**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 – final version |
| Due date | Dec 31, 2022 |
| Delivery date | Mar 21, 2023 |
| Work Package | WP1 |
| Lead Partner for deliverable | Technopolis Group Belgium (TGB) |
| Authors | Paresa Markianidou (TGB)<br>Lena Tsipouri (TGB) |
| Reviewers | Utku Demir (ZSI)<br>Dominique Guellec (HCERES)<br>Joseba Sanmartín Sola (FECYT) |
| Approved by | Jerónimo Arenas García. Technical Manager (UC3M) |
| Dissemination level | Public |
| Version | 3.6 |

**Table 1. Document revision history**

| Issue Date | Version | Comments |
|---|---|---|
| Dec 9, 2022 | 0.1 | First draft of the policy brief (M24) |
| Dec 25, 2022 | 1.0 | Version incorporating all comments from technical partners |
| Jan 9, 2023 | 1.1 | Version ready for reviewers |
| Jan 22, 2023 | 2.0 | Annotated version with all comments from reviewers. |
| Feb 15, 2023 | 3.0 | Updated version of the policy brief with improvements in the AI use case, ready for a new review |
| Mar 1, 2023 | 3.1 | Updated version that addresses the comments from reviewers |
| Mar 2, 2023 | 3.2 | Version for approval |
| Mar 8, 2023 | 3.3 | Version incorporating comments from the Technical Manager |
| Mar 14, 2023 | 3.4 | Annotated version with all comments from the Technical Manager |
| Mar 20, 2023 | 3.5 | Version ready for submission |
| Mar 21, 2023 | 3.6 | Submitted version |

# 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.

Among the data that are easy to access, e.g., publications and patents, those can be retrieved fairly easily at an aggregate level.

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 (across scientific disciplines, research teams and 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, the latter need to be 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 so far has concentrated into the needs and policy questions of STI policy makers at the stages of agenda setting and monitoring and evaluation. For this, global and local knowledge is needed regarding the focus of scientific research (measured in publications and patents), its impact (measured in citations) and the role of collaborations.

To test how well the efforts to address the lower granularity are tackled by IntelComp, the platform is being tested in three domains: artificial intelligence, climate change and health. Domain experts within IntelComp's Living Labs are consulted for the content of each domain and help work out priority STI policy questions.

Datasets were drawn from CORDIS, OpenAIRE Research Graph, Semantic Scholar and PATSTAT.

NLP and machine translation pipelines, as well as other AI technologies, are applied for the study of unstructured textual content of the identified information sources.

For the moment, 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.

2. *Classification service*. In the resulting datasets, documents are automatically classified using predefined taxonomies (e.g., FOS and SDGs).

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.

Results are visualised in the STI-Viewer, which defines and implements IntelComp's visualisation components customised for policy making.

The Climate Change and AI case studies below give an overview of the complexity when trying to address policy questions at the lowest possible level and present the work undertaken so far.

**Using data and AI tools for agenda setting in the case of climate change**

Substantial contribution to climate change adaptation depends mainly on seven areas of economic activity. Among those, the energy sector is selected as a use case and 24 sub-systems of production and storage are created via IntelComp's data enrichment services.

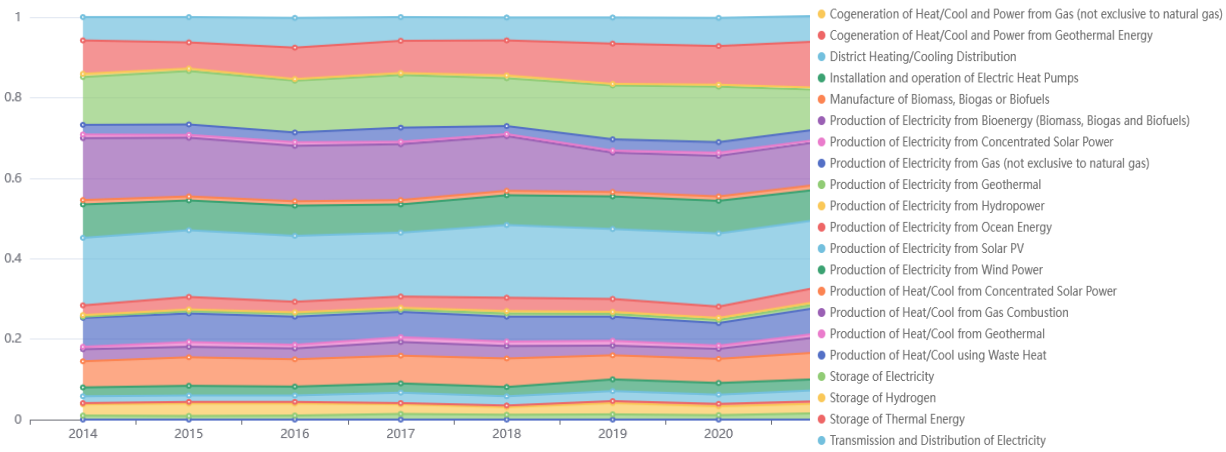
The analysis of the Living Lab started with the European Union and is proceeding with datasets specific to Greece and how policies convert from global to local, demonstrating the value of the potential to compare national with macro-regional data.

The work started with the **number of annual scientific publications in the Energy domain by topic**.[1]

This helps policy makers understand in which topics there is the highest effort in research and decide whether it is worth joining global efforts in the particular topics.

[1] The ontology of topics is taken from the technical annex of the Final Report of the EU taxonomy for sustainable activities https://finance.ec.europa.eu/sustainable-finance/tools-and-standards/eu-taxonomy-sustainable-activities_en#documents

**Figure 1. Publication in the topics of Energy over time (share in total Energy)**



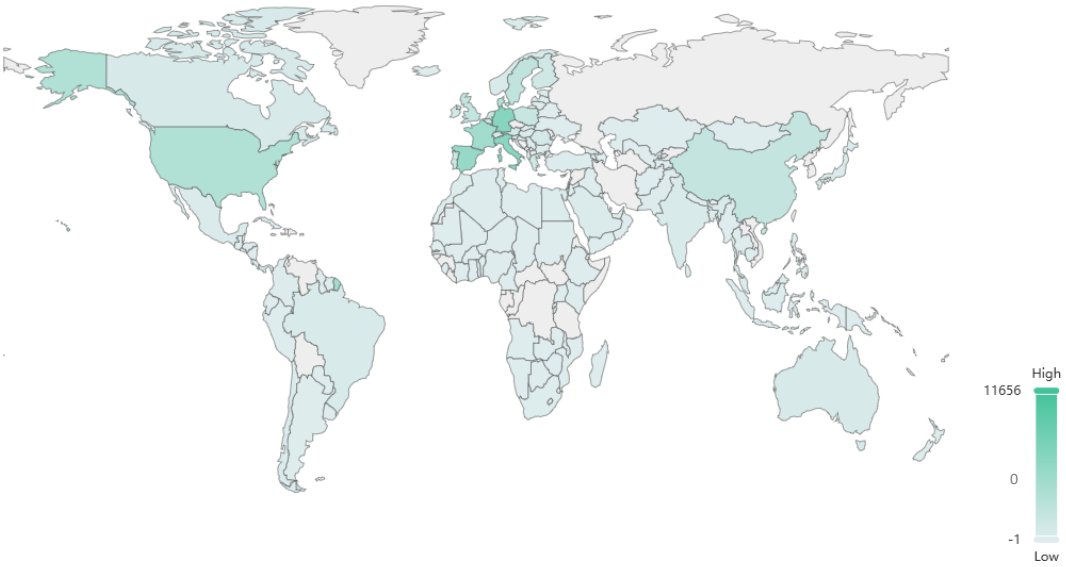

**Trends in Open Access** publications, helping policy makers decide if/when they wish to integrate open access policies into their regulatory framework.

**Figure 2. Number of Open Access publications over time in the Energy domain**


**Description/Source:** The graph shows the evolution in the number of publications that are open access (best available access rights).

**Scientific production origins** via the density of publications by country, helping policy makers to decide where to focus funding for bilateral agreements.

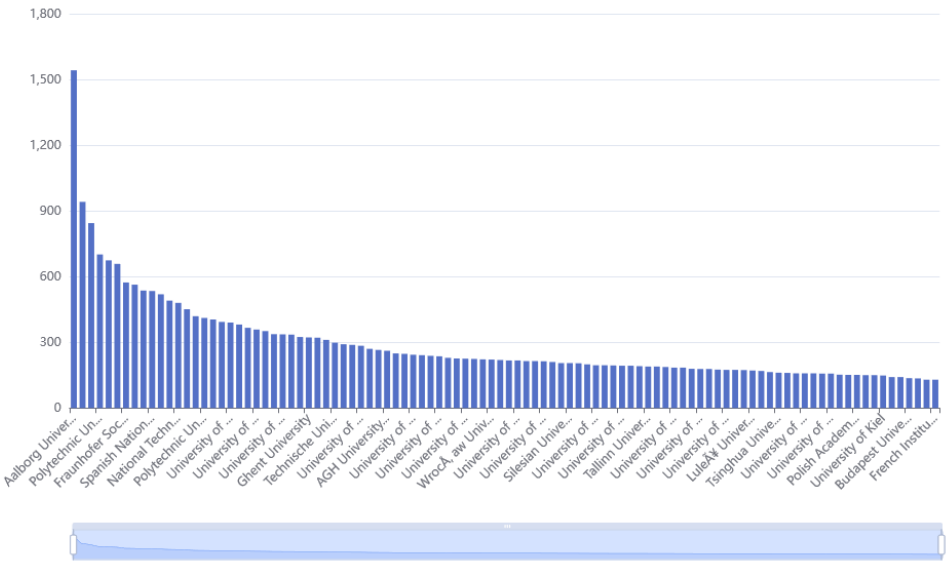
**Figure 3. Scientific production by origin in the Energy domain (2014-2020)**




**Description/Source:** The graph shows the number of publications by the country of the affiliated organization of an author. In the case of multiple authors each organization and corresponding country is counted as a separate publication (e.g., one publication with three authors, where two are from Greece and one is from Spain is counted as two publications in Greece and one in Spain).

**Publications by top 100 organizations**, helping policy makers and individual research teams to identify partnerships with top organisations.

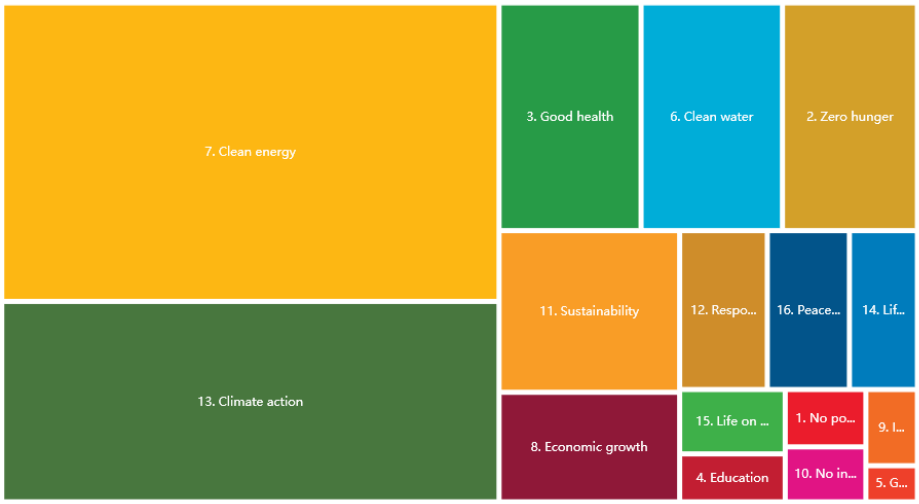
**Figure 4. Scientific production by organisation in the Energy domain (top 100) (2014-2020)**




**Description/Source:** The graph shows the top 100 organizations in terms of number of publications in Energy.

**Scientific production characteristics** applied to breakdown the publications by SDGs and Top 50 Journals and Publishers. The former is relevant for international organisations and NGOs, as well as national policy makers when broken down by country. The latter is highly relevant for researchers to select where they will undertake literature search and/or submit their manuscripts.

**Figure 5. Scientific production in the Energy domain by SDGs**




**Description/Source:** The graph shows the number of publications by UN Sustainable Development Goal (SDG, https://sdgs.un.org/goals). Our SDG classification system uses deep learning techniques to assign a scientific publication to the SDGs that are related to it.

Turning to **impact** citable papers compared to non-citable are shown, followed by total citations over time per topic, average citations per publication over time and a field-weighted citation impact[2] over time. Citations per publication by country, by organization and by funder are a crucial indicator for the quality of research and scientific return on investment.

Finally, **international collaborations** over time are presented showing the evolution in the number of publications with authors affiliated to organisations in at least two different countries. Superimposed is the number of those that are cited by at least one publication.

The underlying data of indicators presented in the STI-Viewer by topic are available for policy makers to download and construct indicators tailor-made to their needs.

While waiting for more datasets to be added and reflected in STI-Viewer some interesting deductions can already be suggested: Interestingly enough, the top topics by citation are not the same as by publication indicating either diminishing returns to scale or shifts in priority topics (more publications suggest emerging fields, more citations more mature fields because of the lags in citing).

When publications are broken down by topic, Production of Electricity from Solar Photovoltaics and Storage of Electricity rank high and production of Heat/Cool from Gas Combustion appear at the top technologies for the first time.

Scientific collaborations by SDGs are concentrated in Clean Energy and Climate Action as in the case of Publications.

[2] Field-Weighted Citation Impact is the ratio of the total citations actually received by the denominator output and the total citations expected based on the average of the subject field.

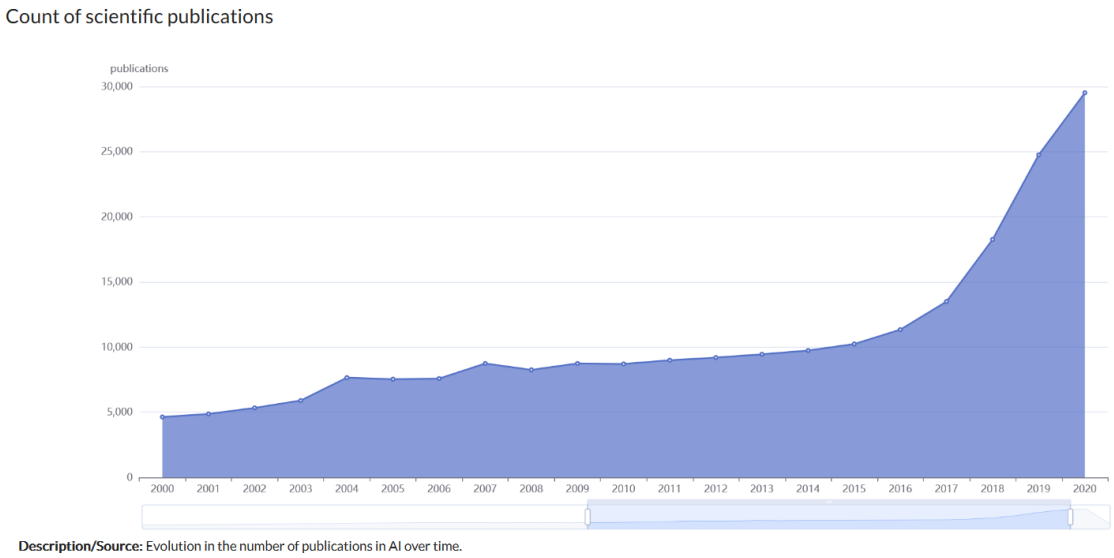
**Using data and AI tools for agenda setting in the case of R&I for AI**

R&D investments in AI aim to foster excellence and help countries achieve economic and social objectives of their policies and strategies. AI applications are diverse making it hard to understand AI use cases and their link to policy making. The aim of the AI Living Lab in IntelComp is to intelligently attribute AI documents to the AI domain and later group them by topics.

The datasets for the use case are documents from research projects of the European Commission (CORDIS data from FP7, H2020 and part of Horizon Europe), the PATSTAT Database, and Semantic Scholar.

The process started with the use of approximately 120 key words known from previous research to be relevant for AI. Such keywords were manually revised to avoid polysemic terms that are not uniquely related to AI (e.g., neural networks). A first set of documents were found using these "past-determined" keywords, selecting documents that contained two or more key words. However, using key words carries the risk of lock-ins and the inability to capture the recently emerging trends. To address this risk machine learning was applied to complement the existing knowledge, using supervised classification and relevance feedback for retraining a transformer-based classifier. As a result, a total of 76,350 (out of 60,296,981) patents and 333,626 papers (out of 99,408,292) papers from Semantic Scholar are included in the AI-related sets. Currently, a similar approach is being undertaken to identify AI-related projects from CORDIS.

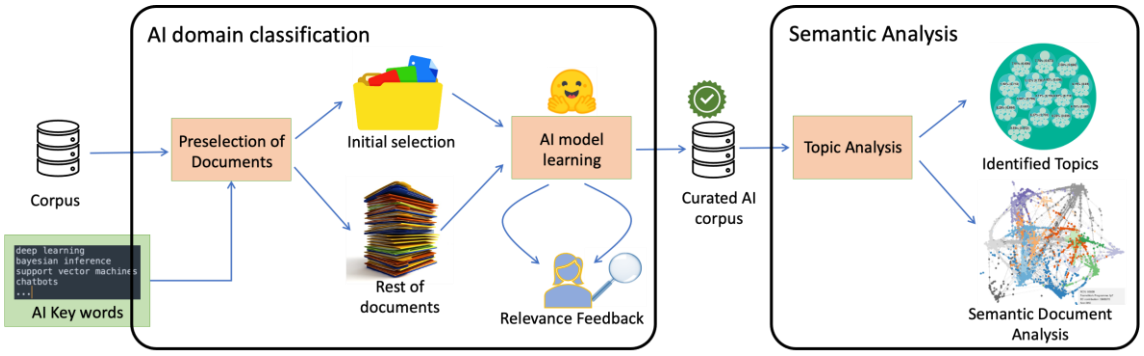
**Figure 6. Scientific production in the AI domain**




**Description/Source:** Evolution in the number of publications in AI over time.

Once the final set of AI-related documents are selected, Latent Dirichlet Allocation is used for topic modelling, identifying sets of words that appear often together in AI documents. Each such set of words characterizes a topic which is then manually labelled by an AI expert such as for example, the ***topics Neuroscience, Robotics, or Applications of AI in Health*** among others.

**Figure 7. Workflow for the identification of documents in the AI domain and the topic model on the selection**




**Next steps**

The first results on Energy as a key element for mitigating climate change and on R&D in AI indicate that IntelComp has successfully produced topic models and automatic classifiers to address the granularity challenge. In the remaining months the limited use cases of Energy and AI in the area of science for agenda setting will be expanded, to include indicators grouped in other areas, e.g. technology, industry, human resources, policy and society. Indicators will be broken down across different topics (granularity) to understand trends and problem areas.