Data Management Plan for IntelComp. D8.2.

Version 2

Description

This document serves as the updated Data Management Plan (DMP) for the IntelComp project that has been created using the OpenAIRE ARGOS DMP service (argos.openaire.eu). It presents an in-depth overview of the project's data management practices, in compliance with the Horizon 2020 policy and FAIR guidelines. The DMP details the types of datasets collected, generated, and used, with a particular focus on the datasets created in the project (output datasets), and describes the management methods implemented in the IntelComp STI Data Space.

Funder

European Commission|| EC

Grant

A Competitive Intelligence Cloud/HPC Platform for AI-based STI Policy Making/ No 101004870

Researchers

 DOI: - 21/12/2023

Organizations

FUNDACION ESPANOLA PARA LA CIENCIAY LA TECNOLOGIA, F.S.P., FECYT, EVERIS SOLUCIONES TECNOLOGICAS SLU, ELLINIKO IDRYMA EREVNAS KAI KAINOTOMIAS, ZENTRUM FUR SOZIALE INNOVATION GMBH, Carlos III University of Madrid, OPENAIRE AMKE, Haut Conseil De L'evaluation De La Recherche Et De L'enseignement Superieur (HCERES), TILDE SIA, BARCELONA SUPERCOMPUTING CENTER-CENTRO NACIONAL DE SUPERCOMPUTACION, Communication & Information Technologies Experts Anonymos Etaireia Symvouleftikon Kai Anaptyxiakon Ypiresion (CITE), TECHNOPOLIS CONSULTING GROUP BELGIUM, MINISTERIO DE ECONOMIA, INDUSTRIA Y COMPETITIVIDAD, ATHINA-EREVNITIKO KENTRO KAINOTOMIAS STIS TECHNOLOGIES TIS PLIROFORIAS, TON EPIKOINONION KAI TIS GNOSIS

  DOI: -
21/12/2023

Datasets

Title: EU-GR Green Skills Dataset

Template: Horizon 2020

The Green Skills dataset, with data analytics on the demand (EU and Greece) and supply (Greece only) of green skills across various sectors, is for the production of the KPIs produced by interlcomp. This dataset provides a comprehensive view of the current landscape and trends in green skills within the workforce. It highlights areas where skills are in high demand but supply is lacking, guiding policymakers in developing targeted educational and training programs. Additionally, by comparing the EU-wide demand with the specific supply context in Greece, this dataset aids in aligning national policies with broader European sustainability goals, ensuring that the workforce is equipped for the emerging green economy.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

1.1.2 What types of data will the project generate/collect?

• derived or compiled (e.g.

• text mining

• 3D models)

1.1.3 What formats of data will the project generate/collect?

Numerical

CSV

1.1.4 What is the origin of the data?

Secondary data

1.1.6 To whom might it be useful ('data utility')?

• Decision makers

- Economy
- The public
- Industry

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
In the case of closed datasets, such as this one, created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

CSV

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

Yes

Comment:
Data for internal project use only

3.1.2.2 Will your data be openly accessible?

none

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

Yes

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

Yes

Comment:
ESCO Green Skills

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

Data for internal project use only

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Dimitris Pappas (orcid:0000-0001-5784-0658)

4.1.4 How do you intend to ensure data reuse after your project finishes?

Data Center Archive Storage

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Comment:
Data for internal project use only

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: All data (input and output)

Template: Horizon 2020

This section outlines the handling and management of input and output data within the project. The datasets are categorized as either new or reused, distinguishing the practices applied by the project team from those applied by others to datasets we received. For datasets created by project partners, detailed information on their FAIRness, the tools and methods used, and data managers is provided. For reused datasets, we describe their components and sources.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- To obtain information
- To make informed decisions
- To develop a product
- To combine with other data

Comment:
Within the IntelComp project, data collection and generation are closely aligned with its core objectives. The data management plan is crucial for efficient information retrieval, aiding stakeholders who engage with the various collections. The structured datasets, enriched with indicators and thematic analyses, are vital in empowering decision-makers in STI policy-making.

This document is complemented by the IntelComp Catalogue which serves as an online platform for operators to explore and search datasets and tools within the IntelComp Data Space and directly access datasets through the data provisioning service. Additionally, a universal model for data abstractions, termed the "Data Mediator" plays a crucial role in the IntelComp data provisioning system and overall architecture. This design enables data to be filtered, pre-processed, and delivered in multiple formats. Both case-specific and general data mediators have been instantiated, illustrating the system's versatility in different data delivery and processing contexts.

To obtain information: IntelComp's STI data space collects and reuses existing or new datasets for various phases of the platform, ranging from raw data to to analytics, indicators, and the back end of visualisations. This structured data collection is primarily aimed at supporting efficient information retrieval, crucial for stakeholders engaging with scientific publications or patent metadata.

To make informed decisions: The datasets, enriched with indicators and thematic analyses, play a pivotal role in empowering decision-makers with essential insights for making informed choices.

To develop a product: The datasets detailed in this document are fundamental to the development of the IntelComp platform. They enable the platform to start with a specific policy question and culminate in a rich Business Intelligence (BI) dashboard designed to provide comprehensive answers. The various end-user tools within the IntelComp platform, such as the STI Viewer, Evaluation Workbench, and Policy Participation Portal, are all underpinned by the datasets presented here. Each tool serves distinct purposes, catering to the varying needs of stakeholders engaged in the domain of STI.

To combine with other data: The methods implemented for storing, classifying, and retrieving data within the IntelComp data space facilitate seamless integration with other datasets and the various analytical tools of the platform. This interoperability enhances the potential for comprehensive insights, providing a thorough understanding of various elements within the STI realm.

1.1.2 What types of data will the project generate/collect?

- text mining
- static
- peer-reviewed data sets
- likely published or curated

• derived or compiled (e.g.

• text mining

• 3D models)

• reference or canonical (e.g.

• static

• peer-reviewed data sets

• likely published or curated

• such as gene sequence databanks or chemical structures)

The project will reuse and further contextualize existing data in order to satisfy the innovative ways of their analysis, such as through AI models. New data will be the outcome of merged, deduplicated and enhanced aggregated or crawled content.

1.1.3 What formats of data will the project generate/collect?

• Text files

• Numerical

• Models

The project has reused and further contextualized existing data in order to satisfy the innovative ways of their analysis, such as through AI models. New data were the outcome of merged, deduplicated and enhanced aggregated or crawled content. Other types include: Metadata - text files: CSV, XML, JSON and Artefacts - PDF, JATS XML, plain text, MS Word, HTML.

1.1.4 What is the origin of the data?

• Primary data

• Secondary data

• Other

In the framework of IntelComp, datasets are classified into three distinct tiers. Tier 1 encompasses raw datasets sourced externally from outside the project. These datasets are unprocessed and serve as the foundational data layer. Tier 2 represents processed datasets and have been subjected to various analytics to extract meaningful insights. Tier 3, on the other hand, emerge from as

aggregations as Key Performance Indicators (KPIs). These Tier 3 datasets are derived from both Tier 1 and Tier 2 datasets.

1.1.5 What is the expected size of the data?

TB (terabyte)

Comment:
The size of all data are around 7.5TB. They are the sum of all the data placed in the gateway node + the HDFS.

1.1.6 To whom might it be useful ('data utility')?

• Research communities

• Decision makers

• Economy

• The public

• Industry

The data produced, reused, and gathered within the IntelComp project offers significant value to a wide array of stakeholders, aligning with the project's focus on Public-Private-People Partnerships and living labs approach. The primary beneficiaries of this data include: Policy Makers: Those involved in shaping STI policies benefit from the data's insights, especially in areas like Climate Change, Health, and Artificial Intelligence (AI). This data helps in informed policy-making, aligning with the evolving needs within these domains. Industry Representatives: The data is crucial for industry stakeholders, providing them with the necessary knowledge to align their strategies and innovations with current scientific and technological advancements. The industry analysis has been conducted on the areas of Energy and Agrifood, both part of the Climate Change domain. Academia and Researchers: Scholars and researchers may find this data valuable for academic studies, research, and development, particularly in understanding and addressing the needs and challenges within AI, Climate Change, and Health sectors, including the evolving topic trends. Civil Society Participants: Organizations and individuals engaged in the collaborative formulation of STI policies use this data to gain nuanced perspectives on the needs and trends within their domains. Education Sector: Educational institutions can utilize this data for curriculum development, research purposes, and in shaping educational programs that are aligned with the latest trends in STI. Economic Analysts and Decision Makers: The data aids in understanding the broader economic impacts of STI policies, helping in forecasting, strategy formulation, and socio-economic planning. General Public: By providing accessible insights into the STI domain, the data empowers the public with a

better understanding of the advancements and changes in this rapidly evolving field. The IntelComp data and the corresponding analyis that can be conducted on it, is instrumental in setting agendas, analyzing policy cycles, and monitoring and evaluating STI policies.

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

Yes

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

Throughout the IntelComp project, a consistent metadata schema was employed for describing all datasets. This schema was based on the OpenAIRE data model, compatible with the CERIF ontology and Datacite/Dublin Core metadata schemas and adhered to the European Union's standards for research documentation. Detailed information on this approach is available at the OpenAIRE Graph Doocumentation Each dataset was integrated into the IntelComp data space in its original form. The primary focus was on ensuring accurate and comprehensive descriptions for each dataset using the established metadata schema, rather than aligning the datasets with each other. This strategy allowed for the preservation of the structure and integrity of each dataset while ensuring that they were uniformly documented for easy access and retrieval. The metadata elements are based on Dublin Core and aligned with the latest OpenAIRE Guidelines for maximum compatibility of practices. It also takes into consideration and applies relevant to the project elements, as found in the EOSC Resource Description The metadata framework included a mandatory set of ten elements: 1) Title, 2) Creator, 3) Description, 4) Publisher, 5) Date, 6) Type, 7) Identifier, 8) Source, 9) Rights, and 10) Access. Furthermore, fifteen optional elements were added to enrich the metadata descriptions: 11) Subject, 12) Contributor, 13) Format, 14) Language, 15) Relation, 16) Coverage, 17) Version, 18) Size, 19) Tags, 20) Main Contact, 21) Standards, 22) Webpage, 23) Logo, 24) Provenance, and 25) Alternative Identifier. By the end of the project, this approach ensured the IntelComp data space comprised a diverse collection of datasets, each described with a uniform yet individualized metadata profile, facilitating ease of access and use while maintaining the unique characteristics of each dataset.

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
All open datasets created in the project, and their associated metadata records will be freely available and deposited in Zenodo. Through OpenAIRE's ingestion of Zenodo content and its integration with the European Open Science Cloud (EOSC) catalogue, these datasets will also be featured in the EOSC catalogue. This dual availability ensures broader visibility and accessibility, allowing users to access the metadata and datasets from multiple platforms without any cost.

In the case of closed datasets created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

Comment:

The metadata for the open datasets created in the IntelComp project is designed to be fully harvestable. By depositing these datasets in Zenodo and integrating them with the European Open Science Cloud (EOSC) catalogue through OpenAIRE, we ensure that their metadata adheres to widely recognized standards like CERIF, Datacite, and Dublin Core. This standardization facilitates the harvesting of metadata using various tools and technologies, enhancing accessibility and integration into broader research and data ecosystems.

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

Yes

Comment:
Reused data are provided in their latest version or in the version of use by the given activity. On the other hand, the Data Space has employed a versioning mechanism for its/our own use.

3.1.1.10 Will you provide persistent identifiers for your data?

Yes

Comment:
Open datasets will be deposited and issued a persistent identifier. Digital Object Identifier (DOI) for data and research outputs, ORCID for researchers, FundRef for funding organizations. For other data sets, we will use any unique reference number that can be associated with the data records, or, when such identifiers do not exist, IntelComp will create them and attach them to the data records.

3.1.1.11 Persistent identifiers

DOI

3.1.1.12 Will you provide searchable metadata for your data?

Yes

Comment:
Subject headings, keywords and tags were used so as to increase searchability of content both internally, in the data space, and externally, by search engines.

3.1.1.13 What services will you use to provide searchable metadata?

Content Management System

OpenAIRE

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

Yes. The IntelComp project recognizes the importance of using standardized formats for data to facilitate interoperability and ease of use. However, it is acknowledged that reused datasets might not always adhere to these standardized formats, depending on their original source and format.

3.1.1.18 Are the file formats you will use open?

Yes

Comment:
Yes, recognizing that reused datasets might be subject to proprietary formats. Not all datasets ingested into the Data Space are in open format.

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

Comment:
Yes. The data in the IntelComp project, including data stored in Elasticsearch, SQL databases, and in Parquet file format, can be accessed using various open-source tools. For instance, Elasticsearch has its own query DSL for data retrieval, SQL databases can be accessed using tools like MySQL Workbench or PostgreSQL, and Parquet files can be read by data processing libraries in Python, such as Pandas or PyArrow.

3.1.1.21 Please describe if data require proprietary tools to access the data?

No proprietary tools are required to access the data in the IntelComp project.

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

Yes

Comment:
In the IntelComp project, dataset accessibility varied across the three tiers of datasets.

Tier 3 (End-User Tools): Primarily open datasets, prioritized for public access to support the functionality of end-user tools.

Tier 2 (Analytics): All datasets are closed, and are meant for internal project use.

Tier 1 (Reused): A mix of open and closed datasets, reflecting their original accessibility status; most are open, but a few remain closed based on their initial terms and conditions.

An Ethics Manager ensured compliance with legal and ethical standards, particularly balancing openness with necessary restrictions for certain datasets.

3.1.2.2 Will your data be openly accessible?

some

3.1.2.4 How will the data be made available?

Other

Open datasets created in the IntelComp project will be available via Zenodo, and through Zenodo's integration with OpenAIRE in the OpenAIRE Graph and the EOSC catalogue as well, enhancing their visibility and accessibility across the scientific community.

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

Comment:
The IntelComp Data Space, located at the Barcelona Supercomputing Center and utilizing Hadoop Distributed File System (HDFS) for storage, adheres to robust security protocols and risk assessment practices. This ensures both

the security of the data and the availability of comprehensive backup and recovery procedures, safeguarding against data loss and maintaining data integrity.

3.1.2.7 Are there any methods or tools required to access the data?

Yes

3.1.2.8 Please provide information about the method(s) needed to access the data

Access methods and tools required for the IntelComp project's data have been previously outlined in response to question 3.1.1.20.

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

Yes

Comment:
For the open Tier 3 datasets in the IntelComp project, while a controlled vocabulary was not uniformly applied across all datasets, common elements do exist. Given that these datasets are aggregates, they share certain standard features, such as indicator, which help in maintaining a level of consistency and interoperability.

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

end of project

3.1.4.4 What internationally recognised licence will you use for your data?

Creative Commons Attribution 4.0 International

Comment:
Tier 3 Datasets will issued under a Creative Commons Attribution 4.0 International.

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of

ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

Yes

Comment:
In supporting the reuse of open (Tier 3) datasets from the IntelComp project, we offer the following:

Comprehensive Metadata: Every dataset is accompanied by detailed metadata, aiding users in understanding the data and facilitating its reuse.

Assistance upon Request: User guides can be provided upon request.

Data Compatibility: The datasets are formatted and structured to ensure compatibility with common data analysis tools, simplifying the process of integration and analysis for various applications.

3.1.4.9 How long do you intend to support data reuse?

Up to 5 years

4.1 Allocation of resources

4.1.1 How will the cost of making your data findable, accessible, interoperable and reusable be covered?

- Use of national infrastructure
- Use of institution infrastructure
- Collaboration with other Projects

Comment:
The costs associated with making data findable, accessible, interoperable, and reusable were covered as part of the IntelComp project budget. It is important to highlight that post the conclusion of the IntelComp project, there will be no further updates or revisions to the datasets, and thus no additional costs incurred for these activities.

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

Comment:
All related consortium partners. We identify the data managers to the new data below, under each dataset's description.

4.1.3 Identify the people or roles that will be responsible for the management of the project data

All related consortium partners

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Comment:
This question has been addressed previously in the response to question 3.1.2.1. To summarize, yes, there are legal issues that impact data sharing within the IntelComp project, particularly regarding the openness of datasets. Different tiers of datasets (Tier 3 open datasets, Tier 2 closed, and Tier 1 reused datasets with varying openness) have specific legal constraints that dictate their sharing and accessibility. These considerations were carefully managed to comply with legal requirements while striving to maximize data accessibility within these parameters. Ethical considerations were also taken into account, particularly in relation to privacy and confidentiality for certain datasets.

6.1.2 What are the methods used for processing sensitive/personal data?

• Privacy constraints and applicable ethical norms

• Privacy policies

Comment:
All applicable methods defined in WP9 of IntelComp project deliverables on data anonymization were considered in the processing of sensitive / personal data.

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: SciNoBo classified scientific publications (AI & Energy domains)

Template: Horizon 2020

To create this dataset the SciNoBo classifier was used to classify scientific publications in the domains of Artificial Intelligence (AI) & Energy.

For more information regarding the algorithm of SciNoBo, please refer to: https://dl.acm.org/doi/abs/10.1145/3487553.3524677

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

Comment:
Classifying scientific publications according to Field-of-Science taxonomies is of crucial importance, powering a wealth of relevant applications including Search Engines, Tools for Scientific Literature, Recommendation Systems, and Science Monitoring. Furthermore, it allows funders, publishers, scholars, companies, and other stakeholders to organize scientific literature more effectively, calculate impact indicators along Science Impact pathways and identify emerging topics that can also facilitate Science, Technology, and Innovation policy-making (the objective of the Intelcomp project).

1.1.2 What types of data will the project generate/collect?

• derived or compiled (e.g.

• text mining

• 3D models)

We collected scientific publications in the year span of 2014-2021 and we used our AI classifier to assign the FoS labels of AI and Energy to them.

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

MB (megabyte)

Comment:
288 MB

1.1.6 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Decision makers

• Education

• Economy

• The public

• Industry

Since the dataset contains scientific publications classified to the two domains (AI & Energy) a lot of interest parties can be benefited. For example, researchers and research communities can use the data to perform different analysis (train new classifiers, provide insights in the research communities of AI and Energy etc.). Furthermore, decision makers can analyse the classified data and identify emerging technologies and make informed decisions (the same goes for Economy and Industry). Finally, from an educational and public aspect, users can easily find scientific publications related to these domains and read them.

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

No

Comment:
In the case of closed datasets, such as this one, created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

JSON

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

Yes

Comment:
Data for internal project use only

3.1.2.2 Will your data be openly accessible?

none

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

Yes

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are

created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Haris Papageorgiou (orcid:0000-0002-7352-2403)

4.1.4 How do you intend to ensure data reuse after your project finishes?

Data Center Archive Storage

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Comment:
Data for internal project use only

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: Topic model of Semantic Scholar documents in the domain of Artificial Intelligence

Template: Horizon 2020

This dataset consists of a subset of Semantic Scholar documents in the AI domain created with the Domain Classifier (https://github.com/IntelCompH2020/domain_classification) and Topic Modeler (https://github.com/IntelCompH2020/topicmodeler) toolboxes. First, we used the Domain Classifier to identify AI-related documents and selected those documents for further analysis. Then, we used the TMT with Mallet to create a topic model of 10 topics and extracted the document-topic distribution for each dataset's document. The resulting dataset includes the Semantic Scholar IDs of the documents used for creating the topic model, their associated lemmas, and a string representing the document-topic distribution in the format 'Topic_i|proportion_i Topic_i+1| proportion_i+1 ...'.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

• To obtain information

• To make informed decisions

Comment:

The Intelcomp project aims to generate topic models for various domains and data sources. As a result, this dataset serves a dual purpose. Firstly, it includes a subset of documents identified as belonging to the AI domain. Secondly, it provides a topic-based representation that enables thematic analysis of the documents in the identified subset. This information enriches Intelcomp's data sources and provides a foundation for other analysis tools, such as measuring impact using semantic similarity graphs.

1.1.2 What types of data will the project generate/collect?

• derived or compiled (e.g.

• text mining

• 3D models)

We identified AI-related documents in the entire Semantic Scholar dataset using the Domain Classifier and generated a Mallet topic model based on those documents with the Topic Modeler.

1.1.3 What formats of data will the project generate/collect?

• Text files

• Models

Parquet, csv, binary files

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

MB (megabyte)

Comment:
264M, but variable depending on the model and size of the data source

1.1.6 To whom might it be useful ('data utility')?

• Researchers

• Research communities

• Decision makers

• Education

• Economy

• The public

• Industry

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
In the case of closed datasets, such as this one, created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

Yes

3.1.1.10 Will you provide persistent identifiers for your data?

Yes

Comment:
We keep the original Semantic Scholar IDs.

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

Yes

Comment:
Data for internal project use only

3.1.2.2 Will your data be openly accessible?

none

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

Yes

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

Data for internal project use only

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Jerónimo Arenas-García (orcid:0000-0003-4071-7068)

4.1.4 How do you intend to ensure data reuse after your project finishes?

Data Center Archive Storage

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Comment:
Data for internal project use only

6.1.2 What are the methods used for processing sensitive/personal data?

Privacy policies

Comment:
No personal data used

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: EU Energy & Agrifood Enriched Patent Set

Template: Horizon 2020

This dataset is a curated collection of patents related to the energy and agrifood sectors. Patents are selected based on specific IPC/CPC codes and, in some cases, relevant keywords. Each patent is assigned to one topic according to standard sets. Further refining ensures inclusion of patents with at least one inventor or applicant from an EU member country. The collection provides insights into energy and agrifood inventions within the EU context.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

Comment:
This dataset helps in identifying emerging technologies, understanding industry dynamics, and guiding policy decisions to foster innovation and address societal needs effectively.

1.1.2 What types of data will the project generate/collect?

sample or specimen data

The dataset is a subset of the PATSTAT 2021B version.

1.1.3 What formats of data will the project generate/collect?

Text files

JSON

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

MB (megabyte)

Comment:
656

1.1.6 To whom might it be useful ('data utility')?

• Researchers

• Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
In the case of closed datasets, such as this one, created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

Yes

Comment:

The **patstat_version** field has the source dataset.

The **created** field has the date that the dataset was created.

3.1.1.10 Will you provide persistent identifiers for your data?

Yes

Comment:
appln_id is a key that uniquely identifies a patent application in the PATSTAT version used to create the dataset.

3.1.1.11 Persistent identifiers

Other

appln_id

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

JSON

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

Comment:
Data for internal project use only

3.1.2.2 Will your data be openly accessible?

none

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

Yes

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

Data for internal project use only

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Dimitris Pappas (orcid:0000-0001-5784-0658)

4.1.4 How do you intend to ensure data reuse after your project finishes?

Data Center Archive Storage

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Comment:
Data for internal project use only.

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: EU-GR Energy & Agrifood Regulations Set

Template: Horizon 2020

Sourced from Eurlex and the Greek National Printing Office, this dataset collates regulations pertinent to the energy and agrifood sectors. Advanced techniques, including crawling and string matching, were utilized to identify and extract relevant regulations along with their metadata. The collection offers a comprehensive insight into the regulatory framework of these sectors across European and Greek jurisdictions.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

Comment:
The regulation analytics dataset is pivotal for agenda-setting among STI policymakers due to its insights into regulatory frameworks and trends. This dataset once aggregated into KPIs empowers policymakers with critical information needed to shape effective science, technology, and innovation policies, ensuring that they are well-informed, relevant, and aligned with current regulatory landscapes and future developments.

1.1.2 What types of data will the project generate/collect?

• derived or compiled (e.g.

• text mining

• 3D models)

1.1.3 What formats of data will the project generate/collect?

• Text files

• Numerical

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

MB (megabyte)

1.1.6 To whom might it be useful ('data utility')?

Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
In the case of closed datasets, such as this one, created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

No

3.1.1.17 Please describe the formats you plan to store your data in, including any URLs to documentation.

Commonly used format

Comment:
JSON

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.2 Will your data be openly accessible?

none

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

Data for internal project use only.

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Ioannis Lyris (orcid:0000-0002-4204-228X)

Data Engineer

4.1.4 How do you intend to ensure data reuse after your project finishes?

Data Center Archive Storage

Comment:
Local data repositories.

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

Use it to produce statistics.

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Comment:
Data for internal project use only.

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: EU-GR Energy & Agrifood ESG Analytics

Template: Horizon 2020

This dataset aggregates ESG (Environmental, Social, and Governance) data specifically from the energy and agrifood sectors within the EU and Greece. Derived from companies' sustainability reports, advanced Language Model techniques (LLMs) were employed to pinpoint and extract specific metric values. The collection provides a comprehensive overview of ESG practices and metrics for companies operating in these sectors in the mentioned regions.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- To share information
- To make informed decisions
- To combine with other data

Comment:
An ESG (Environmental, Social, and Governance) analytics dataset plays a crucial role in agenda setting for STI (Science, Technology, and Innovation) policymakers for several reasons. First, it provides comprehensive insights into sustainability practices, social responsibility, and governance structures, which are vital for formulating policies that align with contemporary global challenges and ethical standards. Second, such a dataset aids in identifying emerging trends and priorities in the ESG domain, enabling policymakers to focus on key areas that drive innovation and sustainable development in the science and technology sectors.

1.1.3 What formats of data will the project generate/collect?

Other

Per company ESG metrics coverage and performance data.

1.1.4 What is the origin of the data?

Secondary data

1.1.6 To whom might it be useful ('data utility')?

Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
In the case of closed datasets, such as this one, created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

Comment:
For internal use in the project

3.1.2.2 Will your data be openly accessible?

none

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

For internal use in the project

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Dimitris Pappas (orcid:0000-0001-5784-0658)

4.1.4 How do you intend to ensure data reuse after your project finishes?

Data Center Archive Storage

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Comment:
Dataset for internal use in the project.

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: NIH Research Portfolio

Template: Horizon 2020

ExPORTER provides bulk RePORTER data that includes: projects, project abstracts, publications citing support, link tables for project to publication associations, patents, and clinical studies. ExPORTER contains research projects funded by National Institues of Health (NIH), Administration for Children and Families (ACF), Agency for Healthcare Research and Quality (AHRQ), Centers for Disease Control and Prevention (CDC), Health Resources and Services Administration (HRSA), Food & Drug Administration (FDA), and Department of Veterans Affairs (VA). More information: https://exporter.nih.gov/about.aspx"

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

• To obtain information

• To make informed decisions

Comment:
This dataset will be used in the Cancer Living lab for the calculation of indicators related to science.

1.1.2 What types of data will the project generate/collect?

• derived or compiled (e.g.

• text mining

• 3D models)

1.1.3 What formats of data will the project generate/collect?

Text files

CSV

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

GB (gigabyte)

Comment:
3.3. GB

2,6 Million projects, 2,6 Million publications, 60K patents, 28K clinical studies

1.1.6 To whom might it be useful ('data utility')?

• Researchers

• Decision makers

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

3.1.1.3 Will your metadata use standardised vocabularies?

No

3.1.1.10 Will you provide persistent identifiers for your data?

Yes

Comment:
Application_ID is the unique identifier for the projects

PMID is the PubMed identifier for the publications

PATENT_ID is the unique identifier for the patents

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.21 Please describe if data require proprietary tools to access the data?

No

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

3.1.2.2 Will your data be openly accessible?

none

Comment:
This dataset is for internal use of the project. The raw data information can be obtained directly from the NIH webpage

3.1.2.4 How will the data be made available?

Other

It will be available at the project Data Space for internal use only

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

This dataset is for internal use in the project. For other uses, the raw data can be obtained directly from the NIH website

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

3.1.4.8 Will you provide any support for data reuse?

No

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Delete at end of project

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.1.2 What are the methods used for processing sensitive/personal data?

National laws

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: Semantic Scholar

Template: Horizon 2020

Semantic Scholar is a search engine for research articles powered by the Allen Institute for Artificial Intelligence. These datasets provide a variety of information about research papers taken from a snapshot in time of the Semantic Scholar corpus. This site is provided by The Allen Institute for Artificial Intelligence as a service to the research community. The site is covered by AI2 Terms of Use and Privacy Policy.

This dataset contains parquet tables for papers, authors, paper_author correspondence, and citations, which are built by merging the following datasets from Semantic Scholar: papers, authors, abstracts, citations.

These are publications from all fields of science available at https://www.semanticscholar.org/. Includes authors, journals and conferences for different publication types since 1931.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

• To obtain information

• To make informed decisions

Comment:
This dataset will be used in all living labs for the calculation of indicators related to science.

1.1.2 What types of data will the project generate/collect?

• derived or compiled (e.g.

• text mining

• 3D models)

1.1.3 What formats of data will the project generate/collect?

• Text files

• Numerical

Parquet

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

GB (gigabyte)

Comment:
211,633,022 papers

81,067,677 authors

2,557,754,271 citations

590,620,314 paper-author relationships

~147GB

1.1.6 To whom might it be useful ('data utility')?

• Researchers

• Decision makers

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

3.1.1.3 Will your metadata use standardised vocabularies?

No

3.1.1.10 Will you provide persistent identifiers for your data?

Yes

Comment:

Unique identifiers internal to semantic scholar are used for the publications and the authors

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.21 Please describe if data require proprietary tools to access the data?

No

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

3.1.2.2 Will your data be openly accessible?

none

Comment:
This dataset is for internal use of the project. The raw data should be requested to the Allen AI Institute.

3.1.2.4 How will the data be made available?

Other

It will be available at the project Data Space for internal use only

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

This dataset is for internal use in the project. For other uses, the raw data can be obtained directly from the Allen AI Institute

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

3.1.4.8 Will you provide any support for data reuse?

No

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Delete at end of project

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.1.2 What are the methods used for processing sensitive/personal data?

National laws

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: AI publications, patents and EU projects

Template: Horizon 2020

This dataset consists of a series of identifiers for publications in OpenAIRE, publications in Semantic Scholar, patents from PATSTAT and projects from CORDIS which are considered relevant for the AI domain, and were used to calculate the AI-related indicators. The criterion for selecting these items has been agreed with members of the SEDIA team, the lead partner for IntelComp's AI living lab, and consists on appearance of at least three AI related keywords in the publications abstract, patent summary, or project description.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

- To share information
- To keep on record

Comment:
This dataset keeps the list of all CORDIS projects, publications, and patents that have been identified as relevant to the AI living lab. This information is compiled and shared to allow projects members to easily access the items that were used for the calculation of indicators in the AI living lab.

1.1.3 What formats of data will the project generate/collect?

Text files

CSV

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

MB (megabyte)

Comment:
50,9 MB

1.1.6 To whom might it be useful ('data utility')?

- Researchers
- Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

No

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
Closed datasets, such as this one, created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

No

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

Yes

Comment:
Unique identifiers from the OpenAIRE, Semantic Scholar, PATSTAT, and Cordis datasets are used to identify the elements in this dataset.

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

3.1.2.2 Will your data be openly accessible?

none

Comment:
This dataset is for project's internal use.

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

This dataset is for internal use in the project.

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Jerónimo Arenas-García (orcid:0000-0003-4071-7068)

4.1.4 How do you intend to ensure data reuse after your project finishes?

Personal Archive

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.1.2 What are the methods used for processing sensitive/personal data?

Not available

Comment:
No sensitive/personal data are included in this dataset.

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: EU AI Science Indicators

Template: Horizon 2020

This dataset provides key indicators about AI scientific production in Europe. It is derived from three main sources: expert-curated topic models using the Interactive Model Trainer (IMT) on OpenAIRE abstracts with AI-focused keywords, OpenAIRE publications with at least one European-affiliated author, and relevant patents from PATSTAT. These indicators have been curated specifically for in-depth analyses within the STI Viewer's Artificial Intelligence dashboards.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

Comment:
This datasets consists of a series of indicators of Science production in the AI domain with at least one author with European affiliation. Its target users are decision makers, to assist them by providing evidence based on data.

1.1.2 What types of data will the project generate/collect?

Other

Aggregated data from other data sources (publications, patents) enriched with topic models

Aggregated data from other data sources (publications, patents) enriched with topic models

1.1.3 What formats of data will the project generate/collect?

• Text files

• Numerical

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

KB (kilobyte)

Comment:
766 KB

1.1.6 To whom might it be useful ('data utility')?

Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
All open datasets created in the project, and their associated metadata records will be freely available and deposited in Zenodo. Through OpenAIRE's ingestion of Zenodo content and its integration with the European Open Science Cloud (EOSC) catalogue, these datasets will also be featured in the EOSC catalogue. This dual availability ensures broader visibility and accessibility, allowing users to access the metadata and datasets from multiple platforms without any cost.

3.1.1.6 Will your metadata be harvestable?

Yes

Comment:
The metadata for the open datasets created in the IntelComp project is designed to be fully harvestable. By depositing these datasets in Zenodo and integrating them with the European Open Science Cloud (EOSC) catalogue through OpenAIRE, we ensure that their metadata adheres to widely recognized standards like CERIF, Datacite, and Dublin Core.

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

JSON

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

3.1.2.2 Will your data be openly accessible?

all

3.1.2.4 How will the data be made available?

• Project website

• Repository of Archive

Zenodo

Couldn't find it? Insert it manually

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

The data can be visualized in the project STI Viewer tool or accessed in raw format

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

end of project

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Jerónimo Arenas-García (orcid:0000-0003-4071-7068)

4.1.4 How do you intend to ensure data reuse after your project finishes?

- Other
- Personal Archive

Zenodo

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.1.2 What are the methods used for processing sensitive/personal data?

Other

Comment:
No personal / sensitive data involved

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: ES AI Science Indicators

Template: Horizon 2020

This dataset compiles key indicators surrounding AI scientific production in Spain. Derived from several primary sources, it includes expert-curated topic models via the Interactive Model Trainer (IMT) on OpenAIRE abstracts with AI-specific keywords, OpenAIRE publications featuring at least one author tied to a Spanish institution, and relevant patents from PATSTAT. These indicators are meticulously crafted for in-depth analyses within the STI Viewer's Artificial Intelligence dashboards.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

Comment:
This datasets consists of a series of indicators of Science production in the AI domain with at least one author with Spanish affiliation. Its target users are decision makers, to assist them by providing evidence based on data.

1.1.2 What types of data will the project generate/collect?

Other

Aggregated data from other data sources (publications, patents) enriched with topic models

Aggregated data from other data sources (publications, patents) enriched with topic models

1.1.3 What formats of data will the project generate/collect?

• Text files

• Numerical

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

KB (kilobyte)

Comment:
705 KB

1.1.6 To whom might it be useful ('data utility')?

Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
All open datasets created in the project, and their associated metadata records will be freely available and deposited in Zenodo. Through OpenAIRE's ingestion of Zenodo content and its integration with the European Open Science Cloud (EOSC) catalogue, these datasets will also be featured in the EOSC catalogue. This dual availability ensures broader visibility and accessibility, allowing users to access the metadata and datasets from multiple platforms without any cost.

3.1.1.6 Will your metadata be harvestable?

Yes

Comment:
The metadata for the open datasets created in the IntelComp project is designed to be fully harvestable. By depositing these datasets in Zenodo and integrating them with the European Open Science Cloud (EOSC) catalogue through OpenAIRE, we ensure that their metadata adheres to widely recognized standards like CERIF, Datacite, and Dublin Core. This standardization facilitates the harvesting of metadata using various tools and technologies, enhancing accessibility and integration into broader research and data ecosystems.

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

JSON

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

3.1.2.2 Will your data be openly accessible?

all

3.1.2.4 How will the data be made available?

• Project website

• Repository of Archive

Zenodo

Couldn't find it? Insert it manually

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

The data can be visualized in the project STI Viewer tool or accessed in raw format

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

end of project

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

3.1.4.8 Will you provide any support for data reuse?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Jerónimo Arenas-García (orcid:0000-0003-4071-7068)

4.1.4 How do you intend to ensure data reuse after your project finishes?

- Other
- Personal Archive

Zenodo

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.1.2 What are the methods used for processing sensitive/personal data?

Other

Comment:
No personal / sensitive data involved

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: EU AI Technology Indicators

Template: Horizon 2020

This dataset aggregates critical indicators pertinent to AI technology within the European Union. It is specifically curated for in-depth analysis within the STI Viewer's Artificial Intelligence dashboards. These indicators emerge from a multifaceted analysis involving an expert-curated topic model using the Interactive Model Trainer (IMT) on PATSTAT patent summaries with AI-specific keywords, AI-related patents granted by the European Patent Office (EPO) as found in PATSTAT, and calculated metrics revolving around AI patents processed and granted by the EPO.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

Comment:
This datasets consists of a series of indicators of Technology production in the AI domain based on patents submitted to the European Patent Office. Its target users are decision makers, to assist them by providing evidence based on data.

1.1.2 What types of data will the project generate/collect?

Other

Aggregated data from other data sources (patents) enriched with topic models

Aggregated data from other data sources (patents) enriched with topic models

1.1.3 What formats of data will the project generate/collect?

• Text files

• Numerical

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

KB (kilobyte)

Comment:
451 KB

1.1.6 To whom might it be useful ('data utility')?

Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
All open datasets created in the project, and their associated metadata records will be freely available and deposited in Zenodo. Through OpenAIRE's ingestion of Zenodo content and its integration with the European Open Science Cloud (EOSC) catalogue, these datasets will also be featured in the EOSC catalogue. This dual availability ensures broader visibility and accessibility, allowing users to access the metadata and datasets from multiple platforms without any cost.

3.1.1.6 Will your metadata be harvestable?

Yes

Comment:
The metadata for the open datasets created in the IntelComp project is designed to be fully harvestable.

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

JSON

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

3.1.2.2 Will your data be openly accessible?

all

3.1.2.4 How will the data be made available?

• Project website

• Repository of Archive

Zenodo

Couldn't find it? Insert it manually

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

The data can be visualized in the project STI Viewer tool or accessed in raw format.

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

end of project

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Jerónimo Arenas-García (orcid:0000-0003-4071-7068)

4.1.4 How do you intend to ensure data reuse after your project finishes?

- Other
- Personal Archive

Zenodo

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.1.2 What are the methods used for processing sensitive/personal data?

Other

Comment:
No personal / sensitive data involved

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: Cancer publications, patents and EU projects

Template: Horizon 2020

This dataset consists of a series of identifiers for publications in OpenAIRE, publications in Semantic Scholar, patents from PATSTAT and projects from CORDIS which are considered relevant for the cancer domain, and were used to calculate the CA-related indicators. The criterion for selecting these items has been agreed with members of the HCERES team, the lead partner for IntelComp's cancer living lab. Publications were selected manually by experts in the field, PATENTS were selected using CPC and keywords criteria, and projects were selected based on their links to pre-selected publications. Additionally, machine learning was used to create domain classifiers for this field and these data sets.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

• To share information

• To keep on record

Comment:
This dataset keeps the list of all CORDIS projects, publications, and patents that have been identified as relevant to the cancer living lab. This information is compiled and shared to allow projects members to easily access the items that were used for the calculation of indicators in the cancer living lab.

1.1.3 What formats of data will the project generate/collect?

Text files

CSV

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

MB (megabyte)

Comment:
201 MB

1.1.6 To whom might it be useful ('data utility')?

• Researchers

• Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

No

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
This is a closed dataset created in the project its metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

Comment:
The metadata for the open datasets created in the IntelComp project is designed to be fully harvestable.

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

Yes

Comment:
Unique identifiers from the OpenAIRE, Semantic Scholar, PATSTAT, and Cordis datasets are used to identify the elements in this dataset.

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

No

3.1.1.17 Please describe the formats you plan to store your data in, including any URLs to documentation.

CSV

Commonly used format

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

Yes

Comment:
This is a Tier 2 (Analytics) dataset: All such datasets are closed, mainly due to Intellectual Property Rights (IPR) constraints.

3.1.2.2 Will your data be openly accessible?

none

Comment:
This dataset is for project's internal use.

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

Comment:
The IntelComp Data Space, located at the Barcelona Supercomputing Center and utilizing Hadoop Distributed File System (HDFS) for storage, adheres to robust security protocols and risk assessment practices. This ensures both the security of the data and the availability of comprehensive backup and recovery procedures, safeguarding against data loss and maintaining data integrity.

3.1.2.7 Are there any methods or tools required to access the data?

No

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

This dataset is for internal use in the project.

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Jerónimo Arenas-García (orcid:0000-0003-4071-7068)

4.1.4 How do you intend to ensure data reuse after your project finishes?

Personal Archive

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.1.2 What are the methods used for processing sensitive/personal data?

Not available

Comment:
No sensitive/personal data are included in this dataset

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: Organizations

Template: Horizon 2020

Organizations Data downloaded from the Crunchbase database using an access key provided by Technopolis-Group.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To obtain information

1.1.2 What types of data will the project generate/collect?

Other

Organization Data

1.1.3 What formats of data will the project generate/collect?

Other

.json files

1.1.4 What is the origin of the data?

Primary data

1.1.5 What is the expected size of the data?

GB (gigabyte)

Comment:
1.2

1.1.6 To whom might it be useful ('data utility')?

- Decision makers
- Industry

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

3.1.1.3 Will your metadata use standardised vocabularies?

No

3.1.1.5 Will you make the metadata available free-of-charge?

No

3.1.1.6 Will your metadata be harvestable?

No

3.1.1.7 Will you use naming conventions for your data?

No

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

No

3.1.1.17 Please describe the formats you plan to store your data in, including any URLs to documentation.

.json files

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

Yes

3.1.2.2 Will your data be openly accessible?

none

3.1.2.4 How will the data be made available?

Repository of Archive

Local Server

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

3.1.2.10 Will you also make auxiliary data that may be of interest to researchers available?

no auxiliary data

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: Funding Rounds

Template: Horizon 2020

Funding Round Data downloaded from the Crunchbase database using an access key provided by Technopolis-Group.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To obtain information

1.1.2 What types of data will the project generate/collect?

Other

Funding Rounds

1.1.3 What formats of data will the project generate/collect?

Other

.json file

1.1.4 What is the origin of the data?

Primary data

1.1.5 What is the expected size of the data?

MB (megabyte)

Comment:
250

1.1.6 To whom might it be useful ('data utility')?

- Decision makers
- Industry

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

3.1.1.3 Will your metadata use standardised vocabularies?

No

3.1.1.5 Will you make the metadata available free-of-charge?

No

3.1.1.6 Will your metadata be harvestable?

No

3.1.1.7 Will you use naming conventions for your data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

Yes

3.1.2.2 Will your data be openly accessible?

none

3.1.2.4 How will the data be made available?

Repository of Archive

Local Server

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

3.1.2.10 Will you also make auxiliary data that may be of interest to researchers available?

no auxiliary data

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.8 Will you provide any support for data reuse?

No

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: EU-GR Energy & Agrifood Industry Analytics

Template: Horizon 2020

The Climate Change Industry Analytics Dataset for Energy & Agrifood (EU/GR) encompasses per company innovativeness indicators and other metadata crucial for the KPIs produced by Intelcomp for agenda-setting in the energy and agrifood sectors, offering insights for both the European Union and Greece specifically.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

1.1.2 What types of data will the project generate/collect?

• derived or compiled (e.g.

• text mining

• 3D models)

1.1.3 What formats of data will the project generate/collect?

Numerical

JSON

1.1.4 What is the origin of the data?

Secondary data

1.1.6 To whom might it be useful ('data utility')?

• Decision makers

• Economy

• The public

• Industry

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
In the case of closed datasets, such as this one, created in the project their metadata will be open and available through the IntelComp Data Catalogue. However, access to this catalogue may be limited.

3.1.1.6 Will your metadata be harvestable?

Yes

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

CSV

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

Yes

Comment:
Data for internal project use only

3.1.2.2 Will your data be openly accessible?

none

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

Yes

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

Yes

Comment:
Fascati Fields of Study, Eurovoc, IPC, NACE

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

never

3.1.4.3 Please describe the reason the data will not be made available

Data for internal project use only

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Dimitris Pappas (orcid:0000-0001-5784-0658)

4.1.4 How do you intend to ensure data reuse after your project finishes?

Data Center Archive Storage

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

Yes

Comment:
Data for internal project use only

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: EU-GR Energy & Agrifood Indicators

Template: Horizon 2020

The EU-GR Energy & Agrifood Indicators dataset provides essential indicators for STI policymakers in the energy and agrifood sectors of Europe and Greece. It covers a broad spectrum of societal sectors including Science, Technology, Industry, ESG (Environmental, Social, and Governance), Human Resources, and Policy. Each indicator set is derived from specific analytics: Science indicators are created from publication analytics, Technology from patent analytics, Industry from company/industry analytics, ESGs from ESG analytics, Human Resources from Green Skills analytics, and Policy from regulation analytics.

This dataset, a product of multiple AI-driven pipelines, is derived by processes on raw datasets (Tier 1) and analytics datasets (Tier 2, as described in this DMP) to produce a comprehensive collection of Tier 3 indicators. These indicators provide a detailed view of the current state and trends within the energy and agrifood sectors, assisting policymakers in agenda setting and policy formulation. The data bridges various aspects of societal development and offers a framework for understanding the intersection of science, technology, industry, and human resource development in these critical domains.

Dataset Description

1.1 Data Summary

1.1.1 What is the purpose of the data collection/generation and its relation to the objectives of the project?

To make informed decisions

1.1.2 What types of data will the project generate/collect?

Other

Aggregated data from other data sources

1.1.3 What formats of data will the project generate/collect?

Numerical

1.1.4 What is the origin of the data?

Secondary data

1.1.5 What is the expected size of the data?

KB (kilobyte)

1.1.6 To whom might it be useful ('data utility')?

Decision makers

2.1 Reused Data

2.1.1 Will you re-use any existing data and how?

No

3.1.1 Making data findable, including provisions for metadata

3.1.1.1 Will you use metadata to describe the data?

Yes

CERIF (Common European Research Information Format)

3.1.1.3 Will your metadata use standardised vocabularies?

Yes

The project used COAR controlled vocabularies (https://www.coar-repositories.org/) as well as ontologies, following the OpenAIRE example for Resource Description Framework (RDF) compatibility. That was important for the project to allow for successful interlinking of resources across all domains.

3.1.1.5 Will you make the metadata available free-of-charge?

Yes

Comment:
All open datasets created in the project, and their associated metadata records will be freely available and deposited in Zenodo. Through OpenAIRE's ingestion of Zenodo content and its integration with the European Open Science Cloud (EOSC) catalogue, these datasets will also be featured in the EOSC catalogue. This dual availability ensures broader visibility and accessibility, allowing users to access the metadata and datasets from multiple platforms without any cost.

3.1.1.6 Will your metadata be harvestable?

Yes

Comment:
The metadata for the open datasets created in the IntelComp project is designed to be fully harvestable. By depositing these datasets in Zenodo and integrating them with the European Open Science Cloud (EOSC) catalogue through OpenAIRE, we ensure that their metadata adheres to widely recognized standards like CERIF, Datacite, and Dublin Core.

3.1.1.7 Will you use naming conventions for your data?

Yes

3.1.1.8 Please provide more details and examples on used naming conversions

In the IntelComp project, we employed basic naming conventions to ensure clarity and consistency in data management. These conventions were straightforward, focusing on facilitating easy identification and organization of the datasets. An example of our naming convention is: [DatasetName]_[VersionNumber]_[Date].

3.1.1.9 Will you provide clear version numbers for your data?

No

3.1.1.10 Will you provide persistent identifiers for your data?

No

3.1.1.12 Will you provide searchable metadata for your data?

No

3.1.1.15 Will you use standardised formats for your data?

Yes

Couldn't find it? Insert it manually

3.1.1.16 Provide information about used standardised formats

JSON

3.1.1.18 Are the file formats you will use open?

Yes

3.1.1.20 Do supported open-source tools exist for accessing the data?

Yes

3.1.1.22 Will you provide metadata describing the quality of the data?

No

3.1.2 Making data openly accessible

3.1.2.1 Are there ethical or legal issues that can impact sharing the data?

No

3.1.2.2 Will your data be openly accessible?

all

3.1.2.4 How will the data be made available?

• Project website

• Repository of Archive

Zenodo

Couldn't find it? Insert it manually

3.1.2.6 Is the storage sufficiently secure for the data and does the storage provide backup and recovery procedures?

secure with backup and recovery

3.1.2.7 Are there any methods or tools required to access the data?

No

The data can be visualized in the project STI Viewer tool or accessed in raw format

3.1.3 Making data interoperable

3.1.3.1 Will you use a controlled vocabulary for your data?

No

3.1.3.2 Will you provide a mapping to more commonly used ontologies?

No

3.1.4 Increase data reuse

3.1.4.1 When do you plan to make your data available for reuse?

end of project

3.1.4.5 Do you have documented procedures for quality assurance of your data?

No

Comment:
While the IntelComp project does not have formally documented procedures for data quality assurance, extensive quality checks and curations are inherently part of the process for all Tier 2 and Tier 3 datasets, which are created within the project. These procedures include various methods of ensuring data accuracy, consistency, and reliability. However, the specific processes and checks implemented are not formally documented but are integral to the workflow of data handling and management in the project.

3.1.4.8 Will you provide any support for data reuse?

No

4.1 Allocation of resources

4.1.2 Will you identify a data manager to manage your data, if not who will be responsible for the management of your data?

Yes

4.1.3 Identify the people or roles that will be responsible for the management of the project data

Dimitris Pappas (orcid:0000-0001-5784-0658)

4.1.4 How do you intend to ensure data reuse after your project finishes?

- Other
- Personal Archive

Zenodo

5.1 Data Security

5.1.1 What do you plan to do with research data of limited use?

Kept on secure, managed storage for limited time

6.1 Ethical aspects

6.1.1 Are there any ethical or legal issues that can have an impact on data sharing?

No

6.1.2 What are the methods used for processing sensitive/personal data?

Other

Comment:
No personal / sensitive data involved

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Title: Input (Reused) Datasets

Template: Horizon 2020

The subsequent raw datasets (Tier 1) were utilized in the IntelComp project. These datasets retain their original metadata as provided by their respective sources and are not shared by the project. They serve as the foundational elements for constructing Tier 2 (analytics) datasets, which are kept internal to the project (and described in detail in the DMP) and remain closed. Additionally, these contribute to the development of Tier 3 (indicators) datasets, which we make publicly accessible and are also described herein.

- OpenAIRE Graph (https://graph.openaire.eu/)
- Semantic Scholar (https://www.semanticscholar.org)

- PATSTAT - EPO (www.epo.org)
- Crunchbase (https://www.crunchbase.com/)

- Orbis (bvdinfo.com/R0/Orbis)
- CORDIS (https://cordis.europa.eu/)
- PubMed (https://pubmed.ncbi.nlm.nih.gov/)
- NIH (https://report.nih.gov/)
- Euraxess (https://euraxess.ec.europa.eu/)
- Drugbank (https://go.drugbank.com/)

- Clinical Trails (https://clinicaltrials.gov/)

Dataset Description

3.1.2 Making data openly accessible

3.1.2.7 Are there any methods or tools required to access the data?

No

7.1 Other

7.1.1 Do you make use of other procedures for data management?

No

Powered by