INTELCOMP PROJECT

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

(GRANT AGREEMENT NUMBER 101004870)

REPORT ON THE SELECTED MEASUREMENT AND DATA COLLECTION. DELIVERABLE D1.2

Deliverable information	
Deliverable number and name	**D1.2 Report on the selected measurement and data collection**
Due date	31 December
Delivery date	
Work Package	WP1
Lead Partner for deliverable:	Technopolis group
Authors	Paresa Markianidou (Technopolis group) Hannah Bernard (Technopolis group) Apolline Terrier (Technopolis group) Lena Tsipouri (OPIX) Jeronimo Arenas Garcia (UC3M) Doaa Samy (UC3M) Ioanna Grypari (ARC) Dimitris Pappas (ARC)
Reviewers	Dietmar Lampert (ZSI) Dominique Guellec (Hcéres)
Approved by	Cecilia Cabello, Project Coordinator (FECYT)
Dissemination level	Public
Version	1.2

Table 1. Document revision history

Issue Date	Version	Comments
11/01/2022	1.0	Document with contributions from internal peer reviewers
11/01/2022	1.1	Version for approval by the Project Coordinator and the Technical Manager
12/01/2022	1.2	Document with contributions from the Project Coordinator

DISCLAIMER

This document contains a 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.

CONTENTS

FIGURES

TABLES

ACRONYMS

AI — Artificial Intelligence

EC — European Commission

FP — Framework Programmes

H2020 — Horizon 2020

HEUROPE — Horizon Europe

IPC — International Patent Classification

LL — Living Lab

NACE — Statistical Classification of Economic Activities in the European Community

NLP — Natural Language Processing

PU — Positive-Unlabeled

R&I — Research and Innovation

SDGs — Sustainable Development Goals

STI — Science Technology and Innovation

TED — Tenders Electronic Daily

TRL — Technology Readiness Levels

1. SUMMARY NOTE

The objective of IntelComp's D1.2 deliverable is to translate its conceptual framework into concrete measurements which serve as basis for the co-creation process in the three science, technology and innovation (STI) domains: AI, Climate Change - Blue growth and Health – Cancer.

In the current version of deliverable D1.2 we report on the progress in defining the **measurements** and **data sources** and briefly explain the tools we foresee for end users and the services which are required for the calculation of the identified measurements. The final listing of measurements and data sources will only be concluded upon finalisation of the living labs' needs.

In section 2 we describe the first set of measurements which correspond to the domain-agnostic (i.e. non domain specific) policy framework. The domain-agnostic set of measurements serves as a catalogue of measurements which require Natural Language Processing (NLP) and Artificial Intelligence (AI) techniques to discover relevant information connected to the target measurements. The list covers measurements: 1) which require AI; 2) for which sufficient data could be sourced (provisional assessment); 3) which are technically feasible (provisional assessment) and 4) which are within the scope of the IntelComp project . The prioritised list of domain agnostic measurements and their corresponding sources to be integrated in IntelComp will be finalised by Month 16.

In section 3 we describe the second set of measurements which correspond to the domain specific policy framework, as expressed by the needs of the living labs. In this report a provisional set of measurements and data sources are provided for the cancer living lab, the first living lab to provide a needs assessment. The inputs of section 3 are subject to prioritisation in the context of the co-creation process with the living labs starting in January 2022. During 2022 the final listing of measurements and data sources for all living labs will be completed.

The data sources are listed in section 4 and Appendix I corresponding to a short and a long list of sources respectively. The data sources list is a result of internal consultations on each of the individual measurements of the policy framework. The prioritised list of sources to be integrated in IntelComp will be finalised by Month 16.

Sections 5 and 6 describe briefly the tools and services provided by IntelComp to calculate those measurements which require processing unstructured text, enriching and extending the evidence-basis for STI policy makers and Public Administrations.

Finally, section 7 provides a gap analysis by comparing the domain agnostic policy framework to the provisional implementation plan in IntelComp. It synthesises the policy questions which cannot be addressed by IntelComp organised by a typology of main reasons for exclusion.

2. DOMAIN-AGNOSTIC MEASUREMENTS

2.1. Measurements

In IntelComp, a distinction is made between statistical indicators and quantitative measurements for policy making.

The OECD glossary of statistical terms defines **statistical indicators** as 'data elements that represent statistical data for a specified time, place, and other characteristics'.[1] The European Statistical System Committee (ESSC) defines **indicators for policy making** as 'a particular subset of statistical information, directly related to a special purpose such as monitoring specific policy objectives' (Eurostat, 2017).

Statistical indicators supporting evidence based policies need to meet stringent quality standards as set in the European Statistics Code of Practice containing 15 principles.[2] To enrich the evidence basis with indicators derived from big data which are trusted by policy makers, the data must meet quality standards as described by the quality dimensions in the UNECE framework for the quality of big data described in the European Statistical System handbook for quality and metadata reports (Eurostat, 2020). This requires for instance sound methodologies applying appropriate statistical procedures to address sample bias.

We are aware that some of the data sources to be exploited by the IntelComp platform do not provide a representative coverage of innovation at either the industry or national level because the data are based on self-selection (e.g. firms that apply for a patent) and the information they provide is often incomplete, covering only one facet of innovation (e.g. company R&D investments). In addition, some data sources are inconsistent in their coverage of innovation activities (e.g. company websites). As a consequence, the measures derived from these data sources may not be considered statistical indicators because of the lack of quality and representativeness of those data sources.

In IntelComp, some measurements are designed even if they do not yet comply fully with quality standards, either because they are geographically restricted to one or a limited number of Member States or because a representativeness analysis is not performed on all possible dimensions of the data (e.g. country, gender, level of education, industrial sectors, etc.). IntelComp data and measurements represent experimental statistics and should be distinguished from traditional statistical indicators compiled by Eurostat and National Statistical Offices. IntelComp data and measurements will be complementary to standard indicators by generating tailor-made measurements aimed for policy making.

Despite their limitations, these measurements serve the purpose of providing relevant information for specific tasks in the policy cycle of a specific strategy, program or call. They are not all designed to inform policy discussions at higher levels, for example to monitor progress

[1] Available here: https://stats.oecd.org/glossary/detail.asp?ID=2547

[2] The 15 principles set in the European Statistics Code of Practice include: professional independence, mandate for data collection, adequacy of resources, commitment to quality, statistical confidentiality, impartiality and objectivity, sound methodology, appropriate statistical procedures, non-excessive burden on respondents, cost effectiveness, relevance, accuracy and reliability, timeliness and punctuality, coherence and comparability, accessibility and clarity.

towards a related policy target. Nor are they all designed for international comparability or benchmarking.

Indeed, during the development of the framework for policy making, an analysis was carried out for each measurement identified, analysing in which cases IntelComp will offer the evidence sought. The use of unstructured text associated with the documents in the datalake and exploitation of AI pipelines constitutes the core of the methodological contribution of IntelComp. As a result of this analysis, four tools are proposed (described in section 5) and a series of minimum services are identified (described in section 6).

Using these tools/services, some of the prioritised measurements (described in section 3) can be directly calculated, while in other cases other correlated information will be obtained. To this aim, IntelComp will enrich the data sets in the datalake with information obtained from the Internet or other available datasets (using the crawling and homogenization services included in the data lake), as well as with the outputs of the AI services.

As an example, the calculation of high-impact publications will involve enriching the desired publication data set with journal quartile information, while topic modelling will allow the analysis to be broken down by areas with the desired level of granularity. This information will be made available to the user in the STI Viewer through a Business Intelligence (BI) panel enriched with the newly calculated data, so that the activation of the corresponding filters will allow the user to obtain the desired information.

2.2. Measurements for Agenda setting

At the start of policy making the problem(s) to be addressed need to be defined. Policy makers need information to understand the array of sectoral/technological/institutional potential for a specific future period, determined by internal and external factors. While policy makers may have solid knowledge of the past performance in their area of competence, emerging changes constitute important information to guide them to the next (usually 5-7 years) policy cycle. Policy needs refer to the decision on priorities and budget allocations.

The information needed is on the current and emerging global societal challenges, the way these challenges are translated into their own context, the way their peers adopt their agendas and the potential of civil society to co-create the agendas but also on opportunities to improve the country's economic benefits in the years to come by identifying sectors or products and technologies with increasing global demand. The outcome can lead to strategic priorities forming Smart Specialisation Strategies as well as lower priority areas to be supported.

Short listed policy questions for agenda setting are described in terms of measurements, data sources and most relevant taxonomies. The list includes measurements which: 1) require AI; 2) sufficient data could be sourced (provisional assessment); 3) are technically feasible (provisional assessment) and 4) are in scope as per the proposal. The final list of measurements to be included in Intelcomp will be provided by Month 16 and is subject to the final list of sources. Equally, the final unit of observation is subject to technical assessment and relevance of each measurement. For instance, in agenda setting measurements, if, for instance, the specific policy is a national strategy, the unit of observation of each measurement would be the country/ies, the scientific discipline (e.g. cancer research) or the technology (e.g. AI) related to that strategy.

Table 1: Entrepreneurial Activity

Policy question	Measurement	Data Sources	Taxonomy
Are national/regional companies adapting to technological transformation trends in their respective sectors? How do they compare with major (foreign or non regional) competitors?	Number of companies developing/adopting transformative technologies per 10,000 companies in the country ['transformative technologies' are defined by living labs]	* Company database compiled from various sources (see data sources for more detail). Representative database for Large R&D investors, Large companies and technology start-ups and scaleups).	* Technological transformation trends/ innovations/corresponding products [LL specific] * NACE * Company type (Largest R&D investors, Large companies, technology start-ups)
What is the composition of emerging technology portfolios of entrepreneurial companies?	* Technology topics supported by venture capital * Technology topics and associated applications by largest R&D investors * Technology topics from companies with highest company valuations	* Crunchbase/ Bloomberg / Thomson Reuters in their news sections * National VC (LL specific) * Largest R&D Investors Websites * Largest R&D Investors Company annual reports * Company valuations from crunchbase and dealroom and news reporting from Pitchbook (market news are published every day or second day)	* Technologies * NACE
Which companies are pioneers in transformative technologies in the country?	* Company types receiving venture capital or other forms of financing for transformative technologies * Companies with highest number of contracts systematically cooperating with top research institutes	* Crunchbase * National VC (LL specific) * OpenAIRE	* NACE * Company characteristics [LL specific]
Who are the companies with persistent innovative activity in the country?	Share of companies with continuous innovative activity in two consecutive periods (periods defined as: every 3 years) Innovative activity is defined according to: 1) patenting activity (at least one) OR 2) trademark applications (at least one) OR 3) design applications (at least one) OR 4) standards (at least one) OR 5) software development (at least one)	* Patstat * EUIPO (for trademarks and design) * Standards (ETSI and ISO micro data) * Github * Company websites (for future temporal analysis)	* NACE
In which technology fields do the persistent innovators invest?	Distribution of technology fields of persistent innovators per sector	* Patstat * EUIPO (for trademarks and design) * Standards (ETSI and ISO micro data) * Github	R&D and innovation topics derived by the data and not predefined
In which technology fields is the highest share of all company R&D investments? [EU & globally]	Listing of technology fields of Top R&D investors captured by different 1) STI outputs: publications, patenting, software; 2) investments: VC and 3) products/services: company websites	* OpenAIRE * Patstat * Github * Crunchbase * Websites	Technologies Scientific Disciplines

Policy question	Measurement	Data Sources	Taxonomy
In which technology fields is the country improving its Revealed Comparative Advantage (RCA)?	RCA on patents and publications: share of an economy's patents/publications in a particular technology field relative to the share of total patents/publications in that economy over time	* OpenAIRE * Patstat * Company websites	* Technologies * Scientific Disciplines

Table 2: Knowledge creation

Policy question	Measurement	Data Sources	Taxonomy
Which scientific fields demonstrate the highest growth in terms of publications/ citations/patents globally?	* Annual Growth in counts of publications/patents by scientific field * Annual Growth in average citations per publication/citations per patent by scientific field/technology	* OpenAIRE * Patstat	* Basic/applied * Interdisciplinarity/ Multidisciplinarity * Technologies (IPC) possibly to check the RISIS classification of patents
Which are the emerging interdisciplinary fields globally (i.e. integrated knowledge from different disciplines)?	* Annual Growth in counts of publications/patents of different topics of interdisciplinary publications * Annual Growth in average citations per publication/patent of different topics of interdisciplinary publications/intertechnological patents (more commonly known as converging technologies, i.e. closely integrated technologies)	* OpenAIRE * Patstat	interdisciplinarity topics
Which are the research teams in the country undertaking research in interdisciplinary fields?	* Ranking of organisations according to counts and citations of interdisciplinary publications (per year, for a period and average annual growth) * Networks of organisations undertaking research in interdisciplinary fields	* OpenAIRE (Interdisciplinary Journals) * Patstat	interdisciplinarity
Are there pockets of excellence for these research areas in the country	Organisations with strong growth and strong system linkages (composite): 1) high growth in cited publications (+10%); high growth in patents filed ; high growth in participation in RDI projects (+10%); 2) participation and involvement in DIH, Cluster organisations; Technology centres; 3) high share of public - private co-publications/co-patenting (+50%) etc.	* OpenAIRE * Patstat * CORDIS * National programmes * CMISA project (pending assessment)	Scientific Disciplines

Table 3: Knowledge Linkages and Diffusion

Policy question	Measurement	Data Sources	Taxonomy
Which knowledge diffusion channels work best in good practices per discipline at international level?	* **International co-publication**: ratio of share of cited International co-publications and share of cited publications of national publications * **Participation in conferences**: ratio of citations by publication in conference proceedings and citations per publication only in peer reviewed journals (excl. those previously in conference proceedings) by scientific area * **Open Access publications**: ratio of citations per Open Access publication and non-open access publications by scientific discipline * **Participation in EU programmes**: ratio of citations per publication from H2020/HEurope and citations per publication of non H2020/HEurope funded research	* O penAIRE * Cordis	* Scientific discipline
What are themes in common between the actors of the ecosystem? What are observed specialisation patterns? What is the evolution of topics among the different actors?	* Topic distribution between industry, science and citizens on scientific disciplines, SDGs and its evolution in time * Concentration measured with Location Quotient which measures the degree that a topic is over-represented in a particular country relative to the topic's overall distribution in Europe	* Industry: websites of actors [LL specific] * Science: OpenAIRE; H2020/HE urope * Citizens: European Media Monitoring	* Scientific discipline * SDGs
Are actors of the ecosystem collaborating? What are forms of collaboration?	* Share of Public-Private collaboration (co-patenting) in total patents * Share of Public-Private collaboration (co-publications) in total publications * Share of Public-Private collaboration (H2020/HE urope projects) in total participations	* Patstat * OpenAIRE * H2020/HE urope	* Scientific discipline * SDGs * Technologies
What are the cross sectoral or cross technological collaborations occurring and among which actors?	* network analysis of topics based on cross technological publications/projects/ patents * network analysis of topics based on cross sectoral publications/projects/ patents * network analysis of actors based on cross technological publications/projects/ patents * network analysis of actors based on cross sectoral publications/projects/ patents	* OpenAIRE * Patstat * H2020/HE urope	* Scientific discipline * SDGs * Technologies

Table 4: Guidance - Contribution to societal challenges

Policy question	Measurement	Data Sources	Taxonomy
To which global societal challenges are research groups contributing to?	* Number of Publications and patents by SDG * Distribution of SDG publications by scientific area * Distribution of SDG patents by technology field	* OpenAIRE * Patstat * H2020/HE urope	* SDGs for publications * SDGs for patents * Scientific disciplines * Technology
To which EU societal challenges are research groups contributing to?	* Number of Publications and patents by EU Mission This indicator would be living lab specific considering the missions (e.g. Adaptation to climate change including societal transformation Cancer; Climate-neutral and smart cities; Healthy oceans, seas, coastal and inland waters) * Share of publications in Missions in total publications * Share of patents in Missions in total patents	* OpenAIRE * Patstat * H2020/HE urope	* EU Missions classifier for publications * EU Missions classifier for patents * EU Missions classifier for Horizon projects
Are there specific national societal challenges?	* Topics from national work programmes and corresponding calls	* National programmes & Calls	* Societal challenges (LL specific)

Table 5: Market formation

Policy question	Measurement	Data Sources	Taxonomy
What is the role of public procurement for transformative technologies (theoretically/practically)? [living lab specific example required]	Rising topics associated to transformative technologies in TED	TED	* Taxonomies in transformative technologies * Topics in focus [living lab specific]
What is the content of policy papers/standards guiding markets?	Topics on technologies in foresight publications and standards (for instance banning of plastics leading to research for biodegradable plastics)	* Set of pre-identified foresight studies * Standards (ETSI and ISO micro data)	* Transformative technologies [living lab specific]

Table 6: Human and financial resources mobilisation

Policy question	Measurement	Data Sources	Taxonomy
What are opportunities for EU financing?	* List of topics financed through national funds which can leverage EU funding	* Living lab specific [EU Public policy documents and national policy documents]	*Topics

Policy question	Measurement	Data Sources	Taxonomy
What are opportunities for EU financing?	* List of Research teams (organisation level) financed through national funds which can leverage EU funding (using Publications of national research teams with acknowledgements to national funding matched to EU funding opportunities in TED)	* OpenAIRE * TED	*Topics
Is there sufficient S&T talent supply?	Number of skilled professionals per technology in total STEM professionals	* LinkedIn (subject to data access rights)	* ESCO * NACE * Technologies
Is there sufficient S&T talent demand?	* Number of skilled professionals demanded per technology in total enterprises * Number of skilled professionals demanded per technology in total enterprises per sector	* Cedefop (subject to the potential for text mining of Cedefop snippets)	* ESCO * NACE * Technologies
Is there a gap between supply and demand?	Derived from the analysis of S&T supply and demand	* LinkedIn * Cedefop	* ESCO * NACE * Technologies

Table 7: Creation of legitimacy/address public concerns

Policy question	Measurement	Data Sources	Taxonomy
What is the public opinion on related topics (old and new ones)	Sentiment analysis: share of positive and negative sentiment in total mentions	European Media Monitoring Parliamentary minutes	* Topics [LL specific]
What is the role of the press in topics addressed in policy objectives? Is resistance expected?	trend analysis: temporal evolution of topics in social media associated to policy objectives	European Media Monitoring	* Policy objectives * Topics [LL specific]

2.3. Measurements for Evaluation

Based on the data generated during implementation, systematic evaluations of efficiency, effectiveness and impact of the policy mix implemented are conducted to help update strategies in the next policy cycle. Policy questions become more complex: Were the targets met? How can we increase efficiency? How did we perform compared to peers? Which results are attributed to which interventions? Evaluations require significant data to check the intervention logic and run counterfactual evaluations. Combining inputs to respond to these questions have always been a challenge because of lack of data and attribution problems. It is mainly in this area where traditional indicators are insufficient that machine learning can add value.

Short listed policy questions for evaluation are described in terms of measurements, data sources and most relevant taxonomies. The list includes measurements which: 1) require AI intelligence; 2) sufficient data could be sourced (provisional assessment); 3) are technically feasible (provisional assessment) and 4) are in scope as per the proposal. The final list of measurements to be included in Intelcomp will be provided by Month 16 and are subject to the final list of sources. In terms of the unit of observation, the specific policy may be a program or a call for funding, and the unit of observation would be the outputs, outcomes or impacts related to that program or call.

Table 8: Knowledge

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
Science	How many scientific publications were published?	Number of scientific publications published	1.output	* Project publications * OpenAIRE	* Scientific disciplines * Technologies * SDGs
Science	How many scientific publications are applied research?	Share of applied research publications in total publications	1.output	* Project publications * OpenAIRE	* Applied research
Science	How many scientific publications are basic research?	Share of basic research publications in total publications	1.output	* Project publications * WoS, Scopus; OpenAIRE	* Basic research
Science	How many scientific publications are interdisciplinary?	Share of interdisciplinary scientific publications in total publications	1.output	* Project publications * WoS, Scopus; OpenAIRE	* Interdisciplinarity
Science	How many presentations were made in top scientific conferences?	Share of conference papers published in top 1% or top 10% of scientific conferences in total conference papers	2.outcome	* Project conference papers * Conference papers classification	* Scientific disciplines * Technologies * SDGs
Science	How many scientific publications were published in top 1% or top 10% of scientific journals?	Share of project scientific publications published in top 1% or top 10% of scientific journals	2.outcome	* Project publications * OpenAIRE * Journal classification	* Scientific disciplines * Technologies * SDGs
Science	How were citations in publications associated to projects compared to scientific discipline average?	Field-Weighted Citation Index of project peer reviewed publications	2.outcome	* OpenAIRE	* Scientific disciplines * Technologies * SDGs

Table 9: Diffusion

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
Science	In which ways has the diffusion of knowledge taken place?	**Towards innovation**: Number of (OS) publications (directly linked to each project result) referenced in non-patents citations of patents	2.outcome	* Project outputs * OS publications - OpenAIRE * Patstat	* NACE * Scientific areas * Technologies * SDGs
Science	In which ways has the diffusion of knowledge taken place?	**Shared knowledge**: Share of research outputs (software, datasets publications) shared through open knowledge infrastructures in total research outputs	2.outcome	* Project outputs * OpenAIRE * GitHub/GitLab	* NACE * Technologies * SDGs
Science	In which ways has the diffusion of knowledge taken place?	**Cocreation:** number and share of projects where EU citizens and end-users contribute to the co-creation of R&I content in total projects [entities are defined by the domain in focus]	2.outcome	* Project periodical/final reports	* SDGs
Science	In which ways has the diffusion of knowledge taken place at programme level?	**Open Science**: Share of open access programme research outputs (publications) actively used/cited after programme in total outputs (publications) OR : average citations of Open Science Research Outputs (i.e. publications in peer-reviewed journals and conferences)	2.outcome	* Project outputs * OpenAIRE * Open science observatory	* Scientific areas
Social	What were dissemination methods used towards the public?	**Events:** Number and share of projects with event participations by type of event in total projects	1.output	* Events in OpenAIRE	* SDGs * Policy objectives * Events typology
Social	What were dissemination methods used towards the public?	**Outreach activities**: Number and share of projects with outreach of scientific results digitally in total projects	1.output	* Newspapers * Social media * Wikipedia	* SDGs * Policy objectives
Social	In which ways has the diffusion of knowledge taken place?	**Engagement:** Number and share of projects with citizen and end-user engagement mechanisms after the project in total projects	2.outcome	* Project descriptions of activities * Open source publications * Social media * Beneficiaries websites	* SDGs * Policy objectives
Social	What were dissemination methods used towards the public?	**General public reach**: Number of people reached through dissemination activities (on topics associated to the project's expected impacts	2.outcome	* European Digital media observatory * Twitter	* SDGs * Policy objectives

Table 10: Innovation/Invention

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
Economy	Has the programme enabled the research activities to reach high technological readiness levels?	**Technology Readiness Level**: Share of outputs with TRL level higher than 6 and above compared to all projects	1.output	* Project outputs/deliverables	* TRLs
Economy	How many patents were produced (applications /grants)	**Patents**: Number of EPO patent applications and grants; Percentage share of patent grants and patent applications [Note: a patent does not signal an innovation, but an invention: i.e. an idea that is demonstrated as operational, but has not necessarily been commercialised]	1.output	* FP/National programme * Patstat	* scientific disciplines * technologies * policy objectives * SDGs
Economy	What innovations were developed?	**Innovations:** Number of innovative products, prototypes, industrial production processes, research datasets, methods, algorithms/software, business models	1.output	* Company websites of beneficiaries f * Project deliverables * Publications of participants * Openaire * Github * Open Access repositories * Classifier of types of innovations	* NACE * Innovations (LL specific) * Technologies
Economy	What were the private returns on investment?	**From innovation to market**: R&D and Innovation products and services brought to market associated to the results of the programme	2.outcome	* Company websites	* NACE * Innovations (LL specific) * Technologies
Economy	What is the uptake of project innovations in the market?	**Company uptake score:** a measure linking the innovations developed in the projects with those taken up by the company beneficiaries after the end of the project lifecycle.	3.impact	* Project deliverables *project publications *company websites	
Economy	Has the programme stimulated the development of transformative innovations necessary for the twin transition of industry?	**Transformative innovations**: Number of projects in transformative technologies; Share of projects in transformative technologies in all projects	3.impact	* Project outputs	* Transformative technologies (LL specific) *TRL
Economy	Has public procurement of innovation produced product/process innovations launched in the market	**Innovations:** Types of Innovations introduced by companies beneficiaries of public procurement (topics)	2.outcome	* National data on public procurement * EU level TED * Companies websites * Companies social media	* NACE * Technologies * SDGs

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
Social	What were the social returns on investments?	**Carbon footprint**: Types and number of innovations on reducing carbon footprint compared to all programme innovations	2.outcome	* Project deliverables	* Carbon footprint innovations (LL specific) * TRL

Table 11: Investments

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
Economy	What were the private returns on investment?	**Private funding**: Private investments raised to exploit or scale up results of the programme (level of organisation) in million euro	2.outcome	* Crunchbase * Companies social media	* NACE * Technologies * SDGs
Economy	What is the total public funding mobilised?	**Public funding:** Amount of public investment mobilised in million euros from EU and National funding	2.outcome	* Framework programme data * National public funding	* NACE * Technologies * SDGs

Table 12: jobs

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
Economy	How many new jobs were created after the project (research and beyond) within the country?	**Temporal evolution**: growth of job offers in the areas of impact	3.impact	* Cedefop online job advertisements snippets (subject to content and volume of text within the snippets published by Cedefop)	ISCED

Table 13: Gender

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
Social	What were the social returns on investments?	**Project participation**: female/male ratio	1.output	* Project participants	* Gender
Social	What were the social returns on investments?	**Research outputs**: Share of research outputs (inc. publications, datasets, software) produced by females in total research outputs	1.output	* Project participants and outputs * Female as first author	* Gender

Table 14: Objectives

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
/	Which societal challenges have been addressed?	* Share of projects by SDG * Share of project outputs by SDG	1.output	* Project outputs * Project descriptions	* SDGs classifier of outputs (publications) * SDGs classifier of projects
/	Which policy objectives have been addressed	* Share of project topics associated to policy documents * Share of project topics associated to parliament discussion minutes	2.outcome	* Overton * Parliament discussion minutes	* SDGs * Policy objectives [LL specific]

Table 15: Other

Objective	Policy question	Measurement	1.output 2.outcome 3.impact	Data Sources	Taxonomy
Leverage	What has been the leverage of national support measures for EU competitive funding?	* Share of EU funding beneficiaries who received national support prior to receiving EU funding (at organisation level) * Share of EU funded project outputs referencing project outputs of Nationally funded projects (at project level)	2.outcome	* OpenAIRE * FP funded projects * National funded projects	* Technologies, * Scientific disciplines * SDGs
Multiplication	What are the multiplication effects of each programme?	Degree of collaborations with other projects within the same programme after the programme measured by the share of co-publications between different project teams in total publications	2.outcome	* Programme/Project outputs i.e. publications	* Scientific areas * Sectors * SDGs
Exclusivity	Are we investing in topics that several other funders are interested in, or supporting a field by ourselves	Number of funders on a specific topic (crowded vs exclusive)	Not applicable	* OpenAIRE * Programme/Project outputs i.e. publications	* Technologies, * Scientific disciplines * SDGs

3. MEASUREMENTS SPECIFIC TO THE DOMAIN OF CANCER

Provisional domain specific measurements are provided for the domain of cancer, focusing on the **analysis of impact of funded research projects** and the **characterisation of 'impact pathways'**. The later focus represents the main area of interest of the cancer living lab. The climate change and AI living labs will equally provide their main areas of interest within 2022.

Three levels of needs have been identified:

1. To characterise the scientific production of funded projects (outputs)
2. To identify and characterise the outcome of funded projects (outcomes)
3. To identify and characteris e the social impact of funded projects (impacts)

In the tables below we describe a set of measurements per level of need identified. These measurements are provisional and will be updated in the course of 2022 in cooperation with the cancer living lab. The final measurements to be implemented in IntelComp will depend on the formulation of narratives on impact pathways defined by the Cancer living lab.

To facilitate understanding of the tables below, we describe shortly the distinction between outputs, outcomes and impacts.

Outputs are the tangibles or intangibles that an organisation or project produces. These could be completed services, products, interventions or other 'deliverables'. They should act to 'spark change' or act as the catalyst for your identified outcomes. They are normally fairly easy to measure and can often be quantified e.g. how many do we do or the number of outputs you create. Outputs in the cancer domain would be e.g. research results, clinical trials.

Outcomes are the intended short to medium effects or the 'step changes', which need to occur in order to achieve your long term or ultimate goal. If you are trying to facilitate change within an individual, you can think of this as the journey your beneficiary needs to go on to reach the change you have identified. They are often more difficult to measure than outputs, as they can frequently relate to an individual's perceptions, emotions or other internal state. So drugs, clinical guidelines and new technologies and treatments are outcomes because we do not know whether they will really reduce mortality rates and improve health.

Impact is your long-term goal or ultimate objective. If you are talking about your organisation's impact, it will likely be closely linked to your mission statement or vision statement. Whether for your organisation or a project, your impact(s) will be what you are ultimately trying to achieve. If you work with individuals, it will be the end state you would like your beneficiary to be in. Your impact should be achieved, as a result of your outcomes. If your outcomes are the journey your beneficiary will go on, your impact is the end destination. Your impact will often be the most difficult to measure, and since it will frequently occur over a long period of time with other influencing factors, it can be challenging to identify whether any changes you do observe are a result of your efforts or something else (attributing causality). Impacts are improved quality of life of individuals with cancer, reduced mortality rates from cancer.

Table 16: Objectives

Objectives	Source	Measurement	Taxonomies
Framework Programmes (H2020 and HEurope)	Cordis	* Topics on expected impacts	* Topics (e.g. tobacco, alcohol, food, pollution) * Technologies and treatments (e.g. genetics, biotherapies, predictive medicine, e-health) * Stages of patient care 1) prevention; 2) early detection; 3) diagnosis and treatment; and 4) quality of life for cancer patients and survivors
National programmes	LL specific		

Table 17: inputs

Inputs	Source	Measurement	Taxonomies
Framework Programmes (H2020 and HEurope)	Cordis	* Funding in million Euro	* Topics (e.g. tobacco, alcohol, food, pollution) * Technologies and treatments (e.g. genetics, biotherapies, predictive medicine, e-health) * Stages of patient care 1) prevention; 2) early detection; 3) diagnosis and treatment; and 4) quality of life for cancer patients and survivors * Beneficiaries (types) * Funders * Applicants (types)
National programmes	LL specific		

Table 18: Outputs (first level needs)

Project outputs	Source	Measurement	Taxonomies
Scientific publications	OpenAIRE	* Number of scientific publications published during the project * Share of scientific publications by taxonomy	* International Classification of Diseases 11th Revision (ICD11) * Orphanet classification * Basic/Clinical * National/International * Scientific discipline * Topics: tobacco, alcohol, food pollution * Technologies and treatments (e.g. genetics, biotherapies, predictive medicine, e-health) * Stages of patient care 1) prevention; 2) early detection; 3) diagnosis and treatment; and 4) quality of life for cancer patients and survivors * Public-Private co-publications
Patents	* Patstat * Programme data	* Number of patents filed during the project * Share of patents filed by taxonomy	* International Classification of Diseases 11th Revision (ICD11) * Orphanet classification * Technologies and treatments (e.g. genetics, biotherapies, predictive medicine, e-health) * Public-Private co-patenting

Table 19: Scientific, medical, and social outcomes (second level needs)

Project outcomes		Sources	Measurement	Taxonomies
Science	Publications Patents citations could feature here	OpenAIRE	* Number of scientific publications published after the project associated to the publications funded during the project * Share of scientific publications published after the project by taxonomy * Field -Weighted Citation Index of project peer reviewed publications	* International Classification of Diseases 11th Revision (ICD11) * Orphanet classification * Basic/Clinical * National/International * Scientific discipline * Topics (e.g. tobacco, alcohol, food, pollution) * Technologies and treatments (e.g. genetics, biotherapies, predictive medicine, e-health) * Stages of patient care 1) prevention; 2) early detection; 3) diagnosis and treatment; and 4)

Project outcomes		Sources	Measurement	Taxonomies
				quality of life for cancer patients and survivors * Public-Private co-publications
	Health data	OpenAIRE	* Number of data objects produced * Number of data objects consumed	* International Classification of Diseases 11th Revision (ICD11) * Orphanet classification * Topics (e.g. tobacco, alcohol, food, pollution)
Medical	Clinical Trials	*Clinicaltrials.gov	* Number clinical trials linked to projects * Type of trial * Phase it ended * Age group targeted * Number of clinical trial references * Citations in same disease or other - cross over * Number of hops before a successful clinical trial * Phase 5 + (new or repurposed) drug or clinical guidelines)	* International Classification of Diseases 11th Revision (ICD11) * Orphanet classification.
	Drugs	* Drugbank	* Number of new drugs linked to projects (through clinical trials) *Number of drug repurposing linked to projects (through clinical trials)	* International Classification of Diseases 11th Revision (ICD11) * Orphanet classification
	Clinical Guidelines	*OpenAIRE *PubMed	* Number of clinical guidelines linked to projects (through clinical trials)	
	New technologies and treatments	* Project deliverables * Patents * Beneficiary websites	* New technologies and treatments from project deliverables, patents and beneficiary websites linked to the projects	* Technologies and treatments (e.g. genetics, biotherapies, predictive medicine, e-health) * TRL
	Clinical guidelines	* PubMed * openAIRE	Clinical guidelines linked to projects	* International Classification of Diseases 11th Revision (ICD11) * Orphanet classification
Social	Social media buzz	* Twitter * European Media Monitor	Reach in Tweets of funded participants related to the outputs/outcomes of the funded project	* Topics (e.g. tobacco, alcohol, food, pollution) * Technologies and treatments (e.g. genetics, biotherapies, predictive medicine, e-health)
	Position papers	* Open Public consultations	Share of positive/negative topics (Sentiment analysis of position papers)	* Technologies and treatments (e.g. genetics, biotherapies, predictive medicine, e-health) * Stages of patient care 1) prevention; 2) early detection; 3) diagnosis and treatment; and 4) quality of life for cancer patients and survivors * Topics: tobacco, alcohol, food pollution

Table 20: Scientific, medical, and social impacts

Project impacts	Sources	Measurement	Taxonomies
Science	OpenAIRE	**World class science**: Number and share of peer reviewed publications from projects that are core contribution to scientific fields in total peer reviewed publications	* Scientific discipline * Topics (e.g. tobacco, alcohol, food, pollution) * International Classification of Diseases 11th Revision (ICD11) * Orphanet classification

Project impacts	Sources	Measurement	Taxonomies
		core contribution: citing top 1% publications in the corresponding subject area	
Medical	* PubMed * openAIRE	**Uptake from practitioners:** Clinical guidelines	* International Classification of Diseases 11th Revision (ICD11) * Orphanet classification
Social	Public health	**Contribution to policy making/Legislation impacting public health:** Share of project topics associated to policy documents and/or Share of project topics associated to parliament discussion minutes	* Topics (e.g. tobacco, alcohol, food, pollution)

4. DATA SOURCES

To address the diverse policy aspects comprised in the scope of IntelComp, we consider a broad variety of potential sources to be ingested and stored in the IntelComp Data Space. The assessment of the sources' feasibility and relevance is made based on six criteria:

1. Text mining potential: The source provides or contains text documents or text sections that can be used for text mining processes. Text mining potential is a qualifier criterion, i.e. if not fulfilled the source cannot be integrated into IntelComp
2. Potential for temporal data and time series data analyses: Sources can be analysed in past and future moments in time allowing time series analyses. Two different issues are important to distinguish: 1) are data sources periodically updated (necessary for future sustainability of the source in IntelComp), and 2) do we have time information for the items in the data source ? (necessary for time analysis)
3. Taxonomy: There are different classifications for the data provided by each source identified. Additionally, we identified classifiers that we intend to use to sort the data (in addition to those already available in the dataset). Both types of classifiers are listed in the tables below under taxonomy
4. Representativeness: The data is derived from the whole population of interest or a representative sample of it. At this stage, this criterion is assessed at a high-level and will be further investigated as well as methods to address biases
5. Open access: The data can be accessed and extracted free of charge. Exceptions apply and are being considered in the framework of the domain specific needs assessment
6. Availability of data for main competitors: Main competitors of the EU are defined as the USA, Japan, South Korea and China. This criterion assesses whether the source also provides data for the cited countries, to allow international comparisons or or homogeneous data from these countries could be gathered from alternative sources

The full list of potential sources under consideration is available in Appendix I – Long list of sources considered, while the current section presents the most promising ones, i.e. the sources that match best the established criteria and that are the most versatile in terms of addressing multiple policy questions. Sources identified belong to various typologies and are sorted accordingly in the tables below.

Table 21: Science & Innovation

Source	Description	Suitability	Taxonomy	Relevant policy questions
OpenAire/ Semantic Scholar	Open access publications platform, with 129M deduplicated publications available	High	Scientific disciplines SDGs Technologies	In which ways has the diffusion of knowledge taken place? In which ways has the diffusion of knowledge taken place at programme level? What was the contribution of the publications to the scientific field? How many scientific publications were published in top 1% or top 10% of scientific journals per discipline? How were citations in publications associated to projects compared to scientific discipline average? How many scientific publications are applied/basic research? How many scientific publications are interdisciplinary? How many scientific publications were published?

Source	Description	Suitability	Taxonomy	Relevant policy questions
				What has been the leverage of national support measures for EU competitive funding? How many people were trained as researchers?
Cordis	Research activities and outputs in the EU framework programmes (public investment). The data available from Horizon 2020 and FP7 is already ingested in IntelComp, through Corpus Viewer. information on countries outside the EU is only available regarding their involvement in H2020 partnerships.	High	Scientific disciplines SDGs Technologies Taxonomy of innovations TRL	Has the programme stimulated the development of transformative innovation? What was the uptake of scientific results in patents? What were the social returns on investments? Has the programme enabled the research activities to reach high technological readiness levels? How many patents were produced (applications/grants)? What innovations were developed? What is the total public funding mobilised? In which ways has the diffusion of knowledge taken place at programme level? How many people were trained as technicians? As researchers? How many presentations were made in top scientific conferences? How many scientific publications are applied/basic research? How many scientific publications are interdisciplinary? How many scientific publications were published? What are the multiplication effects of each programme? What were dissemination methods used towards the public? Which societal challenges have been addressed?
Patstat	Online inventory of patents with complete coverage of patents (more than 100 million patent documents)	High	IPC Technologies SDGs TRL NACE Policy objectives	What is the generation of patentable (appropriable) knowledge? In which ways has the diffusion of knowledge taken place? What was the uptake of scientific results in patents? How many patents were produced (applications/grants)?
Github	Code repositories used by 4+ million companies	High	Technologies	What innovations were developed? In which ways has the diffusion of knowledge taken place at programme level? What were dissemination methods used towards the public?

Table 22: Company websites and financials

Source	Description	Suitability	Taxonomy	Relevant policy questions
Innovative companies' websites	Own compilation of innovative companies from different sources organised by different types of companies. The listing of company websites is expected to rely on several companies repositories: 1) Crunchbase for large companies and tech start-ups; 2) the JRC Scoreboard of the largest R&D innovators (top 2500 worldwide and top 1000 in EU); 3) Bloomberg, Dealroom and/or Pitchbook; 4) Patstat i.e. websites of companies with large number of patents, 5) websites of Unicorns; 6) Framework Programmes for beneficiary companies active in FP7, H2020 and HE urope projects, and 7) the Living Labs will provide insights on the main local innovators.	Medium	NACE Technologies Taxonomy of innovations Policy objectives SDG	What was the contribution of innovations to turnover, profits, market shares? What innovations were developed by companies? What innovations were developed in the project? What is the total funding mobilis ed?

Table 23: Public and private investment

Source	Description	Suitability	Taxonomy	Relevant policy questions
Crunchbase/ Pitchbook	Inventory of worldwide companies with comprehensive information on their funding rounds (private investment) and news items	Medium	Company size Company establishment and funding Industries Technologies NACE	What were the private returns on investment?

Table 24: Legal and policy documents

Source	Description	Suitability	Taxonomy	Relevant policy questions
Eurlex	Online database of European Union treaties, legal acts, consolidated texts, international agreements, etc.	High	SDG Policy objectives Strategic pillars Sectors Technologies Scientific areas	Which policy objectives have been addressed? Are currently available strategies/policies coherent?
Overton	Index of policy literature with comprehensive publication information	Medium		
Own policy documents database	A compilation of various sources: 1) European Parliament (different committees) and the publications from all EU entities and agencies; 2) SIPER and Fteval initiatives; 3) online repositories of EU/OECD countries of R&I policy and technology evaluations ; 4) foresight studies, from the European Commission, the Competence centre on foresight and the OECD strategic foresight work	Medium		
Foresight studies	Compilation of studies shaping R&D future orientations from different institutions	High		

Table 25: Public procurement

Source	Description	Suitability	Taxonomy	Relevant policy questions
TED	Online database of active and past public procurement offers from local, national and European authorities for services, works and supplies. TED has 4,390,327 tenders registered, providing a comprehensive, if not exhaustive, overview of procurement by public authorities in Europe. An expected obstacle is the difficulty to link procurement offers with the technology taxonomy.	Medium	Contract characteristics Technologies Sectors	What are opportunities for EU financing? What is the role of public procurement for transformative technologies (theoretically/ practically)?

Table 26: Social Media

Source	Description	Suitability	Taxonomy	Relevant policy questions
European Media Monitor	The EU Competence Centre on Text Mining and Analysis extracts information from online data, including traditional or social media, or from large public or proprietary document sets	Low		In which ways has the diffusion of knowledge taken place at programme level? What were dissemination methods used towards the public?
Twitter	Twitter activity (tweets) of pre-identified actors: innovative companies, FP projects, beneficiaries. Tweets and their associated reach are considered as dissemination activities and citizen engagement mechanisms.	Medium	Technologies Sectors	Has public procurement of innovation produced product/process innovations launched in the market (lead markets) In which ways has the diffusion of knowledge taken place at programme level? What were dissemination methods used towards the public?

Table 27: Skills demand and supply

Source	Description	Suitability	Taxonomy	Relevant policy questions
LinkedIn[3]	Public profiles of professionals associated to specific skills or to FP programmes' positions	Pending	Industries Skills (ESCO) Scientific disciplines (FOS2) Jobs typology Detailed taxonomies developed with Living Labs	How many new jobs were created after the project (research and beyond) within the country? How many new jobs were created for researchers during the project? What was total employment created? What was the career development of participating researchers?
Euraxess	European Commission's job offers and funding opportunities platform for researchers	Medium		

[3] LinkedIn is considered as the most promising source for skills demand and supply. Access to LinkedIn public profiles is however not confirmed yet.

5. TOOLS FOR STI POLICY ACTORS

IntelComp integrates different underlying technologies capable of providing evidence to answer policy questions relevant for all phases of the policy cycle, addressing the needs of STI policy actors. IntelComp builds upon the components and services from Corpus Viewer and Data4Impact adding newly developed components, exploiting both structured metadata available for the datasets and the output of AI pipelines that build on unstructured text. All these components and services, together with the necessary visualisations, are grouped into four main IntelComp tools:

1. **STI Viewer**: This tool targets mainly the Policy makers and Public Administrations. It offers basic and advanced visualisations based on the back-end components for the analysis of both structured and unstructured data. In addition to STI Viewer, advanced users from this target group, such as policy analysts, will also have the possibility to analyse their own datasets using IntelComp integrated components such as NLP pipelines, machine translation, etc. Also, they can carry out inter-corpus comparisons against publicly available datasets in the IntelComp Data Lake. This tool answers a wider range of policy questions described in greater detail in Sections 2 and 3.
2. **Interactive Model Trainer**: It is a tool provided for technical and advanced users from STI Policy Makers and Public Administrations. The Interactive Model Trainer allows this type of users to use the back-office IntelComp architecture and components to train their own models: either topic models or classification models. It also allows them to play an active role in the creation and validation of these models to ensure "human-in-the-loop" principles and unbiased data selection. In this way, they could customise their analysis, comparisons and visualisations according to the newly trained models. From a technical point of view, this tool could answer questions such as: "How can we make use of IntelComp components to train our own models?" or "How can we validate and interact in the process of creation and training?".
3. **Evaluation Workbench:** This tool targets Public Administrations and Funding Entities to assist in the evaluation process of STI proposals. The Evaluation Workbench will assist in different tasks such as: Identifying possible evaluators whose expertise and profile match the thematic area under evaluation; contextualising the proposal within the STI information space by comparing to existing patents, publications & funded projects, classifying proposals automatically according to available taxonomies and, finally, checking if similar proposals have been already funded. The Evaluation Workbench will assist in answering questions such as: "Who are the experts in a specific area that can act as evaluators?" or "How could proposals be classified according to available taxonomies?" or "Has this proposal or a similar one been evaluated / funded before?".
4. **STI Participation Portal**: The tool targets stakeholders from academia, industry as well as citizens. It allows stakeholders to visualise the general STI panorama and its evolution across the different domains at the national, regional or institutional level. It also links this information with trending topics and provides some insights on the lag between the STI outcomes and the social media impact. Moreover, stakeholders will also be able to interact and share their views and feedback through the Participation Mailbox to guarantee channels for an ongoing co-creation process. In this sense, the Participation Portal will assist in providing answers to the following questions, among others: "What are the thematic domains that have been funded? or In which areas were the STI public funds spent?" or "Which are the emerging areas?" or "Which entities are the most active in the STI panorama?" or "Where do we stand nationally or regionally with respect to other countries, regions, etc.?".

The tools will provide different visualisations and services according to the users' profiles. Examples include:

- "Enriched" business intelligence panels with topic data and graph exploration
- Graphs for recursive information navigation (for large corpora or multi-corpora logical datasets)
- Topic models exploration tools: for static and dynamic models
- Inter-corpus comparison tools
- Bipartite graphs
- Other services supported by back office elements, using unstructured text as input (e.g., classification services, topic inference, machine translation, etc).

6. SERVICES

The identified measurements described in sections 2 and 3 showcase the needs of STI policy makers and public administrators in both structured and unstructured data. IntelComp's technological proposal consists of the development of a platform that brings together a series of data analysis tools to provide the evidence demanded by the proposed policy-making framework. In a very global and probably over-simplified way, the procedure will involve three phases linked to corresponding IntelComp work packages: 1) data acquisition and homogenization, 2) enrichment of the datasets applying state-of-the-art AI and NLP techniques, and 3) visualis ation of results. Users will be involved in all the steps of this procedure through the co-creation activities that will be carried out in the living labs. From an information enrichment point of view, and with the aim of providing data-based evidence that goes beyond traditional metadata-based analysis, IntelComp focuses on applying NLP techniques to unveil relevant information connected to the target measurements. When necessary, we also consider enriching the available information by extracting additional information from other data sources, or using the Internet as a Data Source (e.g., for extracting the quartile of publications, etc), but the main focus of the project and what we will describe in this section is the application of AI pipelines.

In addition to other NLP auxiliary services, the five main services that IntelComp relies on for data enrichment are the following:

1. Service for domain-related subcorpus generation
2. Classification service
3. Advanced topic modelling service
4. Topic-based time analysis service
5. Graph-based impact analysis

Below we briefly describe the listed services, as well as their connection with the information demands and objective measures identified within the framework developed for evidence-based policy making.

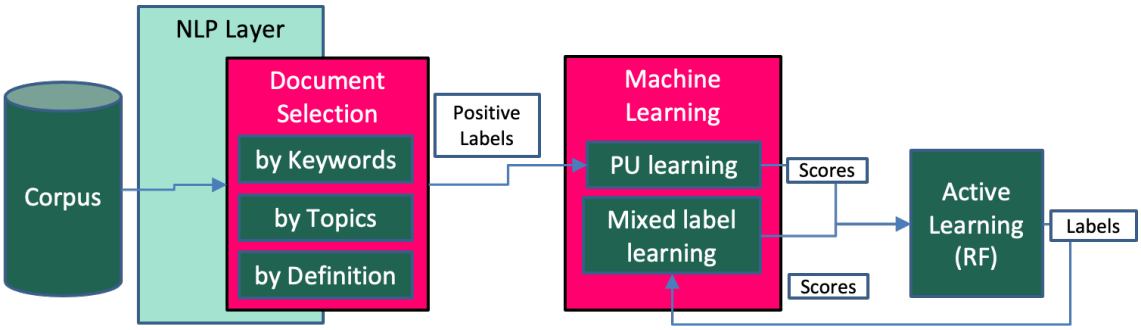
6.1. Service for domain-related subcorpus generation

In IntelComp we build domain agnostic tools, but we are aware that on most occasions the platform will be applied to analyse data of a specific domain. This is indeed the case for the three living labs considered in the project.

Then, and since many of the datasets in the data lake are very wide in scope, we first need to identify the documents that are relevant for a specific domain, which is a question that can in many cases not be answered in a completely objective manner, e.g., do we care just about core AI papers, or do we also wish to include application-related works?

For this reason, we envision a human-in-the-loop-based service for identifying documents relevant for a particular domain using a relevance feedback mechanism. The basic structure of the processing pipeline is shown in the figure below.

Figure 1: basic structure of the processing pipeline to identify relevant documents



The main components in the process are the following:

- Data source: a corpus of STI documents, with some metadata. For some components of the process, it will be assumed that the corpus has been processed with NLP tools and by topic modelling algorithms (see Subsection 5.3).
- Initial document selection. A set of tools that facilitate the selection of a subset of documents from the domain specified by the user. In particular, the user will be allowed to select documents from the subcorpus in three ways:
  - By keywords: the user provides a list of keywords and a set of filters are applied to select a subset of documents highly scored with respect to the given keywords.
  - By topics: the user selects one or several topics from those inferred by the topic modelling service, maybe specifying a weight or importance value of each topic. A set of filters is applied to select a subset of documents highly scored with respect to the selected topics.
  - By definition: the user provides a label identifying a specific domain. Then, a zero-shot classifier is applied to select documents aligned with the label name. To do so, the classifier might use documents defining the category specified by the label (e.g., using related articles from wikipedia).
- Machine learning (classification algorithm): after the document selection, a subset of documents from the target domain is available and used as the training set for a learning

algorithm. Since the training set contains documents from the positive class only, standard supervised learning algorithms are not feasible, and PU (Positive-Unlabeled) learning models will be applied.

- Active learning. The active learning module provides a relevance-feedback mechanism to include a human in the loop. The user will be provided with tools to label specific documents from the positive and negative classes. This will be useful to refine the learning algorithm with a training set containing both positive and negative samples.

6.2. Classification Service

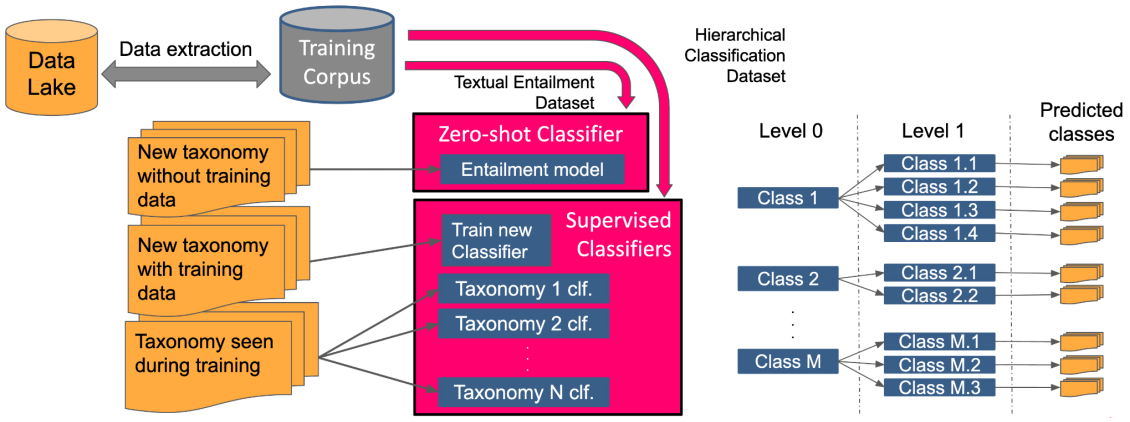
Some of the requested measurements, as well as comparative analysis, need the joint analysis of several datasets. Experts find convenient some of the best known taxonomies which are connected to their intuition, but the issue here is that different datasets include heterogeneous taxonomies, which makes the joint analysis difficult. A second issue is that in many cases labelling is carried out by the author or evaluators of the document (paper, project proposal, etc), which introduces biases.

The classification service aims at producing labels associated with existing taxonomies, so that the comparison can be carried out along these dimensions. It will allow labelling a dataset according to a taxonomy which is not available for that dataset, or even relabelling documents that have not been correctly labelled. The output of the classifiers will allow the end user to objectively compare the similarity between documents from different datasets.

In order to do so, we will train supervised classifiers whenever possible. For this to be done it is necessary to have labelled data with several examples of documents that belong to each target class, so that the classifier can learn to predict them accurately. In the worst case scenario where there is no training data available, the service will resort to a zero-shot text classification approach, even though its performance is known to be far from the state-of-the-art.

The basic structure of the classification service is shown in the figure below.

Figure 2: Basic structure of the classification service



The main components are the following:

The input to the service will be the data to be classified together with the desired taxonomy. At this point there are three possible scenarios:

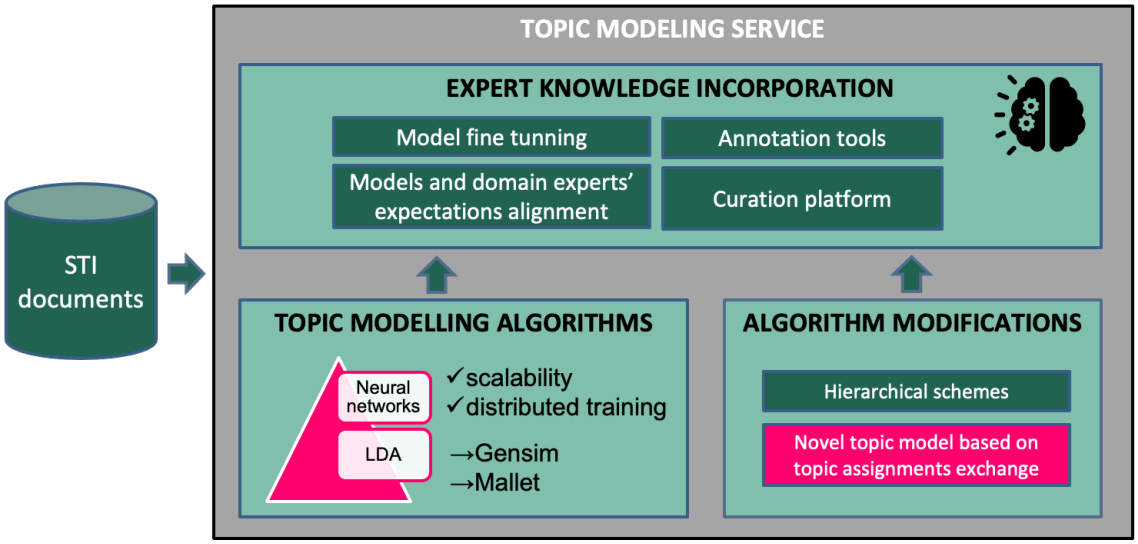
- Taxonomy for which a classifier is already available: In this situation an already trained classifier will be used. Depending on the taxonomy, the classifier will be a single model or will be composed of a cascade of models arranged in a hierarchical manner (as can be seen in the right part of the figure).
- New taxonomy with training data: In this situation the classification service will train a new classifier using the provided supervised data. This pipeline will also allow the user to arrange a set of models in a hierarchical structure. Note that to train a new classifier from scratch it is required to have a reasonably large amount of data for each label, the more the better.
- New taxonomy without training data: This situation should be avoided at all costs, since trying to classify new documents in a never seen taxonomy without any training example is clearly a hard problem, especially in large-scale classification scenarios. However, when this situation cannot be avoided the system will resort to a zero-shot classifier. The idea of this classifier is to use an entailment method to compare the embeddings of the documents to the embeddings of the labels (or a definition of the labels extracted from a database like wikipedia). These models can only be expected to perform reasonably well with the shallow labels of the taxonomies.

6.3. Advanced Topic Modelling Service

Topic modelling will be used to provide an additional dimension for analysis and comparison with respect to existing taxonomies. This makes it feasible to analyse data with different levels of granularity and detect niches that require specific consideration.

A pipeline of the processes involved in the topic modelling service is shown in the figure. The topic modelling algorithms are fed with a corpus of STI documents, maybe after some pre-processing using auxiliary NLP pipelines (lemmatization, stopword removal, N-gram identification, etc).

Figure 3: Topic modelling pipeline



The service includes standard topic modelling services based on efficient and scalable implementations of the Latent Dirichlet Allocation algorithm and, also, algorithms based on neural networks. The service is expected to provide models based on corpora with tens of millions of documents. A topic model will produce two kinds of outputs:

- The topic model itself, which identifies and provides a characterisation of the most relevant themes for a particular dataset
- The assignment of documents to the topics in the model

In this way, we can automatically detect the main topics for a given dataset and include this information for the analysis by the experts. Furthermore, since the number of topics can be varied according to experts' preferences, topic modelling offers a way to analyse data with different granularity levels.

With respect to existing fully automatic topic modelling implementations, IntelComp advanced topic modelling will bring modifications to satisfy the requirements from policy analysts: more stable topics, better alignment with other available metadata, automatic labelling of topics, and the introduction of a set of edition capabilities. This will be provided to the experts inside a tool for model training, to facilitate the construction of high-quality models that are aligned with expert intuition.

The service will also implement hierarchical models that allow providing a topic description with different levels of resolution. The higher level topics provide a broad view description of the corpus, while lower levels provide information about the internal structure of topics as a collection of subtopics.

In IntelComp, the information obtained from the topic models may be exploited jointly with that obtained through the classification modules or taxonomic information available directly for some of the data sources used. That is, the user will be able to simultaneously view the available taxonomies, those inferred through the classification modules, and the topics calculated, or a subset of these, as well as study the relationships between them and other available metadata (eg, temporal or geographic information). In this sense, the information on topics adds value compared to the available taxonomies since, for example:

- allows to analyse the data with different levels of granularity, e.g., by analysing specific topics included within a same category of the taxonomy
- as it is a completely automatic approach, it allows identifying novel topics, not included in a specific taxonomy

- allows a soft assignment of documents in different topics

6.4. Topic-based time analysis service

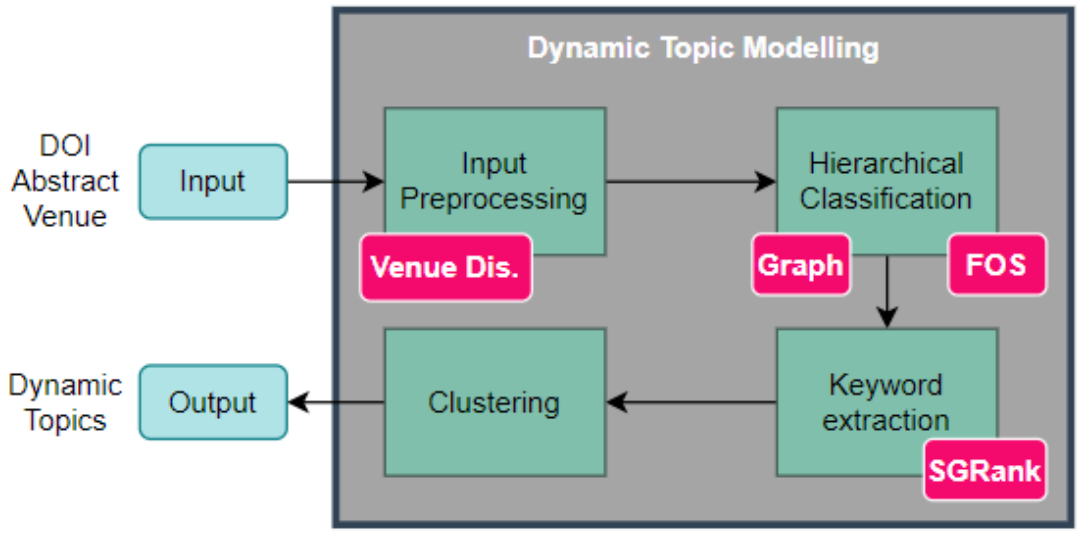
The Dynamic Topic Modelling service assigns one or more topics to a publication using the title, abstract, and venue of the publication. A pipeline of the processes involved is shown in the figure below. Given DOI-venue-abstract triplets collected from scientific articles, input pre-processing transforms the textual data into a useful input for the next parts of the pipeline. At the same

time a disambiguation rule is applied on the name of the venue that the publications were published in.

Hierarchical Classification is applied to detect the fields of science (FOS) of the publication. An extended version of the Frascati manual developed by the Organisation for Economic Co-operation and Development (OECD) is used to detect fields of science in different granularities. The simultaneous hierarchical classification allows a dynamic assignment of topics across the scientific domains rather than assigning a general topic from a universally trained topic model. Graph Analysis is applied to detect sets of venues that form the topics of a specific field of science. The graph is developed using the publication venues and their connections through publication citations from millions of publications.

Further detailed classification is provided using keyword extraction per field-of-study and grouping of keywords allows detection of more fine-grained topics as dynamic topics formed in a specific field of science. Keyword extraction can detect more subtle topics addressed in each publication separately.

Figure 4: Dynamic topic modelling pipeline



Overall, the pipeline consists of pre-trained modules (disambiguation, graphs, KW extraction) and can be applied in collections of publications to detect dynamic topics. The Dynamic Topic Analysis extends previously introduced topic modelling methods by correlating topics to a scientific classification schema with different granularities of detail.

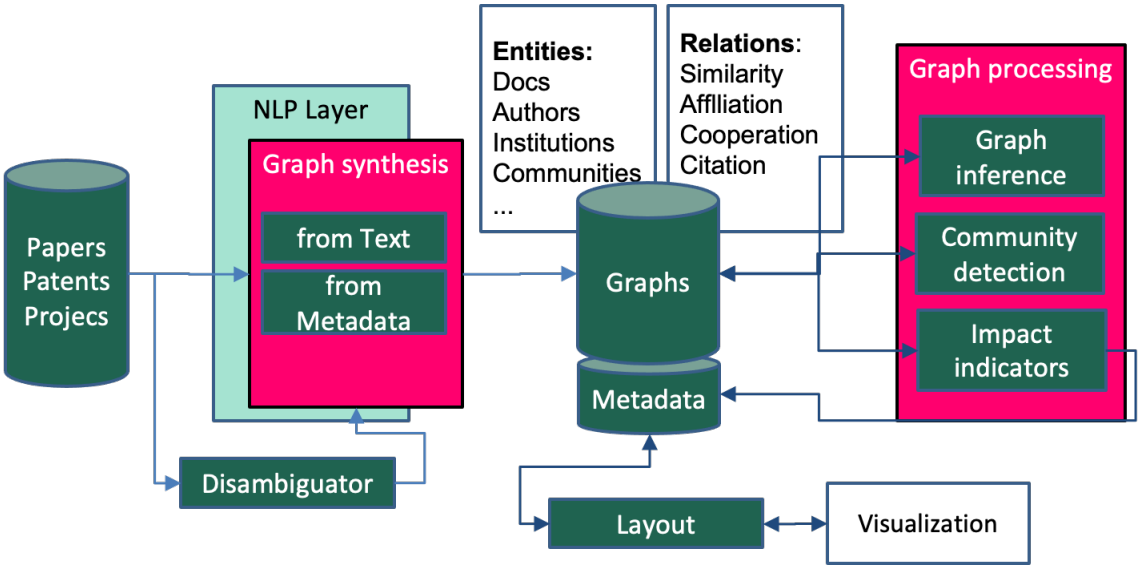
Further, given the dynamic topics and using the year of publication of the input data, we are able to create per-year and per-topic collections and conduct different types of time analysis, such as, but not limited to: detection of emerging topics and lead-lag analysis.

6.5. Graph-based impact analysis

A set of graph analysis tools will be incorporated in IntelComp to facilitate the analysis of the impact of research agents (authors, inventors, institutions, publications) in their respective fields.

To do so, different types of graphs will be generated from the text corpora and the metadata contained in the STI data sources (patents, publications, funding applications, etc.). The general structure of the processing pipelines is illustrated in the Figure below.

Figure 5: Structure of the processing pipelines for graph-based impact analysis



Graphs may be used to encode and represent different types of relations between documents (similarities, citations, co-citations), authors (cooperation, semantic similarity, citation), or institutions (cooperation, etc). Bipartite graphs will be used to connect documents to their authors and their funding institutions, or their clusters or communities

Graph inference methods and community detection algorithms can be applied to identify the cluster structure of documents and agents. This, in combination with graph metrics to analyse the impact or the relevance of nodes in graphs, can be used to extract information about the particular role of each member of the network in its community, or the impact of a specific publication or author in the advancement of a research field.

7. GAP ANALYSIS

A gap analysis is performed to compare the domain agnostic conceptual framework i.e., the identified needs of STI policy stakeholders to the provisional implementation plan in IntelComp. In other words the gap analysis identifies the policy questions which are not possible to address with the tools of IntelComp.

Table 28: Typologies of Policy questions not addressable in intelcomp

Agenda Setting Evaluation	Policy Rationale	Policy question
I. Policy questions which require traditional data		
Evaluation	Skills	How many people were trained as technicians? How many people were trained as researchers?
Evaluation	Taxes	How much tax income was generated?
II. Policy questions which are best analysed through qualitative methods		
Agenda setting	Market formation	What is the regulation globally for these technologies?
Agenda setting	Legitimacy	What are the reasons justifying the political choices made?
III. Policy questions for which no data source is available		
Agenda setting	Knowledge diffusion	Which networks e.g., clusters, hubs, intermediaries operate nationally per discipline?
Agenda setting	Resources mobilisation	Which financial resources were most effectively used in the previous cycle (evidence from the evaluation part of the cycle)? [exclude as question] What is the size of resources needed to become competitive in each emerging technology? What type of resources can be mobilised outside the national public funding (EU, foundations)?
Evaluation	Innovation	How many patents were licensed? How many patents were used in-house? How much royalties did patents produce?
Evaluation	Markets	Has public procurement of innovation created lead markets?
Evaluation	Markets	Has the regulation adopted facilitated the creation/access to new markets?
IV. Policy questions which require statistical analysis or other methods		
Evaluation	Innovation	What was the contribution of innovations to turnover, profits, market shares?
Evaluation	Jobs	What was total employment created?
Evaluation	Cost effectiveness	What was the cost per publication? At scientific discipline level?
Evaluation	Cost effectiveness	What was the cost per patent? At scientific discipline level?
Evaluation	Cost effectiveness	What is the cost benefit ratio of each programme?
V. Policy questions which can only marginally inform the policy question		
Agenda setting	Entrepreneurial activity	Are scale ups leaving the country?
Evaluation	Jobs	How many new jobs were created for researchers during the project?

REFERENCES

- Eurostat. 2020. European Statistical System handbook for quality and metadata reports. Available at: https://ec.europa.eu/eurostat/documents/3859598/10501168/KS-GQ-19-006-EN-N.pdf/bf98fd32-f17c-31e2-8c7f-ad41eca91783?t=1583397712000
- European Commission. 2020. Experimental big data statistics. Available at: https://ec.europa.eu/eurostat/cros/content/Experimental_big_data_statistics_en
- Deloitte. 2016. Big data analytics for policy making. Available at: https://joinup.ec.europa.eu/sites/default/files/document/2016-07/dg_digit_study_big_data_analytics_for_policy_making.pdf
- Eurostat. 2017. Towards a harmonised methodology for statistical indicators. Available at: https://ec.europa.eu/eurostat/documents/3859598/8071770/KS-GQ-17-007-EN-N.pdf/7d34c904-2d07-4e71-bd6f-8fe9ee373b60

APPENDIX I – LONG LIST OF SOURCES CONSIDERED

Typology	Source label	Short description
Company financials/websites	Open corporates	Open database of companies (200 million companies)
Company financials/websites	European e-justice Business Registers	Compilation of business registers in EU, Iceland, Liechtenstein and Norway
Company financials/ websites	RISIS FirmReg	A reference register on private actors, combining the firms from 3 firm datasets (CIB, VICO and Cheetah) with their linkages, enabling actor-level harmonisation at European level. Currently at the prototype stage.
Skills demand	Euraxess	European Commission's job offers and funding opportunities platform for researchers
Skills demand	Cedefop	Toolkit of sources of labour market intelligent, with complete economy coverage
Skills supply	LinkedIn	Public profiles of professionals associated to specific skills or to FP programmes' positions
Innovation	Patstat	Online inventory of patents with complete coverage of patents (more than 100 million patent documents)
Innovation	ETSI - standards	Online IPR database (14826 standards from 352 companies) for the telecommunication sector, hence no coverage of the whole economy
Innovation	ISO micro data - standards	Complete database for European standards, but not informative on other standards
Innovation	Github	Code repositories used by 4+ million companies. More than 200 million codes available Country of repositories to be retrieved from the contributors
Innovation	EUIPO trademarks and design	Inventory of trademarks and designs covering 40 million trademarks and 9 million designs, and used by 200 countries
Science	OpenAire	Open access publications platforms (with 128M deduplicated publications)
Science	Cordis	Research activities and outputs in the frame of H2020 programmes
Investments priv	Crunchbase	Inventory of worldwide companies with comprehensive information on their funding rounds
Investments priv	National VC	Own compilation of venture capital websites
Investments priv	National Investment Laws	Living Lab specific as heterogeneous across countries, e.g. in Greece all investments supported by the State are public
Legislation	EURLEX	Online database of European Union treaties, legal acts, consolidated texts, international agreements, etc.
Policy documents	Overton	Index of policy literature with comprehensive publication information
Policy documents	Parliament discussion minutes	Minutes of European Parliament minutes (different committees)
Policy documents	Government sources	National government's policy documents based on the compilation of national governments' sources, Living Lab specific as heterogeneous across countries
Policy documents	EU publications	Repository of publications by all EU entities and agencies
Policy documents (evaluations and IAs)	SIPER	Repository of research and innovation policy evaluations, EU and OECD countries
Policy documents (evaluations and IAs)	Fteval	Repository of the Austrian Platform for Research and Technology Evaluation. Includes mainly European countries' evaluations.
Foresight studies	EC; Competence centre on foresight; OECD strategic foresight	Compilation of studies shaping R&D future orientations from different institutions (EC; Competence centre on foresight; OECD strategic foresight)
Procurement	TED	Online database of active and past public procurement offers from local, national and European authorities for services, works and supplies. 4,390,327 tenders registered

Typology	Source label	Short description
Procurement	National data on public procurement	Own compilation of national procurement websites, Living Labs specific as heterogeneous across countries
Social media	European Media Monitoring	The EU Competence Centre on Text Mining and Analysis extracts information from online data, including traditional or social media, or from large public or proprietary document sets
Social media	Twitter	Twitter activity (tweets) of pre-identified actors: innovative companies, FP projects, beneficiaries. Tweets and their associated reach are considered as dissemination activities and citizen engagement mechanisms.
Online media	Online news	Press announcements for radical technologies

APPENDIX II – SELECTION CRITERIA FOR POLICY QUESTIONS

Figure 6: Criteria selection for policy questions