated to entities pairs, plus majority voting

Preprocessing (new homogeneity factor)



$$Hom(x) = [\frac{1}{len(x)} \sum_{t_i \in x} (1 - \left(1 - 2 * \frac{count(t_i)}{len(x)}\right)^2)]^2$$

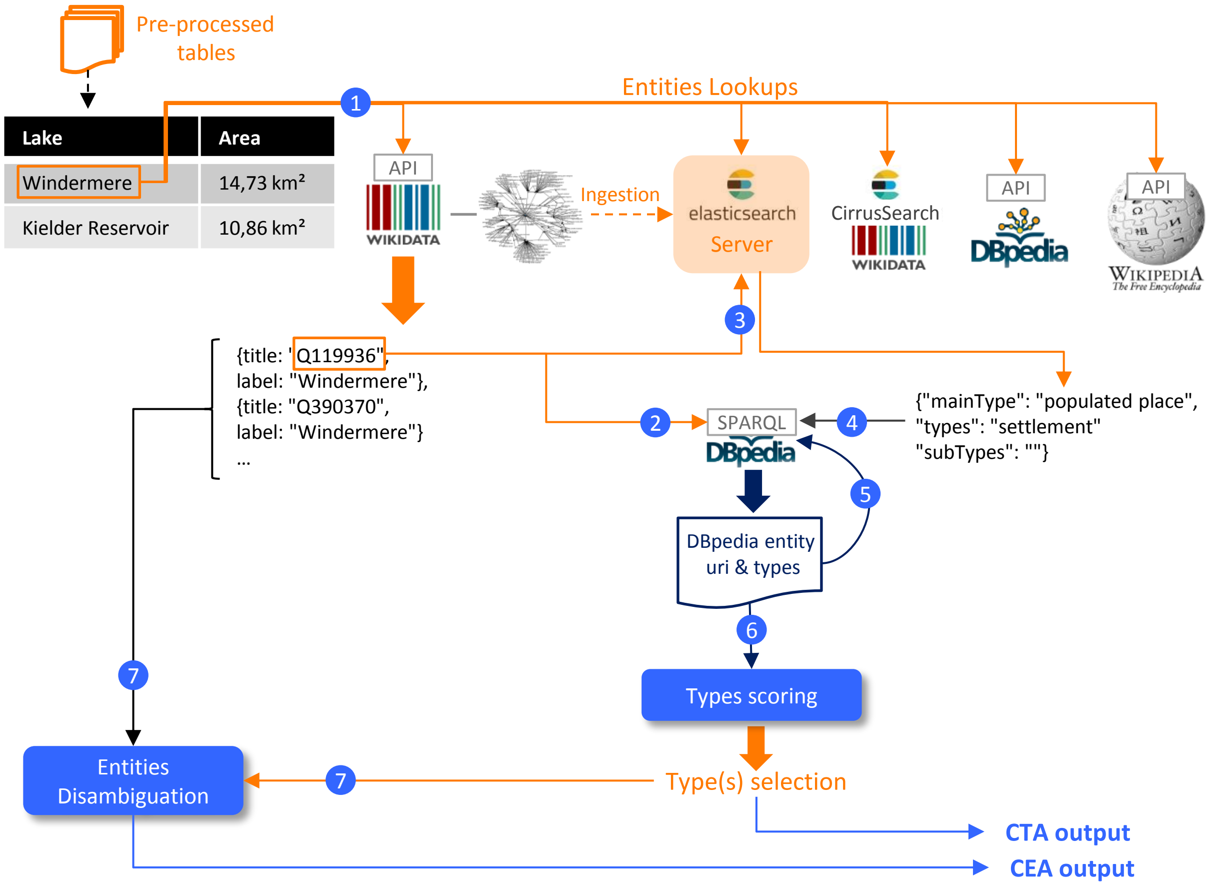
Lake	Area	Depth	Country	Hom. RH
Windermere	String_number	String_number	String unknown	**0.89**
Kielder Reservoir	String_number	String_number	String unknown	**0.89**
Ullswater	String_number	String_number	String unknown	**0.89**
Bassenthwaite Lake	String_number	String_number	String unknown	**0.89**
Derwent Water	String_number	String_number	String unknown	**0.89**
Hom. CH	**0**	**0**	**0**	

$$Mean(CH) < Mean(RH) \rightarrow \boldsymbol{Horizontal}$$

$$\exists\ col\ where\ Hom(col[0:3]) \neq 0 \rightarrow\ \boldsymbol{Header = true}$$

[1] https://subversion.assembla.com/svn/commondata/WDCFramework/tags/1/0/3/

Baseline lookups



1 Lookups from all tables cells (4 external sources + 1 internal Wikidata ES)

3 Wikidata as pivot metadata

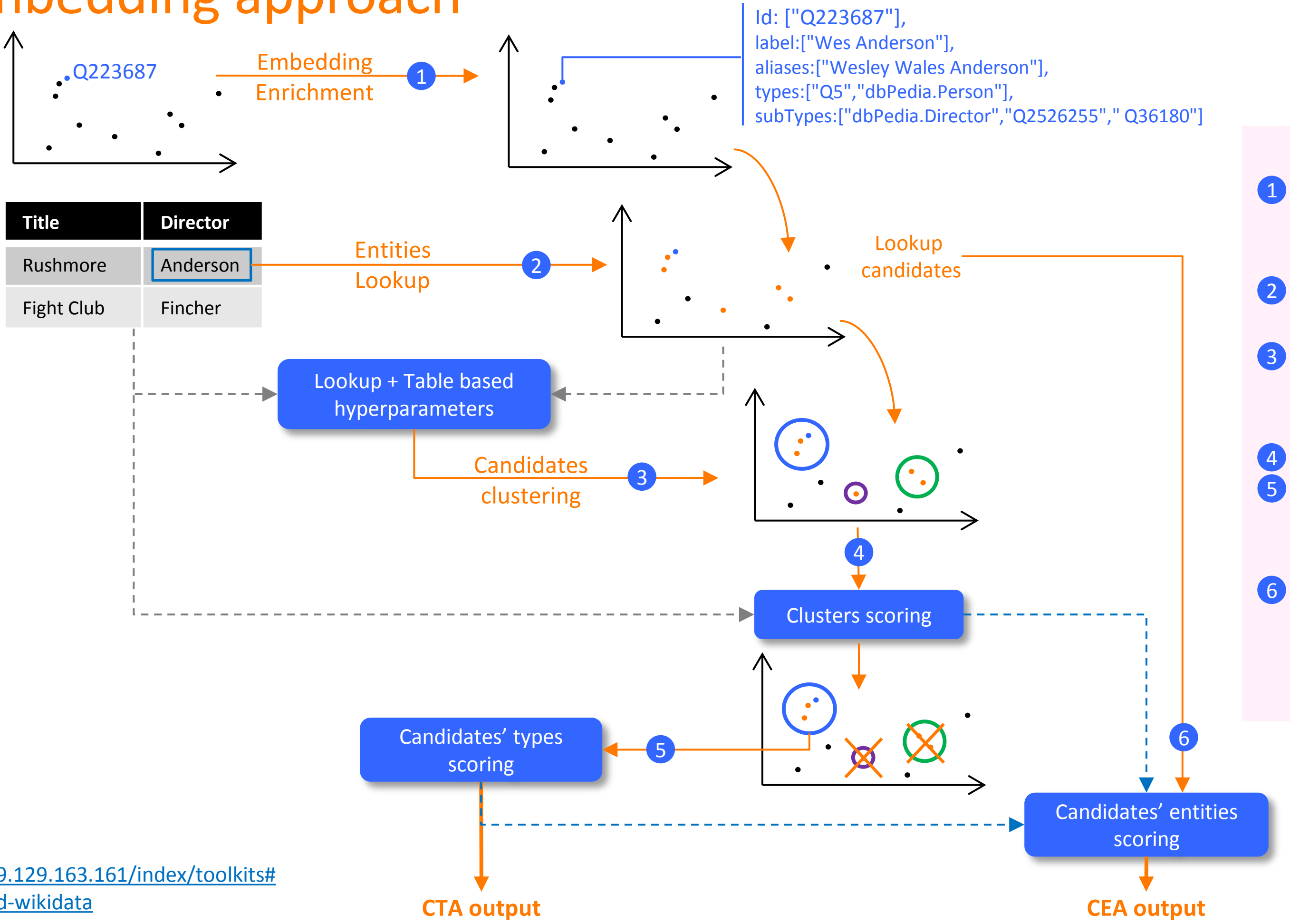
2 DBpedia translation (uri &
4 types)

6 TF-IDF-like types scoring

7 Entities disambiguation with target type(s)

Embedding approach

WIKIDATA EMBEDDING OpenKE [2]

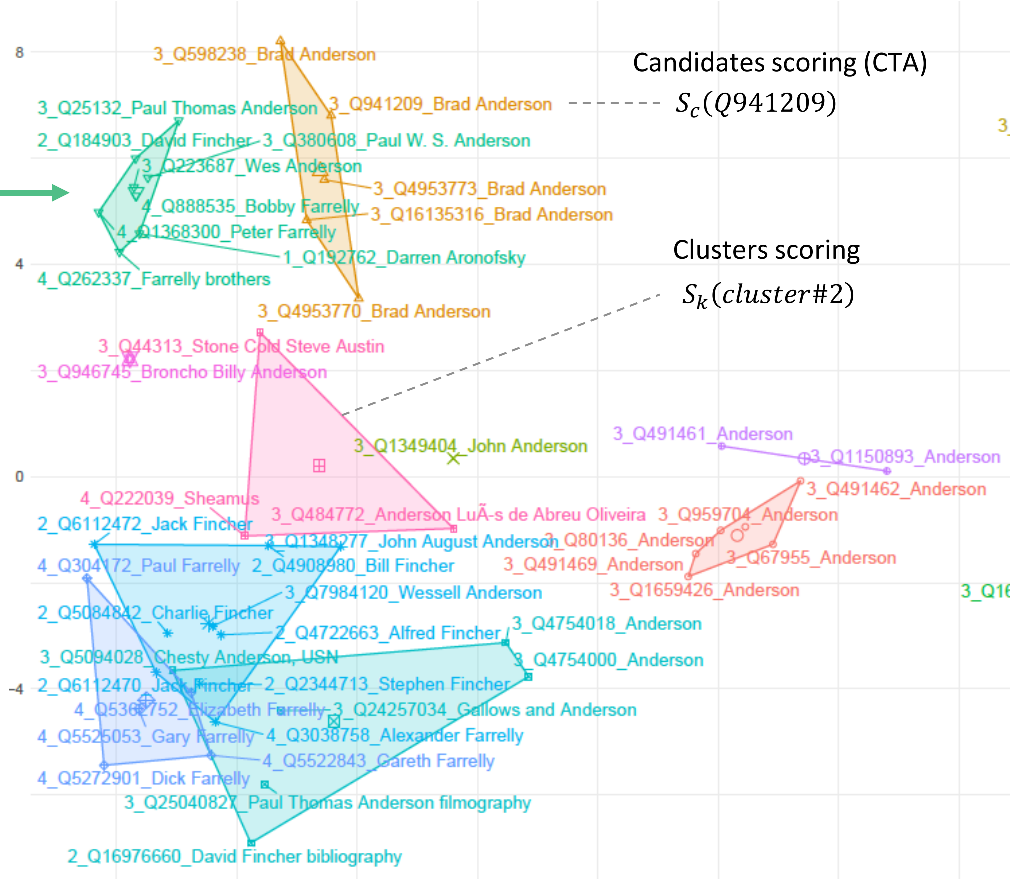


1. Embedding enrichment through Wikidata ES server
2. Regex + Levenshtein lookup
3. K-means clustering over candidates space
4. Scoring algorithm to extract
5. best cluster and deduce target type
6. Candidates disambiguation from clusters, types and entities scores

[2]
http://139.129.163.161/index/toolkits#pretrained-wikidata

Embedding approach example

Title	Year	Director
Requiem For A Dream	2000	Aronofsky
Fight Club	1999	Fincher
Royal Tenenbaums	2001	Anderson
There's Something About Mary	1998	Farrelly



Candidates scoring (CTA)

$S_c(Q941209)$

Clusters scoring

$S_k(cluster\#2)$

Entities scoring (CEA):

$$S_e(i) = 0.25 * S_k(n) + 0{,}5 * R_T + 0.2 * S_c(i)$$

Entities disambiguation:

$$S_e(Wes\ Anderson) > \begin{cases} S_e(Paul\ Thomas\ Anderson), \\ S_e(Paul\ W.S.Anderson) \end{cases}$$

Results

Table1: Preprocessing results

Task/Tool	DWTC	DAGOBAH
Orientation Detection	0.9	**0.957**
Header Extraction	Not evaluated	**1.0**
Key Column Detection	0.857	**0.986**

Table 2: Round 1 results (own evaluator < AI crowd evaluator)

Task	CTA				CEA	
Criteria	F1	Precision	AH	AP	F1	Precision
Baseline	0.517	0.482	NA	NA	0.784	0.814
Baseline++	0.641	0.641	1.108	0.246	**0.881**	**0.890**
Embedding	**0.683**	**0.683**	**1.483**	**0.258**	0.840	0.852

Approach	Pros	Cons
Baseline	▪ High coverage (multiple sources) ▪ Computational efficiency	▪ Lookup-services dependency (reliability) ▪ Blackbox (indexing, scoring…) ▪ Queries volume
Embedding	▪ Lookup strategy independence ▪ Relevant clustering even with few data ▪ Generalization (no tailored cleaning + less heuristics in lookups and scoring)	▪ Computational performances ▪ K optimization ▪ Embedding dependency

Discussion & Future Work

- Performance bottlenecks (due to the challenge context):
  - ✓ Light Data cleaning … on purpose
  - ✓ Basic lookup strategies … on purpose (e.g. no use of dictionary)
  - ✓ Missing Wikidata – DBpedia type mappings
  - ✓ Subset embedding (restricted to baseline candidates)
- Future work:
  - ✓ Test other Wikidata embeddings methods (on the whole space)
  - ✓ Compute joint embeddings with Wikipedia/DBpedia to enhance coverage
  - ✓ Experiment more clustering algorithms and parameters on different datasets
  - ✓ Learn data table embedding and find vectorial transformation(s) with KG embedding space
  - ✓ …

DAGOBAH

Datatable-powered Accurate-knowledge Graph for Outstanding and Beautiful Answers to Humans

Thanks!