Building an AI-Ready Data Infrastructure Towards a SDG-focused Observatory for the Brazilian Amazon

Joao Pita Costa
IRCAI, Jozef Stefan Institute
Ljubljana, Slovenia
joao.pitacosta@quintelligence.com

Mirozlav Polzer
GloCha, Climate Chain Coalition
Klagenfurt, Austria

Leonardo Barrionuevo
MetAmazonia, AMAGroup
Curitiba, Brazil

Joao Paulo Veiga
CIAAM, University of São Paulo
São Paulo, Brazil

ABSTRACT

The effective monitoring and advancement of the United Nations Sustainable Development Goals (SDGs) demand a robust, interoperable, and scalable data infrastructure that can allow research collaboration. This paper introduces an AI-ready SDG data infrastructure, a curated collection of globally sourced datasets mapped to specific SDG indicators. By transforming heterogeneous data into machine-readable, semantically enriched formats, the infrastructure supports advanced analytics and AI-driven insights into sustainability challenges. We explore the design principles behind the system, including metadata standardization and open data practices. Furthermore, we discuss how this infrastructure enables AI applications such as trend detection, policy simulation, and AI agent support. We demonstrate these in the particular context of the Brazilian Amazon rainforest that sets stage to the COP30, the UNESCO Landslides Observatory and the ELIAS Migration Observatory, as pilot observatories empowering decision-makers, researchers, and civil society to accelerate progress toward the 2030 Agenda.

CCS CONCEPTS

• Real-time systems • Data management systems • Education • Document management and text processing

KEYWORDS

Sustainable Development Goals (SDGs), AI-ready data infrastructure, FAIR data principles, Open data, Semantic interoperability, Brazilian Amazon, COP30.

1 Introduction

The United Nations' 2030 Agenda for Sustainable Development outlines 17 SDGs aimed at addressing the worlds most pressing social, economic, and environmental challenges. Achieving these goals requires not only coordinated policy action and resource mobilization but also robust AI-enabled data systems capable of tracking progress, identifying gaps, and informing interventions. However, current efforts to monitor and evaluate the SDGs are often hampered by fragmented, inaccessible, or outdated data that are not designed with advanced analytics or AI applications in mind [1]. As the volume and variety of sustainability-related data continue to grow (ranging from satellite imagery and sensor networks to administrative records and citizen-generated content) there is a critical need to rethink the way data infrastructures are designed. Despite AI-related advancements, the broader ecosystem of SDG data remains siloed, with significant disparities in data availability, quality, and usability across countries and sectors. National statistical offices often lack the infrastructure or capacity to generate real-time, high-resolution data, while non-governmental data sources remain underutilized due to interoperability issues or lack of trust. As a result, policymakers and researchers face substantial barriers when attempting to harness AI for sustainable development monitoring. There is growing recognition that SDG data must be AI-ready: structured, interoperable, machine-readable, and enriched with metadata that allows for automated processing and semantic understanding [2]. AI-ready data infrastructures enable the use of artificial intelligence and machine learning tools for trend detection, predictive modeling, and evidence-based policymaking, accelerating the global effort toward sustainable development. Several initiatives have emerged to bridge the gap between data collection and actionable insights. In this context, the IRCAI SDG Observatory, an open-access data infrastructure developed by the International Research Centre on Artificial Intelligence under the auspices of UNESCO (IRCAI), aggregates and organizes datasets related to SDG indicators, news, policies, educational resources and innovation ecosystems, facilitating their use in AI applications through adherence to open data standards, consistent metadata schemas, and semantic alignment with the SDG framework. It represents a step toward a scalable, reusable AI-ready data architecture that can support both global and local decision-making. This paper presents a conceptual and practical framework for AI-ready SDG data infrastructure, building on the design principles and implementation strategies demonstrated by the IRCAI SDG Observatory, as well as by the preceding NAIADES Water Observatory [3] focusing on AI and Water Sustainability, and the recently deployed UNESCO Landslides Observatory discussed in section 4, both in the intersection of SDG 13 (Climate Action) with SDG 6 (Water Sustainability) and SDG 11 (Resilient Cities and Communities). We follow the discussion in [4] and propose an AI-ready and AI-enabled data and metadata infrastructure that can be leveraged for research purposes in what relates AI and Sustainable Development. Through this lens, we argue for a paradigm shift demonstrating an Amazon-focused SDG data ecosystem built on this new paradigma—moving from static,

indicator-focused reporting systems to dynamic, AI-compatible engine that supports (i) education and training for sustainability; (ii) desinformation monitoring practices in the sustainability discourse; and (i) data-driven decision-making and global collaboration.

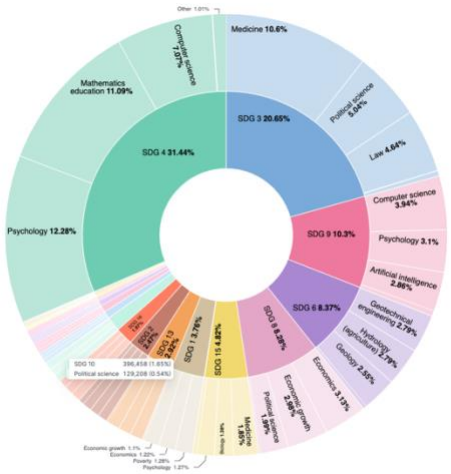

Figure 1: The SDG distribution of the ingested scientific article abstracts and their Amazon-related main concepts.

2 Data and Metadata Architecture

Designing an AI-ready SDG data infrastructure requires more than simply aggregating datasets—it demands a structured, extensible architecture that enables machine interpretability, semantic consistency, and interoperability across domains. The IRCAI SDG Observatory proposes in [5] a data structure incorporating both heterogeneous data and complex preprocessed metadata layers to support automated reasoning, text mining applications, and dynamic sustainability analysis. At the core of the infrastructure lies the data layer, which consists of curated datasets aligned with specific SDG indicators. These datasets are collected from a variety of sources, including international organizations, national statistics offices, worldwide news engines, open government data portals, and research institutions. To ensure consistency and usability, raw datasets undergo a 3-step transformation process:

- **Harmonization:** Raw data is converted into standardized formats (e.g., CSV, JSON, RDF) using predefined schemas (as the official SDG indicator framework defined by the UN Statistics Division [6]).
- **Normalization:** Variables such as geographic units, time periods, and measurement scales are normalized to ensure comparability across countries and regions.
- **Validation:** Data quality checks are implemented to flag missing values, outliers, or inconsistent units, helping maintain reliability and analytical integrity.

IRCAI is engaging domain experts for the different SDGs to explore the most relevant KPIs to monitor, the search terms in the ontology (discussed in the next section) and the outcomes from the analysis. The resulting datasets are thus not only clean and standardized (considering limitations of the data sources, including different types of bias analysed and exposed) but also structured in elasticSearch indices to support downstream AI applications acting over powerful Lucene queries through the native API. Surrounding the data layer is a robust metadata architecture that enables discoverability, semantic enrichment, and AI-readiness. The metadata design is informed by the FAIR data principles (Findable, Accessible, Interoperable, and Reusable) and includes the following key components: (i) **Descriptive Metadata,** including descriptive elements such as title, description, source organization, temporal coverage, geographic coverage, and associated SDG goals, enabling human and machine agents to easily understand the scope and purpose of each data index; (ii) **Structural Metadata,** specifying the internal structure of the dataset, such as data types, column definitions, units of measurement, and relationships between variables, facilitating data parsing and automatic preprocessing by text mining tools; (iii) **Source Metadata,** capturing information about the dataset's origin, transformation steps, update frequency, and quality assurance processes, ensuring transparency, reproducibility, and trustworthiness; and (iv) **Semantic Metadata,** leveraging ontologies and controlled vocabularies to provide machine-readable semantics, linking dataset elements to established knowledge graphs, enabling reasoning across data indices and automated alignment of conceptually related information.

To ensure accessibility and integration with external systems, the infrastructure exposes datasets and metadata through native RESTful APIs, allowing developers and researchers to query and retrieve relevant data programmatically, enabling use in dashboards, modeling pipelines, and decision-support systems. Furthermore, adherence to open data standards such as DCAT (Data Catalog Vocabulary) and JSON-LD (Linked Data) ensure that the infrastructure can interface with other open government data platforms, research data repositories, and semantic web services. The architecture is designed with scalability and modularity in mind, allowing new datasets to be integrated with minimal manual intervention. Through automated ingestion pipelines and schema mapping tools, the infrastructure can accommodate additional data sources while preserving metadata integrity and interoperability. Governance mechanisms, including data quality audits and contributor guidelines, ensure the sustainability and reliability of the system over time.

Figure 2: Visualisation of the SDG distribution of the ingested OECD AI policies according to the SDG ontology built on Wikidata terms defined with SDG topic experts.

To support the in-depth analysis and leverage the availability of multilingual text resources at Wikidata, we have developed a SDG ontology inspired by [7] based on terms that correspond to Wikipedia pages. Currently published in a CSV format on GitHub [8], it defines rows corresponding to SDG entities—such as goals—and maps them to Wikidata Q-IDs. Key columns include: Level (e.g., SDG Goal), Code (e.g., "1", "1.2", "1.2.1"), Wikidata Q-Identifier (e.g., Q23442, Q3048436, Q28146087), label (human-readable name), Description (concise textual summary), and related concepts (optional Q-IDs linking to domains like health, energy, gender equality) Each SDG Goal row includes its code and corresponding Wikidata ID. Targets (e.g. 1.2) are mapped to both their own Wikidata entity and an explicit parent Goal. Indicators (e.g. 13.2.1) reference the relevant Target and define unit, measurement scale, and description. Using the CSV mappings, the ontology is constructed so that:

- sdg:hasTarget links a goal entity to its targets
- sdg:hasIndicator links targets to indicators
- sdg:measuredIn aligns indicator measures to Wikidata units

Additional cross-concept links (sdg:relatedTo) connect indicators to external Q-IDs in domains such as "maternal health" or "clean water". During dataset ingestion, each column bearing an indicator code is annotated using the corresponding Wikidata Q-ID from the ontology, enabling dataset cataloging via sdg:indicator URIs, semantic filtering and query based on concept-level tagging, as well as automatic generation of metadata triples (e.g. linking dataset to indicators and units).

3 The Amazon Observatory and Other Pilots

The prominence of domains such as digital data processing and machine learning illustrates AI's multidimensional capacity to address complex challenges in resource allocation, public health systems, and environmental sustainability. Comparative analysis between global discourses and those specifically oriented toward the Brazilian Amazon—driven by the expertise and coordinated efforts of the MetAmazonia initiative—reveals a pronounced emphasis on environmental preservation, biodiversity monitoring, and climate resilience in the latter. This divergence indicates that AI's contributions to sustainable development are not uniform but instead conditioned by region-specific priorities, ecological constraints, and socio-technical contexts. These findings underscore the necessity of developing adaptive, context-aware AI frameworks capable of aligning with the heterogeneous demands of both urban and rural environments.

The Amazon Observatory delivers outcomes such as the MetAmazonia chatbot, a multidimensional open data platform, and accessible resources for students and researchers to advance knowledge and innovation in the region. The system will be the basis for the planned MetAmazonia Chatbot, leveraging these datasets within the broader SDG AI-agent development, aligned with open education principles and UNESCO collaboration. It aims to make knowledge resources directly useful for learners and professionals engaged with Amazonia and their communities. Table 1 shows the data feeding the system across a diversity of topics from news, science and policies, exposing concerns of the public opinion, the knowledge we hold on priority topics, and part of the regulatory landscape.

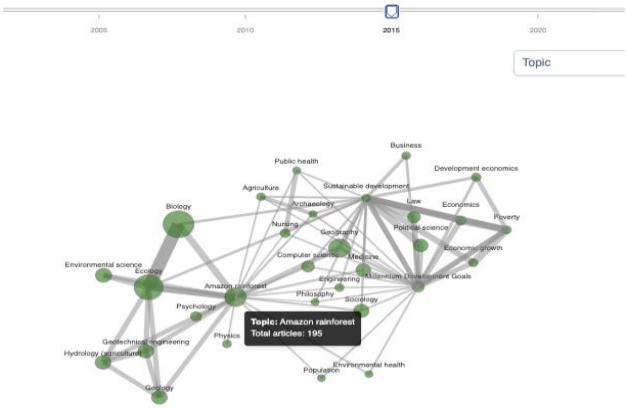

Figure 3: Evolution in time of the relation between research concepts related to the Brazilian Amazon Rainforest.

To illustrate the potential of such engine, five initial modules have been developed and are being made available for COP30 activities in Belem, at the heart of Amazonia: (i) the News Stream with Sentiment provides multilingual coverage of Amazonia-related news, complemented by word clouds of main concepts and sentiment analysis visualized through maps and gauges; (ii) the Data Exploration Dashboard integrates multiple datasets, displaying global research trends, SDG policy coverage, and innovation activity; (iii) the relation between the

concepts (edges) relevant to the Amazonia research and the interconnections between these concepts, being stronger or weaker according to the amount of published articles, where these are topics in common (see figure 3); (iv) in the Education view, the system visualizes open educational resources by mapping Amazonia-related topics to SDGs, highlighting key domains and their relevance to specific goals such as SDGs 11, 13, and 15; and (v) in regards to innovation ecosystems, we depict the different initiatives that relate to priority topics in the Brazilian Amazon context and could help establishing international collaboration to address specific problems with local and global data.

Table 1. Data ingested into the Amazon Observatory from worldwide news (indicating the language coverage), published AI-related scientific articles, and related legal and regulatory landscape

Concepts	2024 News (Lang. Coverage)	Science	Policy
Biodiversity	18083 (100)	44693	3628
Indigenous peoples	8070 (96)	2014	107
Bioeconomy	156 (16)	33	31
Carbon Credits	236 (26)	2127	133
Public Health	26454 (69)	42355	697
Amazon rainforest	3936 (87)	172	115

4 Conclusions and Further Work

As the global community continues to pursue the 2030 Agenda, the importance of robust, interoperable, and machine-actionable SDG data infrastructure has never been greater. This paper has explored the architecture and implementation of an AI-ready data infrastructure for the SDGs, using the IRCAI SDG Observatory and its derived pilots as case studies. Central to this infrastructure is a well-defined metadata schema, semantic alignment with Wikidata entities, and adherence to FAIR data principles—all designed to support automation, reasoning, and integration of data across domains and geographies. By embedding SDG indicators, targets, and goals into a linked-data framework, the system transforms static reporting datasets into dynamic, queryable resources. This enables a wide range of AI applications, from natural language querying to knowledge graph reasoning and real-time decision support. The SDG Ontology—based on mappings to Wikidata Q-IDs—serves as a semantic backbone, enabling interoperability with external datasets and ontologies while enhancing transparency and reusability. Despite these advancements, several challenges remain. Data fragmentation across jurisdictions, lack of standardization in national reporting, and uneven metadata quality continue to hinder full automation and scalability. Furthermore, ethical considerations around data use—particularly in the context of AI-based decision-making—require further exploration.

To improve the Amazon Observatory, future development of AI-ready data infrastructure will focus on several key areas: (i) Automated Ontology Expansion. Leveraging large language models and knowledge extraction tools to automate the discovery and integration of new SDG-related concepts from policy documents, scientific literature, and real-time news streams; (ii) Interoperability with National Platforms. Building tools that support seamless integration of local statistical data with global and local SDG indicators (e.g., focusing Amazonia), using schema mapping and automated alignment with the SDG ontology; (iii) Real-Time Data Ingestion and Streaming Analytics. Incorporating real-time data sources, such as remote sensing, sensor networks, and social media, to enable early-warning systems and near-instant progress monitoring; (iv) AI-Powered Decision Support Tools. Developing interfaces and tools that allow policy-makers to simulate interventions, explore causal relationships, and evaluate trade-offs between SDG targets using AI models trained on the structured data; (v) Community Governance and Open Collaboration. Establishing open, participatory governance models for ontology evolution, dataset curation, and quality assurance to ensure that the infrastructure remains globally relevant and inclusive.

In conclusion, AI-ready SDG infrastructure represents a transformative opportunity for evidence-based policy, global collaboration, and data-driven action on sustainability. By continuing to invest in semantic technologies, metadata standards, and open data ecosystems, we can enable a new generation of intelligent tools that accelerate progress toward the SDGs globally but also locally.

Acknowledgements

We thank the support of the European Commission projects ELIAS (GA101120237) and RAIDO (GA101135800)

References

[1] Bachmann, N., Tripathi, S., Brunner, M. and Jodlbauer, H.(2022). The contribution of data-driven technologies in achieving the sustainable development goals. Sustainability, 14(5), p.2497.
[2] Stahl, B.C., Schroeder, D. and Rodrigues, R., 2022. AI for Good and the SDGs. In Ethics of artificial intelligence: Case studies and Options for addressing ethical challenges (pp. 95-106). Cham: Springer International Publishing.
[3] Pita Costa, J., (2023) Water Intelligence to Support Decision Making, Operation Management and Water Education: NAIADES Report. IRCAI.
[4] Pita Costa J., Barrionuevo L.. Kovič Dine M. (2025) Observing the Impact of AI in the Progress of Sustainable Development Goal 11. Proceedings of the 23rd IADIS International Conference e-Society 2025
[5] Mitja Jermol, Joao Pita Costa and Matej Kovačič (2025) Onwards to an Ethical and Bias Aware Education for Sustainability through AI. Journal of Artificial Intelligence for Sustainable Development (to appear)
[6] Sustainable Development Solutions Network(2015) Indicators and a Monitoring Framework for the SDGs. United Nations.
[7] Joshi, A., Gonzalez Morales, L., Klarman, S., Stellato, A., Helton, A., & Lovell, S. (2019). A Knowledge Organization System for the United Nations Sustainable Development Goals. Proceedings of the 2019 International Conference on Knowledge Engineering and Knowledge Management (EKAW). Springer.
[8] Pita Costa, J. (2025) IRCAI SDG Ontology. GitHub. Available at https://github.com/IRCAI-SDGobservatory/data