Disruptive AI and linked big data against complexity to promote new culture for evidence-based policymaking

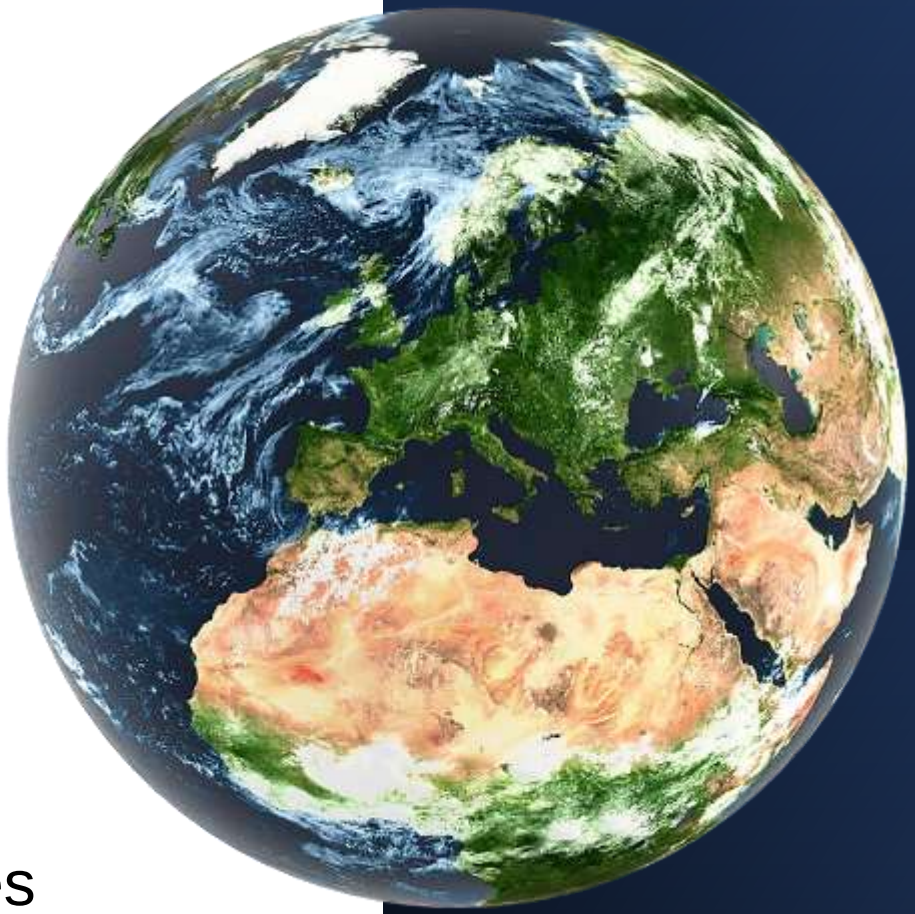
Dr. Sergio DI VIRGILIO
Knowledge and Data Management, RTD GH6
DG R&I European Commission
Brussels, December 14th 2023

Information Day on a new era of STI policy making with AI

Opportunities, Understanding the Limitations and Risks

THE WORLD IS CHANGING – Bringing new policymaking challenges

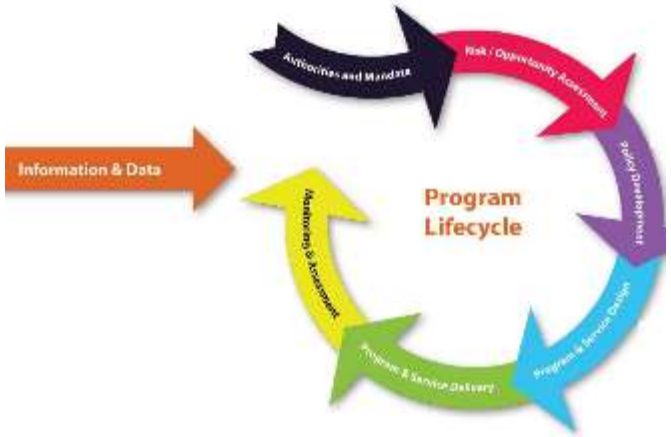
- Climate crisis
- Geopolitical shifts
- Democracies under threat
- War in Europe
- Migration
- Pandemic
- Demographic changes
- New information environnent
- Emerging digital disruptive technologies
- …



European Commission

Why are data so important

Better informed policy-making

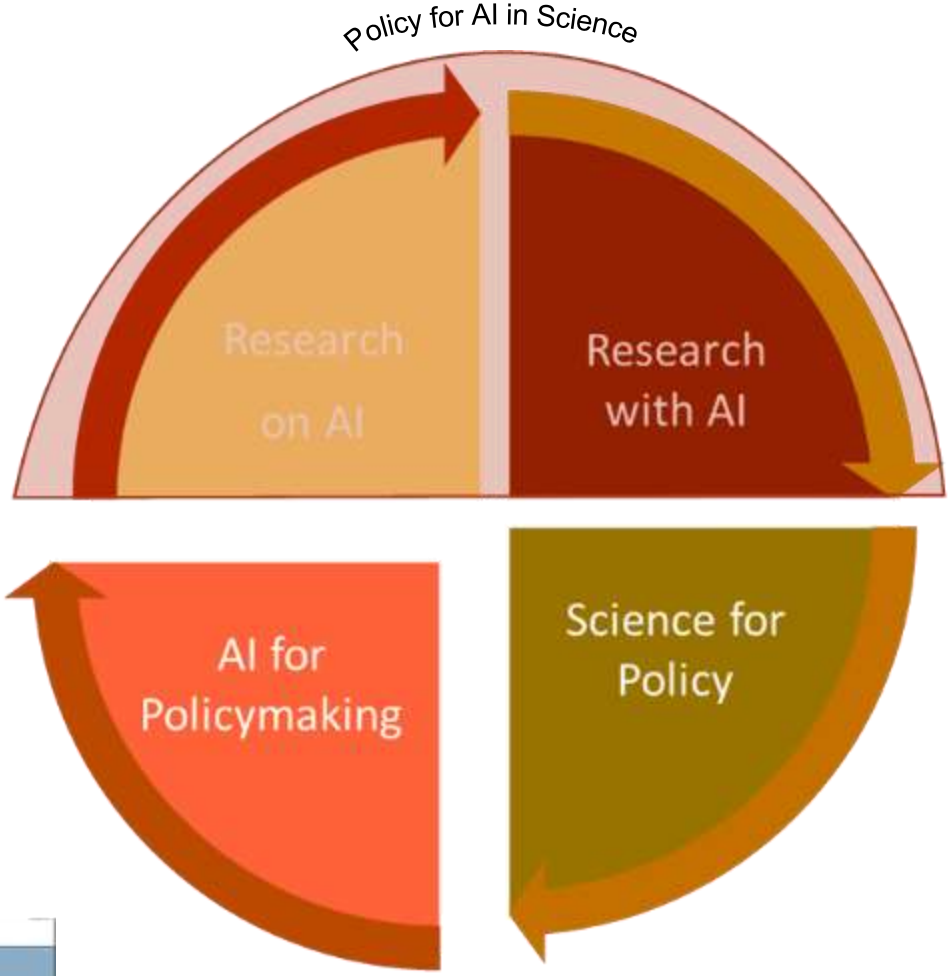


Support the programme lifecycle

Fulfil our legal obligations

European Commission

AI, Research and Policymaking



DOI 10.2777/418191
DOI 10.2777/128148
DOI 10.2777/67092

European Commission

The global chain governance change issue

Graphic Facilitation by THE VALUE WEB
www.thevalueweb.org

World Economic Forum Annual Meeting 2013
Davos-Klosters | 23-26 January 2013

European Commission

Issues to overcome

New data sources may provide new insights for (evidence-based) policy making

This may require **new methodologies**,

will Machine Learning, Natural Language Processing, Topic Modelling Deep Learnig and Large Language Models enhance policy models?

It is a **co-creation** with involvement of various stakeholders in different phases of policy cycle

One of the big challenges include **organizational readiness** and **policy makers'willingness and skills** for using data and data-driven methods for policy making

Project mission:

(1) create new knowledge by applying **big data approaches**

(2) Improve the monitoring of **EU and national R&I** programmes

(3) Better assess the **societal impact** of research funding

1 November 2017
31 October 2019

Five connected phases:

(1) Scoping Phase: Identifies **R&I policymaker needs** and data gaps

(2) Exploration Phase: Takes eight key questions identified from the scoping phase and develops short pilots **using big data and new data analytics** to address these questions

(3) Data Collection and Analysis Phase: Scales up four data pilots with the biggest policy potential enabling the production of **Relevant, Inclusive, Timely, Trusted and Open (RITO) indicators for R&I policy**

(4) Validation Phase: Systemically **validates all of the indicators** generated in the data collection and analysis stage with the goal of building trust around their use

(5) Communication and dissemination Phase: Seeks to enhance the impact and transparency of outputs by disseminating them in a way that is actionable and reproducible, including through **open datasets, open-source repositories, and interactive data visualisations and dashboards.**

1 January 2018
31 July 2021

Four tasks:

(1) design of a methodology for tracking research results,

(2) setting up a repository of data sources and collection of data,

(3) analysis of and reporting on the collected data, and

(4) training of Commission staff.

7 August 2018
1 March 2023

Analytics opportunities and motivations

The big data and AI revolutions have brought with them an explosion in the volume and complexity of the data that are available and in the techniques that can be used to extract useful information from them.

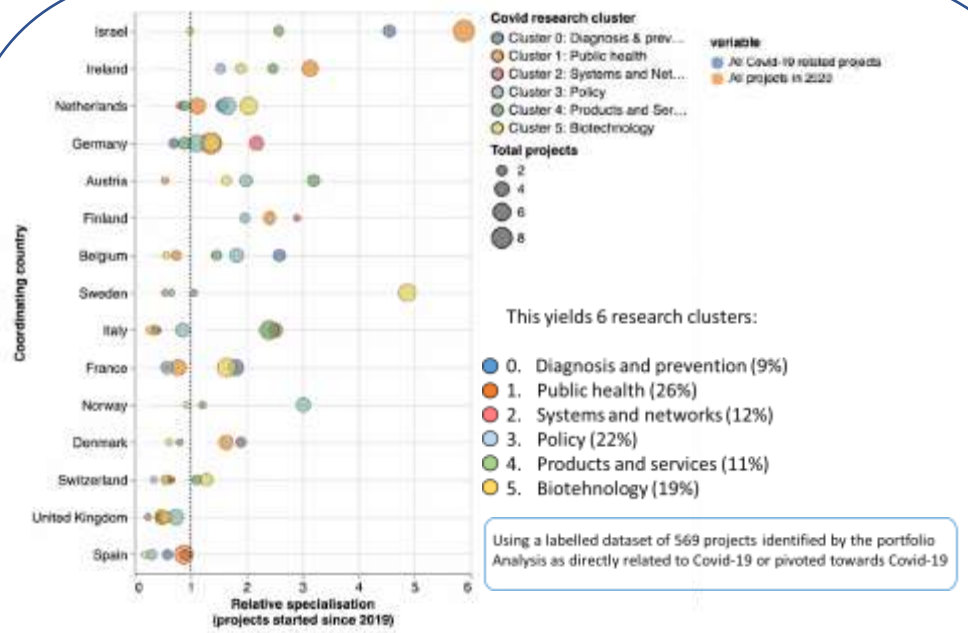
In terms of data, we have seen a transformation in our ability to work with unstructured text data in R&I-relevant documents such as academic papers, patents, grant applications, reporting documents, publications, + policy or legislative documents and descriptions of products and company profiles or job descriptions.

There has also been an explosion in web sources with useful information for understanding innovation in particular sectors such as AI (e.g. open source software and open datasets and models) or biomedical and health research (genomic and proteomic information, drug development and clinical trial databases).

European Commission

The COVID-19 Story

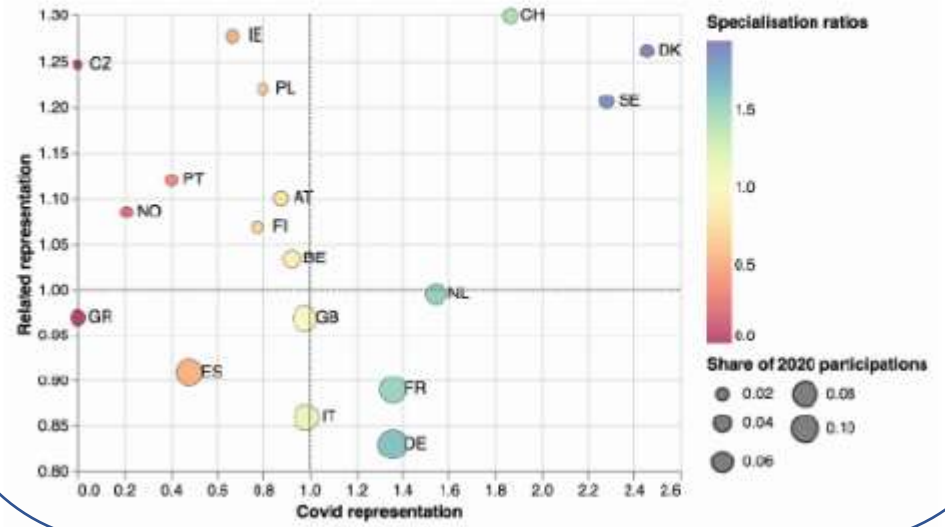
Tracking of Research Results



This yields 6 research clusters:

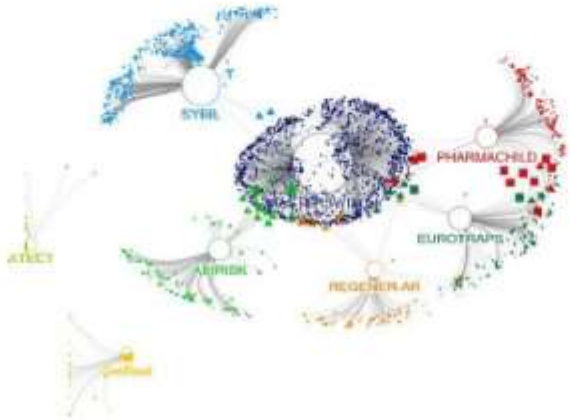
- 0. Diagnosis and prevention (9%)
- 1. Public health (26%)
- 2. Systems and networks (12%)
- 3. Policy (22%)
- 4. Products and services (11%)
- 5. Biotehnology (19%)

Using a labelled dataset of 569 projects identified by the portfolio Analysis as directly related to Covid-19 or pivoted towards Covid-19



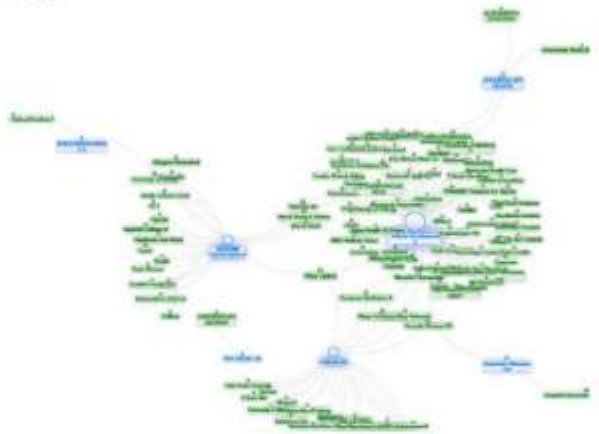
Tocilizumab

An immunosuppresive drug, mainly for the treatment of rheumatoid arthritis but today evaluated in patients admitted to hospital with COVID-19 (RECOVERY)



Gatekeepers

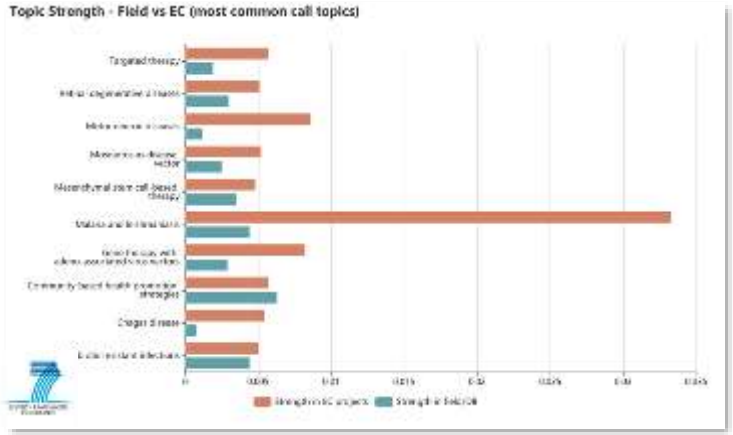
COLLABORATION NETWORKS OF RESEARCHERS LINKED TO PROJECTS RELATED TO TOCILIZUMAB

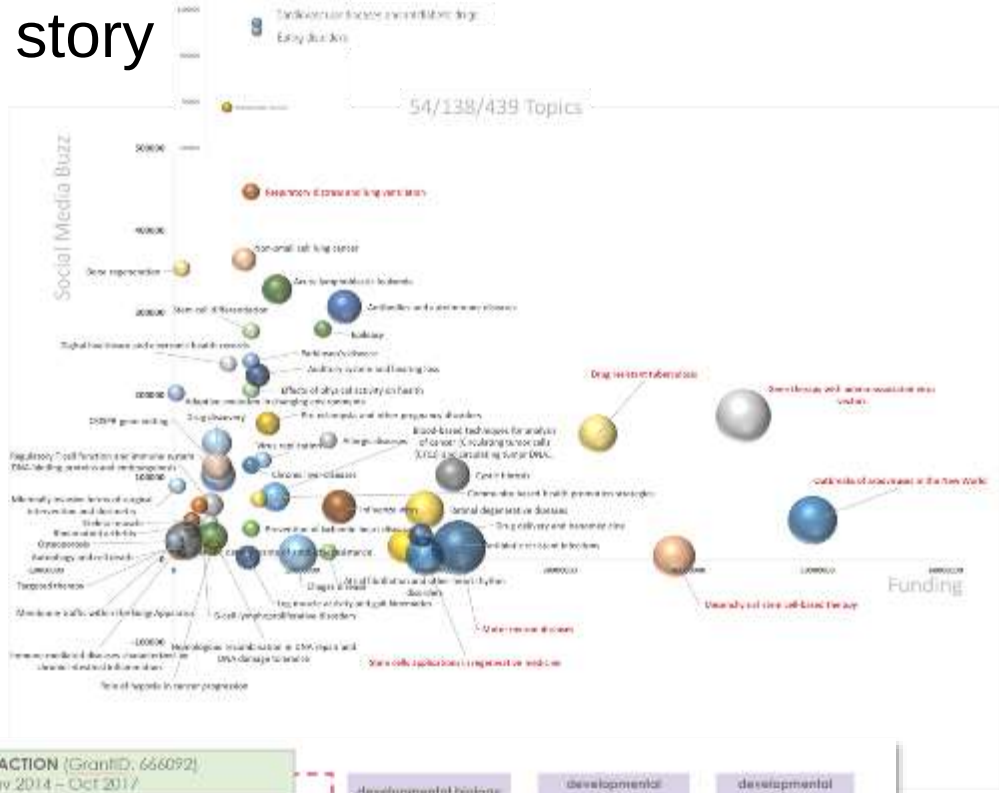


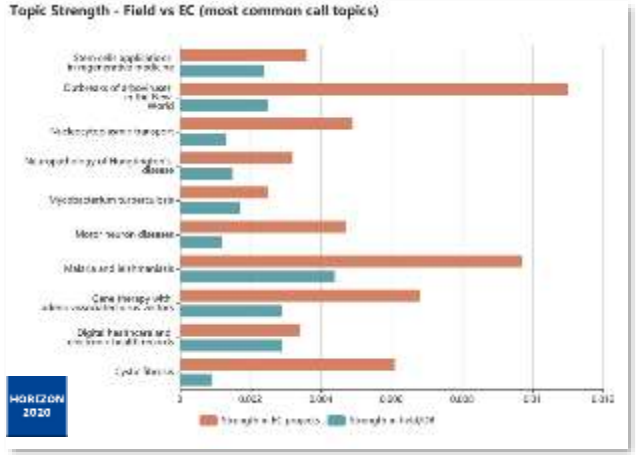
COLLABORATION NETWORKS OF COMPANIES INVOLVED IN PROJECTS ADRESSING TOCILIZUMAB

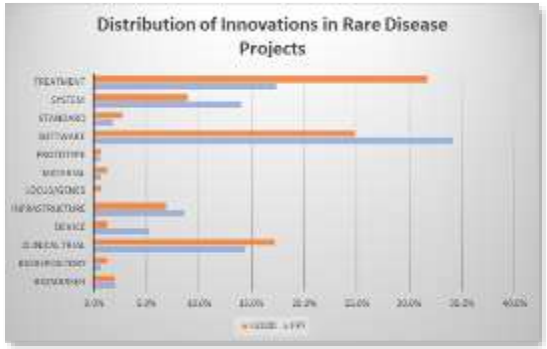
European Commission

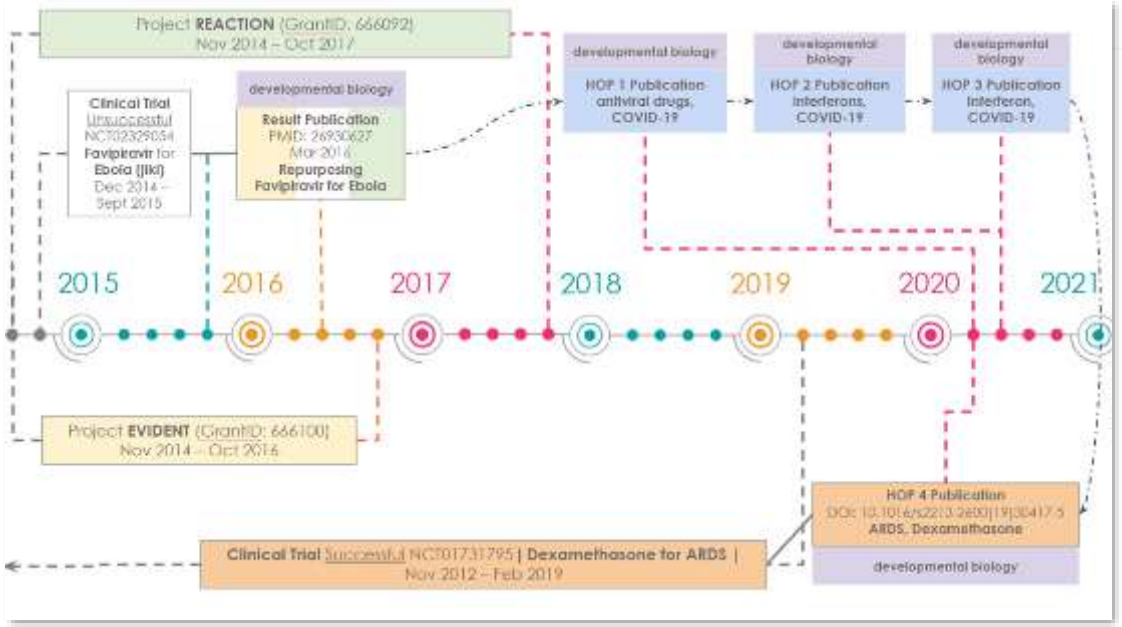
The Rare Diseases story





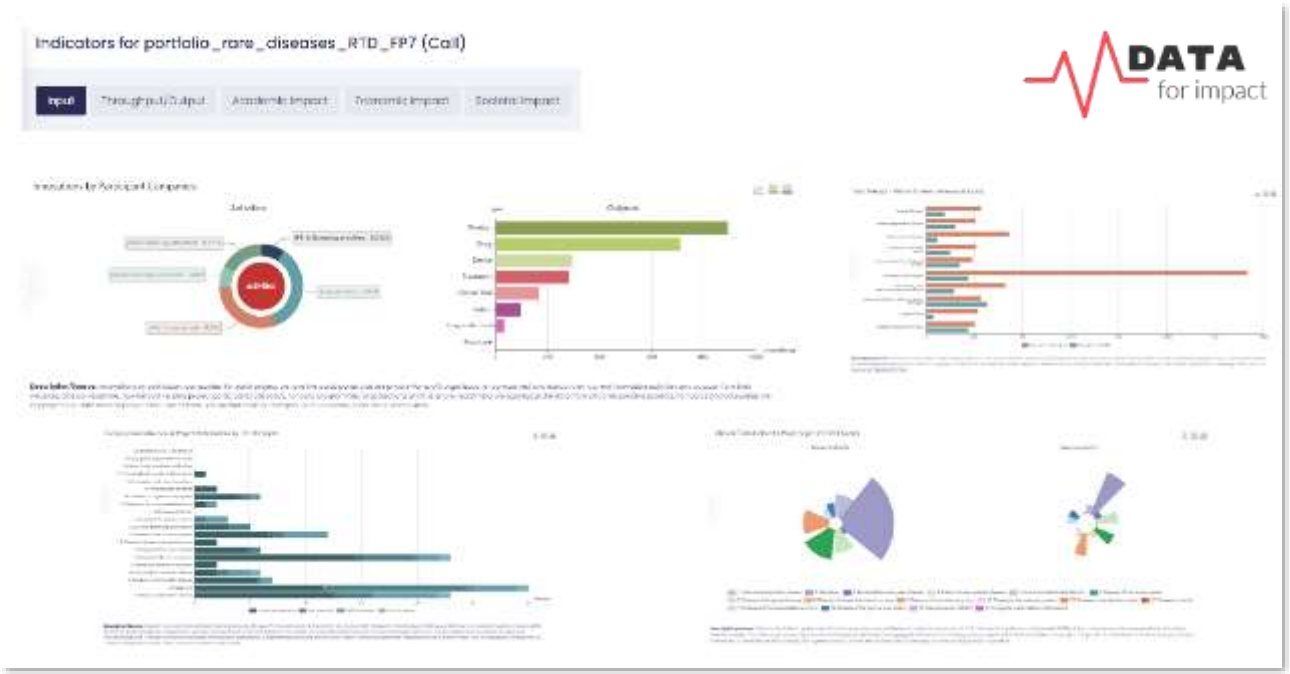


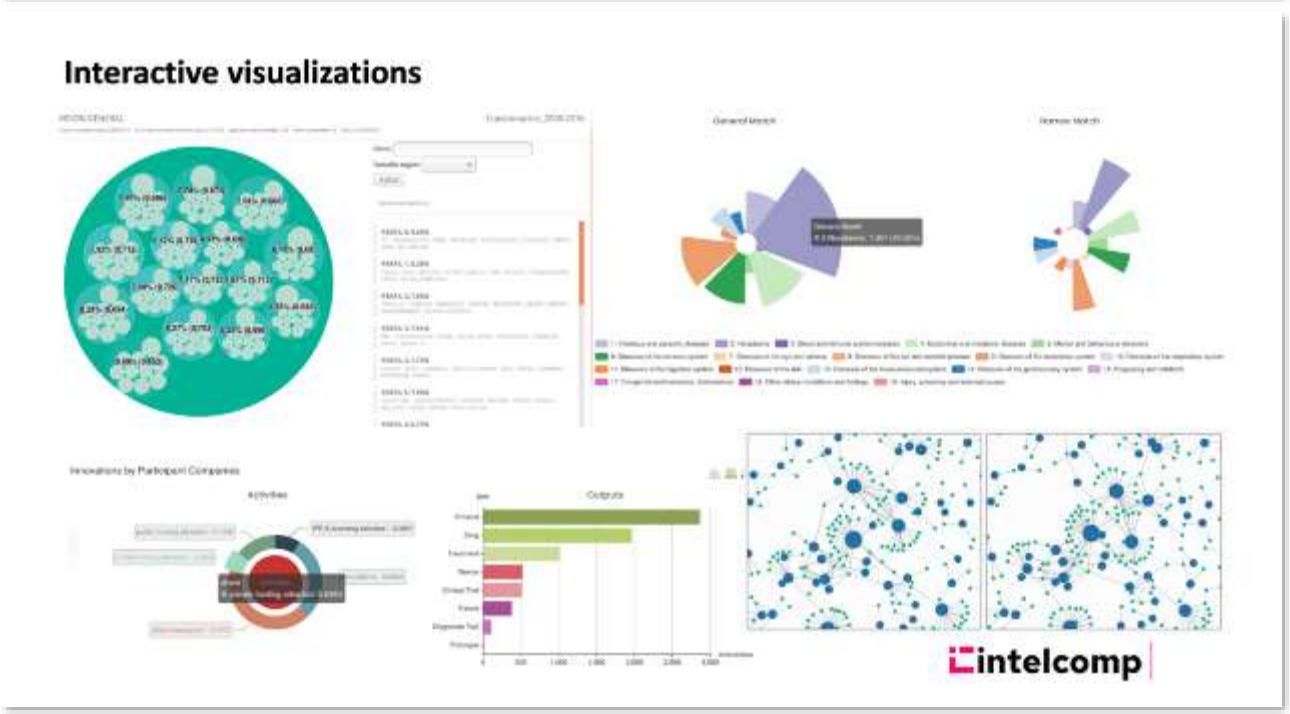




From Data4Impact to Intelcomp

And the importance of the Monitoring Tools





CORDIS
EU research results

HORIZON DASHBOARD

nature navigator

HORIZON RESULTS PLATFORM

TIM

EURITO indicators

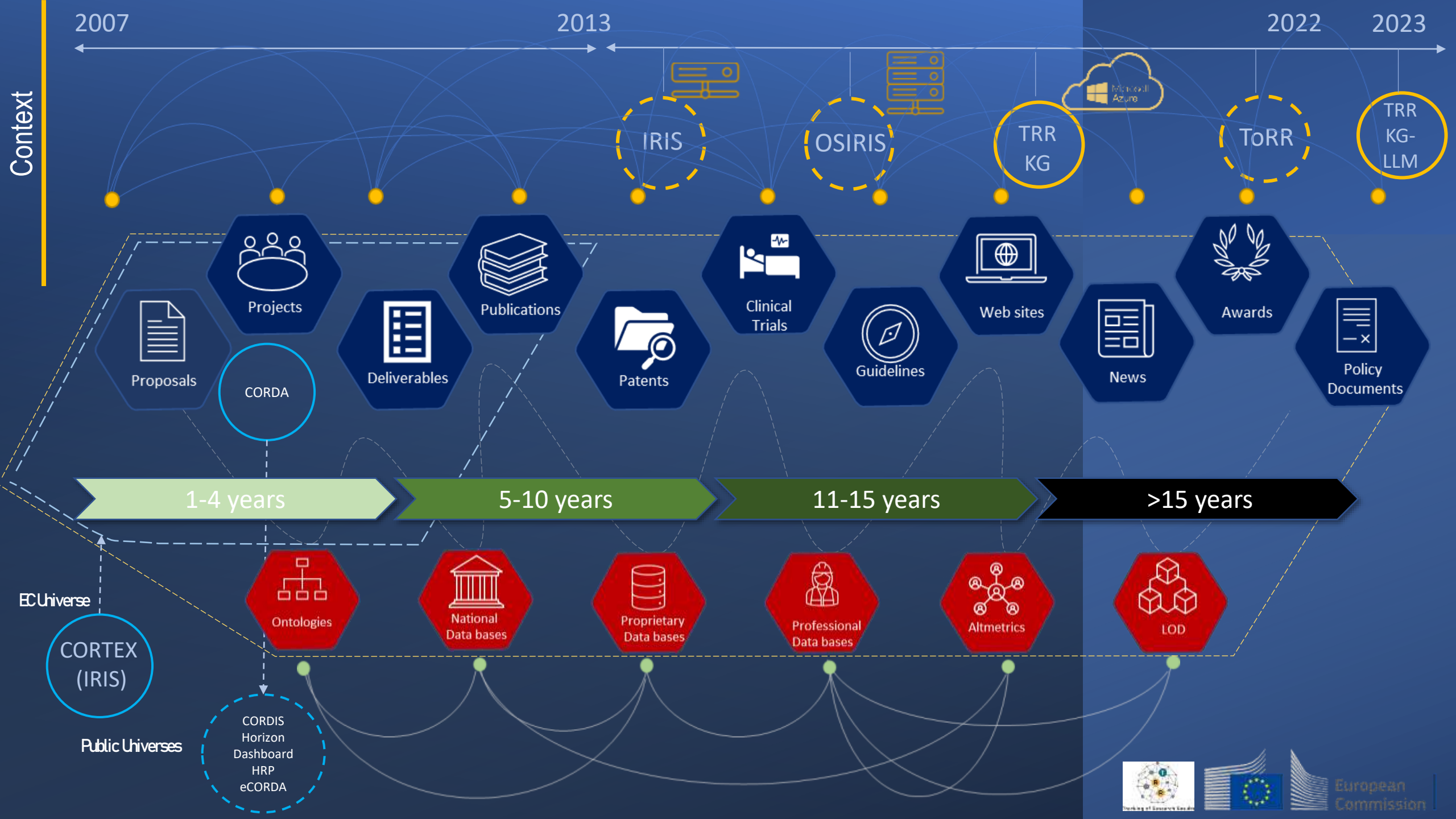
Rare Diseases Monitor

STIPCOMPASS

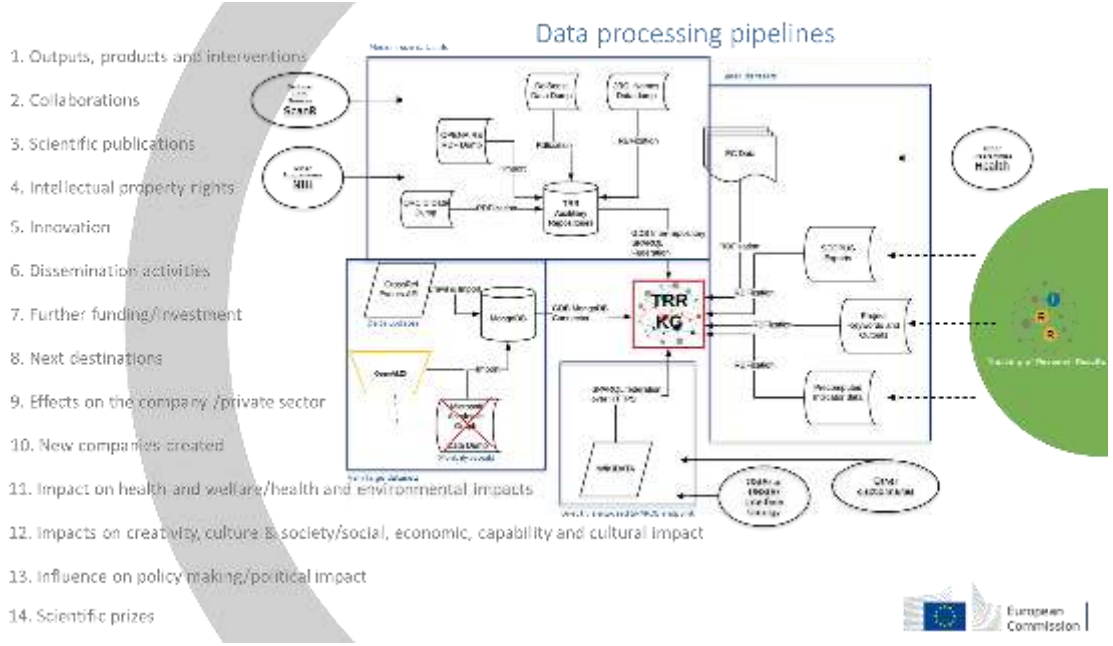
CORTEX

CORDA

The climate change issue


Context
2007
2013
2022
2023
IRIS
OSIRIS
TRR KG
ToRR
TRR KG-LLM
Projects
Publications
Clinical Trials
Web sites
Awards
Proposals
CORDA
Deliverables
Patents
Guidelines
News
Policy Documents
1-4 years
5-10 years
11-15 years
>15 years
Ontologies
National Data bases
Proprietary Data bases
Professional Data bases
Altmetrics
LOD
EC Universe
CORTEX (IRIS)
Public Universes
CORDIS
Horizon Dashboard
HRP
eCORDA
European Commission

An AI-based big data solution



Imput from 40+ primary external data sources

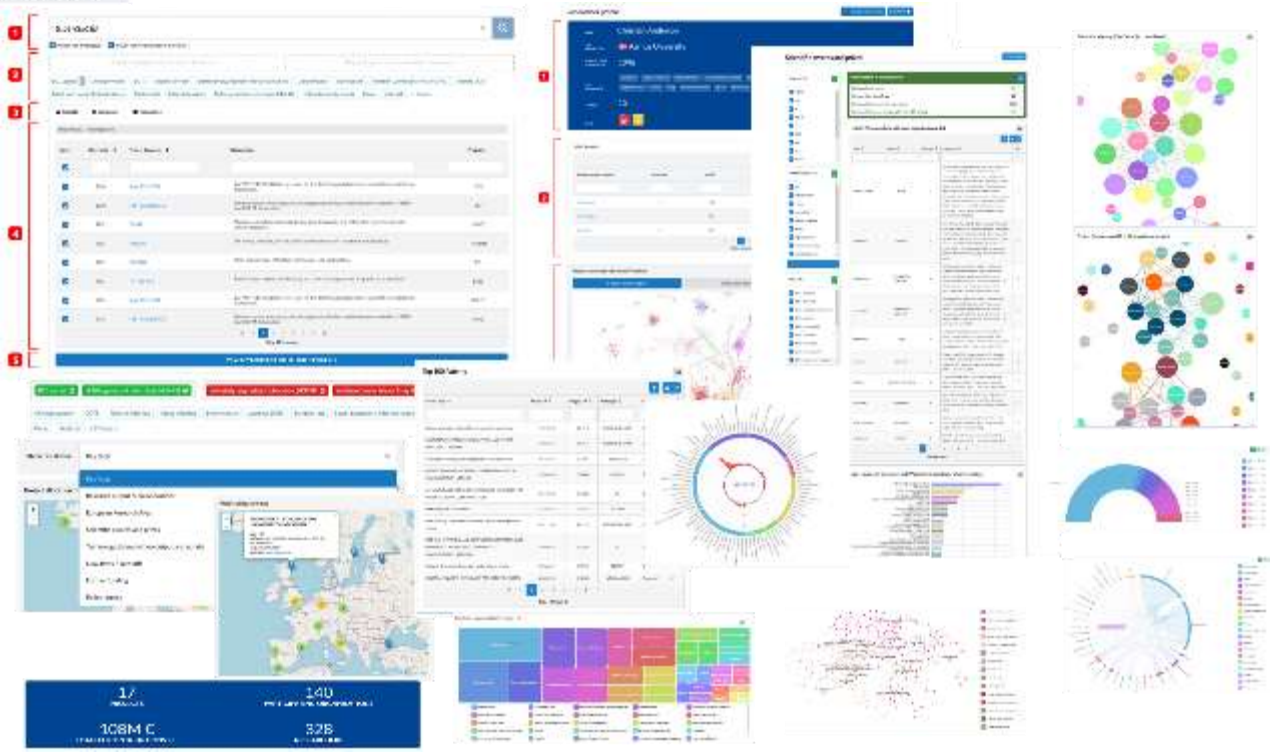
From 14 to 47 Indicators

ENTITY	INDICATOR NAME
PROJECTS	Project keywords
	SDG intensity
	SDG distribution
	Innovation outputs
	Expected impacts
	Final SDG labels
	FET Score
	Number/share of top-1% most cited publications
	Number/share of high tech value patents
	Similar projects
	Similar products by companies
	Collaboration networks
	Clinical studies
	List of diseases addressed
	List of EMA medicines and orphan designations
	Low carbon technologies
	List of circular technologies

ENTITY	INDICATOR NAME
COMPANIES	COMPANY KEYWORDS
	Industry
	Persistent identifiers in other DBs
	FET Score
	SDG intensity
	SDG distribution
	SDG final labels
	Financial performance
	Capital raised
	Investment announcements
	Collaborations
	Similar companies
	Similar EP outputs

ENTITY	INDICATOR NAME
RESEARCHERS	LINK TO EP PROJECT(S)
	Researcher keywords
	SDG labels
	Number/share of top-1% most cited publications
	H-index
	Similar researchers
	Prizes won
	Collaboration networks

ENTITY	INDICATOR NAME
PUBLICATIONS	PUBLICATION KEYWORDS
	Citation count
	Top-1% cited
	SDG label
	Citations in policy documents
	Citation in clinical studies
	Funding acknowledgments
PATENTS	PATENT KEYWORDS
	Estimated technological value
	Triadic patents



TRR: a tool for the many or the few?

The (TRR) User Interface exposes users to a large volume of indicators from which they can draw their own insights and conclusions. One can argue that this democratises the data collection and analysis process.

Any (EC) policy or project officer can build their own analyses via TRR or similar interface.

However, the uptake is not straightforward.

This is particularly the case with more complex indicators where data analysts' researchers had uneven understanding of their meaning and appropriate use.

It requires **extensive training** and active involvement of **data scientists who validate** the findings and conclusions.

(EC) policy officers would face similar challenges with using TRR data. Does it mean that TRR should remain a tool for the few?

The answer depends on which data and indicators are being used. The tables on the right shows examples ranked by their complexity:

Level 1 complexity indicators can be understood by the average researcher/policy officer and as such have a **lower risk of misuse.**

Level 2 and **Level 3** , however, a more specialised expertise and training are needed to **correctly use and interpret the data.**

Level 1:
Core, widely used, easy to comprehend indicators

- EU contribution to specific area (e.g. microplastics, Ebola) by programme area
- Number of researchers supported
- Number of prizes won
- SDGs data

Level 2:
Core indicators requiring specialist expertise

- Share of top-cited publications
- High value patents
- Structuring effect of FP funding

Level 3:
Niche indicators, complex assumptions

- Involvement score of FP researchers in projects
- Control groups
- Innovation uptake

Tracking of Research Results

Problem 1: we don't know much about what happens after or beyond FP funding

- **Solution to problem 1:**
- **By continuously collecting data one can follow the performance of beneficiaries and their control groups without creating additional administrative burden. The only current alternative would be regular surveys.**

Problem 2: time-to-research-results

- **Solution to problem 2:**
- **By having access to organised data, one can significantly reduce the time it takes to prepare the methodology and collect data for analysis.**

Problem 3: constantly changing policy landscape versus rigid monitoring systems

- **Solution to problem 3:**
- **By collecting data in a bottom-up way, one can create a flexible, bottom-up monitoring system that aggregates data upwards to the required levels of analysis and concepts.**

TRR methodology and data have been used in numerous EC studies and evaluations to leverage the TRR database to answer a series of policy questions.

The typical user is an (EC) policy or project officer

knows his/her project portfolio and needs quick access to key metrics, indicators and access to the underlying data.

The curious officer/explorer

wants to build their own portfolio for analysis from scratch (e.g. on SDGs, Missions,…). They use the TRR tool to build and validate a portfolio, following which they can analyze and download the data.

The skilled researcher/data analyst

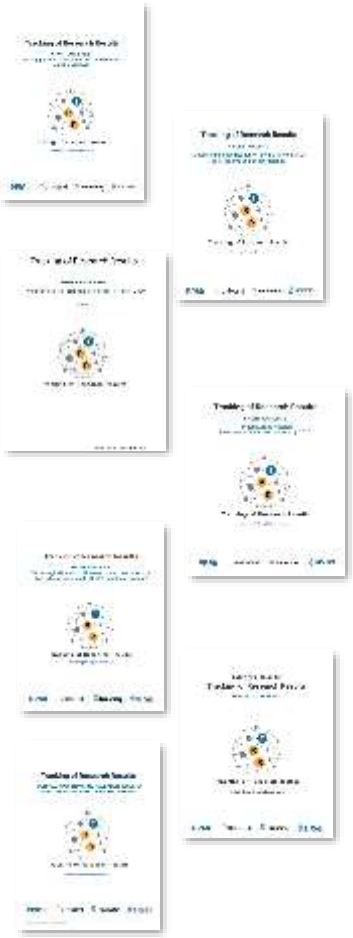
Has a specific policy question to answer.

Knows how to use data to their advantage.

Will not go to the TRR User Interface (ToRR)

Will query the database and build his/her own datasets for analysis.

Study name	Question analysed with TRR data	Short analysis series
Evaluation study of the European Framework Programmes for Research and Innovation for a Resilient Europe – RTD/2021/SC/021	• What is the timeline of FP research in the analysed programmes? • What is the structuring effect of FP funding in the area of anti-microbial resistance? • What has been the societal impact of FP research in the analysed programmes?	
Evaluation study of the European Framework Programmes for Research and Innovation for an Innovative Europe – RTD/2021/SC/019	• What was the economic performance of firms supported by the EU FPs? • What was the effectiveness of the R&I activities in the areas of quantum computing/ semiconductors research/etc.? • What has been the contribution to SDGs in the analysed programmes?	
Evaluation study on the implementation of cross-cutting issues in Horizon 2020 – RTD/2021/SC/009	• What is the gender composition of FP research teams? • What is the level of international collaboration in FP projects? • What is the survival rate of FP SMEs?	
Study to support the monitoring and evaluation of the Framework Programme for research and innovation along Key Impact Pathways – RTD/2019/SC/016	• What is average H-index value of FP researchers? • What is the expected/baseline contribution of R&I to SDGs? • What is the average annual turnover/ employment growth rate of FP SMEs?	
Evaluation study on Excellent Science in the European Framework Programmes for Research and Innovation	• What has been the contribution to SDGs in Pillar 1 of H2020? • To what extent did the programmes contribute to new/emerging research fields? • What are the patenting propensities of SMEs that participated in the analysed programmes (esp. FET)?	



LESSONS LEARNED

- Policy and data expertise must work hand-in-hand to reach acceptance 

Policy Officers will only accept data if they can validate the selection. The process should be participatory (e.g. RD, Researchers, system biology, participatory democracy)

- With big data, can one do without EC monitoring data? 

Big data benefits the most from extensive, well described project activities and results.

- How stable is the framework? 

Can reproduce results on a regular basis using the same methodology (stable algorithm yielding data for indicators eliminated the risk of human error)

What if the wealth of data becomes unavailable? Three developments in four years (e.g. GRID, MAG-OpenAlex, Lens.org)

- Can one really reduce the time-to-research-results? 

Concrete example of the study “Evaluation study of the European Framework Programmes for Research and Innovation for a Resilient Europe – RTD/2021/SC/021”: release in three months of bibliometric analysis, patent analysis, network analysis, clinical studies data, SDG analysis, analysis of economic impact, as well as analysis of contributions to human medicinal products and orphan designations.

- How can one leverage TRR methodology to measure impact? 

Access to some readily available impact-level indicators which measure scientific, technological impact or contributions to SDGs

Find projects that had traces of policy impact in the first place.

Provide data for non-funded entities, i.e. the control groups.

New Challenges
Require new skills

and
a culture change

One multi-faceted, interlinked data infrastructure for all use cases

Because the underlying data is harmonised and curated in a Knowledge Graph repository , it is also *flexible and expandable* in how it can be used, and what it can be used for.
R&I: bottom-up and multidisciplinary
Missions oriented policies are top-down and multidisciplinary

Tracking of Research Results

Discover

ToRR: « Ask anything » semantic queries

Analyse and Report

Metrics and Indicators

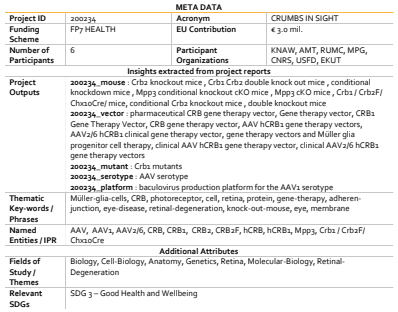
Direct Access to Data

Bulk downloads

GraphDB

ArangoDB

jupyter



META DATA			
Project ID	200234	**Acronym**	CRUMBS IN SIGHT
Funding Scheme	FP7 HEALTH	**EU Contribution**	€ 3.0 mil.
Number of Participants	6	**Participant Organizations**	KNAW, AMT, RUMC, MPG, CNRS, USFD, EKUT
Insights extracted from project reports			
Project Outputs	**200234_mouse** : Crb1 knockout mice , Crb1 Crb2 double knock out mice , conditional knockdown mice , Mpp5 conditional knockout cKO mice , Mpp5 cKO mice , Crb1 / Crb2F/ Chx10Cre/ mice, conditional Crb2 knockout mice , double knockout mice **200234_vector** : pharmaceutical CRB gene therapy vector, Gene therapy vector, CRB1 Gene Therapy Vector, CRB gene therapy vector, AAV hCRB1 gene therapy vectors, AAV2/6 hCRB1 clinical gene therapy vector, gene therapy vectors and Müller glia progenitor cell therapy, clinical AAV hCRB1 gene therapy vector, clinical AAV2/6 hCRB1 gene therapy vectors **200234_mutant** : Crb1 mutants **200234_serotype** : AAV serotype **200234_platform** : baculovirus production platform for the AAV1 serotype		
Thematic Key-words / Phrases	Müller-glia-cells, CRB, photoreceptor, cell, retina, protein, gene-therapy, adheren-junction, eye-disease, retinal-degeneration, knock-out-mouse, eye, membrane		
Named Entities / IPR	AAV, AAV1, AAV2/6, CRB, CRB1, CRB2, CRB2F, hCRB, hCRB1, Mpp5, Crb1 / Crb2F/ Chx10Cre		
Additional Attributes			
Fields of Study / Themes	Biology, Cell-Biology, Anatomy, Genetics, Retina, Molecular-Biology, Retinal-Degeneration		
Relevant SDGs	SDG 3 – Good Health and Wellbeing		

eBriefing Lab

European Commission

The AI change issue

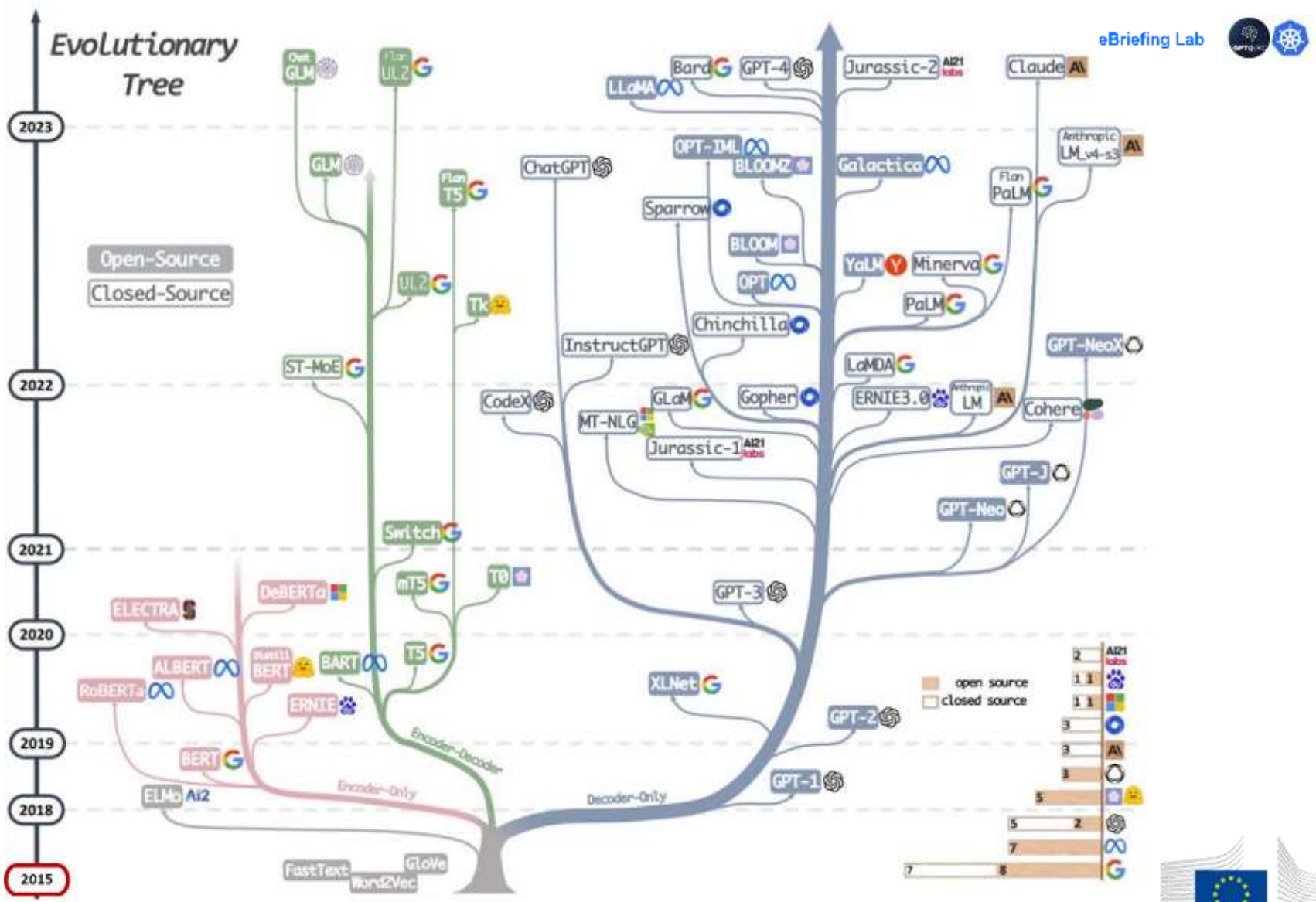
TRR KG-LLM

TRR KG

OSIRIS

CORTEX (IRIS)

IRIS



European Commission

Filling the gap by upskilling colleagues

R&I Data analytics developers

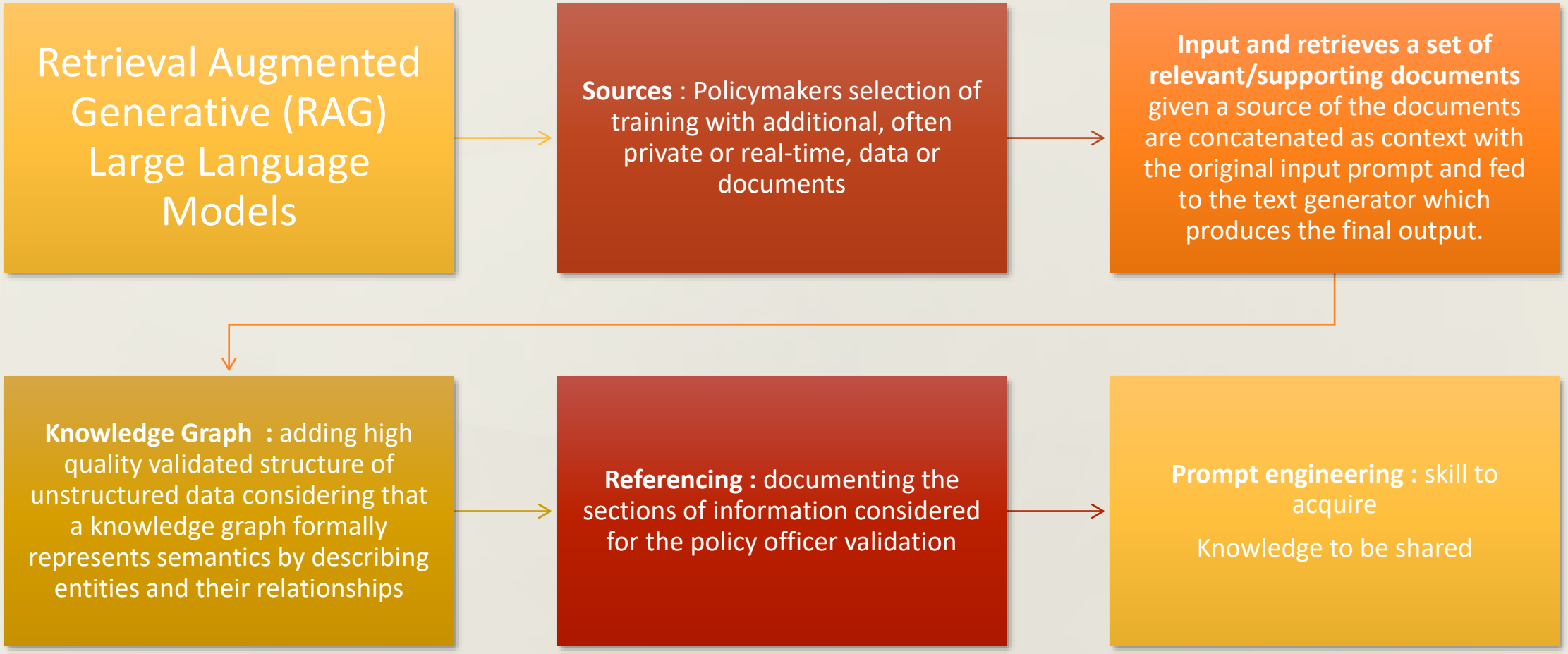
- Need to build up their policy knowledge (e.g. interdisciplinary IT/Business teams)
- Need to move from delivering discrete research reports to evidence production systems (e.g. data pipelines that can be updated over time)
- Need to make their analyses robust and reproducible and as much as possible share code and data (e.g. validations, triangulate novel/experimental results with existing sources)

R&I Policymakers

- Need to build up their data analytics capabilities (e.g. use internal/external teams for riskier and more ambitious R&D projects and develop their prompting skills)
- Need to support infrastructure development and strengthen the data ecosystem (e.g. HPC, Cloud, Open Access,…)
- Need to rigorously experiment with new analytics methods and share the results (e.g. explore new/radical opportunities for data powered R&I policy)

Based on Eurito and TalTech Talk: A vision for data-enabled research and innovation policy. Nov. 2022

Building Trust = human ex-ante & ex-post validation



- Will data enthusiasm and AI overtake scientific policy advice ?
- Will future generation of scientists lose their analytical skills?
- Data-driven outputs vs challenge-driven outputs?
- Open science and open data vs IPR (as open as possible, as closed as necessary)

European Commission

Semantic visualization for decision making

- Policy decision-making processes are often very complex and not always easy to comprehend.
- New dependencies constantly emerge, and policy cycle is accelerating
- The technology is there to clearly and comprehensibly illustrate the information needed for complex decision-making processes.
- To enable policy makers and citizens to visually grasp the impacts of new policies and understand their relationships to other areas.
- The data are useful in that they can show data links and narrow down the search from huge volumes of data to several potential targets, but our view is that the final decision/analytics should remain in the hands of a Policy Analyst or Policy Officer.

What can be the added value of AI based TRR methodology for policy making?

- Anticipate (foresight) the detection of problems by constant monitoring before they become intractable but also retracted papers, withdrawn patents (volatility of sources)
- Can offer a multilingual fruitful involvement of any stakeholder (internal or external) in the policy making activity (expandable to any (open) source contributions)
- Can be the interface of a cooperative platform (bridging data and policy) for multidisciplinary work (SDGs, Missions) (purpose-built teams)
- Surface holistic information and insights across silos of content and data by knowledge sharing (Basket and Alerts) including feedback to policy (human and machine readable for enterprise search)
- Uncover causal relationship behind policy issues (known the unknown)
- Identify and give access to cheaper and real-time proxies for traditional official statistics (reveal or complete the picture)
- Identify key stakeholders or expert networks to be involved or be the target of specific policies (identification of gatekeepers)
- Anticipate or monitor in right-time the impact of policies (societal pulse without the survey gap)

Broader context

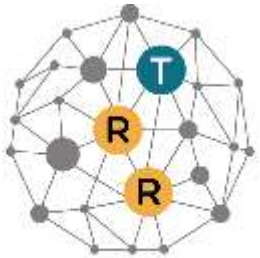
REITER

DATA for impact

STIPCOMPASS
INTERNATIONAL DATABASE ON STI POLICIES

EURITO

intelcomp



Tracking of Research Results

KNOWMAK
KNOWledge in the MAKing in European Society

THANK YOU

stop TESTING and start TASTING

European Commission