Digital Industry Classification: NAICS

July 25th, 2022

Domain to Industry Classification:

Profound leverages Natural Language Processing (NLP) and other statistical methods to map & score industry classification by extracting textual details from the source of truth, the company's own website. A dynamic process indexes over 100 million unique websites and indexes B2B links such as "About Us", "What we do", "Products & Services", and other pages of company (or organizational entity) websites likely to contain the text that describes the business to produce a robust industry classification. In addition, Profound tracks links coming from and pointing to these websites to infer relationships between organizations that frequently cluster around the same industry.

Profound has built a dynamic, fully automated, location independent, multi-language industry classification tool based on website content that covers primary languages representing 85% of global GDP languages. The goal is to cover 97% by Q2 2023.

	Country/Territory	Language	UN Region	GDP per IMF	Year		United Nations	Year	World Bank	Year
	World	-		93,863,851	2021		87,461,674	2020	84,705,567	2020
✓	United States	English	Americas	25,346,805	2022		20,893,746	2020	20,936,600	2020
✓	China	Chinese	Asia	19,911,593	2022		14,722,801	2020	14,722,731	2020
✓	Japan	Japanese	Asia	4,912,147	2022		5,057,759	2020	4,975,415	2020
✓	Germany	German	Europe	4,256,540	2022		3,846,414	2020	3,806,060	2020
✓	India	English;Hindi	Asia	3,534,743	2022		2,664,749	2020	2,622,984	2020
✓	United Kingdom	English	Europe	3,376,003	2022		2,764,198	2020	2,707,744	2020
✓	France	French	Europe	2,936,702	2022		2,630,318	2020	2,603,004	2020
✓	Canada	English;French	Americas	2,221,218	2022		1,644,037	2020	1,643,408	2020
✓	Italy	Italian	Europe	2,058,330	2022		1,888,709	2020	1,886,445	2020
✓	Brazil	Portuguese	Americas	1,833,274	2022		1,444,733	2020	1,444,733	2020
	Russia	Russian	Europe	1,829,050	2022		1,483,498	2020	1,483,498	2020
	South Korea	Korean	Asia	1,804,680	2022		1,637,896	2020	1,630,525	2020
✓	Australia	English	Oceania	1,748,334	2022		1,423,473	2020	1,330,901	2020
	Iran	Farsi	Asia	1,739,012	2022		939,316	2020	203,471	2020
✓	Spain	Spanish	Europe	1,435,560	2022		1,281,485	2020	1,281,199	2020
✓	Mexico	Spanish	Americas	1,322,740	2022		1,073,439	2020	1,076,163	2020
	Indonesia	Bahasa Indonesian	Asia	1,289,295	2022		1,058,424	2020	1,058,424	2020
	Saudi Arabia	Arabic	Asia	1,040,166	2022		700,118	2020	700,118	2020
	Netherlands	Dutch	Europe	1,013,595	2022		913,865	2020	912,242	2020
✓	Switzerland	English;German;French;Italian	Europe	841,969	2022		752,248	2020	747,969	2020
✓	Taiwan	Chinese	Asia	841,209	2022		669,324	2020	668,500	2020
	Poland	Polish	Europe	699,559	2022		596,618	2020	594,165	2020
	Turkey	Turkish	Asia	692,380	2022		720,098	2020	720,101	2020
✓	Sweden	Swedish	Europe	621,241	2022		541,064	2020	537,610	2020
✓	Belgium	French	Europe	609,887	2022		521,861	2020	515,333	2020
✓	Argentina	Spanish	Americas	564,277	2022		383,067	2020	383,067	2020
✓	Norway	Norwegian	Europe	541,938	2022		362,522	2020	362,009	2020
	Thailand	Thai	Asia	522,012	2022		501,795	2020	501,795	2020
	Israel	Hebrew	Asia	520,703	2022		407,101	2020	401,954	2020
✓	Ireland	English	Europe	516,146	2022		425,889	2020	418,622	2020
✓	Nigeria	English	Africa	510,588	2022		429,899	2020	432,294	2020

✓	Primary language is covered
✓	English is an official language in India. Hindi is the official language and queued for development.
✓	Websites in these countries frequently provide an English version. English is the official language of Nigeria

- o Many countries in the long tail use currently covered languages. Profound's 85% coverage is therefore a conservative coverage estimate.

o   Profound will return, where available: the NAICS code for the line of business based on textual extraction from website along with our confidence codes. Several classifications may be provided, and the one with the highest confidence score will always be in the first position.
o   Other languages in development and will be provided when available
o   Classification for languages other than English typically have lower confidence scores due to the fine tuning we implemented for English. As improvements to classification and training data of non-English content are made, the confidence scores are expected to increase.
o   Many websites have a native English version (or other supported language) that is evaluated for NAICS assignment. E.g. If we see a website from Morocco that offers Arabic and French, the system will process the French version, which is supported.
o   The NAICS code field will be titled "naicschoice1" and the corresponding quality or confidence score "choice1quality" which is essentially a percentage. The higher the percentage, the better the confidence. Additional classifications include fields "naicschoice2" and "choice2quality"and "naicschoice3" and "choice3quality" where available.
o   Please note, we include a custom NAICS number "99" which means that the website's content could not be matched to a NAICS. This is different from a NULL value which means that there was no website or content to be matched whatsoever. See table below:
o   NAICS mapping is at a minimum a 2-digit code. NAICS with a second and third NAICS will also be delivered, where available. As we know, some businesses are not limited to a single industry.
o   The NAICS field in your database should accommodate up to a 6-digits to allow for future expansion of NAICS granularity.
o   All improvements in NAICS coverage, granularity, and accuracy are included in our agreement.

NAICS 2 digit code	Description
11	Agriculture, Forestry, Fishing, Hunting
21	Mining, Quarrying, Oil, Gas Extraction
22	Utilities
23	Construction
31-33	Manufacturing
42	Wholesale Trade
44-45	Retail Trade
48-49	Transportation and warehousing
51	Information
52	Finance and Insurance
53	Real Estate, Rental, Leasing
54	Professional, Scientific, Tech. Services
55	Mgmt. of Companies & Enterprises
56	Administrative, Support, Waste Mgmt., Remediation Services
61	Educational Services
62	Health Care and Social Assistance
71	Arts, Entertainment, Recreation
72	Accommodation and Food Services
81	Other Services (except Public Administration)
92	Public Administration
99	*Not enough content to classify*
NULL	Likely no content or website is broken

Batch File NAICS Deliverable with Languages found on website Domain

Data field Name	Definition
Domain	The website or email domain
Naicschoice1	The primary NAICS code with the highest confidence score
Choice1quality	The primary NAICS confidence score ranging from 0 to 1.
Naicschoice2	The secondary NAICS code with the highest confidence score
Choice2quality	The secondary NAICS confidence score ranging from 0 to 1.
Naicschoice3	The tertiary NAICS code with the highest confidence score
Choice3quality	The tertiary NAICS confidence score ranging from 0 to 1.
Languages	The language(s) that the website content is written in.

Optional DBI Enrichment

Company name	As published on the company website
Street address	As published on the company website
City	As published on the company website
State/province	As published on the company website
Postal code	As published on the company website
Country	As published or inferred on the company website
Company phone	As published on the company website
Domain classification	Business or language flag – see Domain Classification table above.
Domain classification score	0 to 1. Similar to a percentage. E.g., ".96" is 96% accurate.
WHOIS date	The date that the domain was registered with WHOIS
WHOIS country	The country that the domain was registered in according to WHOIS
Profound 1st seen date	The date that the domain was first seen by Profound.
Domain IP address	The IP address of the website hosting the domain.
Domain IP address cc	The ISO-3166 2-character Country Code where the IP address of the website is physically hosted.

For additional DBI Technographic attributes, please review:

Global DBI 2022-Q1 report