INTELCOMP PROJECT

A COMPETITIVE INTELLIGENCE CLOUD/HPC PLATFORM FOR AI-BASED STI POLICY MAKING

(GRANT AGREEMENT NUMBER 101004870)

Policy Brief on the Use of AI and Data-Driven Tools for STI Policy Design – final version (D1.5)

Deliverable information	
Deliverable number and name	D1.5. Policy Brief on the Use of AI and Data-Driven Tools for STI Policy Design
Due date	Dec 31, 2023
Delivery date	Dec 31, 2023
Work Package	WP1
Lead Partner for deliverable	Technopolis Group Belgium (TGB)
Authors	Paresa Markianidou (TGB) Lena Tsipouri (TGB)
Reviewers	Joseba Sanmartín Sola (FECYT) Jerónimo Arenas García (UC3M)
Approved by	Jerónimo Arenas García (UC3M)
Dissemination level	Public
Version	1.1

Issue Date	Version	Comments
Dec 22, 2023	0.1	First draft of the policy brief
Dec 25, 2023		Version incorporating all comments from technical partners
Dec 25, 2023	0.2	Version ready for reviewers
Dec 27, 2023		Annotated version with all comments from reviewers.
Dec 29, 2023		Updated version of the policy brief with improvements in the AI use case, ready for a new revies
Dec 29, 2023		Updated version that addresses the comments from reviewers
Dec 29, 2023	1.0	Version for approval
Dec 30, 2023	1.1	Version incorporating comments from Technical Manager
Dec 30, 2023	1.1	Annotated version with all comments from Technical Manager

DISCLAIMER

This document contains description of the **IntelComp** project findings, work and products. Certain parts of it might be under partner Intellectual Property Right (IPR) rules so, prior to using its content please contact the consortium coordinator for approval.

In case you believe that this document harms in any way IPR held by you as a person or as a representative of an entity, please do notify us immediately.

The authors of this document have taken any available measure in order for its content to be accurate, consistent and lawful. However, neither the project consortium as a whole nor the individual partners that implicitly or explicitly participated in the creation and publication of this document hold any sort of responsibility that might occur as a result of using its content.

The content of this publication is the sole responsibility of **IntelComp** consortium and can in no way be taken to reflect the views of the European Union.

The European Union is established in accordance with the Treaty on European Union (Maastricht). There are currently 27 Member States of the Union. It is based on the European Communities and the member states cooperation in the fields of Common Foreign and Security Policy and Justice and Home Affairs. The five main institutions of the European Union are the European Parliament, the Council of Ministers, the European Commission, the Court of Justice and the Court of Auditors.

(http://europa.eu.int/)

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 101004870.

ACRONYMS

AI — Artificial Intelligence

FOS — Field of Science and Technology Codes

FP7 — Seventh framework programme of the European Community for research and technological development including demonstration activities from 2007-2013

H2020 — The European Union research and innovation funding programme from 2014-2020

NLP — Natural Language Processing

NGO — Non-Governmental Organization

R&I — Research and Innovation

SDG — Sustainable Development Goal

STI — Science, Technology and Innovation

Policy brief on the use of AI and data-driven tools for STI policy design

Advancing data-driven policies

How far is the STI Policy Making Community in using AI and data-driven tools in 2022? The granularity challenge

R&I policy making is being revolutionised using text data analytics. The existence of repositories of scientific knowledge, together with the advances in text analytics are increasingly offering the opportunity to research funders to improve their policy making across the policy cycle:

- They can be supported in their agenda setting because they are informed almost in real time about scientific and technological progress, using data on publications, patents, etc. These data are translated into information on the extent of diffusion and cooperation through text analysis.
- They are also supported in monitoring and evaluating their own programmes or policy mixes in terms of scientific, technological, economic and societal impact. The combination of funding with outputs and outcomes helps identifying the attribution or contribution of their funding to grantees, the scientific community, the economy, and society.

The content of open data repositories increases by the day and so do technical expertise of research teams. This tendency represents an opportunity for STI policy makers to overcome access to data limitations and be able to re-use and replicate results.

However, a lot more is required because of the complexity of the system, which is characterised by a multitude of interdependencies at scientific/ economic/ societal levels, at the level of macro-regions/national and subnational policies and last but not least at the level of different actors, namely research organisations, companies, intermediaries and NGOs.

The complexity leads to a granularity challenge. Sophisticated policy questions can only be answered using refined disciplinary, stakeholder and territorial breakdowns, which in turn call for advanced mappings, bottom-up machine-guided approaches and multifaceted visualisations.

For instance, STI policy makers want to understand in which way (funded) research addresses SDGs (topics addressed across scientific disciplines, research teams, funders) and contributes to the achievement of SDGs (technology generation and diffusion).

STI policy makers need AI due to its unique ability to perform certain operations on data: exploit unstructured data (e.g., policy documents); process and make sense of text (e.g., publication or patent abstracts); match various data sources according to particular criteria (e.g., company name); classify data based on rich, multidimensional information (e.g. technologies or SDGs).

Despite the underlying complexity of AI methods, it is possible to simplify and/or make their application more intuitive. This is definitely a need in the context of policy making to make these tools clear and user friendly to be attractive to non-AI experts.

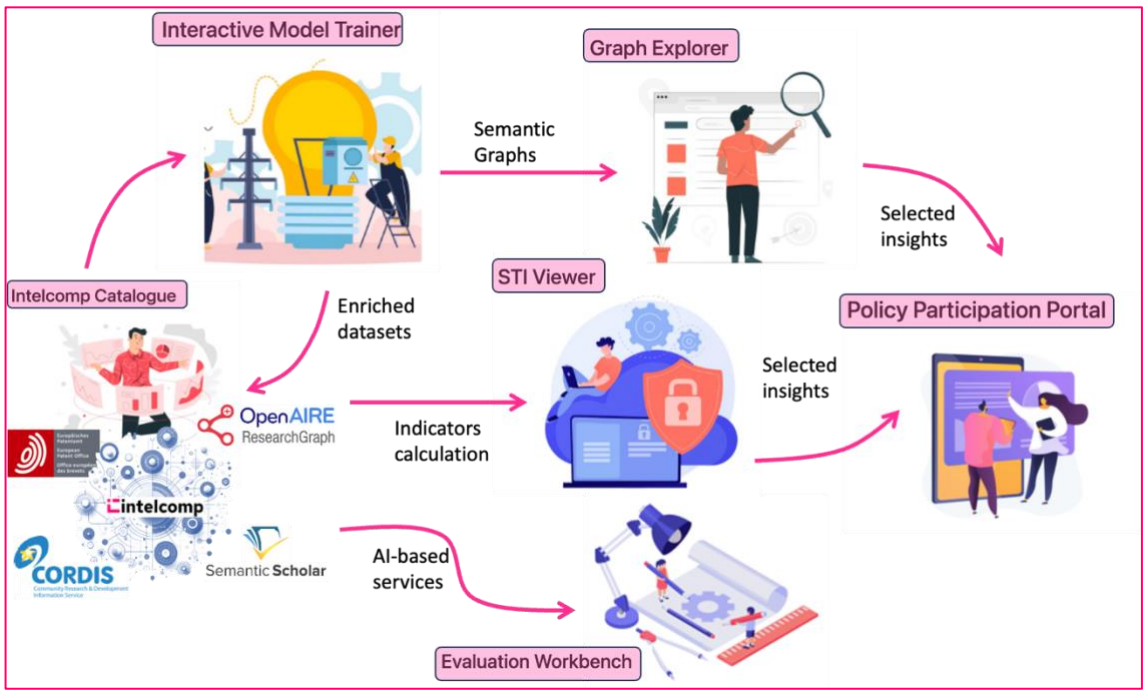
Part of this great potential of AI is not there yet. Its development is hampered notably by the fact that databases are often not sufficiently clean or accessible. And sometimes AI cannot compensate for the lack of clean data. More needs to be done in this regard.

The IntelComp Approach: responding to the granularity challenge

The objective of IntelComp is to deliver an integrated platform that provides tools assisting the whole spectrum of STI policy. The work undertaken has concentrated into the needs and policy questions of STI policy makers at the stage of agenda setting and monitoring and evaluation. To test how well the efforts to address the lower granularity are tackled by IntelComp, the platform was tested in three domains: artificial intelligence, climate change and health.

To offer a comprehensive solution the use of multiple datasets and state-of-the-art AI and NLP techniques is indispensable, underscoring the need for platforms that can decipher complexity and serve as inputs for the formulation of STI policies. The figure below shows the workbench of tools developed. The IntelComp catalogue includes a variety of heterogeneous STI-related datasets including both structured and unstructured data available at the IntelComp Dataspace. These datasets (or subsets of them) can then be enriched using different pretrained text analysis services, or ad-hoc models trained using the Interactive Model Trainer. AI-enriched information is then used to provide data-based evidence through two of the IntelComp end-user tools: the STI Viewer and the Graph Visualizer. The Policy Participation Portal collects the views of STI policy stakeholders as a channel for continuous improvement of IntelComp's services. Finally, the Evaluation Workbench is another end-user tool providing a series of services that can be used during the evaluation of funding proposals.

Figure 1 - IntelComp Workbench of tools



Source: IntelComp, 2023

The three main services that IntelComp relies on for data enrichment are the following:

1. *Service for domain-related subcorpus generation*. The domain related datasets are created by IntelComp using transformer-based classifiers that retrieve the relevant domain documents for any available dataset.

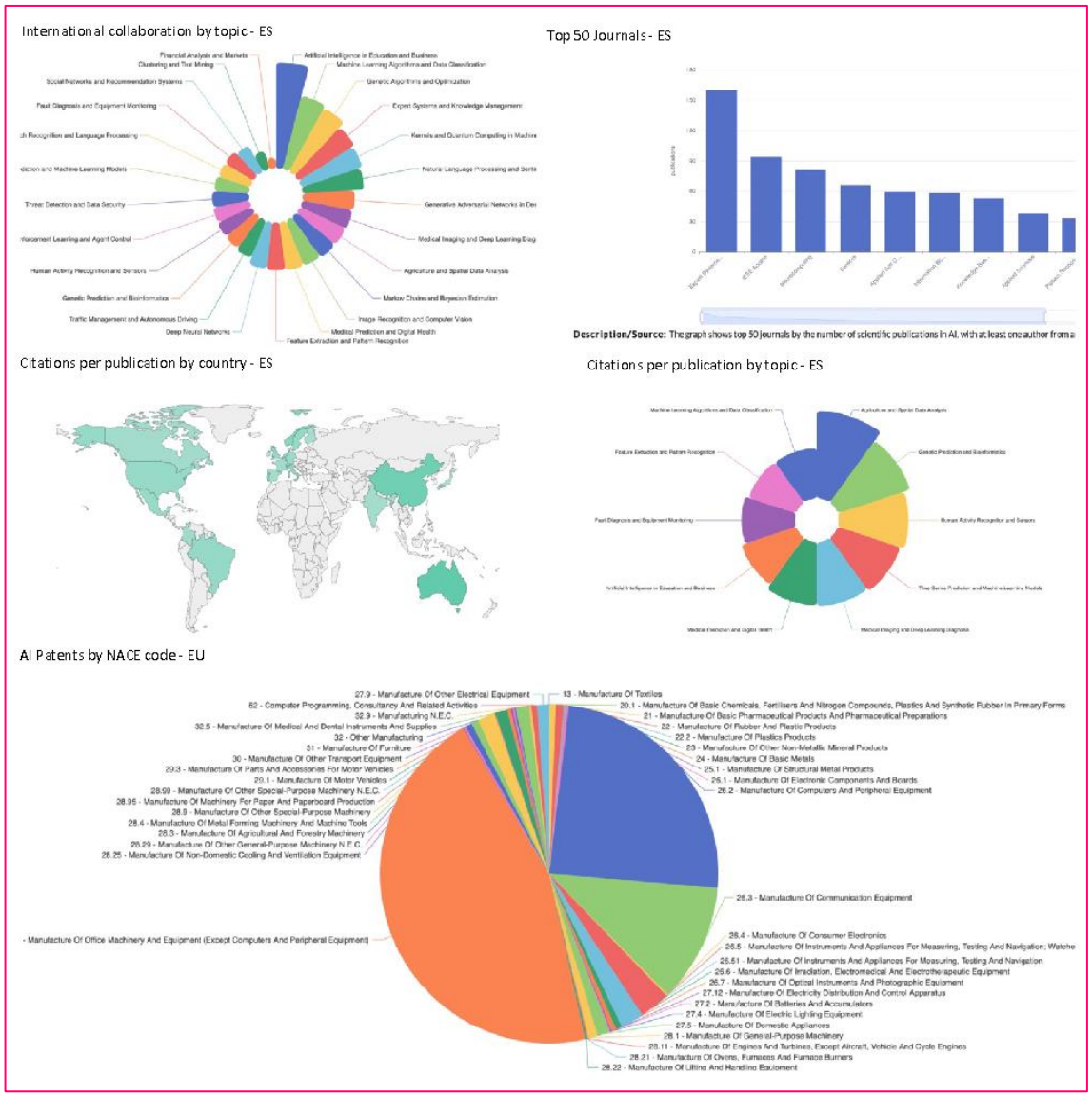
2. *Classification service*. In the resulting datasets, documents are automatically classified using predefined taxonomies (e.g., FOS, IPC, and SDGs). The project has also generated a service to train new classification models for other taxonomies requested by the users.

3. *Advanced topic modelling service*. Topic modelling is used in the domain related datasets to compile appropriate indicators making it feasible to analyse data with different levels of granularity. In IntelComp, a comprehensive approach to scope the artificial intelligence and cancer domains is employed. Data-driven methods are utilised to extract AI-related topics, curated by domain experts and tailored to the specific corpus of each data source (science, technology, industry) corresponding to different functions of the innovation system.

Results are then:

4. Visualised in the STI-Viewer, which defines and implements IntelComp's visualisation components customised for policy making. The STI viewer provides policymakers with adaptable, interactive visualisations collaboratively developed with users. It facilitates the analysis of extensive datasets from various sources, applying a diverse array of filters to disaggregate data based on countries, topics, institutions, funders, etc., in a user-friendly fashion.

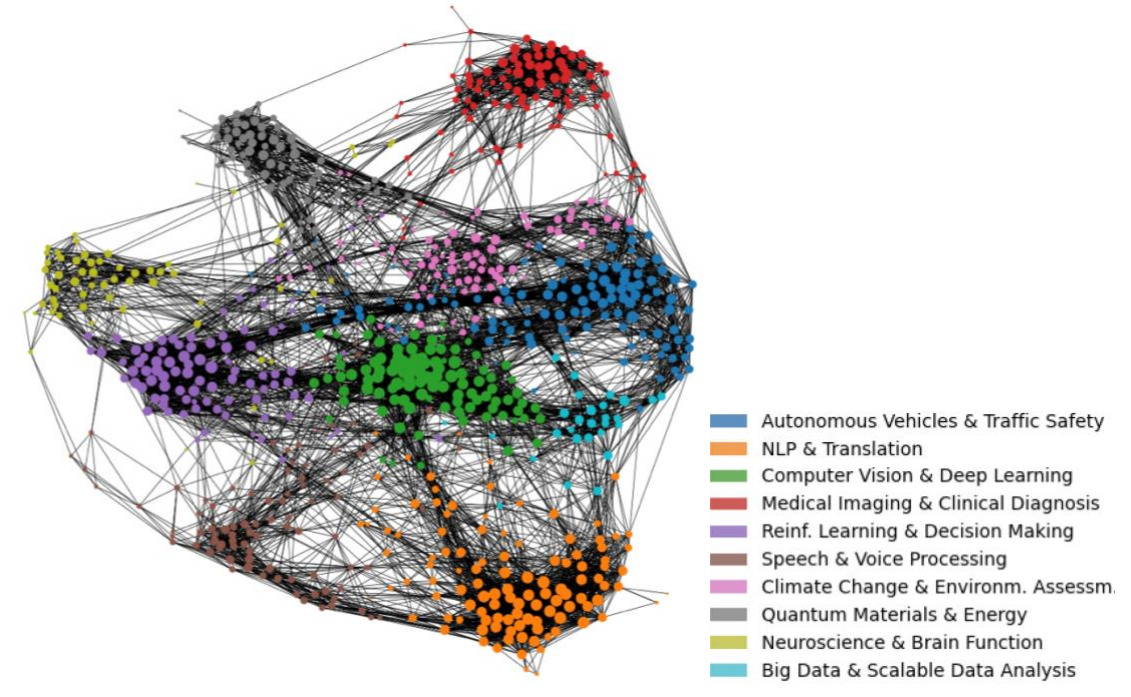
Figure 2 - Examples from STI viewer on Artificial Intelligence



Source: IntelComp, 2023

5. Inserted in the graph visualiser, which provides a one-shot visualisation of the whole collection of documents in the selected dataset, allowing the visualisation of communities, and filtering according to different criteria (e.g., project budget, country information, etc). The visualisations and corresponding measurements enable STI policy makers to analyse the role of participants in the network using a wide range of measurements to benchmark countries, organisations and projects.

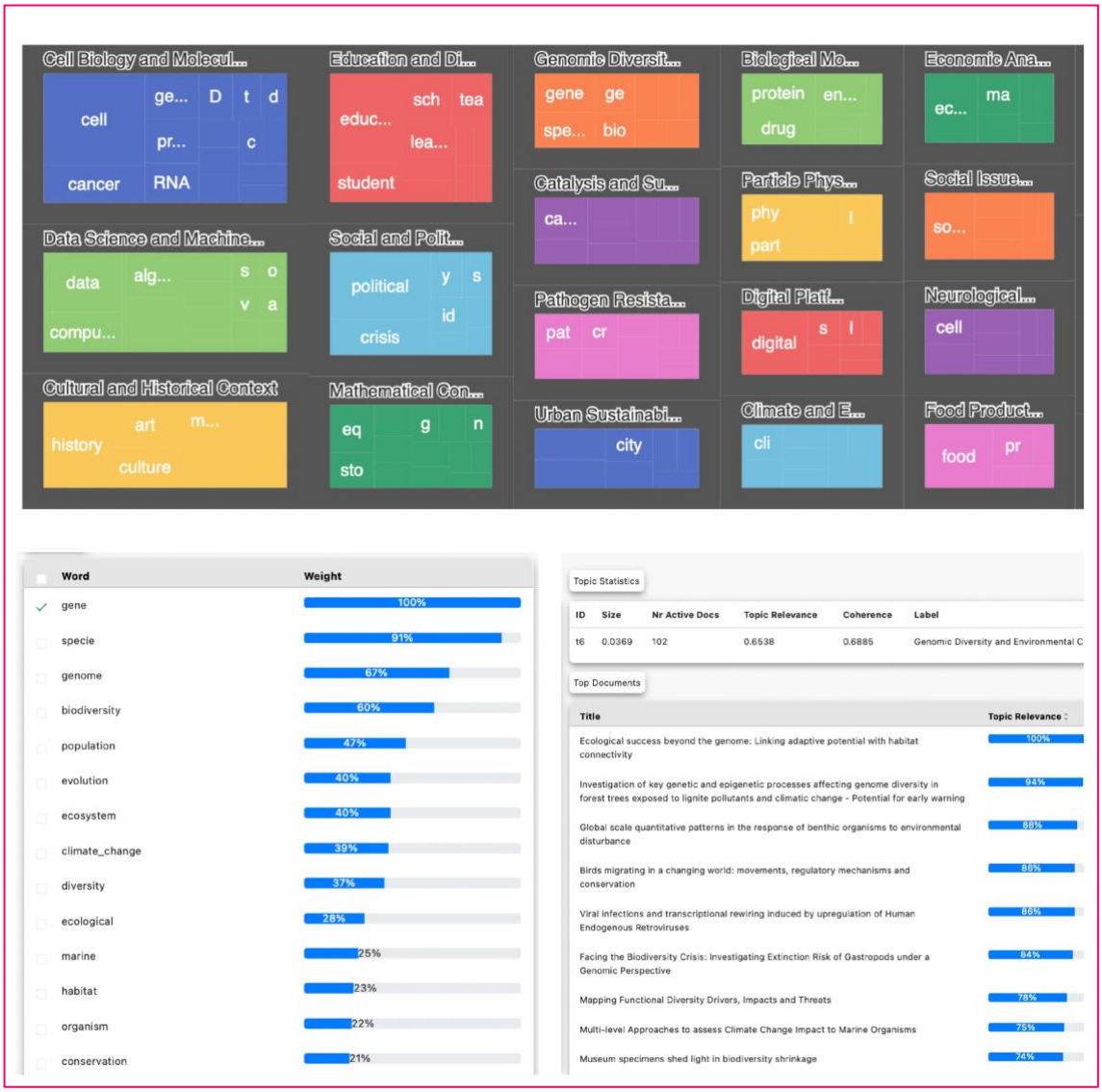
Figure 3 - Example of graph visualizer of Communities in Artificial Intelligence using data of CORDIS covering the framework programmes (FP7, Horizon 2020 and Horizon Europe data available from CORDIS in August 2023)



Source: IntelComp, 2023

6. Exploited to provide a series of AI-powered services through the Evaluation Workbench to assist in the evaluation of project proposals. These services are supported by topic models trained on STI funders' data related to projects and experts, and can be utilised by funders to examine the similarity among projects and experts, as well as to establish taxonomical classifications of project abstracts. This tool empowers users to curate the models, leverage intelligence to assign evaluators, and generate data for monitoring funding distribution. Additionally, it facilitates evaluations of calls or programs.

Figure 4 - Example of Evaluation workbench applied to Hellenic Foundation for Research and Innovation



Source: IntelComp, 2023

Conclusion and way forward

The work in IntelComp showed that topic modelling provides valuable more detailed information, revealing semantic themes within analysed corpora in the Science, Technology and Innovation domain. The Interactive Model Trainer developed in IntelComp proves instrumental in selecting the optimal number of topics, eliminating irrelevant topics, and creating curated vocabularies to build high-quality models . However, following a user-in-the-loop approach is crucial to manage expectations, considering the non-deterministic nature of model outputs and the need for a comparative interpretation of topic sizes.

The processing of a large volume of unstructured data from new, open or proprietary sources necessitates the use of data science tools introducing complexities related to curating topics using domain and technology terminology, multilingualism and ensuring the robustness of AI methods. The STI viewer, graph visualiser and evaluation workbench, serve as valuable tools for STI policymakers, facilitating the management of large datasets, curation of topics, and complex calculations. They enable regularity in the analysis and reduce resource intensity.

The application of Artificial Intelligence (AI) in science, technology, and innovation (STI) policy analysis brings about a transformative paradigm, particularly in managing the time dimension of data. AI facilitates the organisation of data to cater to Hindsight, Now Sight, and Foresight, ensuring that information is not only historical but also current and forward-looking.

The types of data encompass a comprehensive spectrum, including science, technology, economic indicators, and social impact metrics. AI enables the integration and analysis of diverse data types, providing a holistic understanding of the multifaceted implications of STI initiatives.

At the level of detail, AI utilises standard classifications such as NACE, FoS, IPC, SDGs, ESGs, and others at the lowest possible granularity, allowing for precise analysis and benchmarking but also validation of outcomes. AI-driven analyses extend to all levels of aggregation, ensuring flexibility in interpreting data from the micro to macro perspectives.

AI also introduces the capability to uncover hidden issues through the creation of new classifications, thereby illuminating nuances that might escape traditional analytical approaches. In essence, AI revolutionises STI policy analysis by offering a dynamic, multifocal, and insightful approach that harnesses the power of diverse data dimensions and classifications.

In the future, Large Language Models (LLMs) will enhance the intuitive interaction between end users and platforms such as IntelComp. The swift progress in technology will mitigate AI hallucinations, thereby fostering increased trust in the collection, classification, and interpretation of data.

In conclusion the ability to leverage vast amounts of data from various sources (both new and conventional) and the potential to utilise cutting-edge Artificial Intelligence Models (including deep learning, NLP, and LLMs) tailored for STI policy will equip policymakers with a versatile toolbox. This includes flexible, interactive visualisations and dashboards that provide valuable insights to inform decision-making.