DOW JONES

Turning a Billion Articles into Actionable Insights

David Arnold
Director of Product Strategy

Dow Jones Professional
Information

May 2018

Dow Jones Professional Information Business

30,000+
Content Sources

1.3bn
Articles

28
Languages

800+
Regions

DNA

DATA / NEWS / ANALYTICS

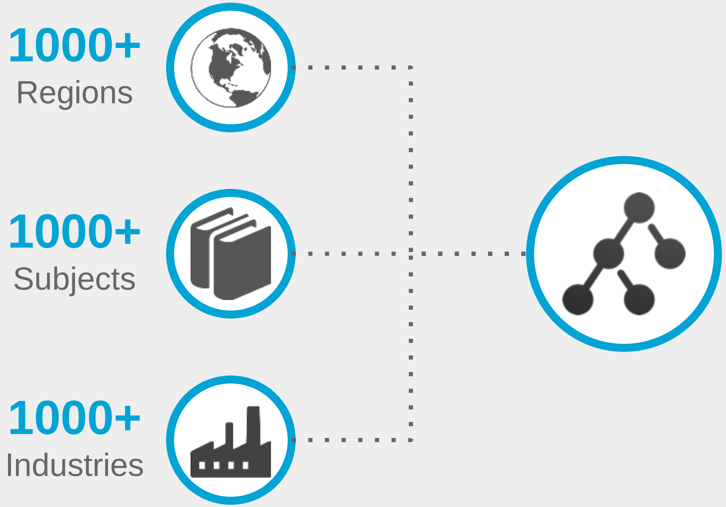
DOW JONES FACTIVA

DOW JONES RISK & COMPLIANCE

May 2018

Dow Jones Intelligent Identifiers

Why a Taxonomy?



1000+ Regions

1000+ Subjects

1000+ Industries

Evolving Taxonomy

- **Not the total number of tags ; We have a dedicated MD team**
- Ensures our taxonomy is **right-sized!**
- Evaluation of new tags (emerging industries, technologies etc…)
- Example: **Drones, Crypto-Currency or AI**
- Evaluation of placement / re-definition
- Evaluation of **tag usage** / removals (VCR technology)
- All part of the 60 day DJID release schedule

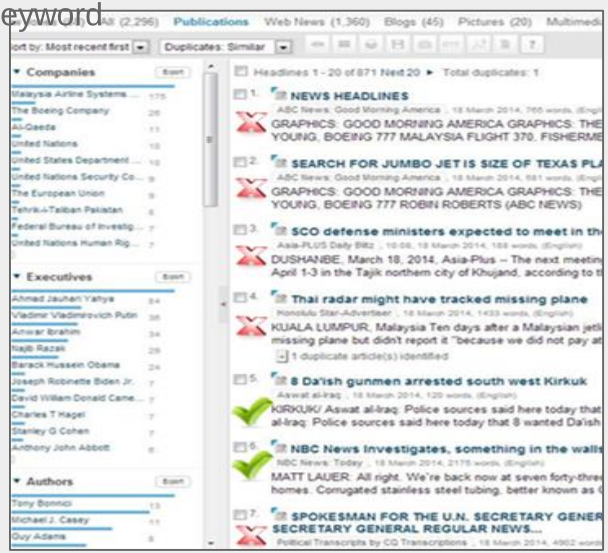
Key Differentiator: *Millions of articles flowing through our pipeline daily, our content is tagged with valuable metadata that defines the article; Harnessing the power of our metadata allows customers to build out models, data analytics, drive visualizations & deep learning.*

May 2018

Keyword Searches vs. Dow Jones Intelligent Identifiers

Keyword search challenges

Retrieves articles with any mention of the keyword



ywords (28) All (2,296) Publications Web News (1,360) Blogs (45) Pictures (20) Multimedi

ort by: Most recent first ▾ Duplicates: Similar ▾

▾ Companies Export

Malaysia Airline Systems ...	175
The Boeing Company	26
Al-Qaeda	11
United Nations	10
United States Department ...	10
United Nations Security Co...	9
The European Union	9
Tehrik-i-Taliban Pakistan	8
Federal Bureau of Investig...	7
United Nations Human Rig...	7

▾ Executives Export

Ahmad Jauhari Yahya	84
Vladimir Vladimirovich Putin	36
Anwar Ibrahim	34
Najib Razak	29
Barack Hussein Obama	24
Joseph Robinette Biden Jr.	7
David William Donald Came...	7
Charles T Hagel	7
Stanley G Cohen	7
Anthony John Abbott	6

▾ Authors Export

Tony Bonnici	13
Michael J. Casey	11
Guy Adams	8

□ Headlines 1 - 20 of 871 Next 20 ▸ Total duplicates: 1

□ 1. NEWS HEADLINES
ABC News: Good Morning America , 18 March 2014, 766 words, (Engli
☒ GRAPHICS: GOOD MORNING AMERICA GRAPHICS: THE
YOUNG, BOEING 777 MALAYSIA FLIGHT 370, FISHERME

□ 2. SEARCH FOR JUMBO JET IS SIZE OF TEXAS PLA
ABC News: Good Morning America , 18 March 2014, 881 words, (Engli
☒ GRAPHICS: GOOD MORNING AMERICA GRAPHICS: THE
YOUNG, BOEING 777 ROBIN ROBERTS (ABC NEWS)

□ 3. SCO defense ministers expected to meet in th
Asia-PLUS Daily Blitz , 10:06, 18 March 2014, 188 words, (English)
☒ DUSHANBE, March 18, 2014, Asia-Plus -- The next meetin
April 1-3 in the Tajik northern city of Khujand, according to t

□ 4. Thai radar might have tracked missing plane
Honolulu Star-Advertiser , 18 March 2014, 1433 words, (English)
☒ KUALA LUMPUR, Malaysia Ten days after a Malaysian jetl
missing plane but didn't report it "because we did not pay at
+ 1 duplicate article(s) identified

□ 5. 8 Da'ish gunmen arrested south west Kirkuk
Aswat al-Iraq , 18 March 2014, 120 words, (English)
✓ KIRKUK/ Aswat al-Iraq: Police sources said here today that
al-Iraq: Police sources said here today that 8 wanted Da'ish

□ 6. NBC News Investigates, something in the walls
NBC News: Today , 18 March 2014, 2175 words, (English)
✓ MATT LAUER: All right. We're back now at seven forty-thre
homes. Corrugated stainless steel tubing, better known as C

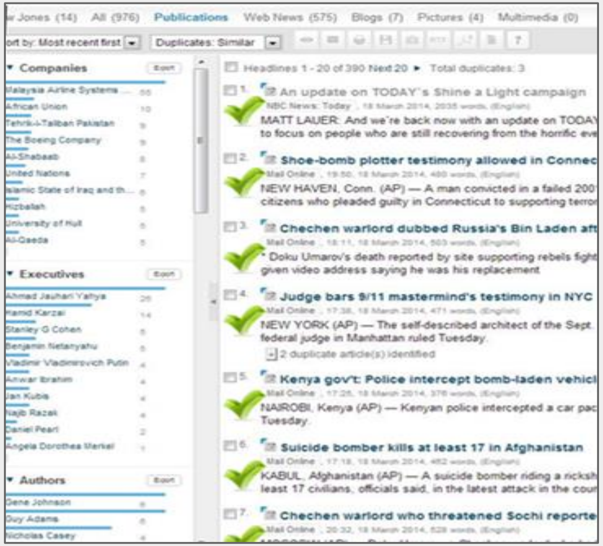
□ 7. SPOKESMAN FOR THE U.N. SECRETARY GENER
SECRETARY GENERAL REGULAR NEWS...
☒ Political Transcripts by CQ Transcriptions , 18 March 2014, 4902 word

Free text = Terroris*

vs.

Solutions using DJID codes

Retrieves relevant content



w Jones (14) All (976) Publications Web News (575) Blogs (7) Pictures (4) Multimedia (0)

ort by: Most recent first ▾ Duplicates: Similar ▾

▾ Companies Export

Malaysia Airline Systems ...	55
African Union	10
Tehrik-i-Taliban Pakistan	9
The Boeing Company	9
Al-Shabaab	8
United Nations	7
Islamic State of Iraq and th...	6
Hizballah	5
University of Hull	5
Al-Qaeda	5

▾ Executives Export

Ahmad Jauhari Yahya	26
Hamid Karzai	14
Stanley G Cohen	5
Benjamin Netanyahu	5
Vladimir Vladimirovich Putin	4
Anwar Ibrahim	4
Jan Kubis	4
Najib Razak	4
Daniel Pearl	2
Angela Dorothea Merkel	1

▾ Authors Export

Gene Johnson	6
Guy Adams	6
Nicholas Casey	4

□ Headlines 1 - 20 of 390 Next 20 ▸ Total duplicates: 3

□ 1. An update on TODAY's Shine a Light campaign
NBC News: Today , 18 March 2014, 2035 words, (English)
✓ MATT LAUER: And we're back now with an update on TODAY
to focus on people who are still recovering from the horrific eve

□ 2. Shoe-bomb plotter testimony allowed in Connec
Mail Online , 19:50, 18 March 2014, 480 words, (English)
✓ NEW HAVEN, Conn. (AP) — A man convicted in a failed 200
citizens who pleaded guilty in Connecticut to supporting terro

□ 3. Chechen warlord dubbed Russia's Bin Laden aft
Mail Online , 18:11, 18 March 2014, 503 words, (English)
✓ * Doku Umarov's death reported by site supporting rebels fight
given video address saying he was his replacement

□ 4. Judge bars 9/11 mastermind's testimony in NYC
Mail Online , 17:38, 18 March 2014, 471 words, (English)
✓ NEW YORK (AP) — The self-described architect of the Sept.
federal judge in Manhattan ruled Tuesday.
+ 2 duplicate article(s) identified

□ 5. Kenya gov't: Police intercept bomb-laden vehicl
Mail Online , 17:25, 18 March 2014, 376 words, (English)
✓ NAIROBI, Kenya (AP) — Kenyan police intercepted a car pac
Tuesday.

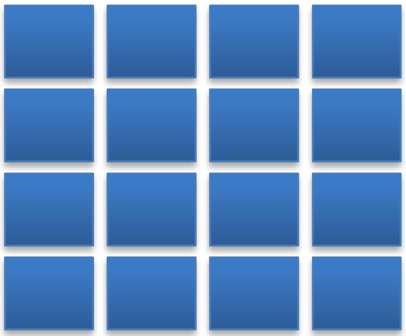
□ 6. Suicide bomber kills at least 17 in Afghanistan
Mail Online , 17:18, 18 March 2014, 452 words, (English)
✓ KABUL, Afghanistan (AP) — A suicide bomber riding a ricksh
least 17 civilians, officials said, in the latest attack in the cour

□ 7. Chechen warlord who threatened Sochi reporte
Mail Online , 20:32, 18 March 2014, 528 words, (English)

DJID code = Acts of Terror

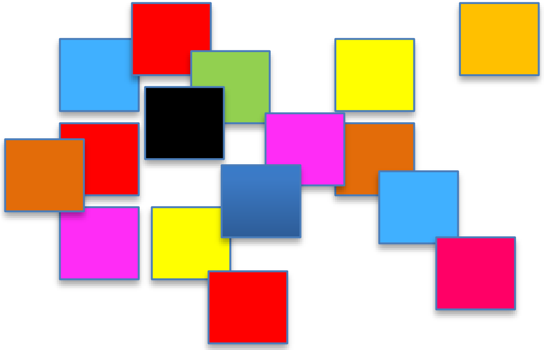
Reading & Understanding Unstructured Text (Our Content)

Structured Data



- Clearly defined
- Expected fields and values
- Easily classified / categorized

Unstructured Data

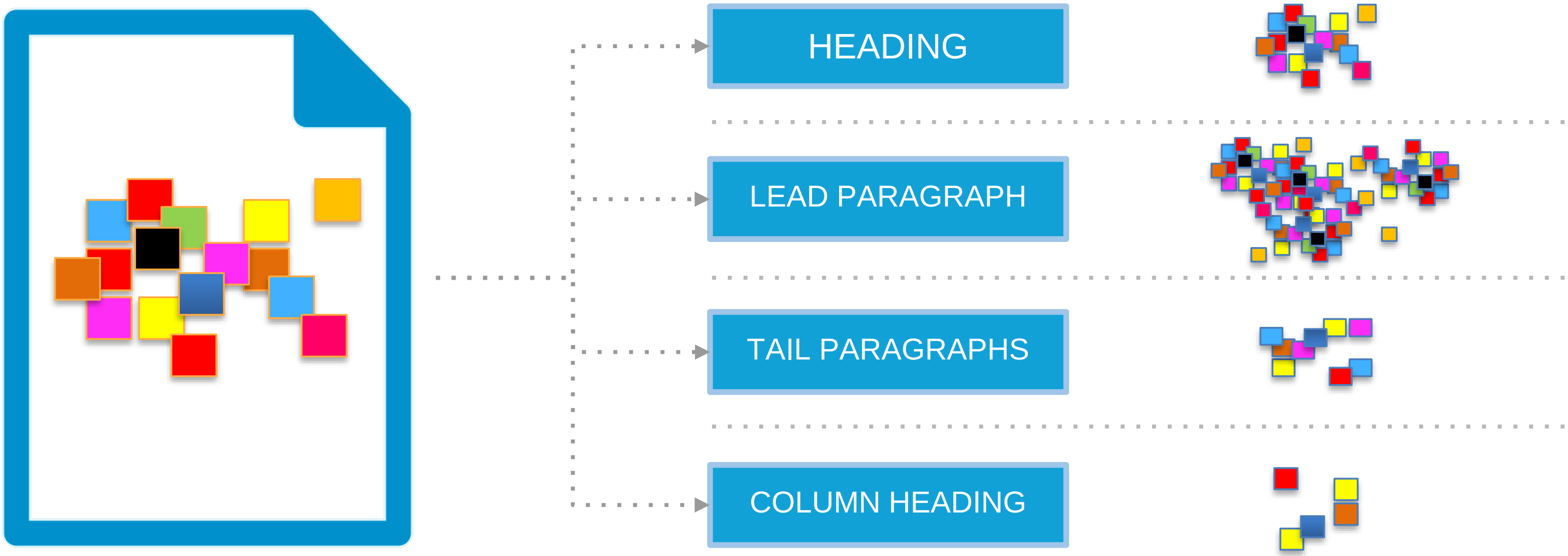


- Data is random / unpredictable
- Data is not defined
- Free form text format
- No expected fields and values

- Classification difficult
- Prone to error
- Requires sophisticated linguistic tools to categorize

Creating structure to a News Article

DISTDOC - Articles are normalized into a fixed format



HEADING

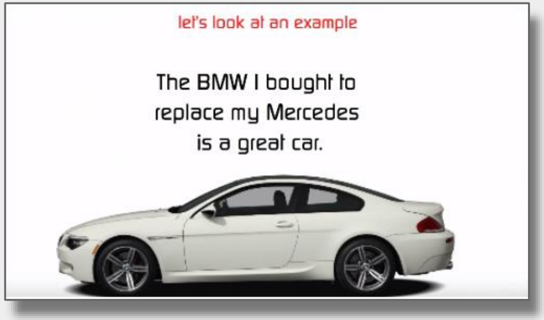
LEAD PARAGRAPH

TAIL PARAGRAPHS

COLUMN HEADING

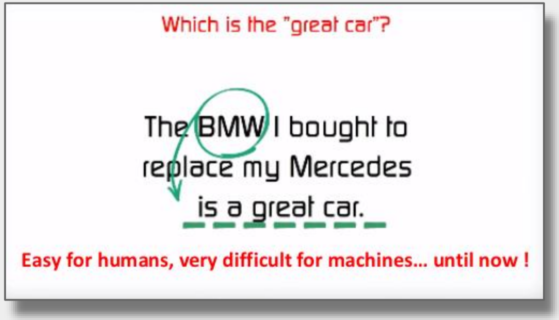
Language is Difficult to Understand

Older systems/machines that only used keywords found it very challenging to correctly interpret the context, the way humans can.



let's look at an example

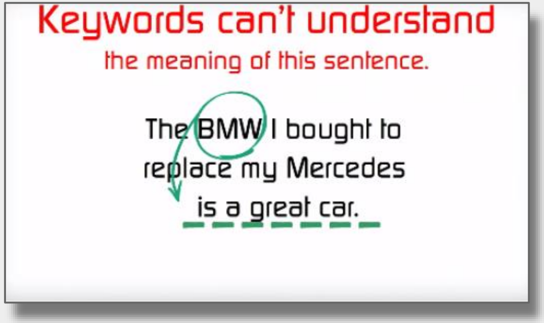
The BMW I bought to replace my Mercedes is a great car.



Which is the "great car"?

The BMW I bought to replace my Mercedes is a great car.

Easy for humans, very difficult for machines... until now !



Keywords can't understand

the meaning of this sentence.

The BMW I bought to replace my Mercedes is a great car.

- Keyword rules based engines
- Rely on proximity and distance to other words
- As a way to collect evidence, what the story was about.
- As you can see here, language is "tricky" !

Language is difficult...

Lexical ambiguity - Homophones

A word that sounds the same but have different meanings.

- “The **bat** slipped from his hand” or a flying mammal?
- “Cinderella had a **ball**” or is ball an event or physical object?
- “Ron **lies** asleep in his bed” or is he telling a lie in his sleep?

“I have four **mouths** to feed at home”

"Mouths" (parts of people) =
"People" (the whole)

“The **strings** were praised for their excellent performance.”

"The strings" (parts of a violin) =
"violins" or "violinists" (the whole)

“Check out my new **wheels**”

[Translation: Check out my new car!]

"Wheels" (parts of a car)
= "car" (the whole)

Imagine **Regional Dialects** (UK English vs. American English) and **Idioms** (“Raining cats and Dogs” and “Barking up the wrong tree”)

What is Autocoding?

- **Autocoding** is the automatic application of Dow Jones Intelligent Identifiers (DJID) codes on all Factiva/DNA content with minimal editorial intervention. (26 languages)

- **Aboutness** is the underlying principle of coding the Dow Jones Intelligent Identifiers taxonomy.

- Autocoding specialists **configure** coding systems in an effort to categorize the item based on what it is about, **not** what sector, country, product or group it may be of interest to (also known as "of-interest-to" coding) or for passing mentions.

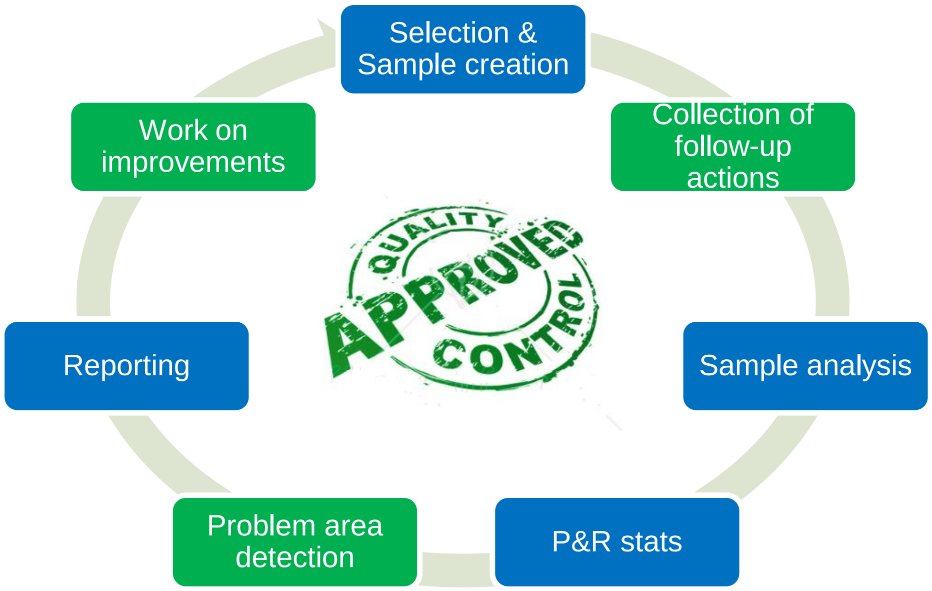
How do we get around these linguistic challenges?

26 Languages

- Hybrid strategic approach - Requiring multiple tools
- **No one tool / system** today can do it all
- Multiple tools fill tackle these linguistic challenges
- Mission: is to hit extremely high **quality** metrics / KPI – No 100%

Quality – A “never-ending” process

- Customers want to hear we “*stay on top*” of our taxonomy and tagging. Remember “*evolving*” DJID?
- We also stay on top of Quality.
- Dedicated Quality team
- Monthly monitoring
- Monitored by staff in Quality, Metadata, Autocoding & external
- Gaps identified assigned for follow-up (by Autocoding)
- Future? More about machine learning and Auto-Feedback loop – *More to come…*



Selection & Sample creation

Collection of follow-up actions

Sample analysis

P&R stats

Problem area detection

Reporting

Work on improvements

Recent GDPR Business Case

- **GDPR** - General Data Protection Regulation
- EU regulation on **data protection and privacy** for all individuals within the European Union.
- It aims primarily to **give control** to citizens and residents over their personal data
- It becomes enforceable on **25 May 2018**

What is a SIP?

Financial

Terrorism

Organized Crime

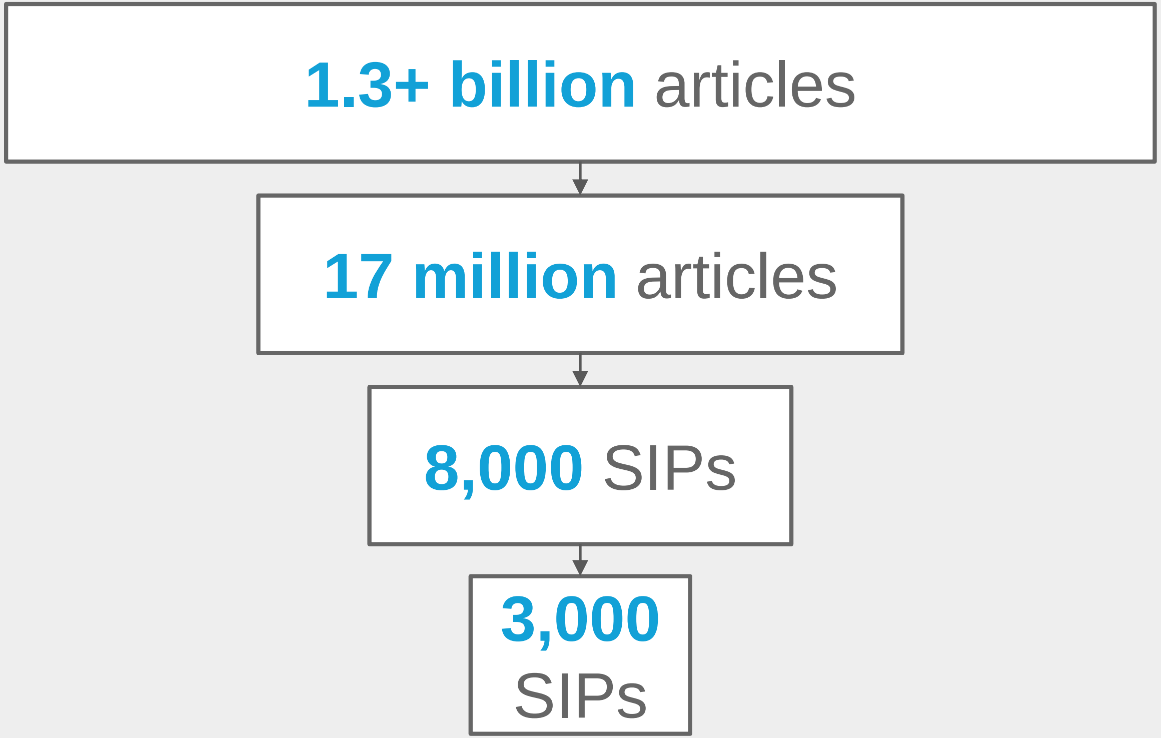
Trafficking

New GDPR Regulation

- Many SIPs were **AQUITTED** of the crime

- **200k SIPs** needed to be reviewed

Solution to the Business Challenge



1.3+ billion articles

17 million articles

8,000 SIPs

3,000 SIPs

Benefit

- Saved **11 research years**
- **GDPR Compliant** Data

Conclusion

- ❖ AI powers Dow Jones **Professional Information Business**
- ❖ **Autocoding** uses AI to apply codes to articles
- ❖ AI utilized to read over **1.3 billion** articles