DAGOBAH

Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data


Viet-Phi Huynh


Jixiong Liu


Yoan Chabot
@yoan_chabot


Frédéric Deuzé


Thomas Labbé
@tau_labbe


Pierre Monnin
@piermonn


Raphaël Troncy
@rtroncy

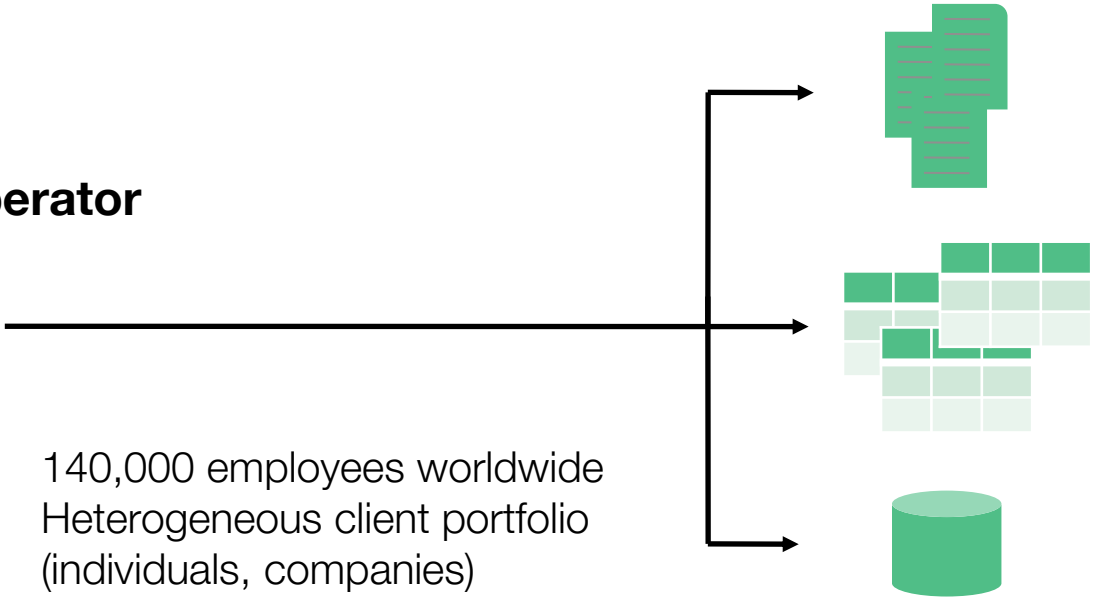
orange™

EURECOM
Sophia Antipolis

IC
29/06/2022

On the Importance of Interpreting Tabular Data

Multiservice operator



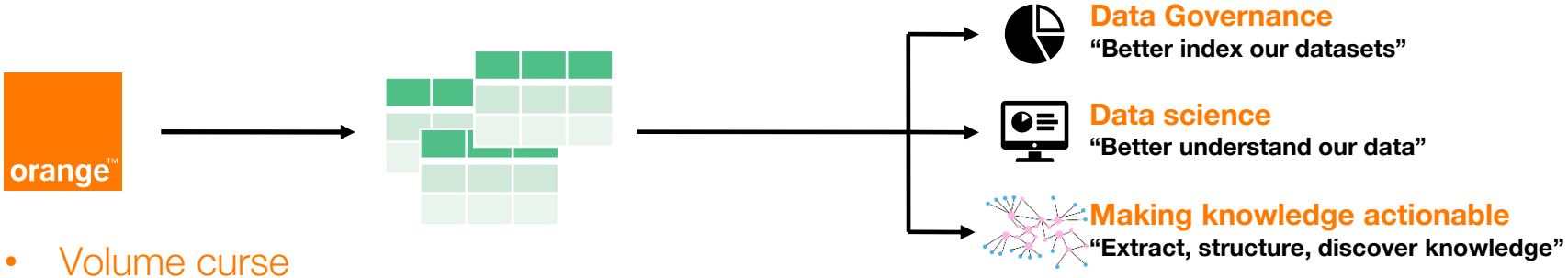
- Telecommunications (services + infrastructure)
- Video On Demand, Music
- Bank
- Cybersecurity
- …

- 140,000 employees worldwide
- Heterogeneous client portfolio (individuals, companies)

https://www.orange.com/en
https://hellofuture.orange.com/en/

Huge amount of generated data

On the Importance of Interpreting Tabular Data



- Volume curse
- 7 languages (FR, EN, PL, ES, …)
- Heterogeneously structured tables
- Little context

Cell types		Deployment environment	Max. number of users	Output power (mW)	Max. distance from base station
5G NR FR2	Femtocell	Homes, businesses	Home: 4–8 Businesses: 16–32	indoors: 10–100 outdoors: 200–1000	tens of meters
	Pico cell	Public areas like shopping malls, airports, train stations, skyscrapers	64 to 128	indoors: 100–250 outdoors: 1000–5000	tens of meters
	Micro cell	Urban areas to fill coverage gaps	128 to 256	outdoors: 5000–10000	few hundreds of meters
	Metro cell	Urban areas to provide additional capacity	more than 250	outdoors: 10000–20000	hundreds of meters
Wi-Fi (for comparison)		Homes, businesses		indoors: 20–100	

Frequency (MHz)	Cell radius (km)	Cell area (km^2)	Relative cell count
450	48.9	7521	1
950	26.9	2269	3.3
1800	14.0	618	12.2
2100	12.0	449	16.2

Réseaux en fibre optique jusqu'à l'abonné (FttH)
- déploiement et mutualisation

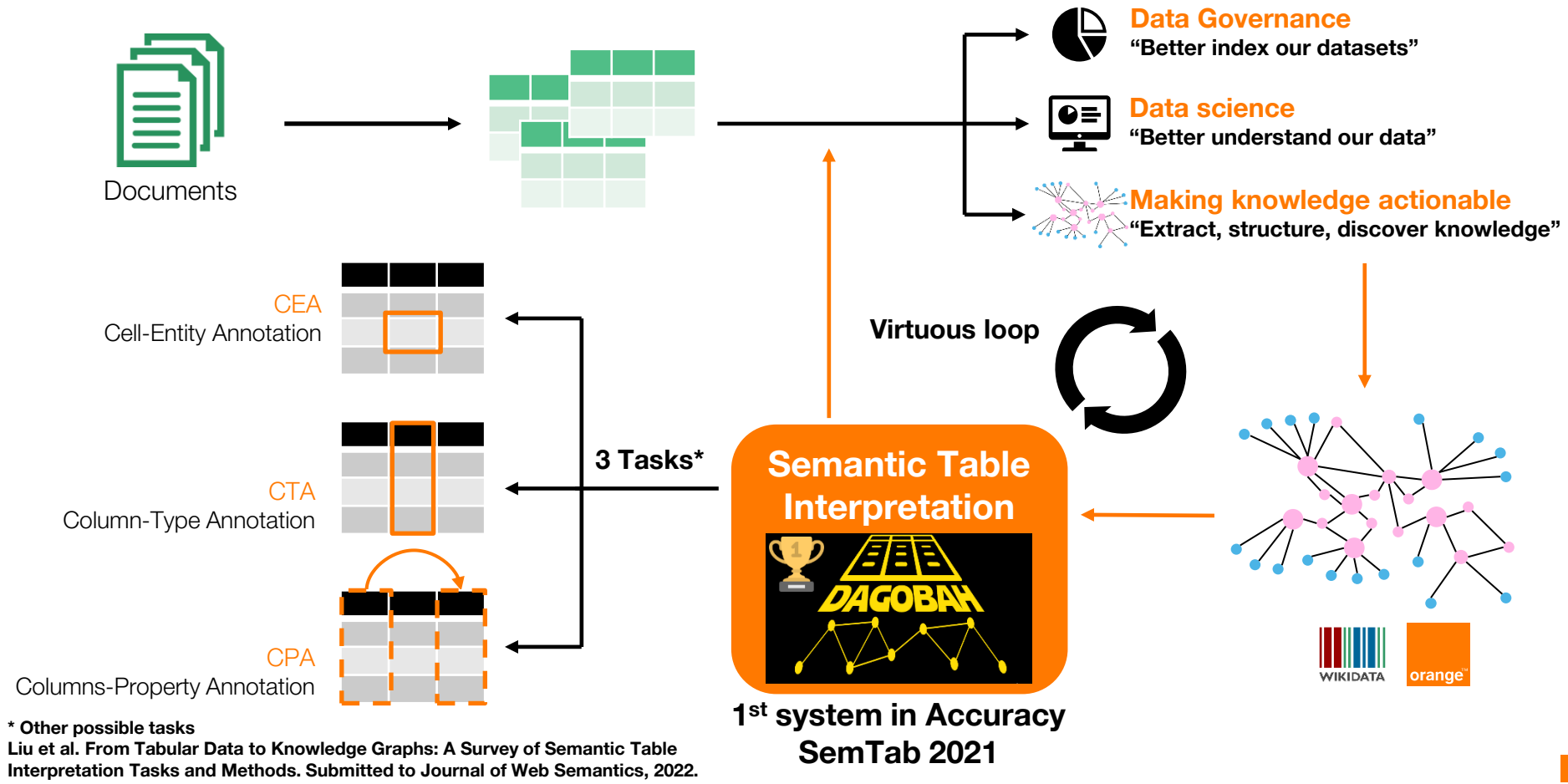
Remarque liminaire
Le total de locaux raccordables par zone ne correspond pas à la somme des lignes raccordables de la zone. Le nombre de lignes raccordables par opérateur d'infrastructure (OI) prend en compte l'ensemble des lignes raccordables alors que le total des locaux raccordables tient compte des déploiements multiples qui sont déduits de la somme.
Dans le cas d'un opérateur verticalement intégré, le nombre de lignes raccordables déployées par son activité d' opérateur d'infrastructure (OI) sur le marché de gros est différent du nombre de lignes éligibles aux offres de son activité d' opérateur commercial (OC) sur le marché de détail

	2012 T1	2012 T2	2012 T3
	2012		
	T1	T2	T3
France			
Total des locaux raccordables et détail des lignes raccordables par opérateur	**1 776 000**	**1 873 000**	**2 157 000**
Orange (OI – initiative privée)	875 000	918 000	1 094 000
Altice France (OI – initiative privée)	369 000	388 000	449 000
Free Infrastructure (OI – initiative privée)	207 000	218 000	221 000
Autres OI	325 000	349 000	393 000
Lignes éligibles par nombre d'opérateurs présents au point de mutualisation :			
Au moins un opérateur présent au PM via la mutualisation passive	1580 000	1749 000	1960 000
	704 000	785 000	333 000
	311000	349 000	539 000
	30 000	39 000	93 000

fres clés de l'activité des opérateurs de communications électroniques

		2015
Emplois (champ : ancien cadre réglementaire)	*Unités*	
Emplois directs	*Unités*	**118 470**
Source ARCEP - Enquêtes annuelles 1998 à 2018 ; enquêtes trimestrielles 2020		
Investissements (champ : ancien cadre réglementaire)	*Millions d'€*	
Investissements au cours de l'exercice*	*Millions d'€*	**10 630**
dont investissements hors achats de fréquences mobiles	*Millions d'€*	7 831
(*) Cet indicateur intègre les montants payés pour l'achat des fréquences mobiles.		

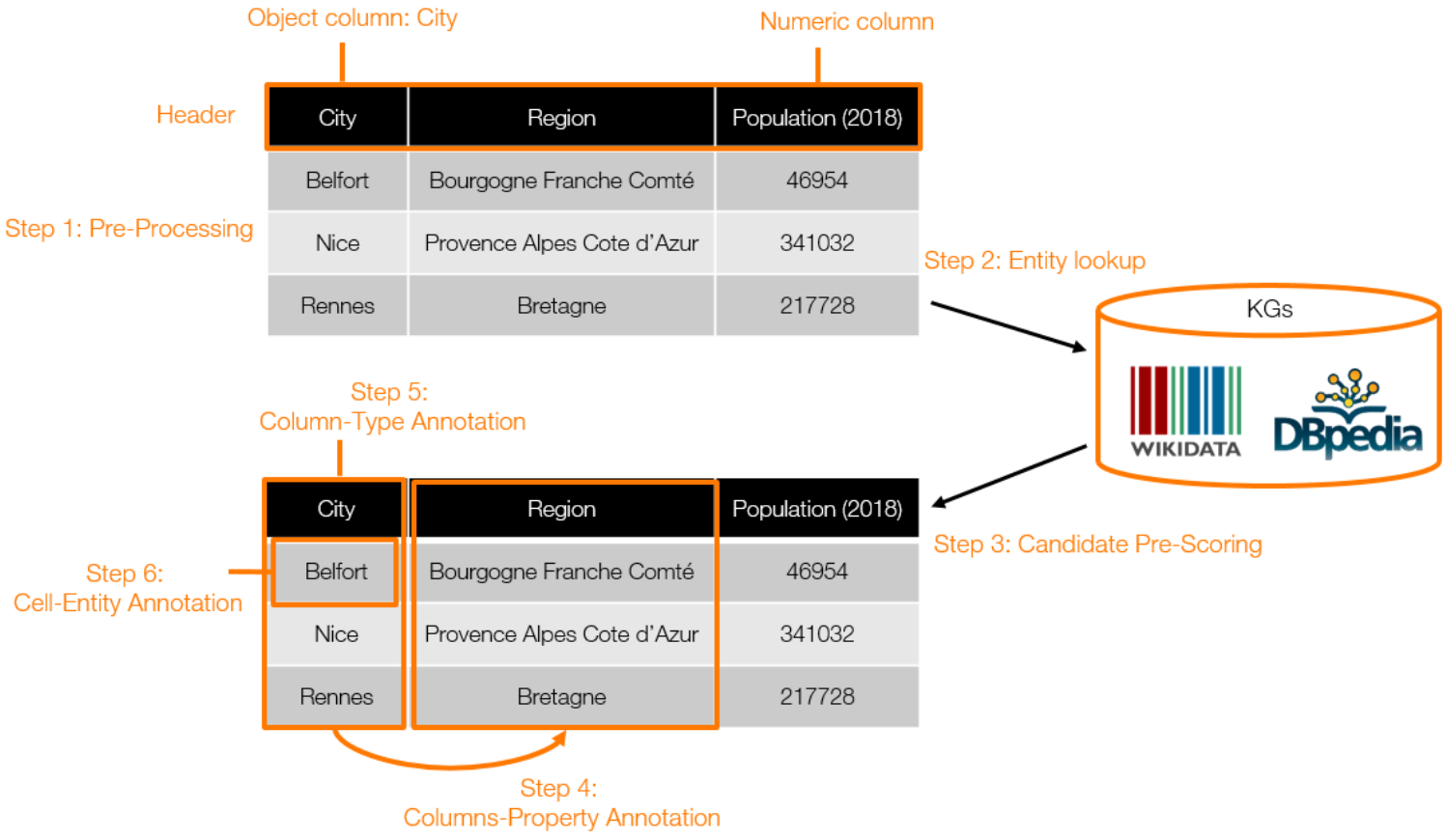
On the Importance of Interpreting Tabular Data



*** Other possible tasks**

Liu et al. From Tabular Data to Knowledge Graphs: A Survey of Semantic Table Interpretation Tasks and Methods. Submitted to Journal of Web Semantics, 2022.

DAGOBAH: Annotation Workflow



Making Sense of Tables: DAGOBAH

- Raw table:

Year	English title	Original title	Director(s)	Ref
2000	Lijmen/Het Been	Robbe De Hert	[22]	
2001	No Man's Land	NiÃ\u008Dija zemlja	Danis TanoviÃ\u008	[23]
2002	The Son	Le Fils	Jean-Pierre and Luc Dardenne	[19]
2003	On the Run	Cavale	Lucas Belvaux	[24]
2003	An Amazing Couple	Un couple Ã©patant	Lucas Belvaux	[24]

Pre-processing

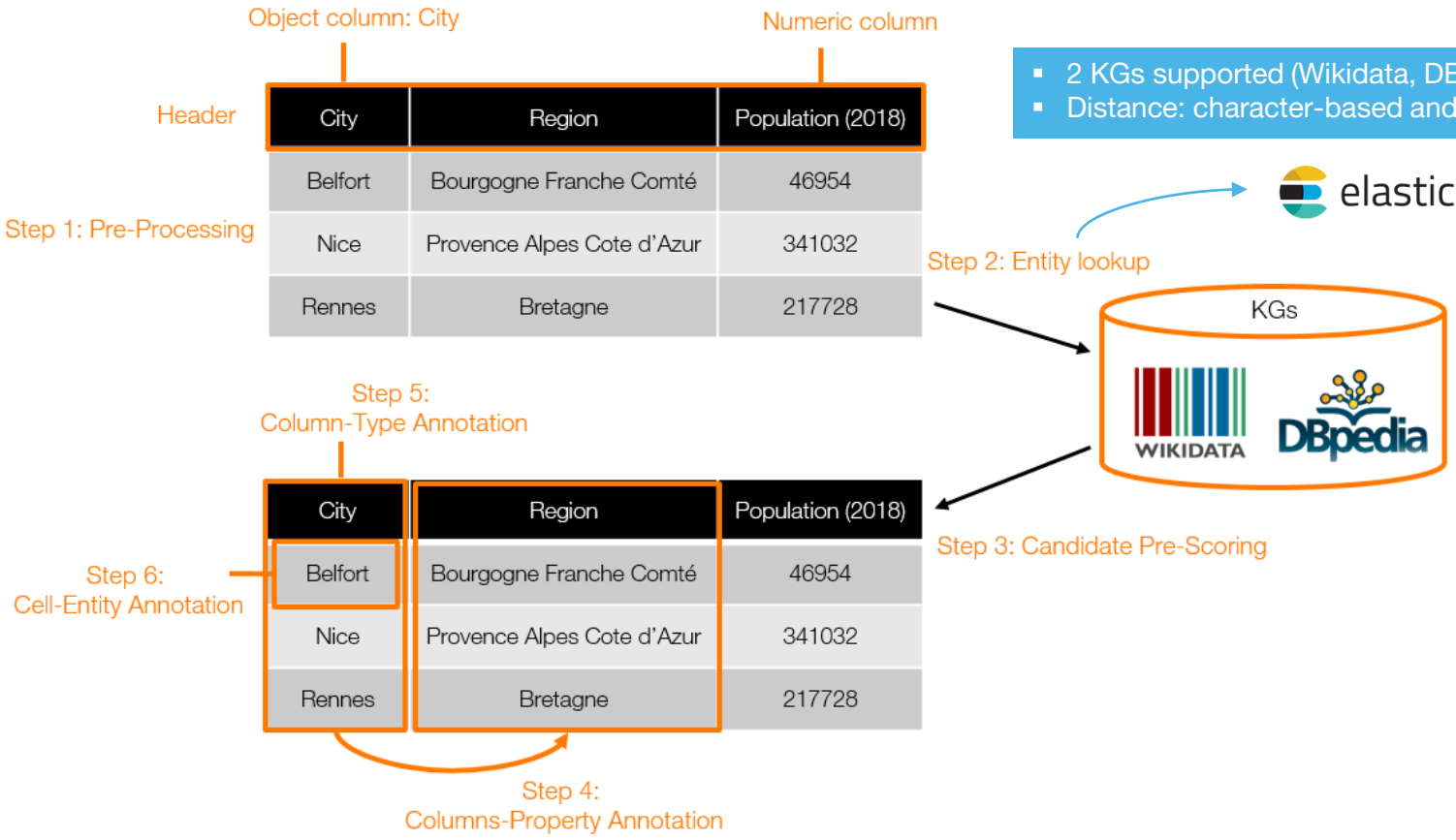
Ad-hoc rules, spaCy NER

- Re-structured table:

Year	English title	Original title	Director(s)	Ref
2000	Lijmen/Het Been		Robbe De Hert	[22]
2001	No Man's Land	Ničija zemlja	Danis Tanović	[23]
2002	The Son	Le Fils	Jean-Pierre and Luc Dardenne	[19]
2003	On the Run	Cavale	Lucas Belvaux	[24]
2003	An Amazing Couple	Un couple épatant	Lucas Belvaux	[24]

- **Title:** {}
- **Orientation:** HORIZONTAL, *Conf.*=1.0
- **Header**: ['Year', 'English title', 'Original title', 'Director(s)', 'Ref']
  Conf = 0.6
- **Key column**: index=1, *Conf.*=1.0
- **Data Type**: [DateTime (*Conf* 1.0), String (1.0), String (1.0), String (1.0), String_Number(1.0)]
- **Primitive typing**: [DateTime (1.0), Unk, Unk, Person (1.0), Unk]
- Possible row re-alignement.

DAGOBAH: Annotation Workflow



- 2 KGs supported (Wikidata, DBpedia)
- Distance: character-based and token-based Levenshtein

Making Sense of Tables: DAGOBAH

City	Region	Population (2018)
Belfort	Bourgogne Franche Comté	46954
Nice	Provence Alpes Cote d'Azur	341032
Rennes	Bretagne	217728

e_m

Context -$\mathcal{N}_{table}(e_m)$

$e_c \in \mathcal{E}_c$

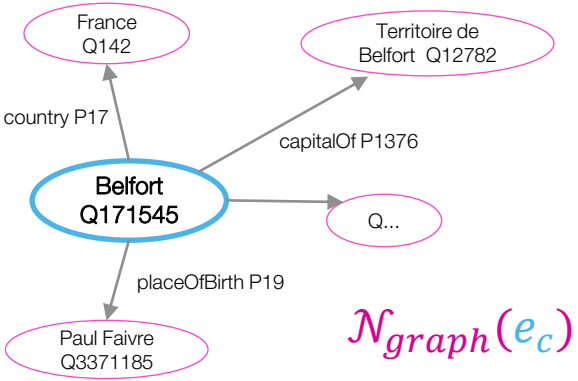
Q171545 : (Belfort) city in France
Q4991979 : (Belfort) Neighborhood in Maastricht

Candidate pre-scoring

$\mathcal{E}_c$ set of candidates for cell e_m
➔ confidence score for each $e_c \in \mathcal{E}_c$:

$$PSc(e_c, e_m) = Sc_{context}(\mathcal{N}_{table}(e_m), \mathcal{N}_{graph}(e_c)) * Sc_{sim}(e_c, e_m)^x$$

Context score of a candidate entity
➔ **quantifying the correspondance of neighboring cells with neighboring entities**



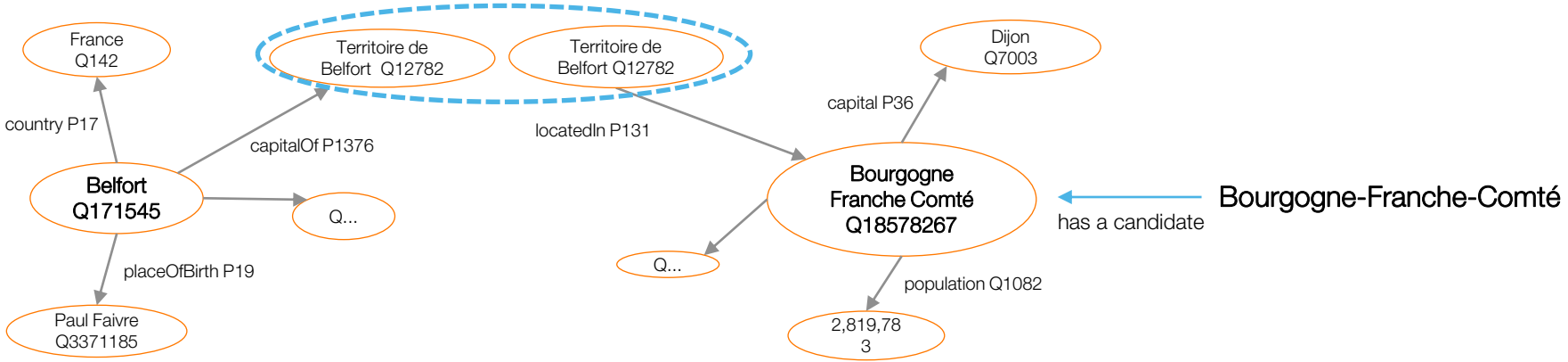
Effective Candidate Scoring

Relevance of cell Bourgogne-Franche-Comté w.r.t. Q171545 (Belfort) is determined through graph contexts discovery

City	Region	Population (2018)
Belfort	Bourgogne Franche Comté	46954
Nice	Provence Alpes Cote d'Azur	341032
Rennes	Bretagne	217728

e_m

Context -$\mathcal{N}_{table}(e_m)$



2-hop predicate path: (Q171545) Belfort is the capital of (Q12782) Territoire de Belfort which is located in (Q18578267) Bourgogne-Franche-Comté

Candidate Soft Scoring

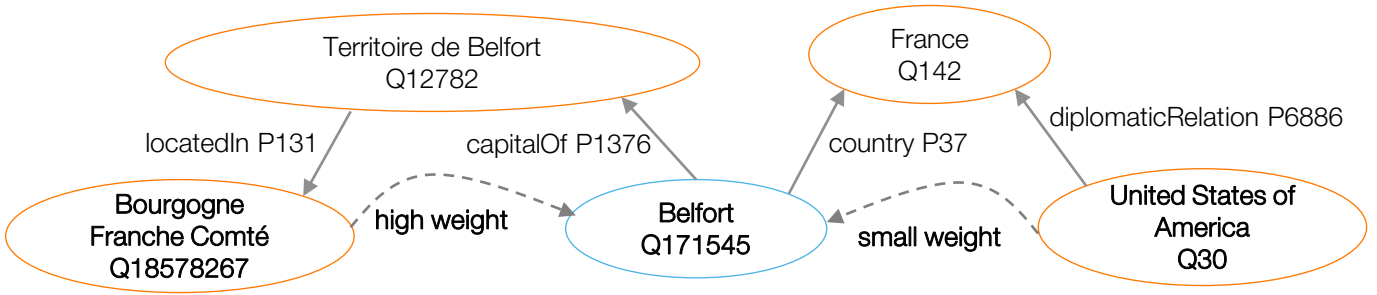
- **Each neighbor does not have an equal contribution**

City	Region	Population (2018)
Belfort	Bourgogne Franche Comté	46954
Nice	Provence Alpes Cote d'Azur	341032
Rennes	Bretagne	217728

e_m

Context -$\mathcal{N}_{table}(e_m)$

- ✓ **Object neighbors** are more important than literals (datetime, unit, plain number…)
- ✓ **Cells in the left side** of the table should have more weight (i.e. key column)
- ✓ **Columns highly connected** to target column shoud have more weight
- ✓ KG predicate paths provide **different information content**



$generality(Q12782) \ll generality(Q142)$ → Q18578267 more informative

Making Sense of Tables: DAGOBAH

City	Region	Population (2018)
Belfort	Bourgogne Franche Comté	46954
Nice	Provence Alpes Cote d'Azur	341032
Rennes	Bretagne	217728

e_m

$e_c \in \mathcal{E}_c$

Context - $\mathcal{N}_{table}(e_m)$

Q171545 : (Belfort) city in France
Q4991979 : (Belfort) Neighborhood in Maastricht

Candidate pre-scoring

$\mathcal{E}_c$ set of candidates for cell e_m
➔ confidence score for each $e_c \in \mathcal{E}_c$:

$$PSc(e_c, e_m) = Sc_{context}(\mathcal{N}_{table}(e_m), \mathcal{N}_{graph}(e_c)) * Sc_{sim}(e_c, e_m)^x$$

Columns-Property Annotation

- The **most frequently occuring property** between best entities pre-scored in the two columns
- Supports multiple relation types,
  e.g. {s,p,o}, {s, p1, p2, o}, {s, p1, (-)p2, o}

Column-Type Annotation

- Majority voting based on pre-scored entities of the same column
- Type hierarchy considered with levels

Cell-Entity Annotation

- Pre-scoring = local information
- Final scoring = local information (pre-scoring) + global information (CTA and CPA)

$$Sc(e_c, e_m) = \frac{PSc(e_c, e_m) + \alpha \times \overline{score_{CPA}} + \beta \times score_{CTA}}{1 + \alpha + \beta}$$

α = CPA_coverage / 2
β = CTA_coverage / 2

Performance: the SemTab 2021 challenge

1st system in Accuracy SemTab 2021

Settings

Setting	Entity's KG context	Scoring
1	1-hop predicate paths	Hard scoring
2	2-hop predicate paths	Hard scoring
3	2-hop predicate paths	Soft scoring
4	2-hop, but using only uni-directional paths, i.e. sub $\overrightarrow{p_1}$ $\overrightarrow{p_2}$ obj	Soft scoring

100 candidates / cell: 49s / table
250 candidates / cell: 92s / table

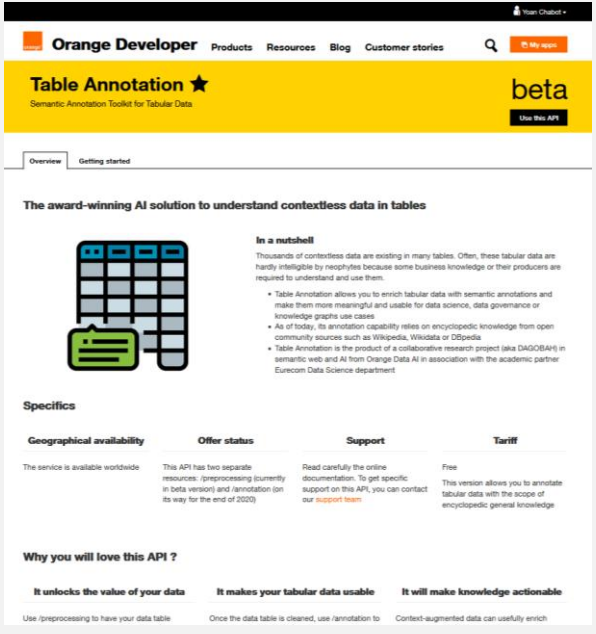
LeaderBoard

Task	Setting	CTA		CEA		CPA	
		F1	Precision	F1	Precision	F1	Precision
Round 1 – WD Table	2*	**0.832**	**0.832**	**0.923**	**0.923**	-	-
	Top 1 SemTab 2021	0.728	0.73	0.907	0.907	-	-
Round 1 - DBPTable	2*	0.422	0.424	**0.954**	**0.946**	-	-
	Top 1 SemTab 2021	**0.46**	**0.468**	0.692	0.692	-	-
Round 2 - BioTable	4*	0.916	0.916	**0.970**	**0.970**	**0.899**	**0.899**
	Top 1 SemTab 2021	**0.956**	**1**	0.964	0.964	**0.899**	**0.899**
Round 2 - HardTable	3*	0.976	0.976	0.975	0.976	0.996	0.996
	Top 1 SemTab 2021	**0.977**	**0.977**	**0.985**	**0.985**	**0.997**	**0.998**
Round 3 - BioDivTable	4*	0.381	0.382	0.496	0.497	-	-
	Top 1 SemTab 2021	**0.593**	**0.595**	**0.602**	**0.611**	**0.947**	**1**
Round 3 – HardTable	3*	**0.99**	**0.99**	**0.974**	**0.974**	0.991	**0.995**
	Top 1 SemTab 2021	0.984	0.984	0.968	0.968	**0.993**	0.994

Research results transferred into usable APIs/UIs

DAGOBAH API

For developers and data scientists

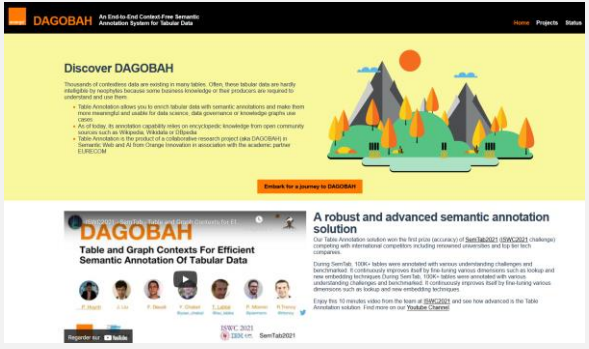


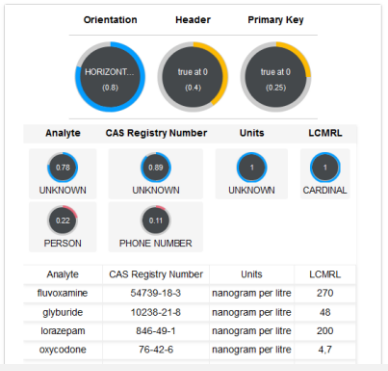
https://developer.orange.com/apis/table-annotation/api-reference

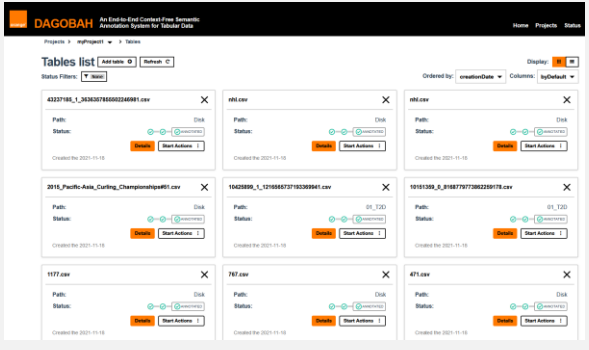
Open for testing upon invitation

DAGOBAH UI

For all Orange collaborators







https://tinyurl.com/dagobah-ui-demo

Sarthou-Camy et al. DAGOBAH UI: A New Hope For Semantic Table Interpretation. Posters & Demo. ESWC 2022

DAGOBAH Science Health Use Case

Context

A **healthcare practitioner** exploring **two datasets** containing information about **drugs**.

health_drugs_class.csv

NOM	PRODUIT	Code EphMRA	Classe EphMRA	Classe ATC
YELLOX 0,9 MG\u002fML COLLYRE FL 1\u002f5 ML	YELLOX	S01R	ANTIINFLAMMATOIRES NON STEROIDIENS OPHTALMIQUES	BROMFENAC SODIUM SESQUIHYDRATE
ESOMEPRAZOLE BIOGARAN 20 MG CPR GASTRORESISTANT 28	ESOMEPRAZOLE BIOGARAN	A02B2	INHIBITEURS DE LA POMPE A PROTONS	ESOMEPRAZOLE
AVONEX 30 MCG\u002f0,5 ML SOL INJ STYLO 4\u002f0,5 ML	AVONEX	L03B2	INTERFERONS, BETA	INTERFERON BETA‰\u0080 \u0093A
ESOMEPRAZOLE BIOGARAN 40 MG CPR GASTRORESISTANT 14	ESOMEPRAZOLE BIOGARAN	A02B2	INHIBITEURS DE LA POMPE A PROTONS	ESOMEPRAZOLE
VALSARTAN\u002fHYDROCHLOROTHIAZIDE CRISTERS	CGP-48933	C09D1	ASSOCIATIONS ANTAGONISTES DE	VALSARTAN ET DIURETIQUES

health_chemical_components.csv

Analyte	CAS Registry Number	Units	LCMRL
oxycodone	76-42-6	nanogram per litre	4,7
sulfamethoxazole	723-46-6	nanogram per litre	6,5
trimethoprim	738-70-5	nanogram per litre	3,5
valsartan	137862-53-4	nanogram per litre	7,2
verapamil	52-53-9	nanogram per litre	7,8

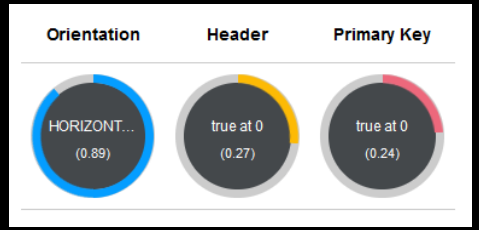
DAGOBAH Science Health Use Case

Context

A **healthcare practitioner** exploring **two datasets** containing information about **drugs**.

DAGOBAH Goals

1. **Clean datasets** automatically (e.g. encoding issues, misalignement, etc.



health_drugs_class.csv

NOM	PRODUIT	Code EphMRA	Classe EphMRA	Classe ATC
YELLOX 0,9 MG\u002fML COLLYRE FL 1\u002f5 ML	YELLOX	S01R	ANTIINFLAMMATOIRES NON STEROIDIENS OPHTALMIQUES	BROMFENAC SODIUM SESQUIHYDRATE
ESOMEPRAZOLE BIOGARAN 20 MG CPR GASTRORESISTANT 28	ESOMEPRAZOLE BIOGARAN	A02B2	INHIBITEURS DE LA POMPE A PROTONS	ESOMEPRAZOLE
AVONEX 30 MCG\u002f0,5 ML SOL INJ STYLO 4\u002f0,5 ML	AVONEX	L03B2	INTERFERONS, BETA	INTERFERON BETA‰\u0080\u0093A
ESOMEPRAZOLE BIOGARAN 40 MG CPR GASTRORESISTANT 14	ESOMEPRAZOLE BIOGARAN	A02B2	INHIBITEURS DE LA POMPE A PROTONS	ESOMEPRAZOLE
VALSARTAN\u002fHYDROCHLOROTHIAZIDE CRISTERS	CGP-48933	C09D1	ASSOCIATIONS ANTAGONISTES DE	VALSARTAN ET DIURETIQUES

NOM	PRODUIT	Code EphMRA	Classe EphMRA	Classe ATC
YELLOX 0,9 MG\u002fML COLLYRE FL 1\u002f5 ML YELLOX 0,9 MG/ML COLLYRE FL 1/5 ML	YELLOX	S01R	ANTIINFLAMMATOIRES NON STEROIDIENS OPHTALMIQUES	BROMFENAC SODIUM SESQUIHYDRATE
ESOMEPRAZOLE BIOGARAN 20 MG CPR GASTRORESISTANT 28	ESOMEPRAZOLE BIOGARAN	A02B2	INHIBITEURS DE LA POMPE A PROTONS	ESOMEPRAZOLE
AVONEX 30 MCG\u002f0,5 ML SOL INJ STYLO 4\u002f0,5 ML AVONEX 30 MCG/0,5 ML SOL INJ STYLO 4/0,5 ML	AVONEX	L03B2	INTERFERONS, BETA	INTERFERON BETA‰\u0080\u0093A
ESOMEPRAZOLE BIOGARAN 40 MG CPR GASTRORESISTANT 14	ESOMEPRAZOLE BIOGARAN	A02B2	INHIBITEURS DE LA POMPE A PROTONS	ESOMEPRAZOLE
VALSARTAN\u002fHYDROCHLOROTHIAZIDE CRISTERS VALSARTAN/HYDROCHLOROTHIAZIDE CRISTERS	CGP-48933	C09D1	ASSOCIATIONS ANTAGONISTES DE L'ANGIOTENSINE II ET ANTIHYPERTENSEURS (C2) ET\u002fOU DIURETIQUES ASSOCIATIONS ANTAGONISTES DE L'ANGIOTENSINE II ET ANTIHYPERTENSEURS (C2) ET/OU DIURETIQUES	VALSARTAN ET DIURETIQUES
VALSARTAN\u002fHYDROCHLOROTHIAZIDE CRISTERS VALSARTAN/HYDROCHLOROTHIAZIDE CRISTERS	CGP-48933	C09D1	ASSOCIATIONS ANTAGONISTES DE L'ANGIOTENSINE II ET ANTIHYPERTENSEURS (C2) ET\u002fOU DIURETIQUES ASSOCIATIONS ANTAGONISTES DE L'ANGIOTENSINE II ET ANTIHYPERTENSEURS (C2) ET/OU DIURETIQUES	VALSARTAN ET DIURETIQUES

DAGOBAH Science Health Use Case

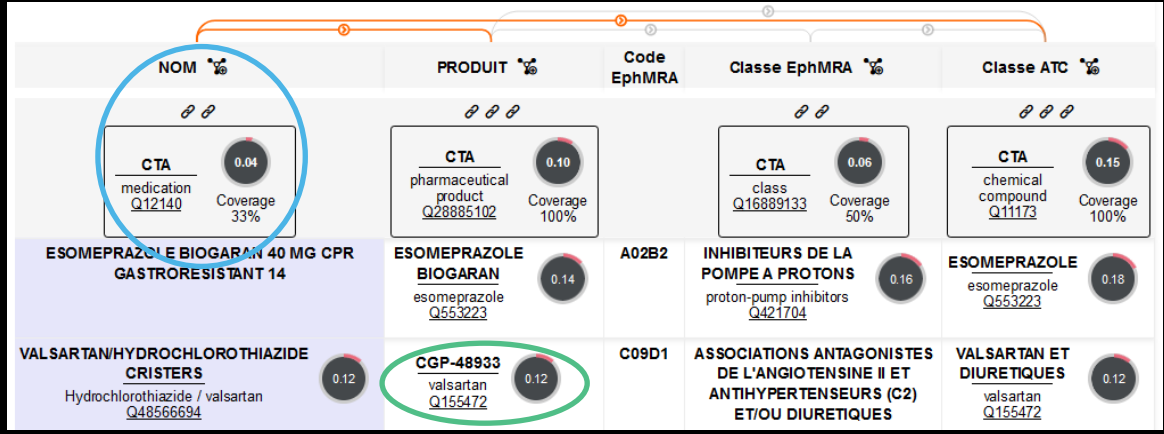
Context

A **healthcare practitioner** exploring **two datasets** containing information about **drugs**

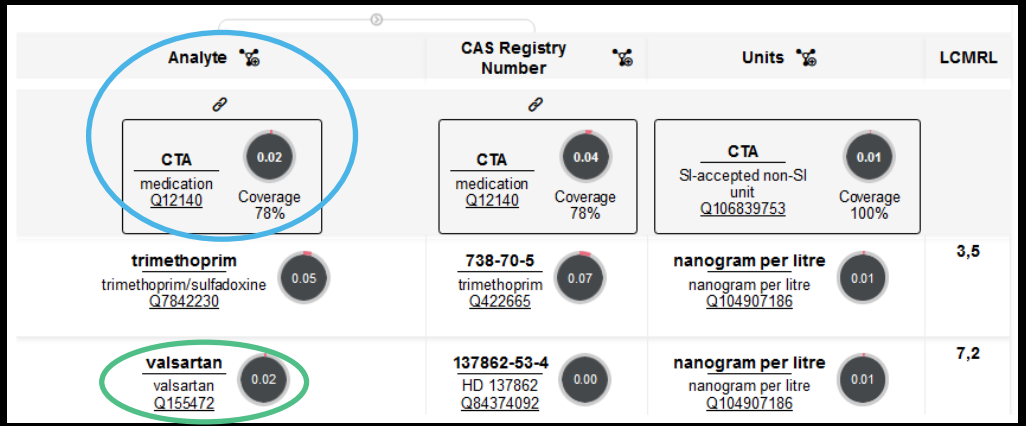
DAGOBAH Goals

1. **Clean datasets** automatically (e.g. encoding issues, misalignement, etc.)
2. Make table **semantics explicit** and **align datasets**

health_drugs_class.csv



health_chemical_components.csv



DAGOBAH Science Health Use Case

Context

A **healthcare practitioner** exploring **two datasets** containing information about **drugs**.

DAGOBAH Goals

1. **Clean datasets** automatically (e.g. encoding issues, misalignement, etc.)
2. Make table **semantics explicit** and **align datasets**
3. **Enrich tables** with new information from the Knowledge Graph (fill missing values, new columns)

health_chemical_components.csv

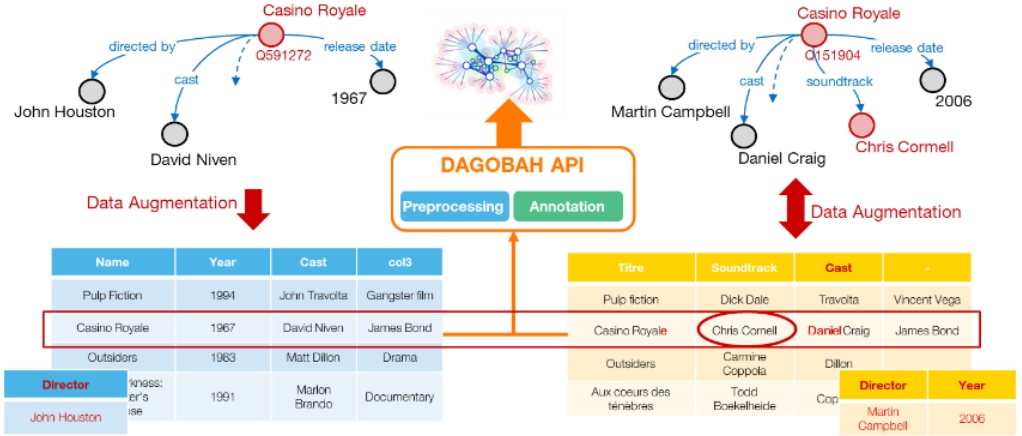
Analyte	CAS Registry Number	Units	LCMRL
oxycodone	76-42-6	nanogram per litre	4,7
sulfamethoxazole	723-46-6	nanogram per litre	6,5
trimethoprim	738-70-5	nanogram per litre	3,5
valsartan	137862-53-4	nanogram per litre	7,2
verapamil	52-53-9	nanogram per litre	7,8

Analyte	CAS Registry Number	Units	LCMRL	mass	chemical formula
sulfamethoxazole Sulfamethoxazole / Trimethoprim Q48566821 0.05	**723-46-6** sulfamethoxazole Q415843 0.08	**nanogram per litre** nanogram per litre Q104907186 0.01	6,5	Knowledge CEA 253.052 253.052253.052	Knowledge CEA $C_{10}H_{11}N_3O_3S$ $C_{10}H_{11}N_3O_3SC_{10}H_{11}N_3O_3S$
trimethoprim trimethoprim/sulfadoxine Q7842230 0.05	**738-70-5** trimethoprim Q422665 0.07	**nanogram per litre** nanogram per litre Q104907186 0.01	3,5	Knowledge CEA 290.138 290.138290.138	Knowledge CEA $C_{14}H_{18}N_4O_3$ $C_{14}H_{18}N_4O_3C_{14}H_{18}N_4O_3$

Future work

- **Annotation enhancement**
  - ✓ Generate global tokens dictionaries (abbreviations, acronyms…) from web data
  - ✓ Leverage embeddings-based approach (clustering, language models applied to table…)

- **Data augmentation: from tables to KGs**
  - ✓ Enrich Knowledge graphs leveraging CTA and CPA to add new entities



- **Towards corpora with new challenges**
  - ✓ SemTab 2021 challenging corpora (new domains, new target KG)
  - ✓ But scores still very high → call to arms to build hard corpora for the community

- **Towards a KG targeting Orange business areas**
  - ✓ Current effort to build such a KG supported with DAGOBAH
  - ✓ Challenges: heterogeneous domains and vocabularies

DAGOBAH

DAGOBAH: Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data

Thanks!

Sarthou-Camy et al. (2022). DAGOBAH UI: A New Hope For Semantic Table Interpretation. In Demo Track ESWC2022.
Huynh et al. (2021). DAGOBAH: Table and Graph Contexts For Efficient Semantic Annotation Of Tabular Data. SemTab@ISWC (pp. 19-31)
Chabot et al. (2021). A Framework for Automatically Interpreting Tabular Data at Orange. In Industry Track ISWC2021.
Huynh et al. (2020). DAGOBAH: Enhanced Scoring Algorithms for Scalable Annotations of Tabular Data. In SemTab@ ISWC (pp. 27-39).
Chabot et al. (2019). DAGOBAH: An End-to-End Context-Free Tabular Data Semantic Annotation System. In SemTab@ ISWC (pp. 41-48).
DAGOBAH: Make Tabular Data Speak Great Again : https://hellofuture.orange.com/en/dagobah-make-tabular-data-speak-great-again/