>> FAO Statistics Division

DataLab

Paris21: Measuring Use of Food and Agricultural Statistics in Policy Making

17th June 2020

Outline

- Methodology
- Architecture, data, and workflow
- Results and web application to explore the data (ShinyApp and FAOLEX search)

Objective

Measuring use of food and agricultural statistics in public policy documents

History of the indicator

2010

Busan Action Plan for Statistics

To assess impact of international statistical capacity building activities in developing countries.

2015

Machine Learning

The manual analysis was partially replaced with data mining and machine learning.

2018

Updating the keyword list

PARIS21 improved the indicator by updating the keyword list and scoring methodology

2019

PARIS21 and FAO

PARIS21 and FAO are working together to apply the scope of the indicator to **agricultural policy making**.

Measuring the use of statistics

How the presence of statistical concepts is represented?

Measuring the use of statistics

Levels: sub-indexes

➔ The policy documents are split into sentences, and then their sentences are classified ***into three mutually exclusive levels based on a set of keywords.***

COMPONENT	DESCRIPTION	EXAMPLE
Basic (**Level 1**)	The introduction of a measurable concept for indicating the state of and progress towards a specific development outcome	*The Government implemented a number of social protection interventions such as the social cash transfer, school feeding and the **food security** pack.*
Diagnosis and quantification (**Level 2**)	The depiction of the state of such measurable concept and the presumed causes for its variation.	*The **mortality rate of 5** per **1 000 adults**, improvement of **5%** in **adult literacy**.*
Statistical analysis (**Level 3**)	It includes use of statistical methods to depict the state of the measurable concept and the causes of its variation	*The improvement of 5% in adult literacy rate is correlated to improvements in life expectancy or the income of the lower **quartile**.*

Measuring the use of statistics

Levels: sub-indexes

➔ In addition to the three levels, our analysis also searches for the use of ***disaggregation*** when a statistical term is referenced.

COMPONENT	DESCRIPTION	EXAMPLE
Disaggregation	Use of disaggregation while referencing to statistics.	---

➔ The structure and keywords of disaggregation originates from the stands for data disaggregation of UNSD[1].

[1] https://unstats.un.org/sdgs/files/Overview%20of%20Standards%20for%20Data%20Disaggregation.pdf

Measuring the use of statistics

Levels: sub-indexes

Disaggregation indicators	Keywords		Disaggregation indicators	Keywords
Sex	Man, Woman, Men, Women, Girls, Boys, Female, Male		**Agroecological zone**	Climate variables, Type of Soil, Geomorphology
Age	Youth, Adults, infants		**Water management**	Irrigated, Non-irrigated
Indigenous status	Indigenous, indigenous goup		**Type of Conservation Facility**	Medium-term, Long-term
Type of products:	Crop, livestock, mixed		**Level of risk**	Critical, critical-maintained, endangered, endangered-maintained
Type of products-subgroups:	wheat, rice, maize, millet, sorghum		**Type of government authority**	Agriculture , Industry Sector, Service Sector
Crop group, Food group	Cereals, Pulses, Fruits, Vegetables, Roots, Tubers, Oil-Bearing crops, Animals Products, Fish		**Economic Sector**	High income, low income, middle income
Geographical	Urban, Rural, National, Sub-national		**Poor and vulnerable population**	Poor, rich, poverty, wealth
Geographical Location	River basin, watershed, administrative unit		**Level of Food Price Anomaly**	Moderately high, High
Land Cover	Forestland, Cropland, Grassland, Wetlands, Settlements		**Type of Tenure**	Customary, Freehold, Leasehold
Agricultural holding type	Household, Non-household		**Hydrological Units**	River Basins , Aquifers
Type of Enterprise	Farming, Pastoral, Forestry, Fisheries, Hunting		**Points of the Value Chain**	Farm , Transport , Markets , Processors , Small-holders , large-commercial farms, firms
Size of Enterprise	Small, Medium, Large		**Mountain Elevation Classes**	Kapos classes

Keywords

FAOSTAT

- 938 keywords

- Languages:
  - English
  - Spanish
  - French

SDGS

- 25 keywords

- Languages:
  - English
  - Spanish
  - French

Disaggregation

- 24 indicators

- There is a list of keywords for each indicators.

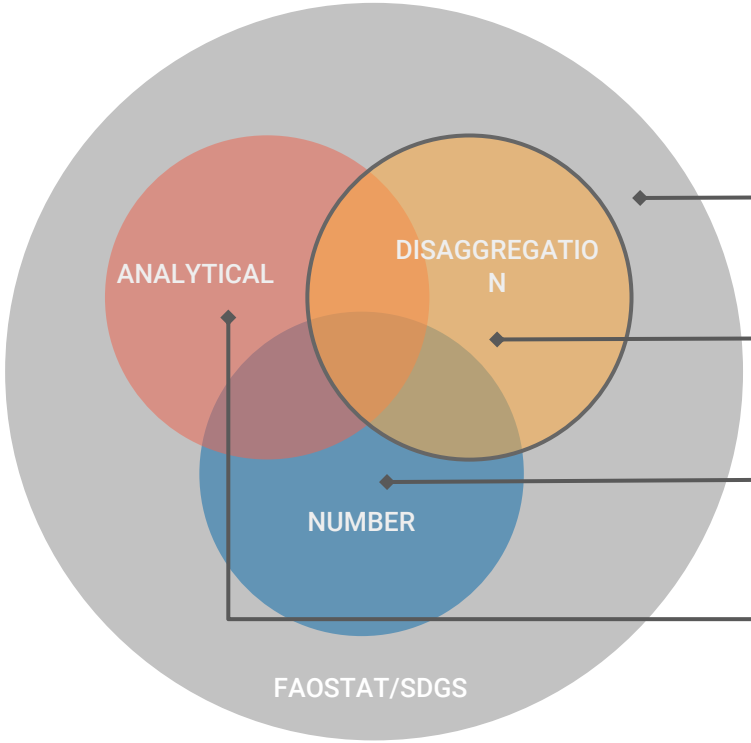
- Languages:
  - English
  - Spanish
  - French

Analytical (math)

- 115 keywords

- Languages:
  - English
  - Spanish
  - French

Classification

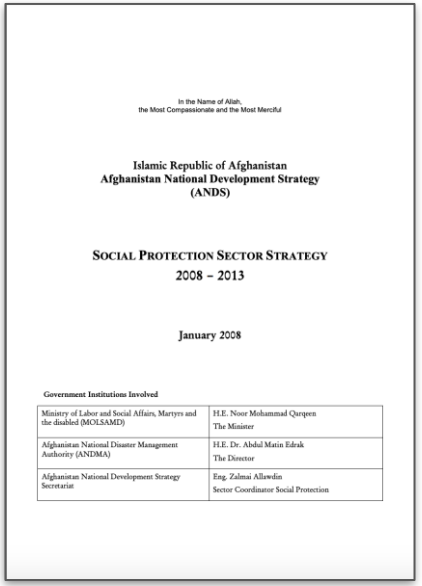


LEVEL 1	FAOSTAT/SDGS
DISAGGREGATIO N	Disag. + FAOSTAT/SDGS
LEVEL 2	FAOSTAT/SDGS + Figures
LEVEL 3	FAOSTAT/SDGS + Analytical

Example

Country: Afghanistan
Title: Social Protection Sector Strategy
Year: 2008
PDF file: afg152314.pdf
Sentence classified at level 3

"Food poverty has been estimated to be even bigger: around 45 percent of the Afghan population was not able to purchase a basic food basket to provide 2100 calories consumption per day."



In the Name of Allah,
the Most Compassionate and the Most Merciful

Islamic Republic of Afghanistan
Afghanistan National Development Strategy
(ANDS)

SOCIAL PROTECTION SECTOR STRATEGY
2008 – 2013

January 2008

Government Institutions Involved

Ministry of Labor and Social Affairs, Martyrs and the disabled (MOLSAMD)	H.E. Noor Mohammad Qurqeen The Minister
Afghanistan National Disaster Management Authority (ANDMA)	H.E. Dr. Abdul Matin Edrak The Director
Afghanistan National Development Strategy Secretariat	Eng. Zalmai Allawdin Sector Coordinator Social Protection

NUMBER | DISAGGREGATION | FAOSTAT/SDGS | ANALYTICAL

Algorithm (simple version)

LOOP OVER DOCUMENTS:

Step 1: it splits the document into sentences;

Step 2: for each sentence checks if it belongs one of the three levels;

Step 3: for each level computes the proportion of sentences;

Step 4: normalise each level to the scale min-max (0 - 1);

At this step we have the three sub-indexes: basic; diagnosis and quantification; statistical analysis.

Weighting

1. Principal Component Analysis

Weighting: Principal Component Analysis

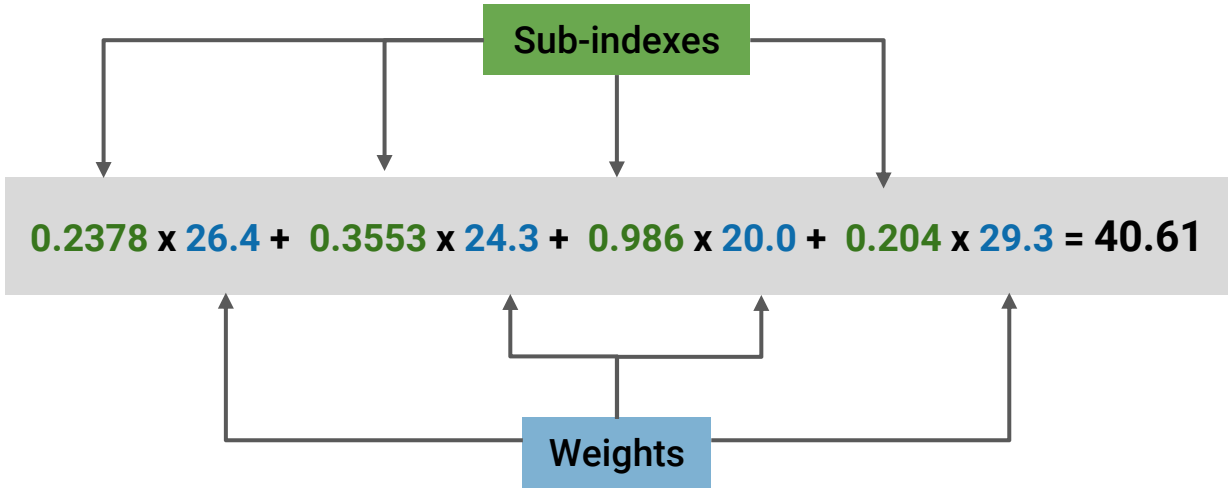
DOCUMENT	LEVEL 1	LEVEL 2	LEVEL 3	DISAGG
doc_01	I_{11}	I_{12}	I_{13}	I_{1D}
doc_02	I_{21}	I_{22}	I_{23}	I_{2D}
. . .	. . .	. . .	. . .	. . .
doc_N	I_{n1}	I_{n2}	I_{n3}	I_{nD}

- Where I_{xy} is the sub-index for the level *y* and the document *x;*
- The PCA method is applied on the matrix above to extract the weights of each level and the PCA ***(Index Overall)***.

Weighting: Principal Component Analysis

SUB-INDEX	WEIGHTS *PCA*
Level 1	26.4
Level 2	24.3
Level 3	20.0
Disaggregation	29.3

ID	Country	Year	# sentence	Level 1	Level 2	Level 3	Level Disagg.	Score PCA
mne180387.pdf	Montenegro	2016	3873	0.2378	0.3553	0.986	0.204	40.61



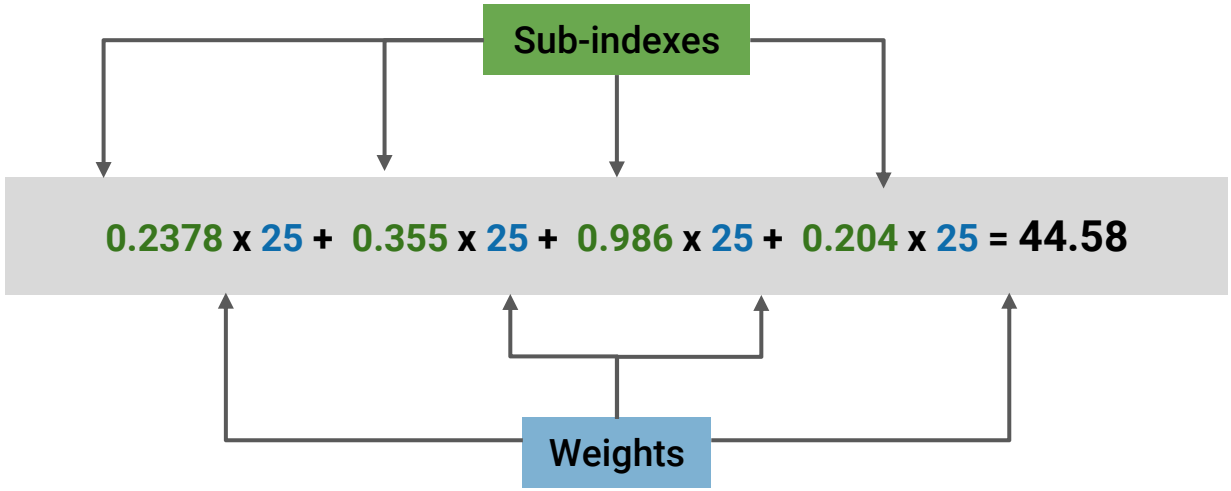
Weighting

2. Simple linear combination: equal weights

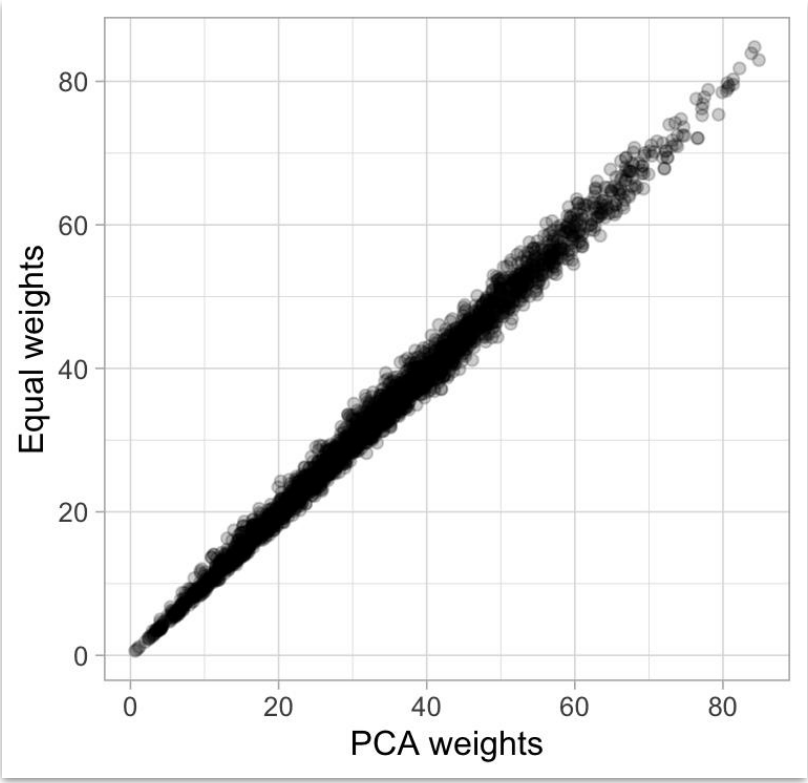
Weighting: equal weights

SUB-INDEX	WEIGHTS *PCA*
Level 1	25
Level 2	25
Level 3	25
Disaggregation	25

ID	Country	Year	# sentence	Level 1	Level 2	Level 3	Level Disagg.	Score Simple
mne180387.pdf	Montenegro	2016	3873	0.238	0.355	0.986	0.204	44.58



Weighting



Weighting

3. Weighting by experts?

FAOLEX database

"FAOLEX is a comprehensive and up-to-date legislative and policy database, one of the world's largest online repositories of national laws, regulations and policies on food, agriculture and natural resources management."

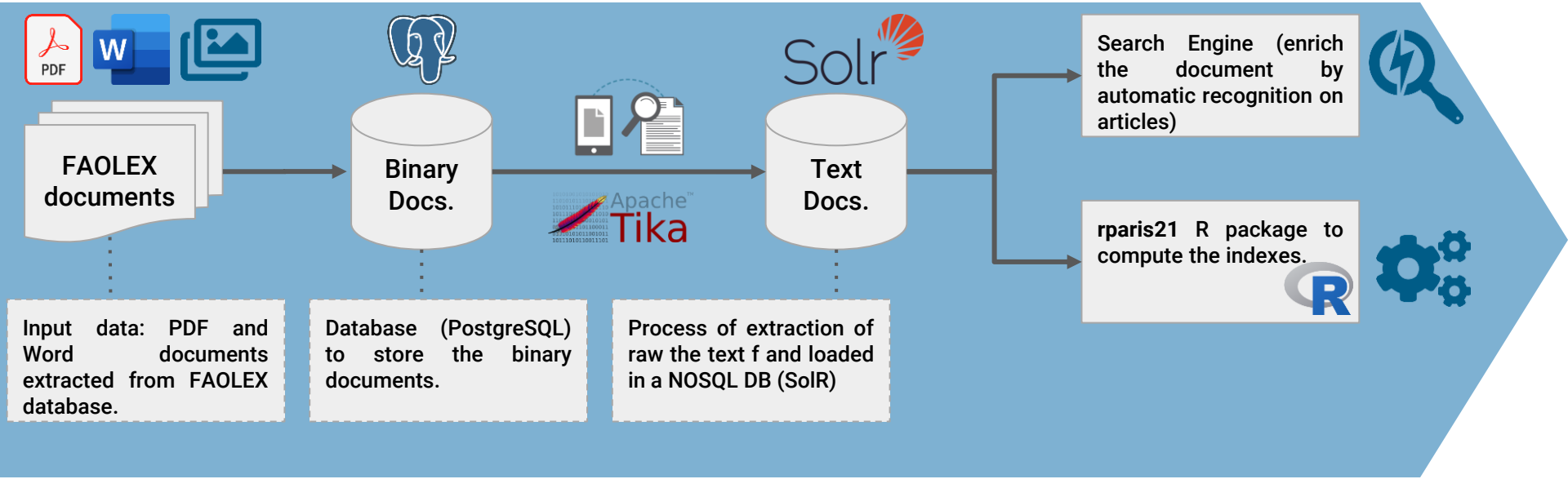
4295
documents

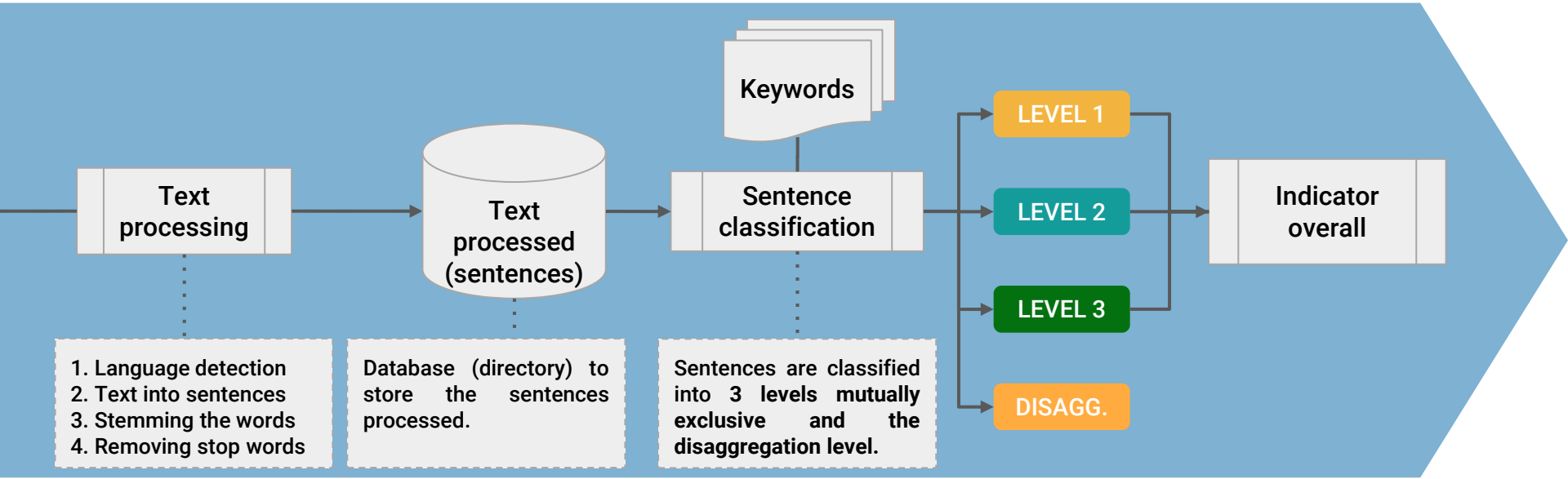
214
countries

1985 - 2019
years

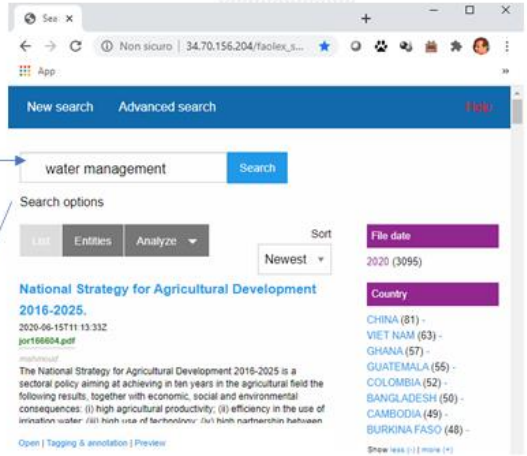
44
languages

Pre-processing the texts




Keywords
Text processing
Text processed (sentences)
Sentence classification
LEVEL 1
LEVEL 2
LEVEL 3
DISAGG.
Indicator overall
1. Language detection
2. Text into sentences
3. Stemming the words
4. Removing stop words
Database (directory) to store the sentences processed.
Sentences are classified into 3 levels mutually exclusive and the disaggregation level.

Search engine: user interaction



Search engine
(Fulltext search)

Easy full text search in multiple data sources and many different file formats: Just enter a search query (which can include powerful search operators) and navigate through the results.

Solr

Solr Features

Solr is a standalone enterprise search server with a REST-like API. You put documents in it (called "indexing") via JSON, XML, CSV or binary over HTTP. You query it via HTTP GET and receive JSON, XML, CSV or binary results.

Search engine: other facilities

Thesaurus & Grammar (Semantic search)

Based on a thesaurus the multilingual semantic search engine will find synonyms, hyponyms and aliases, too. Using heuristics for grammar rules like stemming it finds other word forms, too.

Interactive filters (Faceted search)

Easy navigation through many results with interactive filters (faceted search) which aggregates an overview over and interactive filters for (meta) data like authors, organizations, persons, places, dates, products, tags or document types.

Exploration, browsing & preview (Exploratory search)

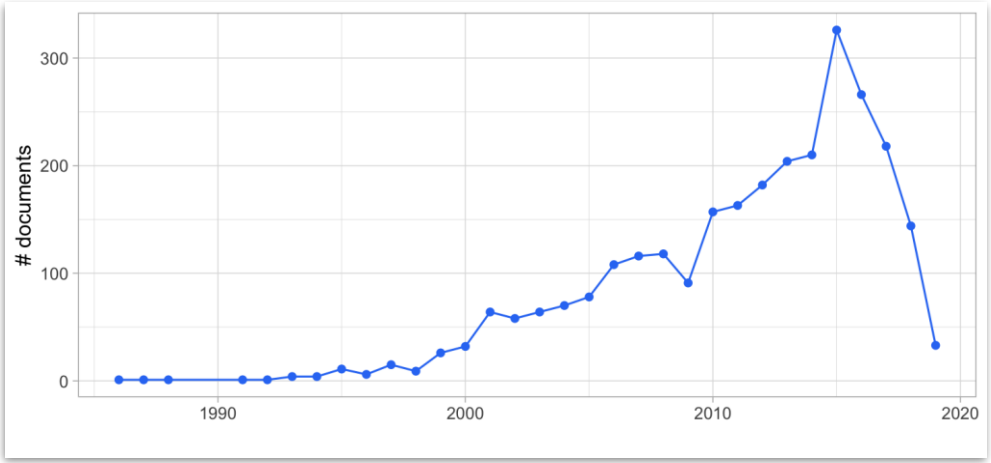
Explore your data or search results with an overview of aggregated search results by different facets with named entities (i.e. file paths, tags, persons, locations, organisations or products), while browsing with comfortable navigation through search results or document sets.
View previews (i.e. PDF, extracted Text, Table rows or Images).
Analyze or review document sets by preview, extracted text or wordlists for textmining.

Collaborative annotation & tagging (Social search & collaborative filtering)

Tag your documents with keywords, categories, names or text notes that are not included in the original content to find them better later (document management & knowledge management) or in other research or search contexts or to be able to filter annotated or tagged documents by interactive filters (faceted search).

Or evaluate, value or assess or filter documents (i.e. for validation or collaborative filtering).

Results: data analysed



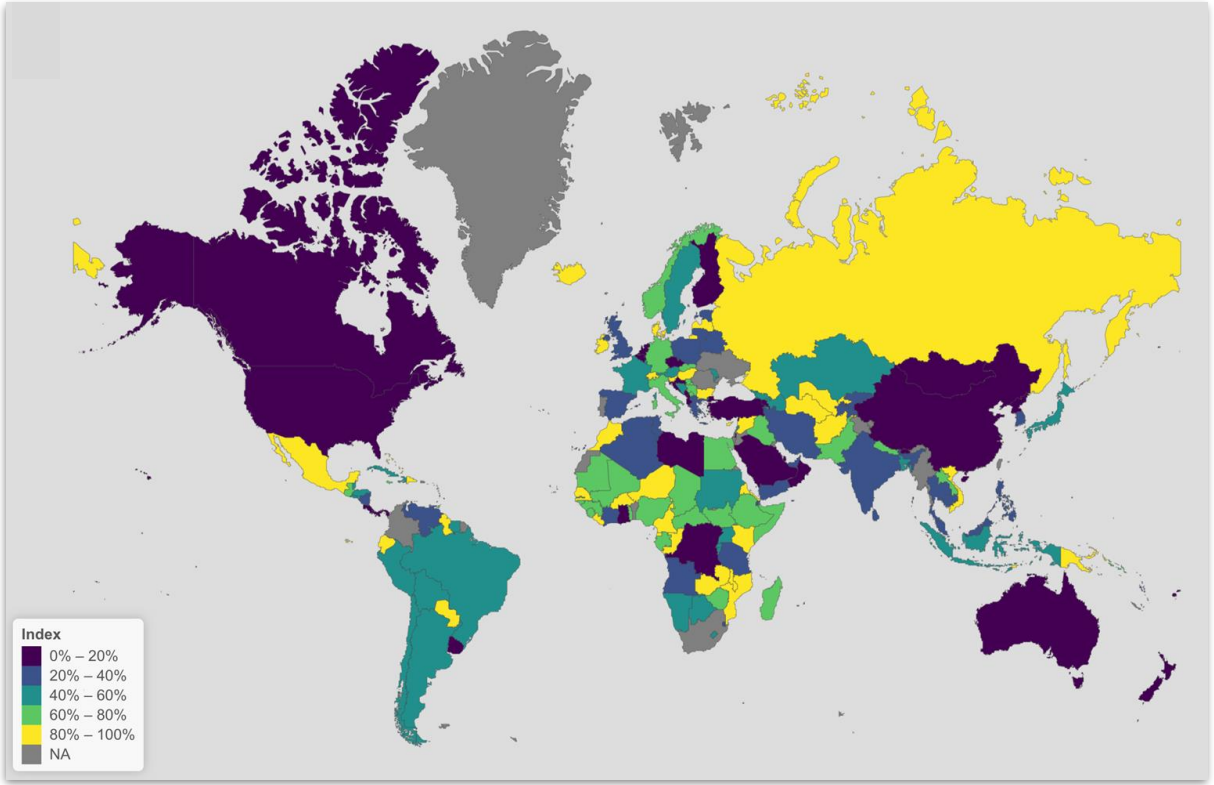
2790 / 4295
documents

200 / 214
countries

1986 - 2019
years

3 / 44
languages

Map: Index based on PCA



ShinyApp

https://foodandagricultureorganization.shinyapps.io/paris21/

Next steps

- ➔ Expert review and feedback
- ➔ Improving the aggregation method for the countries, and years level.
- ➔ Improving the keywords
- ➔ Platform to explore and visualize the data and results