The S3 Observatory

Methology and quality review

S3 Community of Practice

Table of Contents

1. The S3 Observatory

The S3 Observatory, hosted within the S3CoP webpage, is intended to provide a core and accessible set of information for S3s around the EU, as it allows users to compare intuitively the specialisations areas of EU regions and members states, as well as providing contact points and links to the strategy.

The S3 Observatory has been built in collaboration among DG REGIO G1, DG REGIO country-desks and the S3CoP Secretariat. The project team identified S3 priorities directly from regions' and member states' S3 documents and deployed tools from Artificial Intelligence (AI), including Generative AI and entity extraction and disambiguation for automatic topic tagging and classification. AI supported the following exercises:

- Categorisation of the strategies and priorities across the different taxonomies (NACE Sections and divisions, NABS codes (1 and 2 digits), and Industrial Ecosystems
- Identification of keywords at the level of S3 strategy and S3 priority

An essential disclaimer for this analysis is that AI is not error-free, as AI models have limitations in contextual comprehension. To address this limitation, we incorporated human supervision and quality review to maximise the accuracy and consistency of the results. This synergy between generative AI and human expertise enhances the system's reliability. Nevertheless, as a pioneering exercise, the S3 Observatory may result in eventual misclassifications or inaccuracies due to the inherent complexity of natural language understanding. For the reasons above, the S3 Observatory will undergo periodic reviews.

Against this backdrop, the project team prioritised transparency and explainability to identify and address potential misclassifications in future revisions. Regional/national institutions or bodies, responsible for the management of the smart specialisation strategy can contact us at contact@s3-cop.eu to rectify or integrate the information provided.

2. Classifying and extracting keywords from S3s through AI

2.1. Structure of the dataset

The S3 Observatory database is structured as indicated in Table 1 below.

Our approach to classify the startegies across different taxonomies, namely NACE Sections and divisions, NABS codes (1 and 2 digits), and European Industrial Ecosystems, is based on generative AI and LLMs (Large Language Models).

Table 1: S3 Observatory database schema

Category	Info
General info	NUTS Level NUTS CODE/MS Country name Region name (national language) Region name (english) Flag Coute of Arms Area Population GDP GDP per capita % of National GDP % of Unemployment Regional GERD (%)
S3 Strategy	Name of strategy Types of strategy Languages Link to the S3 website Document 2021 - 2027 Strategy approved (Y/N)
Thematic platforms and partnership	Thematic platforms name Partnership name Role in the partnerships Partnership link
S3CoP activities	S3CoP activities in which the region is involved
S3 Priorities	Priorities (1 line per priority) Keywords (Linked to priorities and gave context to those priorities European Innovation Ecosystem Economic classification (Nace Section) Economic classification (Nace division) Scientific classification (NABS digit 1) Scientific classification (NABS digit 2) Keyword search (integrated into the filtering system)
Contact points	Name and surname Organisation Email
Other	Events News Other relevant documents

The initial approach involved screening all the regional S3 strategies available, using text mining to retrieve content related to the specified priorities. This was done to clearly identify sections of the text where the priorities were mentioned or where the phrasing of the priorities could be located.

Following that, we made use of PEGASUS to transform the responses into succinct paragraphs.

Such input was then used for two purposes, both pursued through a Zero-shot classification approach with the OpeanAI API (text-davinci-003 mode), namely:

- The classification of priorities by NACE, NABS and European Industrial Ecosystems.
- The identification of succint passwords per each S3 priority.

The graph below summarises the approach:

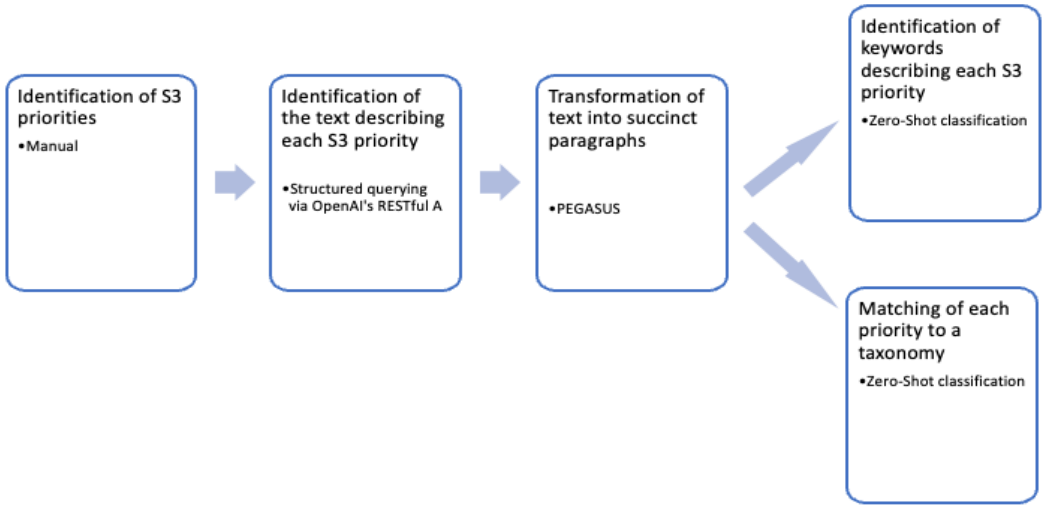

Figure 1 Constructing the S3 Observatory database

2.2. Identification of keywords

In order to develop the *search function* of the Observatory and to provide a definition of the different S3 priorities, a set of keywords were developed, through both AI and manual processes.

2.2.1. Identification of general keywords by S3 strategy through AI

An AI exercise was developed to identify keywords referred to the S3s as a whole, combining Large Language Models and frequency analysis. To this aim, we proceeded in 4 different steps:

1. **Identification of the main topics:** we used a Large Language Model which is daily trained on Wikipedia data to cover a large set of languages that allows us to work in a multilingual setting. In this context, a topic is identified as a wiki concept that is available on Wikipedia. We cover therefore a gigantic corpus, which is a landmark dataset for semantic analysis.

2. **Computation of the frequency of words associated to each topic.**

3. Focus of the analysis on the **best predicted topics measured by relevancy score** (here set at minimum 40%)

4. **Restriction to the most frequent words (top 20)**

The resulting list was later manually explored, to remove exceedingly generic words, which would dilute the search process.

2.2.2. Keywords by S3 priorities (description)

As well as the general kyewords by S3 Strategy, we identified a set of keywords, corresponding to each individual S3 priority. As described above, this was done through the set of steps described under figure 1.

2.2.3. Manual integration of keywords

The list of keywords identified through AI was complemented by manually by integrating a set of terms semantically related to (1) the priorities (2) the Industrial Ecosystems. As S3s are heterogenous in their writing style and use of terms. The AI approach, by itself, cannot ensure full disambiguation and harmonisation of terms.

In future releases of the S3 Observatory, the search function will be improved to increase disambiguation and harmonisation of keywords.

3. Quality review

3.1. Human intervention

Our quality review methodology focuses on employing a two-tiered artificial intelligence (AI) system, supplemented by human intervention, to ensure accuracy and precision in the results from the classification and keywords extraction functions.

The classification of all S3 priorities has been revised manually, by the S3CoP Secretariat, to ensure consistency between the title of the priority and the classifications by NACE, NABS and European Industrial Ecosystem.

In particular, the human supervision involved two options:

a) Confirm the AI's output if it meets the standards and is coherent.
b) Reject and relable the output if it doesn't align with the document's intent or the classification guidelines.

3.1. Interaction with the users

An discussed in the introduction, an essential disclaimer for this analysis is that AI is not error-free, as AI models have limitations in contextual comprehension.

Regional/national institutions or bodies, responsible for the management of the smart specialisation strategy are kindly invited to contact us at contact@s3-cop.eu to rectify or integrate the information provided.

Annexes

Table 2: Summary of AI tools and methods for keyword extraction and classification.

Component	Description
****Model/Libraries****	
spaCy	Library for advanced Natural Language Processing in Python
NLTK	The Natural Language Toolkit, or more commonly NLTK, is a suite of libraries and programs for symbolic and statistical natural language processing for English written in the Python programming language. Employed for tokenization and stopword removal.
pandas	Facilitated data manipulation and I/O operations.
OpenAI GPT-3	Generative Pre-trained Transformer 3 (GPT-3) is a large language model released by OpenAI in 2020.
PEGASUS	Used for text summarization to condense large texts.
****Methods****	
`clean_text`	Method for text preprocessing, removing special characters and stopwords, and performing lemmatization.
`extract_keywords`	Leverages OpenAI's API to extract keywords related to the specified strategic priority.
****Approach****	
PEGASUS	Employed PEGASUS summarization of prompt replies.
OpenAI	Zero-shot classification with OpenAI API for classifiying priorities in accordance to taxonomies.
OpenAI	Used Davinci model with specific prompts for strategic priority-based keyword extraction.