Why Table Understanding Matters

Practical solutions to real-life problems

Vincenzo Cutrona

IT Consultant | R&D OpenLab @ Corvallis SRL (Tinexta Spa Group) Ph.D. Student | University of Milano - Bicocca





<u>Glossary</u>

CEA: Cell to Entity

2





Discover Business Data

#1

<u>Challenge</u> Collecting and aggregating information about a business entity from public sources

<u>Goal</u> Providing a **data marketplace** based on a **highly interconnected graph** of company-related information (a.k.a. *the business KG*)



Scenario #1: KG Construction

- Table schema to KG schema
 - Inferred with CTA (e.g., ColNet) and CPA
 - Manually provided (mapping languages like RML)
 - Human in the loop is acceptable, because the number of columns is usually under control
- Records assumed to be unique
 - NO deduplication: rows representing the same object (i.e., same ID), are collapsed into a single entity (i.e., same URI)
- Main challenges:
 - Help the user in defining the schema mapping (e.g., type and property suggestion)
 - Speed up the **triple construction process** (separable problem -> high parallelism)



KG Construction in euBusinessGraph

- CTA and CPA with user in the loop
- Optional: CEA to reuse existing entities



STEP 1: CTA + CPA (+ CEA, optional)





KG Construction in euBusinessGraph (cont.)





Supporting KG Construction with DataGraft

- **Common data model**: the euBusinessGraph ontology
- **DataGraft**: a "one-stop-shop for *hosted data management*" with two integrated tools:
 - Grafterizer 2.0
 - Tabular Transformation
 - RDF Mapping (with Grafter)
 - o ASIA
 - Tabular annotation (CTA and CPA)
 - Vocabulary suggestion with ABSTAT
- **Big Data Environment**: to execute the RDF mapping on massive amount of data

_							
							CUPDATE FOR MAPPINOS 0 SE
	AdGroupid 🧷 👩 🖬	Keywordd 🧷 🙍 🖬	Category / g 🖬	Category-label 🧷 🛊 🖬	Company Campaign Name // 😦 💷	Keyword 0 💿 🖬	match-id 🤌 🚳
2	Typetat Advisore Concern Prep: ketingsTaldOroup	Typede Admon	Papettit Company Prisc Company BourceCol AdDrived	Owstupe: string Prop. Mont BourseCot. Coloury	Description and the second sec	Destype string Prop. Moel Septement Kernsteld	Typest Advertises Prep Sedests Descended Revertied
4	SourceCol. Keywarahi						
1	18572500607	5154040950	Hoard 1144	Noakitata	LER DE DO P HELEINE L'HADO SS O'S APPE	innobeet eistegen	1007/200007-51540/09/03-0004599-2276-20
2	18572503667		Picarii 1144				1867 PECKEP SISAHI7070 1004979 2276-2
	THE PROPERTY AND		Production	Production of			
-	10073302007	51682020	Provincial and	Bastister			THE THE OWN OF A REPORT OF A DOMESTIC AND A DOMESTI
	1847134CHAT	SNORAFING.	Encoder 2014	Baselitter	THE PE PO E BANKSTON 100000 OD OF 5 AFTER	active space in attendance	THE PERSON PARTY AND ADDRESS OF THE
7	18172300467	\$195222810	Production	Poststate	SER DE DO P RIMENTAGE L'EDDOG OD OF & AF285	withoutsten in seconda	1007 FRC667 119 022812 1004852 2276-20
8	18572350567	8258223490	DealEstote	RealEstate	089_08_00_P_Nex85446_170000_00_01_5_44255	Nuser aur miete hamburg	18172160567-5258223490-100488 2015010100000
9	18572160667	5258223490	PealEstote	RealEstate	0EP_DE_00_P_ResEstate_L70000_rm_eL_5_A4255	Nikser zur miete hamburg	18673160647-5258223450-100490 2016/010/00000
10	18672160667	5258223490	PealEstate	RealEstate	GER_DE_00_P_ResEnters_LTG000_co_cl_5_A#255	Niuser zur miete hamburg	10673/60667-5256223490-103496 2016/30/000000
11	18572160667	5430920102	Pearlistee	Realization	GER_DE_DO_P_ResEntane_LTG000_co_pt_S_A#255	häuser mieten in nrw	18573160667-5420620122-1028816-2276-20
12	18572350567	5460947608	Pearlistee	RealEstate	66P_26_00_P_RodExtre_1.15030_oo_01_5_AP255	wg zimmer is greifowaid	18673950567-5460947608-904830 2099707000099
13	18173360667	6729637878	Pearlicture	Problemate	GER_DE_DO_P_NonEstate_L10000_oo_ct_8_A#288	dreamight soundcits	10073102007-072032972-0240533-2270-22
14	18972360667	8729637973	Pearlietuna	PloadEstate	EER_DE_DD_P_NoAEstate_L10020_oo_c4_8_AF288	dreamight traumbold	18873760667-570333973-9048866-2276-20
18	18672360667	\$738335010	DealEstola	Realisiate	089_06_00_P_RevEstanc_170000_oo_eC_5_A4255	volvurgsneid rotivel	16673160647-6731038010-6048670 20160101000000
16	18573160667	5737934730	RealEstota	RealEstate	GER_DE_00_P_RevEstate_LT0000_os_eL_5_A4255		18673190957-573194730-9048709-2276-20
17	18573160667	5731256790	RealEstote	RealEstate	GER_DE_00_P_RevEstate_LT0000_os_eL_5_A4255		18673190957-5731256710-9048750-2276-20
19	19573160667	5731367670	PealEstote	RealEstate	GER_DE_00_P_ResEntrie_L10000_col_cl_5_A4255	mietvohrungen zvenkau	1067310067-573236150-5046830-2276-20 10673160667-5731367870-9046885

Different outputs: **RDF data** (small data) **Executable file** (massive data)



Supporting KG Construction with DataGraft (cont.)

Executable File

(RDF Mapping)

h						
S Obereitan 22	100010001000	1011010				
						Contract and a second second
Approvant 🖉 🖨 🗎	Novembo 2 0 0	самаруу 🖉 🖨 🖪	Chegory-store 2 0 1	Cancegrittene 🧭 🖨 🗈	Payword 2" 🔹 🖬	1413-1 2 • B
	Tables Holders					Page Indigedition Page Indigedition InstanCel Reported
1007200061						WATER COLD AND ADDRESS TO ADDRESS 2274 20223 - 0074 2
1002/002667	SIGAACCO'S	Realitation				BETTERSED ANALYSINE BOURS 2016 20205 SED 34
188/12/00067		NUCCO				10072900687-0104080900-100407-2279-20225-089-25
942/240661	CRENCENSO	Ballulate	Insilvate		have been an adverted	INCOMPANY AND INCOMESTICATE THE SCORE CERTS OF
						THE PRODUCT DISABILITY OF AN ALL TIME STORE AND A
1002100/0012	Amploident	built new	huthers	NAME OF ON OR REALIZING A DAMAGE AND AN OF S AND A		WATHINGST REPORTED AND AND ADDRESS OF A DESCRIPTION OF A
0070100001	6268000490	Inclusive	Barlbare	000,00,00,0,0,0mBases(70000,m,r1,0,0400	Numeror ministrations	NR T28040*129422495-004895 2715-202 201009/000000
1007200067	525522440	horbors	hattara	688,00,00,0,00,00,000,00,00,00,000	Name of the Netton	1867/296083/1028223430/100480/2274-2823 201608/800000
106727020617	\$158223490	NUCLEO	Northern	088,01,01,P,MoR1093,170000,00,01,5,4429	NAME OF THE OWNERS	INCOMPANY ADDRESS ADDRESS STORE STORE
467960667	5400009832	Realition of	keathdata	000,00,00,0,0,0,000,00,00,00,00,00,0000		INCOMPANY CONSIDERATION ADDREES TO A CONCORD OF
WATHOOKT	6400H4700B	Basilistation	Beallytate	000,00,00,00,00,000,000,00,00,0000,000,0000	ing simone in publicasis	1007290082* 548094*906 5048206-223%-282 2016082900000
WAT/ROOMET		Beallstate	Institute	088,08,00,7,8ee0x4ee,170000,ee,e1,8,34318		INCOMENTATION AND ADDRESS STORE STORE OF
967260667		holber Industr	Barthara Barthara	CER (H. (H. P. Joseff Arm.) (10000, pp., of C. Anton CER (H. O. P. Baselbare (10000, pp. of C. Anton		NETWOORT CONTRACT AND ADDRESS 2014 2014 2014 2014
						2010/01/00000
100720-0001		hutbers	Reflore	THE OF ON & BRATCHER LEVELOW BE ALL & ARTS		AND INCOMES ADDRESS IN AN ADDRESS OF THE DESIGN OF
1007200061	STREETSO	Burbler				INTERNET STREETS IN ADVANCE 2276 20238-009-2
1007200001	SCHOOL STREET	Northday	NUMBER	088,98,98,9,9,96,9,96,97,000,00,00,0,0,0,04,000	midworrunger pearstas	MATTHEORY ATTICK MATS COMMAN 2216 2020 2010/01/02/000



- Pipeline split into **Processing Units**
- Processing Units are deployed to execution nodes (parallel execution)





KG Construction - Results from euBusinessGraph

Project results

- Release a **marketplace** to publish company data <u>http://marketplace.businessgraph.io</u>
 - 4 data providers (~1.6 billion triples)
- 6 business cases supported

Additional results

- SIRENE challenge: publishing the official database of French enterprises and establishments
 - More than 10 million units (16 GB)
 - The resulting RDF dataset counts ~3 billion triples (.nt)



EWSHOPP

Supporting Event and Weather-based Data Analytics

#2

<u>Challenge</u> Data collected by individual companies provide a **partial view** on the customer journey

<u>Goal</u> Provide tools to develop analytical services that consider events that impact on customer decisions (a.k.a. *the EW-Shopp Toolkit*)



Scenario #2: Enrich Tabular Data

- CEA is the only key task
 - CTA and CPA may support CEA (e.g., CTA used to filter CEA results)
- Main challenges:
 - Deal with **multi-valued properties** in the enrichment
 - Tables are not ready to include such data
 - CEA on "custom" KGs
 - A few properties are assumed to exist in STI algorithms (e.g., rdfs:label)
 - Speed up CEA
 - Property usually overlooked (not evaluated in recent challenges)
 - Crucial in industry (how to enrich a large dataset with millions of records)



Weather Service

Ge
Names

gn:2950157

Tabular Data Enrichment in EW-Shopp

#im

REGION

Date

64 Thuringia 11/03/2017

KEYWORD

194906

Google

AdWords

Step 1 Business Data

Step 2 CEA (from AdWords labels to GeoNames IDs

Step 3 Extension from service (from <GeoNames ID, Date> to temperature)

	-	- J-					
517827	50	Bavaria	12/03/2017			6	gn:2822542
459143	42	Berlin	12/03/2017				
891139	36	Bavaria	11/03/2017			gn	n:2951839
				-			j.
KEYWORD	#im	REGION	Date	geold.			
194906	64	Thuringia	11/03/2017	gn:2822542			
517827	50	Bavaria	12/03/2017	gn:2951839			
459143	42	Berlin	12/03/2017	gn:2950157			
891139	36	Bavaria	11/03/2017	gn:2951839		:X	ENSION
KEYWORD	#im	REGION	Date	geold.	C°/+0	C°/+1	
194906	64	Thuringia	11/03/2017	gn:2822542	18	20	
517827	50	Bavaria	12/03/2017	gn:2951839	17	19	
459143	42	Berlin	12/03/2017	gn:2950157	17	20	
891139	36	Bavaria	11/03/2017	gn:2951839	19	23	



Table Annotation for Data Enrichment



Declarative and interactive approach

- Each step annotates the table and enables new actions (Data Discovery)
- Annotations steps are converted to data transformation steps
- Integration under user's control (intermediate results)



Supporting Tabular Data Enrichment in EW-Shopp

• ASIA

- Tabular annotation (CEA, CTA and CPA)
- Different reconciliation service for CEA
- Different extension service (weather and events)
 - Different aggregation strategies!
- STI to Data Transformation Pipeline
 - Allow us to reuse DataGraft/Grafterizer 2.0
- **Big Data Environment**: to execute the data transformations on massive amount of data
 - Custom deployment to support several reconciliation services and speed up the process

	🚯 Grafterizer 2.0						
							CUPDATE KDF MAPPINGS 0 SETTIN
	Addrouptd 🧷 😦 🖬	Keywordd 🖉 🚳 🖽	Category / g 1	Category-label 🧷 🗿 🖪	Compaignhiame 🧷 😝 💷	Keyword 0 🗧 🖬	matchid 🖉 🧔 🖬
	Pypetit Address Concern Prep: AntonysTableDroop	Typede Admon	Typettit, Celepiny Prop. Gategory BourceCol, AdDrought	Owarupic string Prop. Ideel BaarueCal Calegory	bestype iting Proc. Isleet ServiceCat. Addressed	Darsetable: string Prop. label BiogramCal. Keywarshi	Typesze Administration Press Badhaba BoarseCot Reymented
	GourceCol. Keyenesisi						
1	18672100067	5154089990	Productive	1402.8.1122.0	LER DE DO P HELEINE L'ILODO 30 D'S AP255	innobies essigen	\$677606675556089901004599327632275
1	1841350003	************	Personal States	revalidade		energy of the second	New 2 November 2019/01/10/10/01/19/9 2276-20225
÷	10073300007	5459775850	Production of	Baskitire			THE PARTY OF A REAL PROPERTY OF A DATE
-	10073300007	53(6825)(10)	Providence and a	Baseline			In These Low Press Control Provide Line and Line
	18972300667	SN6845IPO	Propilicante	Postitute	SER DE DO P ROMENTADE L'ISODO OD OF & AF255	withoutpers in attrendom	1887/1905617-5168445/10-100-9041-0276-20235-
7	1817230067	\$195222810	Production	Postitute	SER DE CO P Realitable L10000 oo of 5 Ar255	NORMAN IN SECOND	1007 210 C 007 107 22200 10 COMB32 2270-20207-
8	18973300067	8298223490	PealEstote	Realisiate	089_08_00_P_Nex85446_110000_e0_41_6_44255	hauser aur miete handung	18472N096742582234901004880-223 20190701000000
9	18673160667	5258223490	PealEstote	RealEstate	GER_DE_00_P_ResEstate_LTG000_co_cl_5_A4255	Niuser zur miete hamburg	19673160647-5256223450-1004901-227 205601010000000
10	18672160667	\$258223490	PosiEstate	RealEstate	GER_DE_00_P_ResExtans_LTG000_co_cl_5_A#255	häuser zur miete hamburg	10673/60667-5258223490-100-4961-227 2016/310000000
11	18573160667	5430920192	RealEstate	Realization	GER_DE_00_P_ResEstane_LTG000_oo_ol_S_A#255	hiuser mieten in nrw	18572160667-5420522122-1028216-2276-20225-
12	18672560667	5460947608	Pearlistate	Peralitation	66P_06_00_P_NoxEstate_L16030_oo_44_5_A#265	ug zimmer is greifousid	18672500607-5460947608-9048306-22 20590300000000
13	18172360667	572537873	PearlictMa	Peallitate	DER_DE_DO_P_NonEstate_LTEDDD_oo_c4_%_AF255	dreavinght traunicks	10673762667-570537978-9348533-2276-23227
14	18672360667	8709637873	Pearlicture	Photol State	GER_DE_DD_P_NoAEx1206_LTDDDD_00_EL3_AP283	dreavaght traumbolit	1887/260467-5701637873-0048866-3276-20238
	19572350567	\$739338010	DealEstola	Realisiate	089_08_00_P_RealEstate_170000_eo_et_5_A4255	wohrungsmaski ratioval	20160101000000
16	19573160667	\$23894330	RealEstote	RealEstate	GEP_DE_00_P_RevExtate_L16000_oo_eL5_A4255		18673190957-573194730-9048709-2276-20236
1	19572160667		see Colora	HeelEstate		metvonnungenhermeskei	ee ramues ris raizoe rid-do48750-2276-20234
19	18572160667	573(367870	RealEstote	RealEstate	GER_DE_00_P_RealExters_LT6000_co_ct_5_A4255	mietvohrungen zwenkau	1867360667-5731307870-0048830-2276-20236 1867360667-5731307870-0048893-227 206070000000

Different outputs: CSV data (small data) Executable file (massive data)



Supporting Tabular Data Enrichment in EW-Shopp (cont.)

A02-1.61							
A0010485							CONCERSION ADDRESS
		**************************************	CINERY /	CREATIVE / 0 3	Cancelonhone Z 🖨 🔋	Paymont / 0 (1	140-4 × 0
		Special Address					Tarente Mathematican Prog. Section Description Represente
100		575+010900					INCOMENTATION PROFESSION OF A DESCRIPTION OF A DESCRIPTIO
46	7260667	SIGAACCO'S	Realitate				BRITERSED AREASTONE BOARDS 2006 20205 GED IN
10	72100061	104090900	NUMBER				1007290087-00400000-00409-2270-20220-089-20
94	7240061	GRENOENSO	Include		GE0_00_00_P_Beathware_170000_pm_p1_0_Addds		INTERCERT EXERCISES FOR A 1116 20205 CER 20
15	72900661	Storements	Real Date	RefErro	SEX.26.26.27.NovEntra_170000.x0.26.3.A4350		WK7200007 010887800 100485-2276 20235-009-20
140			Beallulate				INCOMPANY ENGINEERING TO ADD TO ADD TO THE TERM OF
							46-26061 STYCERO (054652 2216 2011) 604.2
940	PRODUCT	6260223490	Bailute	Inclusion	000,00,00,0,0,0,0,0,0,0,0,0,0,0,0,0,0,		30800200000
- 16	2340667	5258223490	kathara	kathara	088,08,08,0,0,8648,449,170000,10,01,5,34055		304000-000000
19	72900661	5898229490	NUMBER				2010/07/07/07
16	7000067	5400000180	Reality of the local section o	Real Marco			METHODAY CAPPEDONS INSIGN 2215 JOSTS OF 21
100	7240961	Eastonations	Bachdate	Inclusion	000,00,00,0,0000,000,00,00,00,0000	ing simon in pull-social	10072NO483 5480947N08-8048308-2234-2822 20140874800800
144							INCOMENTATION FOR SOMELY 2274 20227-2014-2
16			institute .				10070030013-070809113-0040005-2074-20206-003-0
100	7260667	67203006	Bealboarte	Inclusion	000,00,00,0,000000,00000,00,0000000000	anter-represent estimat	10072900821072008070-90480190-22310-20221 2010/02/000000
240	7260661	ETTIH(T)O	Beallstate	Beallstate	GER_DE_DD_P_Realbase_170000_ex_ol_1_A4318		18672800601 67379-0730 8048709-2276 28238 689-3
46	2960663		beatters.				-662/800001 02/82502/0 8048260 5026 58234 609 9
	7200861	STREAM OF	habbara	habbara	ORK, DL. OL P. BARDARA, 170000, AL ALSA		MATTACHET ATTEMPTS REAMEN 22% 2023 SUP 2 MATTACKET ATTEMPTS REAMEN 22% 2023



- Enrichment steps to Transformation Pipeline (using **Grafter**)
- Pipeline split into **Processing Units**, deployed to **execution nodes**
- ASIA Services (for CEA and extensions) deployed next to the data





Tabular Data Enrichment - Results from EW-Shopp

Project results

- Release of the EW-Shopp Data Preparation and Enrichment tool (part of the EW-Shopp Toolkit), with comes with different features:
 - Tabular Data Transformation (Grafterizer 2.0)
 - Tabular Data Annotation and Extension (ASIA)
- 4 business cases supported
 - The EW-Shopp Data Preparation and Enrichment tool used to transform and enrich digital marketing campaign data with geographical and weather-related data (~100 GB)



Journey Experience Data Improve

#3

<u>Challenge</u> Tourism relevant data are scattered (high variety, due to different providers, different formats) and continuously updated

<u>Goal</u> Provide a Data Manipulation Platform to support the **Knowledge Extraction** from **Tourism Big Data** (a.k.a. *the DMP*)



Scenario #3: KG Population

- Deduplication **is fundamental**!
 - The same object may have different IDs in different sources
- Incremental update?
 - Typical approach: creating a new versioned KG from scratch (e.g., DBpedia 2014, 2016)
 - Not suitable in industry (imagine a graph containing daily bank transactions)
- Deal with "different" sources, e.g., **streams** (infinite sequence of rows)
 - "Record" understanding? (a table with only one row)



KG Population in JEDI





KG Population in JEDI (cont.)

STI for KG Population

Done on "well-formed" entities but:

- Variety: equal entities may be modeled differently (depending on the underlying data model)
- Velocity: how to perform STI on streaming data? (open point)





Supporting KG population in JEDI



Q&A

vincenzo.cutrona@corvallis.it vincenzo.cutrona@unimib.it





Find out more!

EW-Shopp https://www.ew-shopp.eu/

euBusinessGraph

https://www.eubusinessgraph.eu/

JEDI <u>https://corvallis.it/open-lab/progetti/</u> (ITA only)

Icons made by <u>ultimatearm</u>, <u>prettycons</u>, and <u>Freepik</u> from <u>Flaticon</u>