Open Web Search
And its Use to Find Arguments Against Populist Claims

Johannes Kiesel
Bauhaus-Universität Weimar
Webis

Open Web Search: The Issue
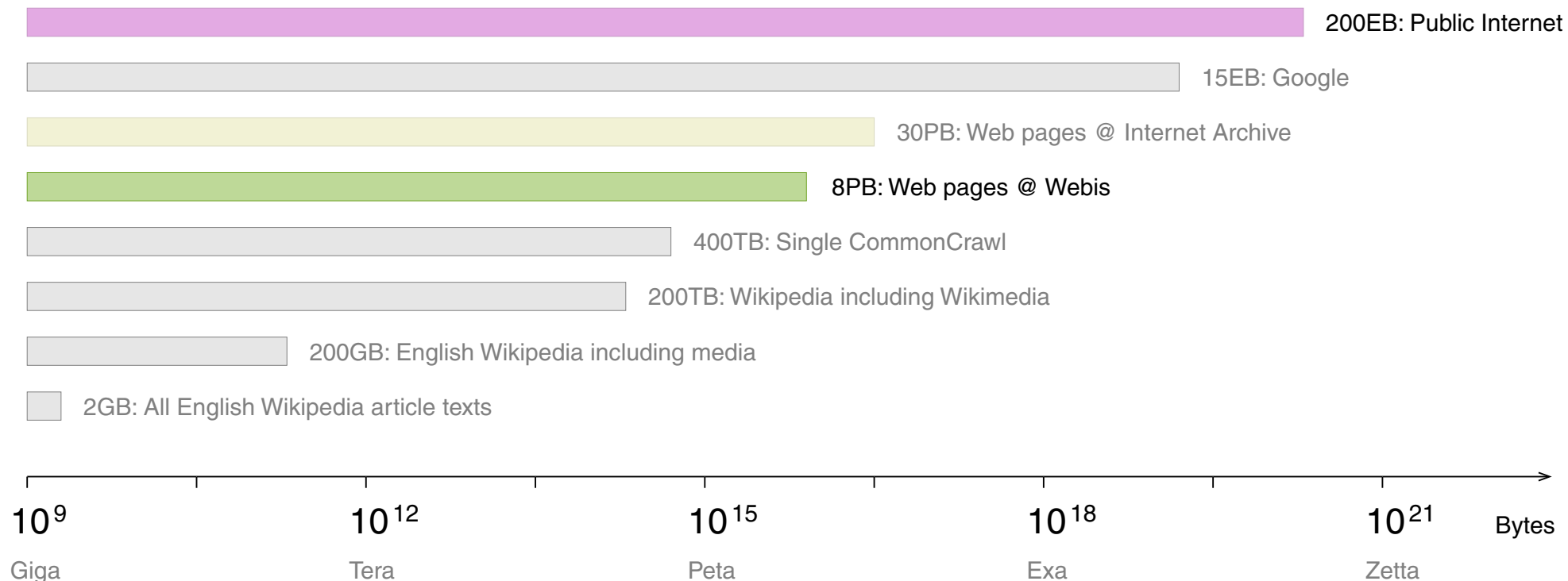
200EB: Public Internet

15EB: Google

30PB: Web pages @ Internet Archive

8PB: Web pages @ Webis

400TB: Single CommonCrawl

200TB: Wikipedia including Wikimedia

200GB: English Wikipedia including media

2GB: All English Wikipedia article texts

10^9	10^{12}	10^{15}	10^{18}	10^{21}	Bytes
Giga	Tera	Peta	Exa	Zetta	

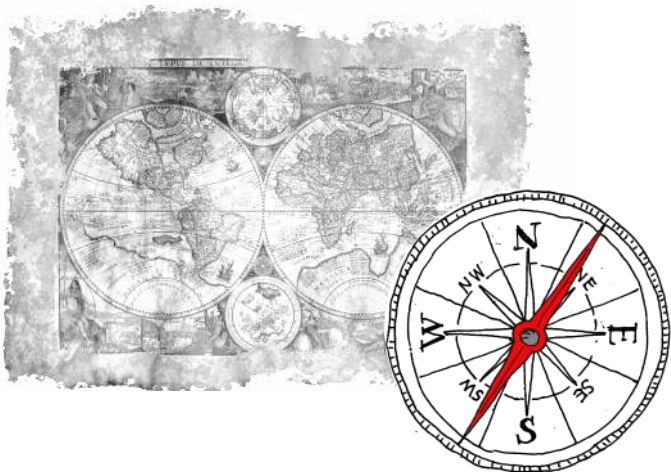$\Rightarrow$ Gigantic gap between open actors and Google

Orientation = Power + Control
(economic, strategic, geographic, cognitive, political …)

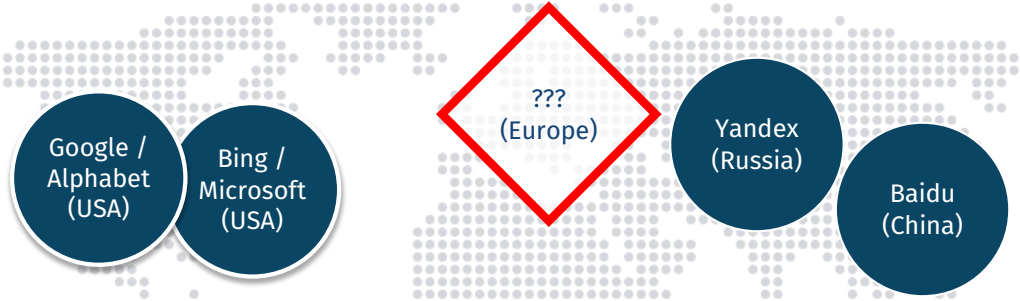**Orientation in the geographic sphere used to be exclusive knowledge and a tool of power**

This is still the case in the digital sphere

→ **Europe needs a Programme like „Galileo" or „Copernicus" for sovereignty in Web Search and Web Data Services**

Illustration: www.atelier-anne-rieken.de

Why does Europe need an independent "Navigation System" for the web?

Google / Alphabet (USA)

Bing / Microsoft (USA)
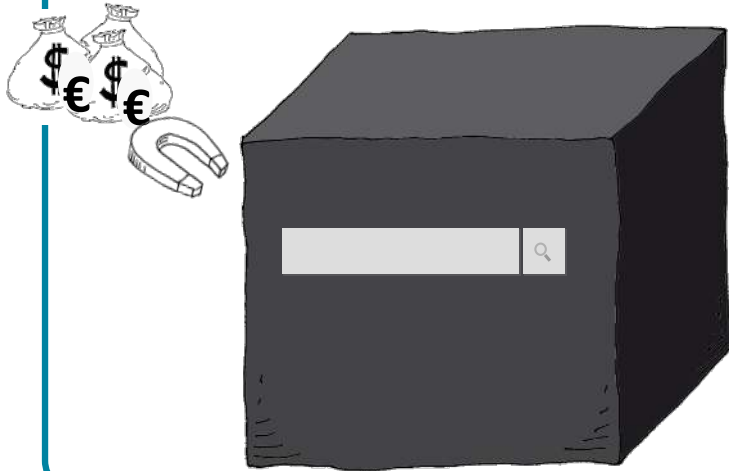
??? (Europe)

Yandex (Russia)

Baidu (China)

→ A web index is the **key element of web search and many further web data services** such as Large Language Models (LLMs).

→ **Globally, only four comprehensive indices exist.** They are either purely commercially driven or state-controlled.

→ **Europe does not have its own web index**. More than 90% of all web search is done via Google.

→ **Europe depends completely on US-American search providers** and their commercial interests.

→ **An Open Web Index for Europe will**

• **strengthen the strategic sovereignty** and technological autonomy through an independent and transparent web access and

• **essentially contribute to the European digital targets for 2030** by building a sustainable digital infrastructure

An Open Web Index will enable transparent and unbiased access to Web Content

From a closed and opaque internet search ...

**... to an open, transparent and auditable web search ecosystem**

Illustration: www.atelier-anne-rieken.de
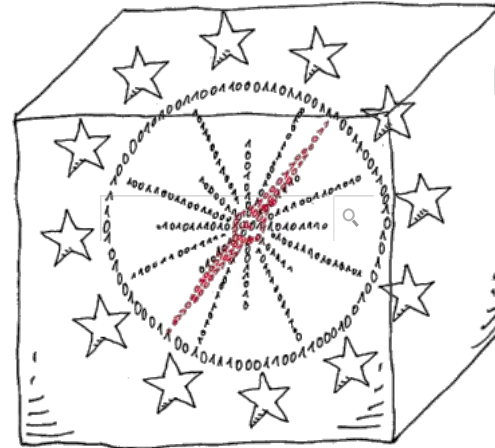
An Open Web Index will enable new public and private web services, boosting innovation in Europe

Open Web Search

Mobility Services

Artificial Intelligence
Large Language Models like ChatGPT

AI

Science and R&D

Online News & Info

NEWS

Web Statistics

e-Commerce

Cyber Security

Industry 4.0

Internet of things

IoT

Business Intelligence

Digital humanities

Geo spatial services

Big earth data processing

WEB Data Pool

Open Web Index

+

The technical set-up:

A fully distributed network of European computing centres to create and host the index for public use by start-ups, SMEs, industry and academia

● Main hubs
● Sub hubs

Goal
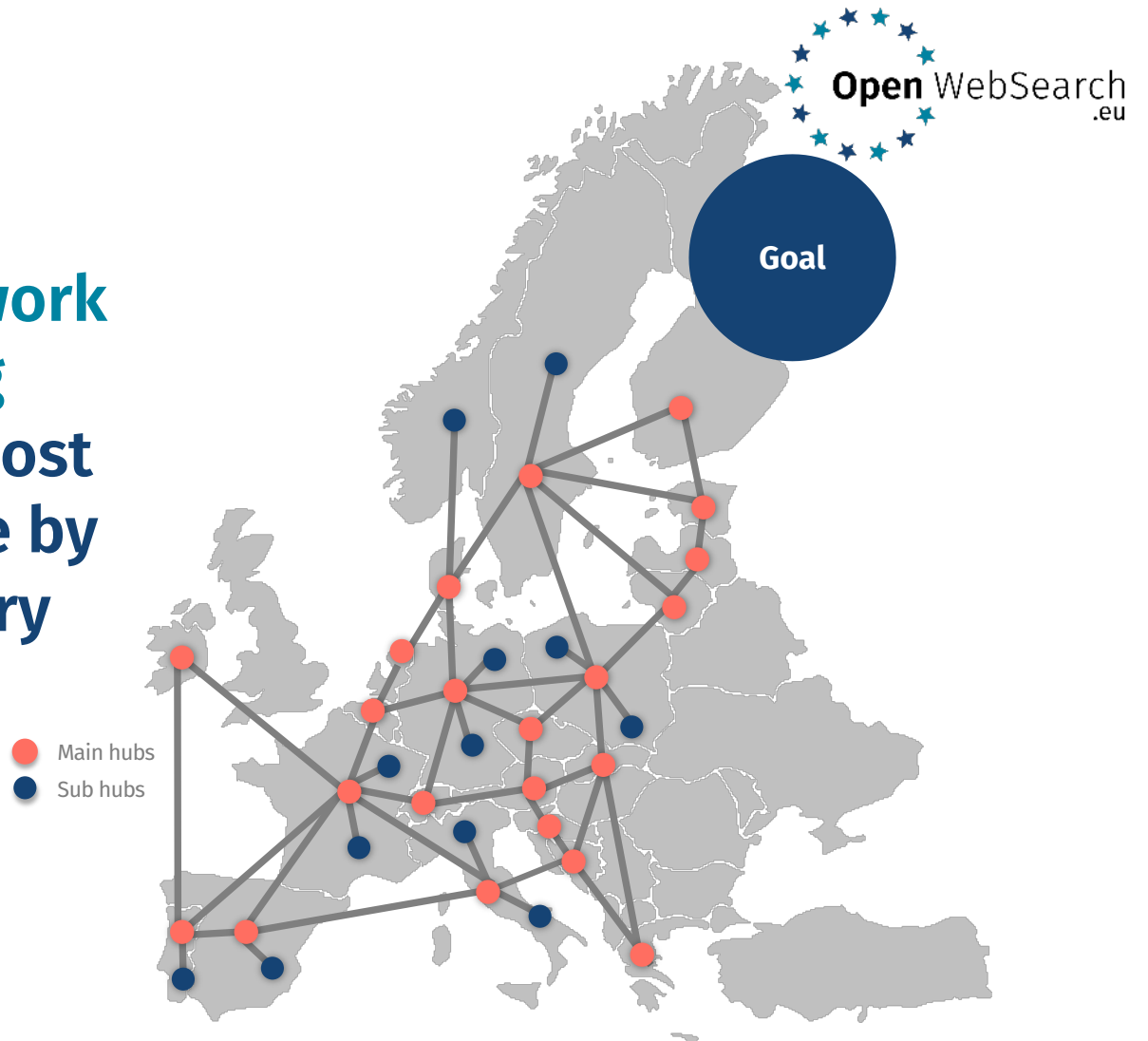
Open WebSearch.eu

Possible road map for Open Web Search 2023-2028

OpenWebSearch.eu

Today

Implementation and
scaling of a distributed European
"Open Web Search" ecosystem in a European Digital
Infrastructure Consortium (EDIC) / Multi Country Project (MCP)

Ensure that OWS is ear-marked with sufficient budget within the next **European Multi-annual Financial Framework** (MFF)

OpenWebSearch.eu Project

New NGI Call (2024) and Project on Open Search (2025)

| 2023 | 2024 | 2025 | 2026 | 2027 | 2028 |

Open Search Initiative - yourinfo@opensearchfoundation.org

OpenWebSearch.eu will create an open European infrastructure for internet search, based on European values and jurisdiction

What?

Restore an open search engine market as a basis for a new Internet Search

→ lay a foundation for a new Internet search

→ contribute to Europe's digital sovereignty

→ empower Europe's researchers, innovators and businesses to systematically tap into the Web as business and innovation resource

Why?

1. Web search is dominated and limited by a few gatekeepers like Google, Microsoft, Baidu, Yandex.

Resulting situation:
→ unilateral, biased, opaque access to information
→ locked-in effects
→ a closed search engine market

2. Tapping the Web as resource is challenging for innovators and researchers

Who?

14 renowned European universities + institutions will pool their expertise and resources.

→ including some of the largest research and computing centres in Europe

→ e.g. IT4Innovations, Leibniz Supercomputing Centre, CSC, European Organisation for Nuclear Research CERN

How?

→ develop the core of a European Open Web Index

Four Objectives
1. Open Technology Stack
2. Resource provision by a network of infrastructure providers
3. Added value services
4. Bootstrapping the ecosystem

NGI OPENWEBSEARCH.EU

Make the core of a web search engine a public good

Web Data as a key ingredient for web search engines, AI and hyperscalers

Enablers in Web Search and Artificial Intelligence

- Web Data and an infrastructure to process the data at scale
- Biggest challenges for (smaller) innovators, companies and researchers to catch up

Opening up web data and providing a corresponding infrastructure

- Opening the core of commercial search engines: the Web Index
- Data structure for fast access to web documents / sites
- Beyond web search engine
- Web data for artificial intelligence
- Data-centric petabyte and compute infrastructure
- Not only technical challenges, but also legal and societal challenges
- The need is also recognized by the open source AI Community

➔ **In short: Breaking down the petabyte of web data in bits and pieces, consumable by Europe's companies, researchers and innovators**
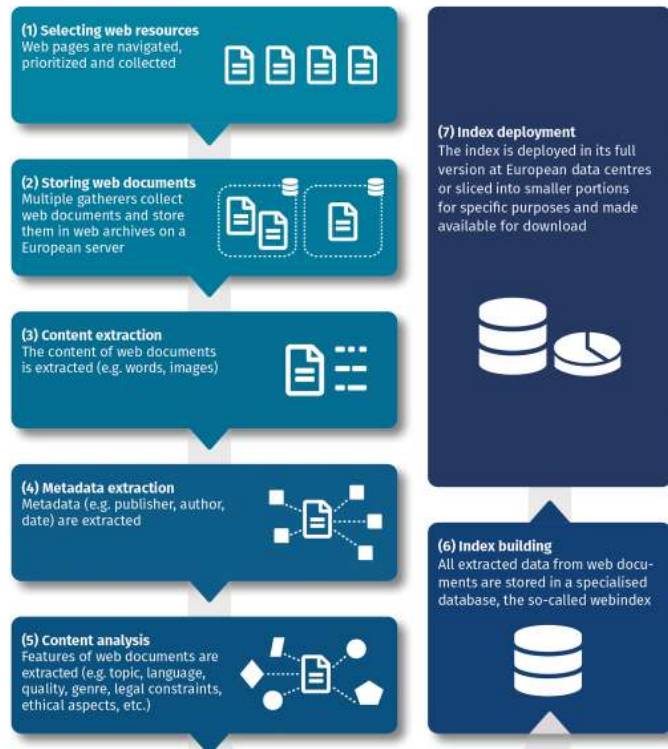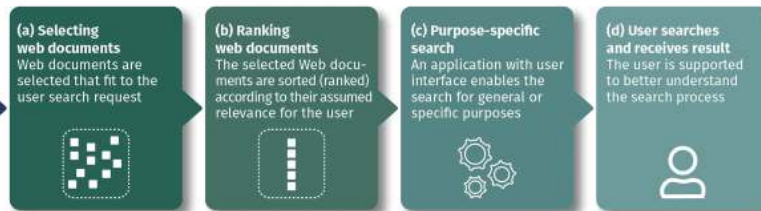
The Infrastructure at a glance

Index Generation
Web resources are selected and retrieved, their content and metadata are analysed, and all data stored in the index database.

(1) Selecting web resources
Web pages are navigated, prioritized and collected

(2) Storing web documents
Multiple gatherers collect web documents and store them in web archives on a European server

(3) Content extraction
The content of web documents is extracted (e.g. words, images)

(4) Metadata extraction
Metadata (e.g. publisher, author, date) are extracted

(5) Content analysis
Features of web documents are extracted (e.g. topic, language, quality, genre, legal constraints, ethical aspects, etc.)

(7) Index deployment
The index is deployed in its full version at European data centres or sliced into smaller portions for specific purposes and made available for download

(6) Index building
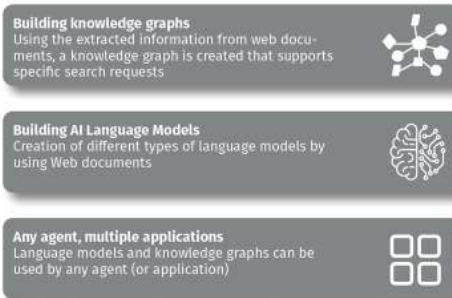All extracted data from web documents are stored in a specialised database, the so-called webindex

Search Applications
A user search request will be answered by a search application that makes use of the open web index.

(a) Selecting web documents
Web documents are selected that fit to the user search request

(b) Ranking web documents
The selected Web documents are sorted (ranked) according to their assumed relevance for the user

(c) Purpose-specific search
An application with user interface enables the search for general or specific purposes

(d) User searches and receives result
The user is supported to better understand the search process

Data Products
Knowledge representation models will be created using the open web index, in order to be used by any agent and for many applications

Building knowledge graphs
Using the extracted information from web documents, a knowledge graph is created that supports specific search requests

Building AI Language Models
Creation of different types of language models by using Web documents

Any agent, multiple applications
Language models and knowledge graphs can be used by any agent (or application)

· · ·

Web-scale Platform for heavy-lifting

Applications and Innovations as Multiplicator

Distributed Infrastructure as Enabler

LUMI@CSC

KAROLINA@IT4I

lrz Leibniz Supercomputing Centre
of the Bavarian Academy of Sciences and Humanities

CERN

DLR

Basis for an Ecosystem with researchers, innovators and businesses

Data Contributions
(e.g. provision of crawls, content push)

Content Curation
(e.g. science, education, languages)

Technology Contributions
(e.g. enrichment,)

Standards and ELSA Clearance
(e.g. license, metadata formats)

Web-scale Platform
for heavy-lifting

New Search Paradigms
(e.g. Argument search, conversational search)

Vertical Search Engines
(e.g. Open Science / Mobile Location Search)

Language Models
(e.g. language specific, search specific)

Benchmarking
(e.g. search engines, language models)

Web Analytics
(e.g. Content distribution, social media)

Applications and Innovations
as Multiplicator

Compute plus Storage Infrastructure

Search as a Service

Distributed
Infrastructure as Enabler

Application: Argument Search

Get the pros and cons for a topic from the web over years.

(manually curated: `kialo.com`)
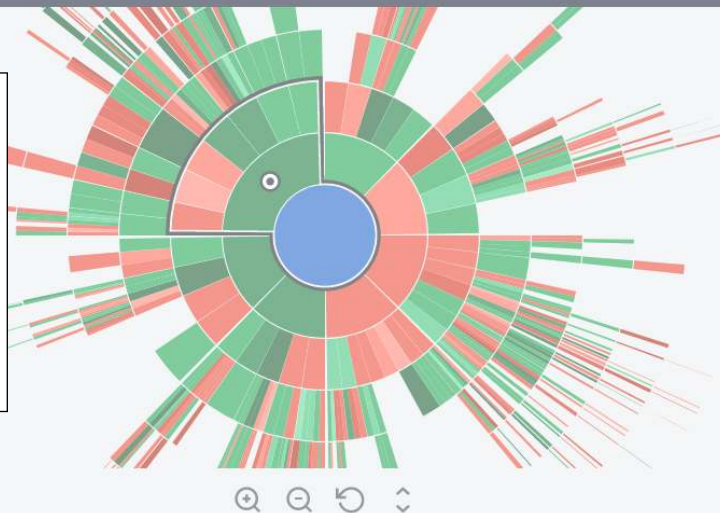
Wealthy countries should provide citizens with a universal basic income (UBI).

A UBI would improve the lives of many people.

Pros		Cons	

A UBI would improve working capabilities and conditions.

A UBI can improve the lives of many members of society in the

@KieselJohannes
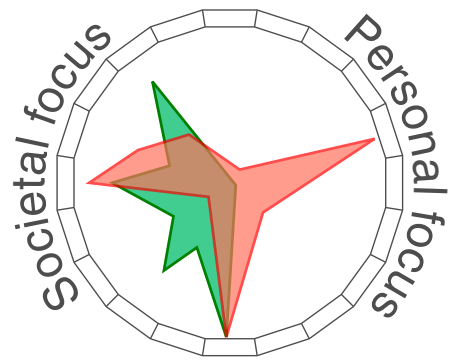
Application: Argument Search

"Epistemological Why"

Why is this true?

"Ethical Why"

Why is this good?

A universal basic income would make the lives of many people more secure
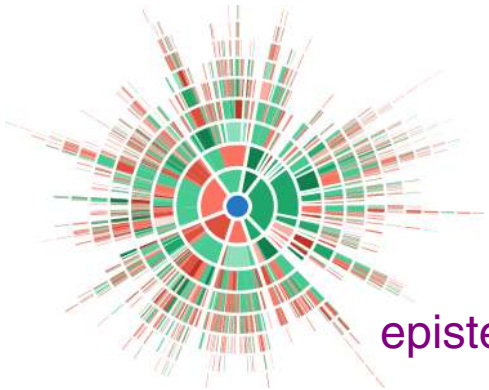
epistemological support

ethical support

Such income would improve working capabilities and conditions

It is good when people have *personal security*

Application: Argument Search
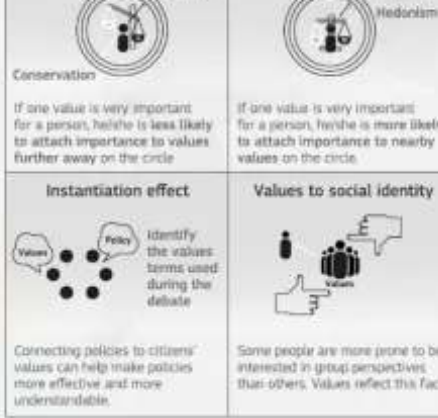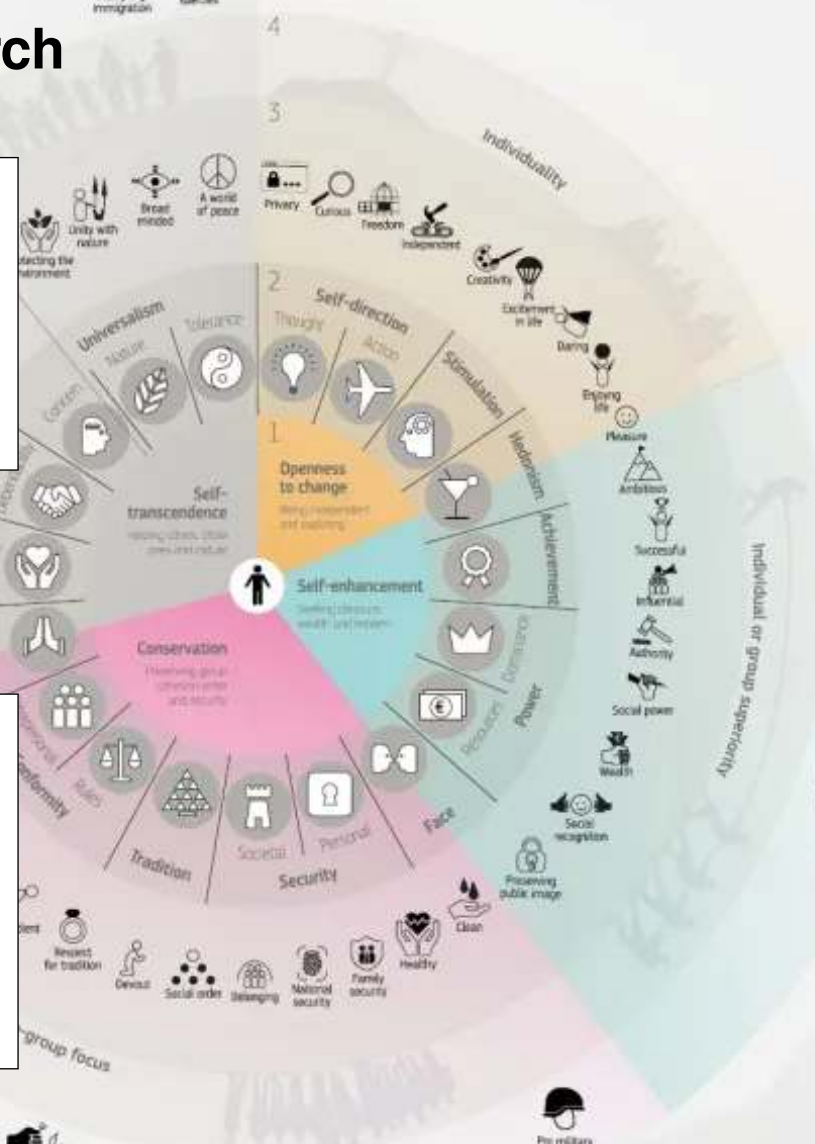
ValueEval'23

39 participating teams

Online demo:

https://values.args.me

ValueEval'24 (joint with JRC)

Data: Europe Media Monitor
(8 languages)

@CLEF'24, Grenoble, Sep. 2024
https://valueeval.webis.de

5th International Open Search Symposium #ossym23

4-6 October 2023

Keynote Speakers

CERN and online

Registration free

Věra Jourová

Vice President of the European Commission for Values and Transparency

Keynote | 5 October | 9:00 |
Keynote

Ricardo Baeza-Yates

Director of Research
Institute for Experiential AI, Northeastern University, USA

Keynote | 5 October | 15.30 |
"Bias in Search and Recommender Systems"

Angella Ndaka

The Centre for Africa Epistemic Justice
University of Otago, Newzealand

Keynote | 6 October | 9.00 |
"Inclusion by whose terms? When being in doesn't mean digital and web search inclusion"