Food and Agriculture Organization
of the United Nations

FAO’s Data Lab approach to topic- and classification- based indexing of articles

Data Lab

FAO Statistics Division

The Data Lab approach to harmonization

Our sources are unstructured and heterogeneous

- By scraping and crawling, we may get a lot of irrelevant content & data
- To make analysis & retrieval possible, these content & data need to be indexed in a harmonized way

Topic definition to:

- automatically **filter** harvested & web-scraped text and data to **get only content relevant to the topics**
- automatically **classify / tag content** against the topics for more granular analysis

Harmonization with standard classifications to:

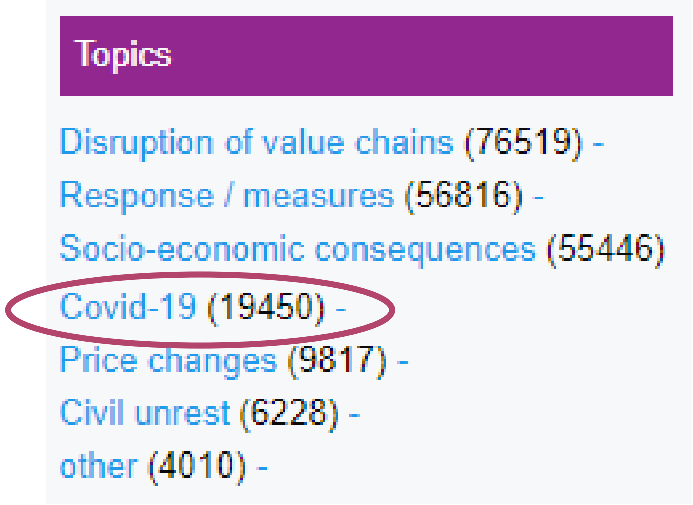
- **Compare / reconcile** results with other standardized data
- Make results **reusable** and interoperable

Harmonization: Topic definition

The platform analyses data on the Covid-19 impact on food value chains.

- > the scope is defined around the following core topics:

- Covid-19
- Food value chain disruptions
  (2 topics: food&agriculture + value chains)
- Socio-economic consequences
  More specific topics:
  - Prices
  - Social unrest
- Government response



Topics

Disruption of value chains (76519) -
Response / measures (56816) -
Socio-economic consequences (55446)
Covid-19 (19450) -
Price changes (9817) -
Civil unrest (6228) -
other (4010) -

Topics & keywords - human input and machine work

Limit manual work / human input for scalability

STEPS

Manual identification of topics and key concepts around topics

- (Optional manual suggestion of important keywords)

Automatic identification of keywords in different languages around the key concepts:

- Translations > variants & synonyms in different languages

Concepts (human input) vs. words (machine-findable)

- CONCEPTS
  the key concepts that define a topic
  selected by humans

e.g. shortage

A concept identified only once in ONE form, preferably by ***experts***

- WORDS (variants, synonyms, translations)
  the different words or lemmas or derivations that can represent concepts
  found by machines

e.g. lack, scarcity, shortages, pénurie…

> variants, synonyms and translations found by ***machines***

Topics & keywords: - Human input

The Wordnet concept, used later by machines to find “words” (lemmas, synonyms and translations)*

Topics and key concepts

	Sub-topics	Key concepts	Normalized
COVID-19			
	covid-19	covid-19	animal virus
		pandemic	pandemic
	lockdown	quarantine	isolation
Disruption of value chains			
	Value chains	value chain	supply
			chain
	upstream	input	input
	downstream	products	commodity
			output
			yield
		distribution	distribution
		retail	retail
		shops	shop
	transport/logistics	transport	transport
	trade	trade	trade

Optional: human-provided keywords

Just flat list; the algorithm will cluster them around topics

Keywords	Sample sentences
contraction	*there has been a contraction in trade*
recession	
remittances	
disruption	
unrest	
crisis	
adverse	
poor	*the number of poor is increasing*
stock	
gap	
spike	*there has been a spike in prices*
cost	
hike	*there has been a price hike*
labour	
unemployment	

Optional, to help machines disambiguate multi-sense words

* *Wordnet is a lexical database that defines “senses” (concepts), relations between senses, and lemmas for each sense.*

Keywords - Machine work

Automatic keywords

Algorithms add **machine-extracted keywords** to human-suggested keywords. Algorithms extract keywords from relevant text corpora.

- keyword extraction methods based on frequency, distance, co-occurrence
- topic mapping techniques like LDA.

Clustering

Algorithms use existing **lexical and semantic resources** and their similarity algorithms to calculate to which degree keywords can be clustered around the key concepts under predefined topics.

- **Wordnet*** database through Python NLTK interface
  - Wordnet “senses” & lemmas
  - Wordnet relations (synonyms, hyponyms…)
  - Wordnet similarity algorithm

Variants, translations…

Algorithms use existing **lexical and semantic resources** to add synonyms, hyponyms, derivations and translations to all clustered keywords.

- Wordnet database + **Open Multilingual Wordnet** (OMW) through Python NLTK interface
  - English Wordnet sense > English lemmas
  - English Wordnet sense > OMW language versions > lemmas in *n* languages

Sentiment

Algorithms use existing **lexical and semantic resources** to get the sentiment or “polarity” of words.

- Wordnet database + **SentiWordnet** Python extension
- Polyglot sentiment (polarity)

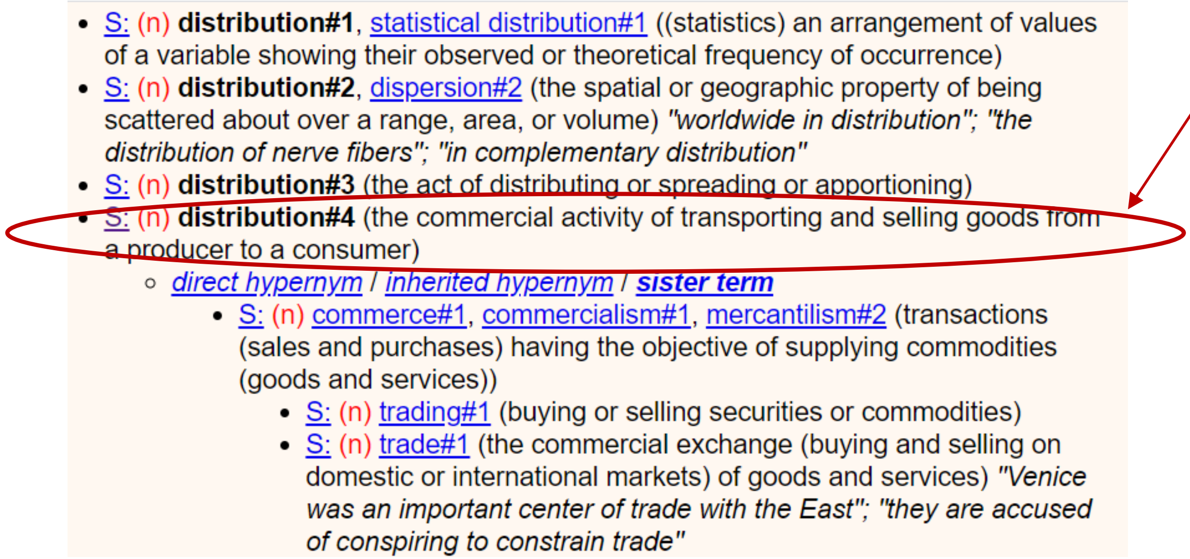
** Wordnet is a lexical database that defines “senses” (concepts), relations between senses and lemmas for each sense.*

Keyword “senses” and translations: Wordnet

Wordnet

Wordnet is an English lexical database that defines **“senses” (concepts)**, **relations between senses**, and lemmas for each sense

Importance of ***senses****: for food value chains we’re interested only in sense n. 4 of “****distribution****”:*

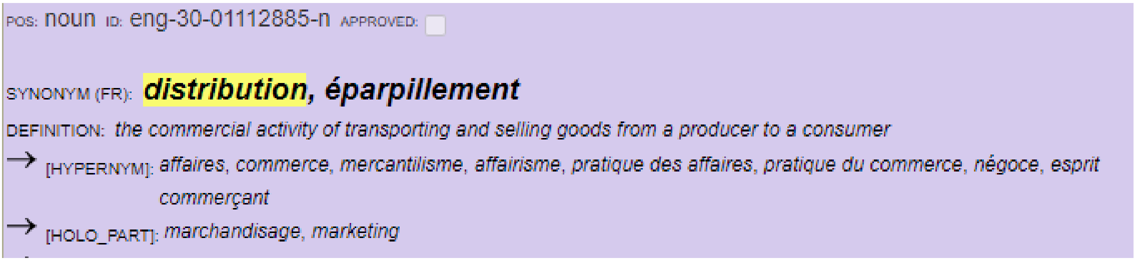


- S: (n) **distribution#1**, statistical distribution#1 ((statistics) an arrangement of values of a variable showing their observed or theoretical frequency of occurrence)
- S: (n) **distribution#2**, dispersion#2 (the spatial or geographic property of being scattered about over a range, area, or volume) *"worldwide in distribution"; "the distribution of nerve fibers"; "in complementary distribution"*
- S: (n) **distribution#3** (the act of distributing or spreading or apportioning)
- S: (n) **distribution#4** (the commercial activity of transporting and selling goods from a producer to a consumer)
    - *direct hypernym* / *inherited hypernym* / ***sister term***
        - S: (n) commerce#1, commercialism#1, mercantilism#2 (transactions (sales and purchases) having the objective of supplying commodities (goods and services))
            - S: (n) trading#1 (buying or selling securities or commodities)
            - S: (n) trade#1 (the commercial exchange (buying and selling on domestic or international markets) of goods and services) *"Venice was an important center of trade with the East"; "they are accused of conspiring to constrain trade"*

Open Multilingual Wordnet

Wordnets have been created in different languages and connected to the original Wordnet “senses” (so that relations between senses don’t have to be defined again)

e.g. French Wordnet



POS: noun ID: eng-30-01112885-n APPROVED: □

SYNONYM (FR): ***distribution*****, *éparpillement***

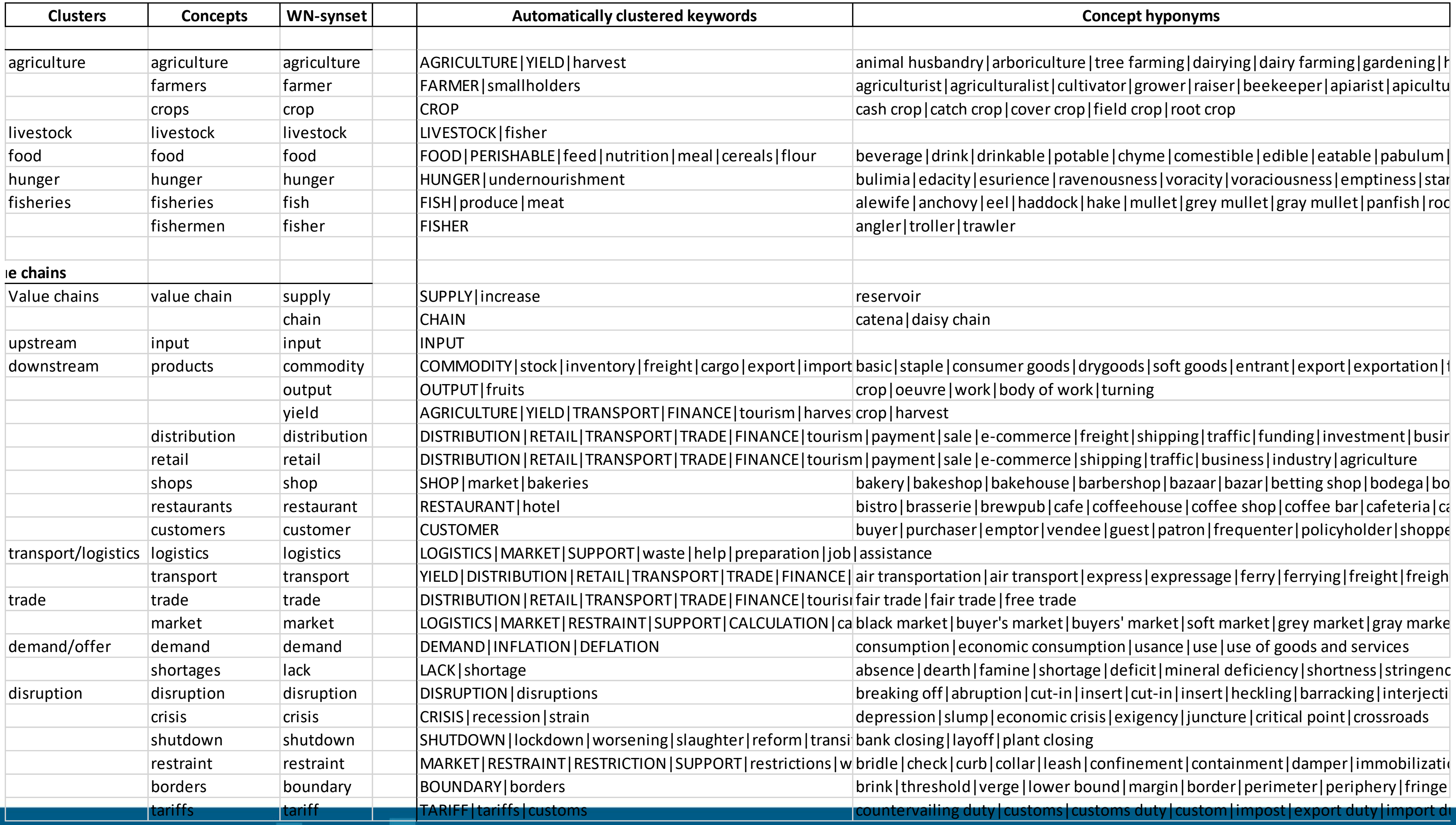
DEFINITION: *the commercial activity of transporting and selling goods from a producer to a consumer*

→ [HYPERNYM]: *affaires, commerce, mercantilisme, affairisme, pratique des affaires, pratique du commerce, négoce, esprit commerçant*

→ [HOLO_PART]: *marchandisage, marketing*

Sample results from Wordnet similarity and synonyms functions

Automatic clustering of keywords around concepts by Wordnet

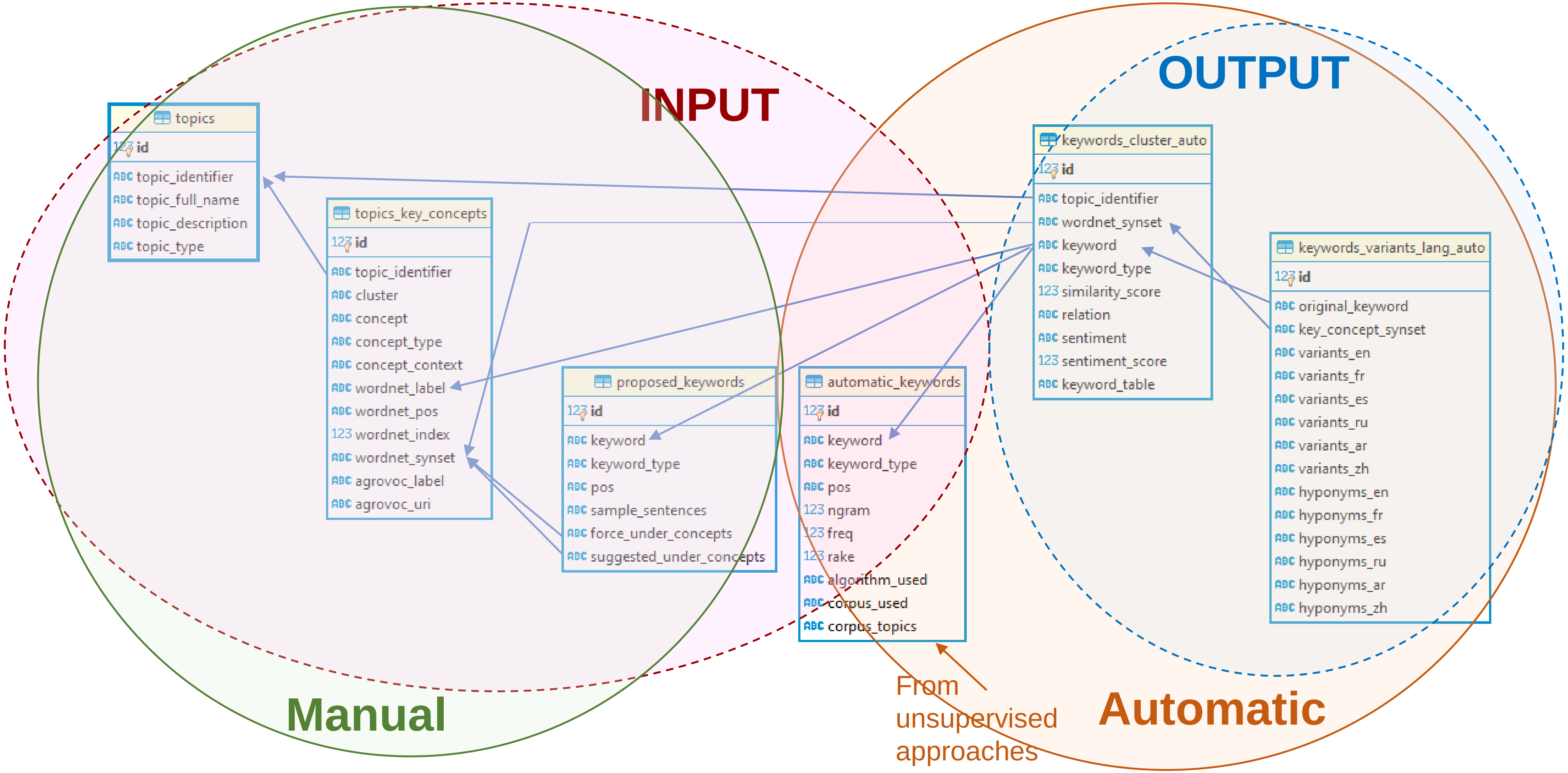


Clusters	Concepts	WN-synset		Automatically clustered keywords	Concept hyponyms
agriculture	agriculture	agriculture		AGRICULTURE|YIELD|harvest	animal husbandry|arboriculture|tree farming|dairying|dairy farming|gardening|h
	farmers	farmer		FARMER|smallholders	agriculturist|agriculturalist|cultivator|grower|raiser|beekeeper|apiarist|apicultu
	crops	crop		CROP	cash crop|catch crop|cover crop|field crop|root crop
livestock	livestock	livestock		LIVESTOCK|fisher	
food	food	food		FOOD|PERISHABLE|feed|nutrition|meal|cereals|flour	beverage|drink|drinkable|potable|chyme|comestible|edible|eatable|pabulum|
hunger	hunger	hunger		HUNGER|undernourishment	bulimia|edacity|esurience|ravenousness|voracity|voraciousness|emptiness|star
fisheries	fisheries	fish		FISH|produce|meat	alewife|anchovy|eel|haddock|hake|mullet|grey mullet|gray mullet|panfish|roc
	fishermen	fisher		FISHER	angler|troller|trawler
ıe chains					
Value chains	value chain	supply		SUPPLY|increase	reservoir
		chain		CHAIN	catena|daisy chain
upstream	input	input		INPUT	
downstream	products	commodity		COMMODITY|stock|inventory|freight|cargo|export|import	basic|staple|consumer goods|drygoods|soft goods|entrant|export|exportation|f
		output		OUTPUT|fruits	crop|oeuvre|work|body of work|turning
		yield		AGRICULTURE|YIELD|TRANSPORT|FINANCE|tourism|harves	crop|harvest
	distribution	distribution		DISTRIBUTION|RETAIL|TRANSPORT|TRADE|FINANCE|tourism|payment|sale|e-commerce|freight|shipping|traffic|funding|investment|busir	
	retail	retail		DISTRIBUTION|RETAIL|TRANSPORT|TRADE|FINANCE|tourism|payment|sale|e-commerce|shipping|traffic|business|industry|agriculture	
	shops	shop		SHOP|market|bakeries	bakery|bakeshop|bakehouse|barbershop|bazaar|bazar|betting shop|bodega|bo
	restaurants	restaurant		RESTAURANT|hotel	bistro|brasserie|brewpub|cafe|coffeehouse|coffee shop|coffee bar|cafeteria|ca
	customers	customer		CUSTOMER	buyer|purchaser|emptor|vendee|guest|patron|frequenter|policyholder|shoppe
transport/logistics	logistics	logistics		LOGISTICS|MARKET|SUPPORT|waste|help|preparation|job|assistance	
	transport	transport		YIELD|DISTRIBUTION|RETAIL|TRANSPORT|TRADE|FINANCE|	air transportation|air transport|express|expressage|ferry|ferrying|freight|freigh
trade	trade	trade		DISTRIBUTION|RETAIL|TRANSPORT|TRADE|FINANCE|touris	fair trade|fair trade|free trade
	market	market		LOGISTICS|MARKET|RESTRAINT|SUPPORT|CALCULATION|ca	black market|buyer's market|buyers' market|soft market|grey market|gray marke
demand/offer	demand	demand		DEMAND|INFLATION|DEFLATION	consumption|economic consumption|usance|use|use of goods and services
	shortages	lack		LACK|shortage	absence|dearth|famine|shortage|deficit|mineral deficiency|shortness|stringenc
disruption	disruption	disruption		DISRUPTION|disruptions	breaking off|abruption|cut-in|insert|cut-in|insert|heckling|barracking|interjecti
	crisis	crisis		CRISIS|recession|strain	depression|slump|economic crisis|exigency|juncture|critical point|crossroads
	shutdown	shutdown		SHUTDOWN|lockdown|worsening|slaughter|reform|transi	bank closing|layoff|plant closing
	restraint	restraint		MARKET|RESTRAINT|RESTRICTION|SUPPORT|restrictions|w	bridle|check|curb|collar|leash|confinement|containment|damper|immobilizati
	borders	boundary		BOUNDARY|borders	brink|threshold|verge|lower bound|margin|border|perimeter|periphery|fringe
	tariffs	tariff		TARIFF|tariffs|customs	countervailing duty|customs|customs duty|custom|impost|export duty|import du

Synonyms and variants found by Wordnet

word	variants
agriculture	farming|agriculture|husbandry
food	food|nutrient
farmer	farmer|husbandman|granger|sodbuster
crop	crop|harvest
livestock	livestock|stock|farm animal
hunger	hunger|hungriness
fisher	fisherman|fisher
chain	chain|concatenation
disruption	break|interruption|disruption|gap
commodity	commodity|trade good|good
output	output|yield
shop	shop|store
travel	travel|traveling|travelling
market	market|marketplace|market place
lack	lack|deficiency|want
shutdown	closure|closedown|closing|shutdown

Keywords - clustering database tables



topics
- id
- topic_identifier
- topic_full_name
- topic_description
- topic_type

topics_key_concepts
- id
- topic_identifier
- cluster
- concept
- concept_type
- concept_context
- wordnet_label
- wordnet_pos
- wordnet_index
- wordnet_synset
- agrovoc_label
- agrovoc_uri

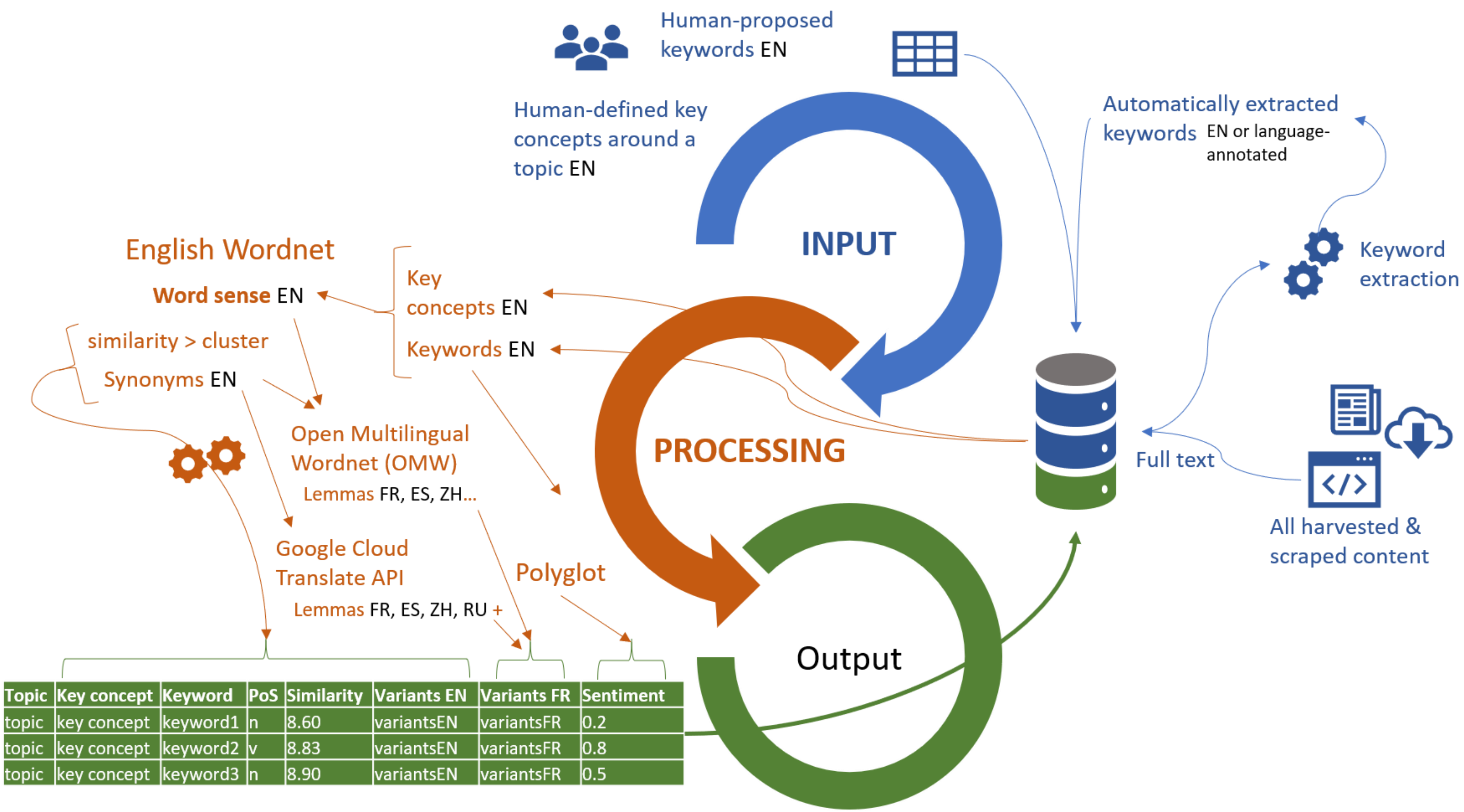
proposed_keywords
- id
- keyword
- keyword_type
- pos
- sample_sentences
- force_under_concepts
- suggested_under_concepts

automatic_keywords
- id
- keyword
- keyword_type
- pos
- ngram
- freq
- rake
- algorithm_used
- corpus_used
- corpus_topics

keywords_cluster_auto
- id
- topic_identifier
- wordnet_synset
- keyword
- keyword_type
- similarity_score
- relation
- sentiment
- sentiment_score
- keyword_table

keywords_variants_lang_auto
- id
- original_keyword
- key_concept_synset
- variants_en
- variants_fr
- variants_es
- variants_ru
- variants_ar
- variants_zh
- hyponyms_en
- hyponyms_fr
- hyponyms_es
- hyponyms_ru
- hyponyms_ar
- hyponyms_zh

Keyword management workflow


Human-proposed keywords EN
Human-defined key concepts around a topic EN
Automatically extracted keywords EN or language-annotated
INPUT
Keyword extraction
Full text
All harvested & scraped content
English Wordnet
Word sense EN
Key concepts EN
Keywords EN
similarity > cluster
Synonyms EN
Open Multilingual Wordnet (OMW)
Lemmas FR, ES, ZH...
PROCESSING
Google Cloud Translate API
Lemmas FR, ES, ZH, RU +
Polyglot
Output

Topic	Key concept	Keyword	PoS	Similarity	Variants EN	Variants FR	Sentiment
topic	key concept	keyword1	n	8.60	variantsEN	variantsFR	0.2
topic	key concept	keyword2	v	8.83	variantsEN	variantsFR	0.8
topic	key concept	keyword3	n	8.90	variantsEN	variantsFR	0.5

Harmonization: Use of standard classifications

Additional tagging for cross-topic analysis: **commodities, geopolitical**.

For commodities and geopolitical entities standard classifications exist.

- Need to use **variants in different languages** to match as many documents as possible
- Need to **consolidate tagging under the standard code / label**

COMMODITIES

- Reference: **CPC 2.1**
  (only most traded commodities from FAOSTAT)
- Synonyms and translations taken from mapped classifications (FCL, HS, ICC) and Yandex
  - ➢ DB table with all synonyms and translations associated with the official code in the CPC classification and the FAOSTAT name
  - ➢ all tagging is consolidated under the CPC 2.1 code
  - ➢ an additional aggregation tag by commodity group is added using the CPC 2.1 “groups” level

GEOPOLITICAL

- Reference: **M49**
- Variant names, demonyms and translations taken from **Wikidata** *(also capitals and admin units)*
  - ➢ DB table with all variants and translations associated with the official code and official name in the M49 classification
  - ➢ all tagging is consolidated under the M49 code and label
  - ➢ an additional aggregation tag by sub-region and region is added using M49 aggregations

Example for geopolitical entities (only displaying EN and FR)

<table>
<tr><th>M49</th><th>ISO2</th><th>ISO3</th><th>M49 name EN</th><th>M49 name FR</th><th>Variants EN</th><th>Variants FR</th><th>Demonym EN</th><th>Demonym FR</th><th>Capital EN</th><th>Capital FR</th></tr>
<tr><td>784</td><td>AE</td><td>ARE</td><td>United Arab Emirates</td><td>les Émirats arabes unis</td><td>United Arab Emirates|Emirates|United Arab Emirates|UAE|U.A.E.|the United Arab Emirates|the UAE|the U.A.E.|the Emirates|Emirates|AE|ae</td><td>Émirats arabes unis|Emirates|Émirats arabes unis|E.A.U.</td><td>Emirian|Emiri|Emirati</td><td></td><td>Abu Dhabi</td><td>Abou Dabi</td></tr>
<tr><td>32</td><td>AR</td><td>ARG</td><td>Argentina</td><td>la République argentine</td><td>Argentina|Argentina|Argentine Republic|AR|ar|ARG|AR</td><td>Argentine|Argentine|République argentine</td><td>Argentinian|Argentine</td><td></td><td>Buenos Aires</td><td>Buenos Aires</td></tr>
<tr><td>854</td><td>BF</td><td>BFA</td><td>Burkina Faso</td><td>le Burkina Faso</td><td>Burkina Faso|Burkina Faso|BF|bf</td><td>Burkina Faso|Burkina Faso|Burkina</td><td>Burkinabè|Burkinabe</td><td>Burkinabé</td><td>Ouagadougou</td><td>Ouagadougou</td></tr>
<tr><td>124</td><td>CA</td><td>CAN</td><td>Canada</td><td>le Canada</td><td>Canada|Canada|Dominion of Canada|CA|British North America|CAN|can|CDN|ca|CA</td><td>Canada|Canada|CA</td><td>Canadian</td><td>Canadienne|Canadien</td><td>Ottawa</td><td>Ottawa</td></tr>
</table>

Reuse of keywords

Keywords are saved in the database for further reuse

with Part-of-Speech, similarity score, sentiment; with lemmas in all languages

Same process for commodity and geographic tags

Queries to filter relevant content

- Select the topic(s) (e.g. covid19 + ag/food + value chains)
- Topics joined by AND (at least one keyword from each has to be present)
- Keywords within the topic joined by OR
- Iterate over languages
- Adjustments depending on constraints of the query engine
  (e.g. limit of keywords in Google: use similarity score to limit to most relevant)

Sample Google query

```
+(~supply OR 'value chain' OR market
OR trade OR ~transport OR import OR
export OR distribution OR customs OR
borders OR ~shortage OR ~retail OR
vessels OR ~trucks) +(coronavirus OR
covid OR pandemic OR lockdown) +(wheat
OR grain))
```

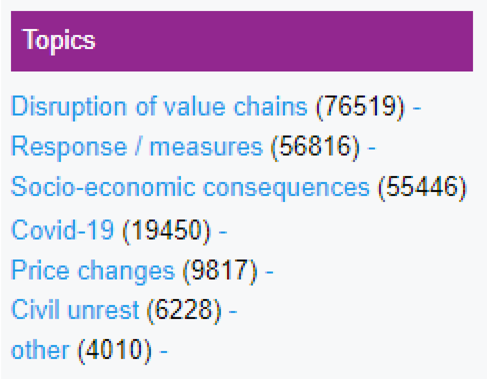
Tagging to allow for faceted search and further analysis

- Check for the presence of lemmatized keywords and variants in lemmatized text according to text language
  (Optionally set minimum number of keywords, or filter keywords above a certain similarity threshold)

➢ Tag under the keyword and under related key concept and associated topic, in all languages

➢ Tags as Solr fields
➢Open Semantic Search facets
➢Further analysis, ShinyApps

Faceted search



Topics

Disruption of value chains (76519) -
Response / measures (56816) -
Socio-economic consequences (55446)
Covid-19 (19450) -
Price changes (9817) -
Civil unrest (6228) -
other (4010) -

Reuse of keywords: tagging

Keywords are saved in the database for further reuse

with Part-of-Speech, similarity score, sentiment; with lemmas in all languages

Same process for commodity and geographic tags

Queries to filter relevant content

- Select the topic(s) (e.g. covid19 + ag/food + value chains)
- Topics joined by AND (at least one keyword from each has to be present)
- Keywords within the topic joined by OR
- Iterate over languages
- Adjustments depending on constraints of the query engine
  (e.g. limit of keywords in Google: use similarity score to limit to most relevant)

Sample Google query

```
+(~supply OR 'value chain' OR market
OR trade OR ~transport OR import OR
export OR distribution OR customs OR
borders OR ~shortage OR ~retail OR
vessels OR ~trucks) +(coronavirus OR
covid OR pandemic OR lockdown) +(wheat
OR grain))
```

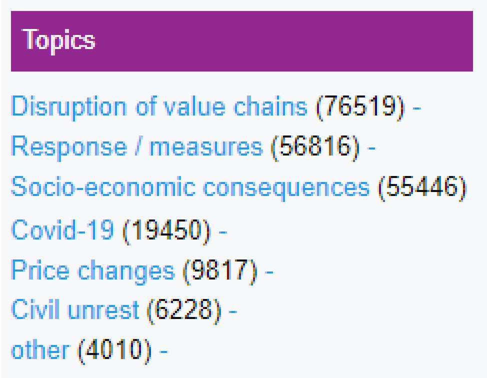
Tagging to allow for faceted search and further analysis

- Check for the presence of lemmatized keywords and variants in lemmatized text according to text language *(Optionally set minimum number of keywords, filter keywords above a certain similarity threshold, calibrate to text length or topic breadth…)*

➢ Tag under the keyword(s) and under related key concept and associated topic
or under the keyword(s) and corresponding standard code for geo and commodities

➢ Tags as Solr fields
➢Open Semantic Search facets
➢Further analysis, ShinyApps

Faceted search



Topics

Disruption of value chains (76519) -
Response / measures (56816) -
Socio-economic consequences (55446)
Covid-19 (19450) -
Price changes (9817) -
Civil unrest (6228) -
other (4010) -

Harmonization > NLP to tag content

With topic keywords and standard classification keywords defined:

- Full text in the DB > tokenized and lemmatized, language detected
- Keywords (lemmas) matched against lemmatized full text using language-specific models

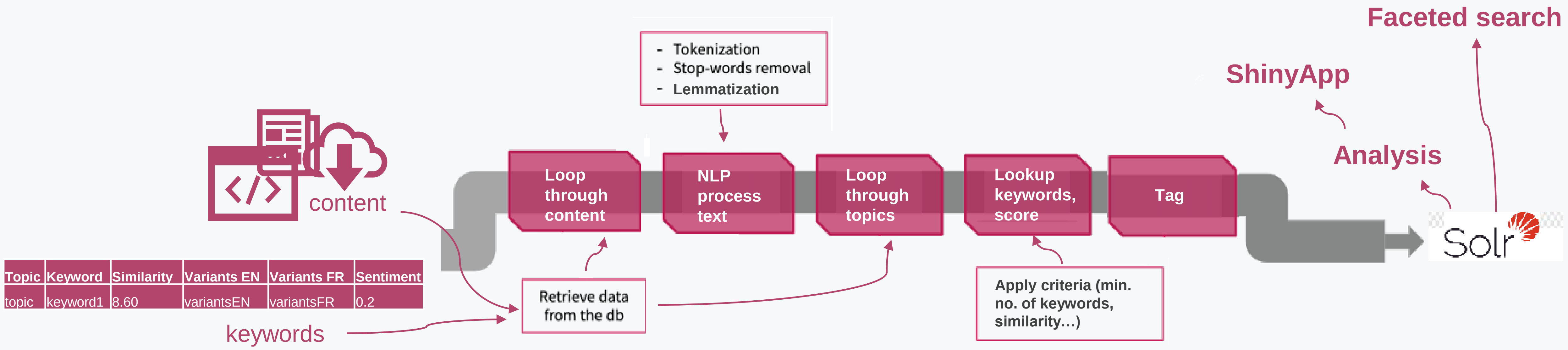
SpaCy

→ Content tagged against keywords and classified **under topics / classification terms**

- *flexible scoring criteria: minimum no. of keywords, only keywords with highest similarity scores, calibrated to full text length and topic breadth*

Polyglot

→ High-sentiment (polarity) keywords → content **sentiment-tagged**



Topic	Keyword	Similarity	Variants EN	Variants FR	Sentiment
topic	keyword1	8.60	variantsEN	variantsFR	0.2

Thank you for your attention

Valeria Pesce (*valeria.pesce@fao.org*)

for the Data Lab team (*ESS-datalab@fao.org*)