Towards building a monitoring platform for a challenge-oriented smart specialisation with RIS3-MCAT.

Enric Fuster[1], Tatiana Fernández[2], Hermes Carretero[1], Nicolau Duran-Silva[1,3], Roger Guixé[1], Josep Pujol[1], Bernardo Rondelli[1], Guillem Rull[1], Marta Cortijo[2], Montserrat Romagosa[2]

[1]SIRIS Lab, Research Division of SIRIS Academic
[2]Ministry of Economy and Finance, Government of Catalonia
[3]LaSTUS Lab, TALN Group, Universitat Pompeu Fabra, Barcelona, Spain

Abstract

In the new research and innovation paradigm, aimed at transformation towards more sustainable, inclusive and fair pathways, it is essential to provide monitoring systems and tools to map and understand the contribution of research and innovation policies and projects to address societal and environmental challanges and generating new patterns of specialisation and new trajectories for socioeconomic development. To address these problems, we present the RIS3-MCAT platform, result of a line of work to explore the potential of open data, semantic analysis, and data visualisation, for monitoring challenge oriented smart specialisation in Catalonia. RIS3-MCAT is an interactive platform that facilitates access to R&I project data in formats that allow for sophisticated analysis of a large volume of texts, enabling the detailed study of thematic specializations and challenges beyond classical classification systems. Its conceptualisation, development framework and use are presented in this paper.

1. INTRODUCTION

The challenges posed by globalization, technology, climate change, and the COVID-19 pandemic require significant changes in our way of living. Although large transition costs are associated with a successful attainment of all those challenges, the potential opportunities brought about are enormous (Bigas *et al.*, 2021). The European Commission aims to accelerate the green transition by implementing the Green Deal (European Commission, 2019) and by allocating funds in the cohesion policy framework and the Horizon Europe (European Commission, 2021) program to mobilize European research and innovation ecosystems toward tackling outstanding societal challenges, such as the Sustainable Development Goals (SDGs) defined by the UN (United Nations, 2019). However, reaching these goals requires changes in the forms of cooperation between governments, companies, academia, and other societal stakeholders, new ways of combining knowledge from diverse disciplines, and new tools for evaluating the impact of public policy and research and innovation (R&I) (Bigas *et al.*, 2021).

In this same direction, strategies for smart specialisation (S3) (European Commission, 2014), which are dynamic agendas for economic and social transformation based on research and innovation and articulated through entrepreneurial discovery processes (EDP), are becoming extraordinarily important. Through the EDP, governments, companies, research and innovation stakeholders and civil society organisations and associations collaborate to identify challenges and priority areas for action and engage in transformative action towards more sustainable development pathways. In this context, it is key to develop new monitoring systems and tools

that help to understand how different actors in the research and innovation ecosystem are contributing to the SDGs and to accelerate transition towards more sustainable development pathways and, therefore, towards a smarter specialisation.

In Catalonia, smart specialisation has been defined "as a progressive, open process of concretion in which, through initiatives, collaborations and investments, stakeholders in the research and innovation system identify opportunities and define sub-sectors of specialisation or prioritisation" (Generalitat de Catalunya, 2022). In the current 2021-2027 programming period, the RIS3CAT monitoring system focuses on understanding transformative processes; in other words, understanding how the actions framed in this strategy contribute to:

- articulating sustainable value chains
- promoting business models aimed at generating shared value
- transforming goods and services delivery systems (socio-technical systems)
- fostering the creation of digital- and technology-based industry
- moving towards a greener, more digital, more resilient and fairer socio-economic model

These transformative processes are complex, as they involve interrelated changes in very different areas (such as the production systems, technologies, markets, regulations, user preferences, infrastructure, and cultural expectations). Accordingly, the monitoring system needs to include and combine different sources of information and types of analysis. Interactive visualisation tools integrating data from different sources are key to identify and analyse emerging areas of specialisation and collaboration networks (within the region and at EU level) in the RIS3CAT priority areas.

Today, EDP, policy implementation and monitoring may be greatly helped by taking advantage of the wider transformative trends in the fields of Open Government and Open Science (European Commission, 2016), which are making data potentially relevant for the public good increasingly available in open and usable formats (Fuster *et al*., 2020). Data on research and innovation activities is made available by a series of initiatives. The availability of this data is helpful for the identification of R&D niches and key actors within territorial R&I ecosystems that might be embarked in those transformative processes. At the same time, the exploitability of this data is increasing thanks to the advances in data science, artificial intelligence, and, particularly, in natural language processing techniques, which are being applied to scientometrics to characterize and analyse the textual content of R&I documents, and combine them with different sources of data (Fuster *et al*., 2020b).

The European Commission led the way by publishing the CORDIS database of European R&I projects. Since then, public administrations have promoted multiple initiatives in the same direction, and initiatives such as the European Open Science Cloud, OpenAire, or Zenodo already link projects and their funding with the results they generate (reports, publications, patents, software, etc.). In any case, unfortunately, the provision and maintenance of open data are highly unequal and do not cover the full range of needs of public policymakers. The availability of open data with sufficient granularity and richness remains a challenge, although it is becoming less limiting as science and technology databases grow in number, size, coverage, quality, interconnection, and content richness. Some major challenges faced at a policy level arise because many of those data sources are not openly available (undermining therefore the participatory processes), they are not interoperable in terms of data classification schemes and

institutional identification (therefore limiting transversal analyses) and they are hardly manageable by non--expert users.

In this framework, a line of work has been established to explore the potential of integrated open data, semantic analysis and data visualisation with the aim of developing methodological proposals for monitoring smart specialisation. This experiment, which tackled the challenges linked with the definition of indicators for monitoring emerging areas, territorial patterns of specialisation and collaboration dynamics between different stakeholders and areas of knowledge, led to the development of the RIS3-MCAT interactive platform[1], whose conceptualisation, development framework and use are presented in this paper.

This paper is organised as follows. Section 2 introduces policy objectives and main functions of the RIS3-MCAT monitoring platform. Section 3 presents the five-year co-design and development process. Section 4 presents data sources, system architecture and presents main results of components and features. Section 5 illustrate the main use cases. Finally, Section 6 draws conclusions and recommends future work directions.

2. POLICY OBJECTIVES AND MAIN FUNCTIONS

The RIS3-MCAT Platform is an interactive tool aimed at visualising, exploring and analysing the specialisation and collaboration patterns of research and innovation projects financed with European funds in Catalonia. It is an open government, artificial intelligence and data visualisation project that integrates and makes openly accessible and interoperable data from science and innovation projects, with the aim of contributing to the following objectives:

- understanding the impact of European funds on the specialisation of the R&I ecosystem of Catalonia,

- identifying opportunities to maximise the collective impact of R&I in Catalonia, based on synergies and the coordination of efforts

- providing new evidence that facilitates decision-making by stakeholders in the R&I ecosystem of Catalonia, promoting new dynamics of collaboration and inspiring new public policies;

- raising the profile of Catalan public and private actors that participate in R&I European networks;

- understanding the contribution of European funds to innovative responses to regional priorities, emergent thematics and the Sustainable Development Goals (SDGs).

Apart from the interactive visualisation tools, RIS3-MCAT provides all its curated and enriched data as open data, via dump downloads and via a 5* open data SPARQL Console.

This has facilitated the elaboration of several complementary policy monitoring as well as transversal and thematic analytical reports, published under the "Monitoring RIS3CAT" document collection[2].

---

[1] Temporary available at: http://sirislab.com/lab/ris3/mcat/ (this link will be updated with the final link).

[2] Available at: https://fonseuropeus.gencat.cat/ca/ris3cat/2030/monitoratge/index.html#googtrans(ca|en)

Figure 1: Screenshot from RIS3-MCAT monitoring platform - Network view.
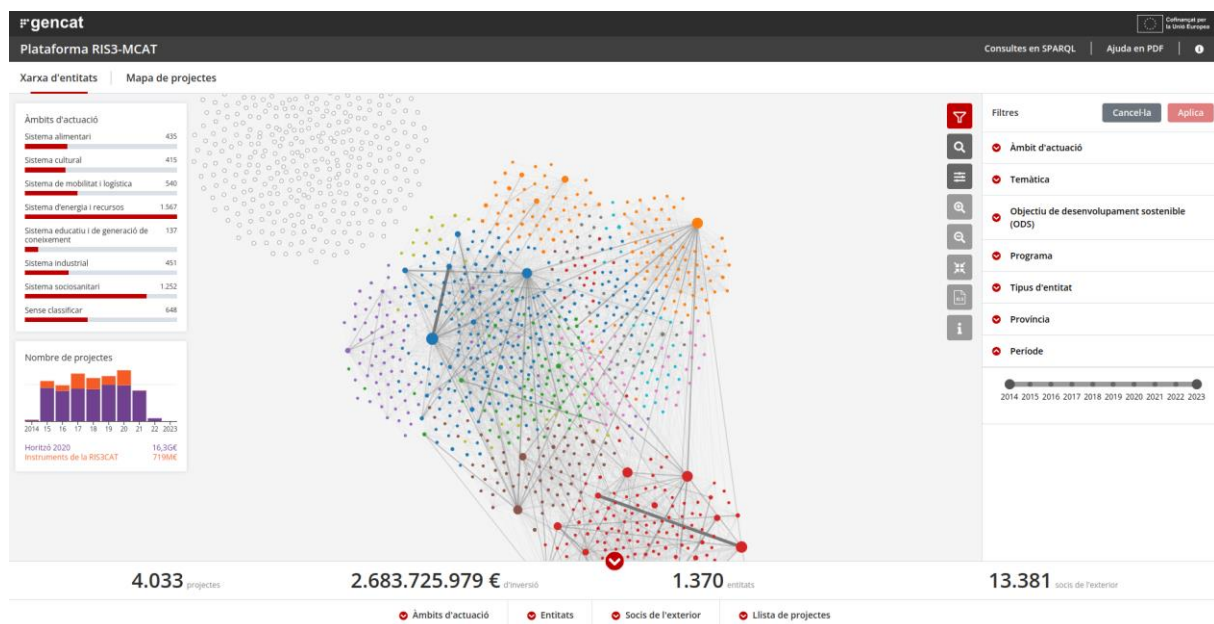
3. A FIVE-YEAR CODESIGN AND DEVELOPMENT PROCESS

As of March 31st, 2023, more than 5,000 unique users from 73 countries have accessed the platform, with an average session duration of over 4 minutes. RIS3-MCAT is currently undergoing a major redesign, with new functionalities being introduced to facilitate R&I portfolio analysis, an essential requirement in the new programming period. The new version is set to be published on April 30, 2023. This is just another step in a long process of co-design, review, evaluation, and iterative development that began in 2017, as outlined in the timeline below:

- 2017: Requirement analysis and feasibility study, focusing on data availability and quality, as well as prospective front-end and back-end technologies and solutions.

- 2018: Proof of Concept development, based on a wide co-design process within the R&I related departments of Generalitat de Catalunya. This led to the first officially published live version, with manually classified Horizon Europe and RIS3CAT R&I projects, focusing on S3 priority analytics and collaboration networks.

- 2019-2020: Consolidation of the Proof of Concept into a fully finished product. Development and inclusion of the SDG project classification. First development of an analytical report, based on the RIS3CAT taxonomy and an automatic identification of main themes via topic modelling (machine learning), to support the decision making around the evolution of the RIS3CAT regional priorities for the new programming period. Provision of data for the S3 monitoring official reports. Participatory development of thematic analysis in three domains: Circular Bioeconomy, Artificial Intelligence, and Plastic Waste Reduction.

- 2021: Participatory review and requirement analysis with Catalan R&I stakeholders , leading to the development of new functionalities, particularly: supporting the identification and analysis of international and inter-regional collaboration and improving the bulk download of the underlying data.

- 2022-early 2023: Adaptation to RIS3CAT 2021-2030, the new policy framework, by automatically reclassifying all projects into the new RIS3CAT regional priorities via deep learning classifiers. Major review and redevelopment of the platform, updating the front-end technology, and streamlining the design patterns. First integration of an emergent classification, the "Thematics," based on topic modelling (deep learning). Integration of the first Horizon Europe projects. Addition of the "Thematic project mapping" platform view, which presents all RIS3MCAT R&I projects in a single visualisation, organised by semantic similarity. This being the first new view introduced in the platform since 2018, it opens the door for further possibilities, such as purpose-built benchmarking tools or geographic maps.

4. TOOL OVERVIEW

4.1. Data sources

CORDIS. Data and metadata related to R&D projects and related organisations which have received funding by the European Commission under the H2020 and FP7 framework programs. The data are accessible through Open Data license, and monthly updated, afterwards provided in the format of CSV, XML and Linked Open data in the CORDIS website. We collect the CORDIS records from UNiCS (Giménez, 2018), an open data platform based on semantic technologies for science and innovation policies which include data cleaning and improved geographical identifications of participants, not always correct in the original datasets.

SIFECAT. The information system used by Generalitat de Catalunya to manage ERDF operation, and the main data source for regionally-managed R&I projects. Most of the data is useful from the raw document and is much less error-prone, specially on location, but some transformations have to be derived to fit the data integration ontology and the front-end requirements.

4.2. RIS3-MCAT architecture

Data integration & cleaning. CORDIS records and projects funded by the Catalan region and ERDF are integrated into different database schemas that are then homogenised into a unified structure with a domain ontology[3]. Access to the data is then done through the ontology by means of the Virtual Knowledge Graph system Ontop (Calvanese *et al.*, 2017), which translates input queries formulated over the ontology into executable queries formulated over the underlying database. In order for these different datasets to be integrated properly, the project beneficiaries must be given an identifier that homogenises differences in spelling. The goal is that a same organisation must have the same identifier, even if its name is written differently in each dataset. To this effect, a process of semi-manual disambiguation has been performed by our experts, using the OpenRefine tool[4]. Additionally, for SIFECAT data, each beneficiary has been annotated with the corresponding type of organisation.

---

[3] Ontology schema available here: https://s3-eu-west-1.amazonaws.com/ontology-documentation-ris3cat/index.html

[4] https://github.com/OpenRefine/OpenRefine

Automatic classification of projects according to the regional priorities. This makes it to capture "top-down" domains when cases where a deeper understanding of the regional dynamics within a specific research area defined a priori was needed. We take advantage of pretrained language models for training textual classifiers based on title and project description, one per each of the seven priority "systems" or areas of application by the RIS3CAT 2030 (Generalitat de Catalunya, 2022). A training set is built for each domain and annotated based on weak-supervision and active learning paradigms, which allow weak annotation of projects from a general sample of R&I projects from both FP7 and H2020 frameworks based on some of their metadata[5]. We take advantage of the tool Argilla (Vila & Aranda, 2023) for annotation and label improving, which allow keep human-in-the loop and interactive feedback from experts to improve label quality. Our best models are based on Specter (Cohan *et al.*, 2020), and it was implemented with Hugging Face Transformers library (Wolf *et al.*, 2020). From an evaluation on a sample of 500 Catalan projects, predicted labels are compared with human labels from 2 experts, obtaining a macro-averaged f1 of 88.1% of accuracy.

Topic modelling. Topic Modelling (TM) is an unsupervised classification problem in machine learning that aims at *discovering* the unknown topics linked with a specific collection of texts, presenting a "bottom-up" picture of the thematics tackled within a specific R&I community. This component takes the project title and abstract, encodes the semantic representation of them based on the Specter sentence-transformer model (Cohan *et al.*, 2020) which was pretrained on scientific documents, and from clustering vectorial representation of documents based on k-means, we obtain groups of similar projects. Each document is linked to a cluster, and the number of clusters is decided by qualitatively choosing the best trade off between the semantic "richness" of the topics and the overall number of topics (in order not to have neither too large topics nor too little ones) on different runs. Names of clusters are added manually based on keywords frequency and by exploring samples of projects.

SDG classification. We identify SDG-related research by using a collection of SDG keywords (a controlled vocabulary) openly available in Zenodo (Duran-Silva *et al.*, 2019), based on a hybrid approach that use automatic methods for enriching human-crafted keywords. R&I projects are tagged with VocTagger tool[6] on their title and project description.

Web front-end. RIS3-MCAT UI and its visualisations were implemented following the W3C standards and using a combination of HTML, CSS and JavaScript, also exploiting third-party libraries (e.g., the D3.js and React.js libraries). The projects and participants information data enrichment is retrieved by querying the SPARQL endpoint.

4.3. RIS3-MCAT features

The RIS3-MCAT front-end is composed of four main parts: navigation bar, main visualisation canvas, operations toolbar and statistical modules. Each of the parts is fully reactive to the user interaction. The main visualisation canvas has 2 different representations of the data (that can be explored by the navigation bar), an collaboration network of institutions, and a semantic cartography of projects.

Search & Filters. To facilitate data exploration, the platform offers users various filtering and search options, allowing them to generate customized visualizations. Initially, the Platform

---

[5] Taking advantage of: EC Area, ERC Panel, EC programme, Topic, and Field of Study.

[6] https://github.com/sirisacademic/VocTagger

displays all the integrated data. However, this visualization can be restricted to subsets of data by applying filters and search parameters or by directly manipulating the network. Filtering features include search by: keyword, participant name, institution type, year province, instrument and programme name, area of action, emerging topic, and SDG. Search features include search by participant and project search based on keyword or text search on title and abstract. All filters are multi selection filters, and all filters work combined with searches to allow users to define a set of exploration of interest.

Network analysis. The network of participant shows the collaboration of Catalan R&I actors. Each node of the network represents a R&I actors with its legal headquarters in Catalonia, and the size of the node is proportional to the volume of the entity's investment in the projects. When two entities collaborate on projects, the nodes that represent them are joined with a line. The size of this line is proportional to the number of projects they share. Force directed graph (network) showing the relationships between participants and defined by the collaborations on projects.

Semantic map of projects. A presentation of the R&I projects in a 2D canvas based on semantic similarity resulting in a topography of the R&I activity, which facilitates the identification of similar projects and a view of the proximity between thematics and the relationship and overlap between taxonomies. It is a T-SNE dimensionality reduction of the embeddings obtained in the topic modelling module, and its clusters and names.

Analytical/Statistical modules. The information modules, presented at the bottom of the tool, extend the textual and statistical information of projects and participants and their classifications, displaying distributions and relationships. They are reactive to the filtering operations. The present types are: summary indicators, rankings of participants, external partners, and projects table view. Different project information modules show extended information of the projects, and they are available from different parts of the application.

Data download & SPARQL Endpoint. The platform offers the possibility of downloading the data filtered interactively as a CSV file, or of making queries about all the data included using SPARQL. Data download is available (XLS format) for the current state of the project's and participants visualisations as well as the regional partners and international partners in the statistical modules.

5. ILLUSTRATIVE USE CASES

We have identified interesting scenarios of use by different target users/actors in the territory, with short descriptions for the four illustrative use cases.

- **Use case 1: Search by actors and projects in similar topics (search for expertise and possible collaborations).** Target users for this use case could be R&I stakeholders, such as researchers or private companies.

Figure 2: Screenshot from RIS3-MCAT monitoring platform - Use case 1
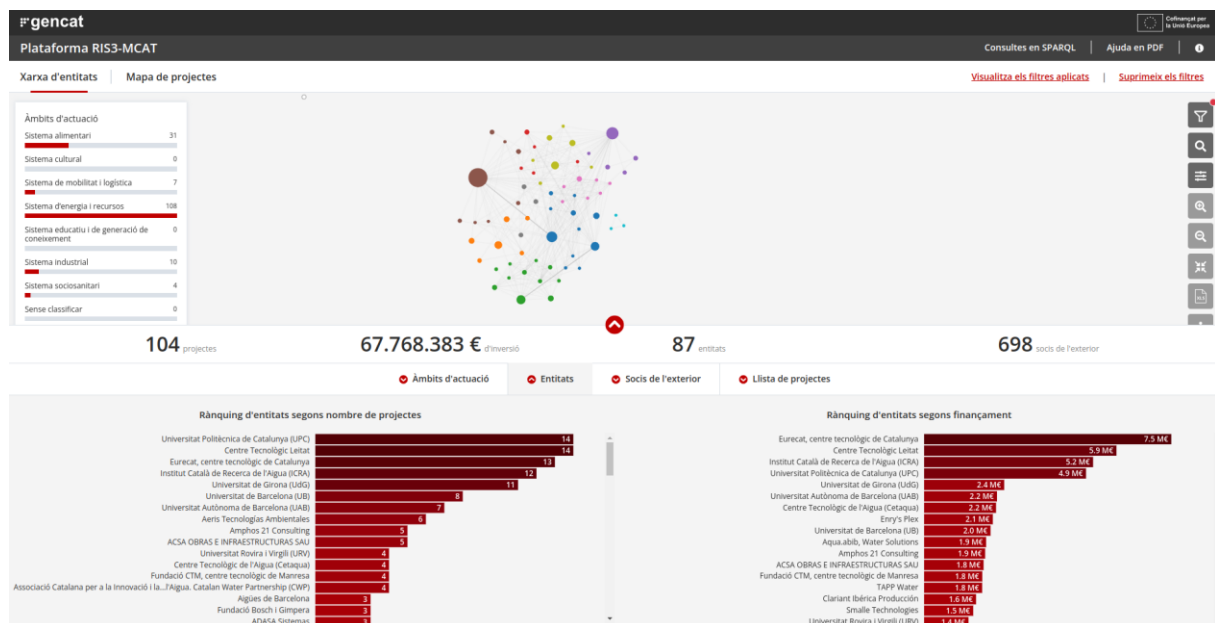
- **Use case 2: Collaboration network for the identification of actors within a thematic and geographic scope.** Target users for this use case could be a R&I policy-makers, with transversal, or thematic / territorial responsibilities.

Figure 3: Screenshot from RIS3-MCAT monitoring platform - Use case 2
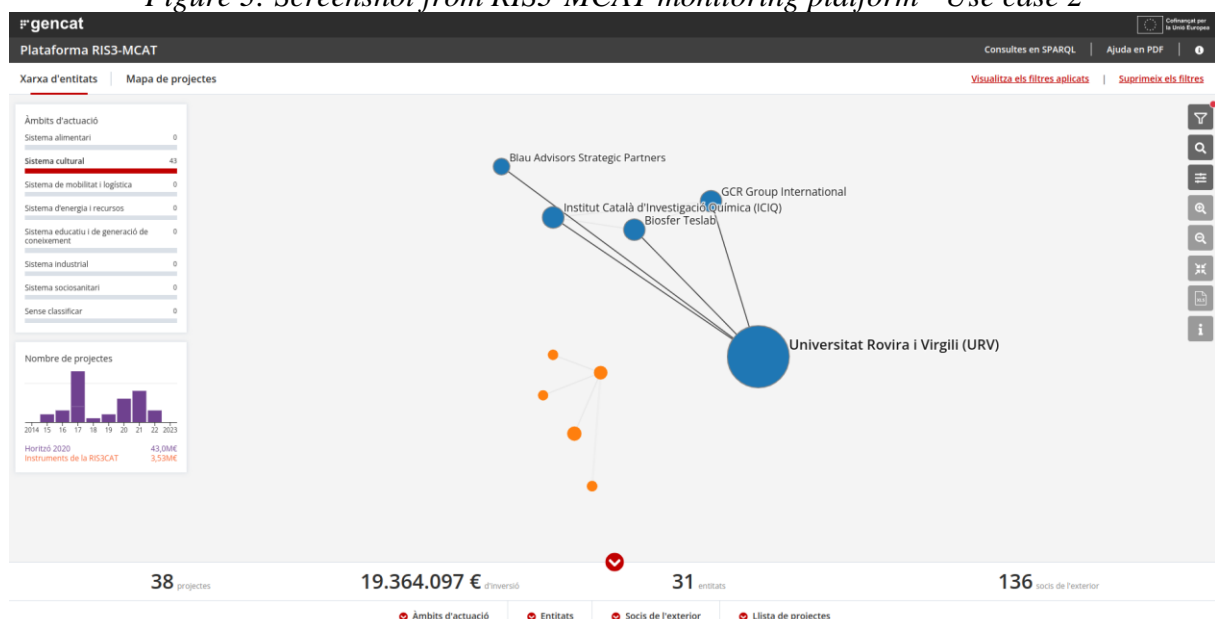
- **Use case 3: State of research for priority area.** This use case allows the study of the intersection between the top-down priority area and emerging topics, in order to find projects in the "core" of the priority, as well as those that are more interdisciplinary/intersectoral. The target user could be a R&I challenge-oriented initiative, leader, applying systems thinking approaches to explore the full perimeter of a societal priority or challenge. We can find in the "core" of the priority projects such as "METROFOOD-RI Preparatory Phase Project" or "Connecting the dots to unleash the innovation potential for digital transformation of the European agri-food sector"; and, in the periphery, projects like "Empowering consumers to PREVENT diet-related diseases through OMICS sciences" in health domain, or "Advanced Multi-

Constellation EGNSS Augmentation and Monitoring Network and its Application in Precision Agriculture" in satellite navigation. Figure 4 captures an example of this use case.

Figure 4: Screenshot from RIS3-MCAT monitoring platform - Use case 3
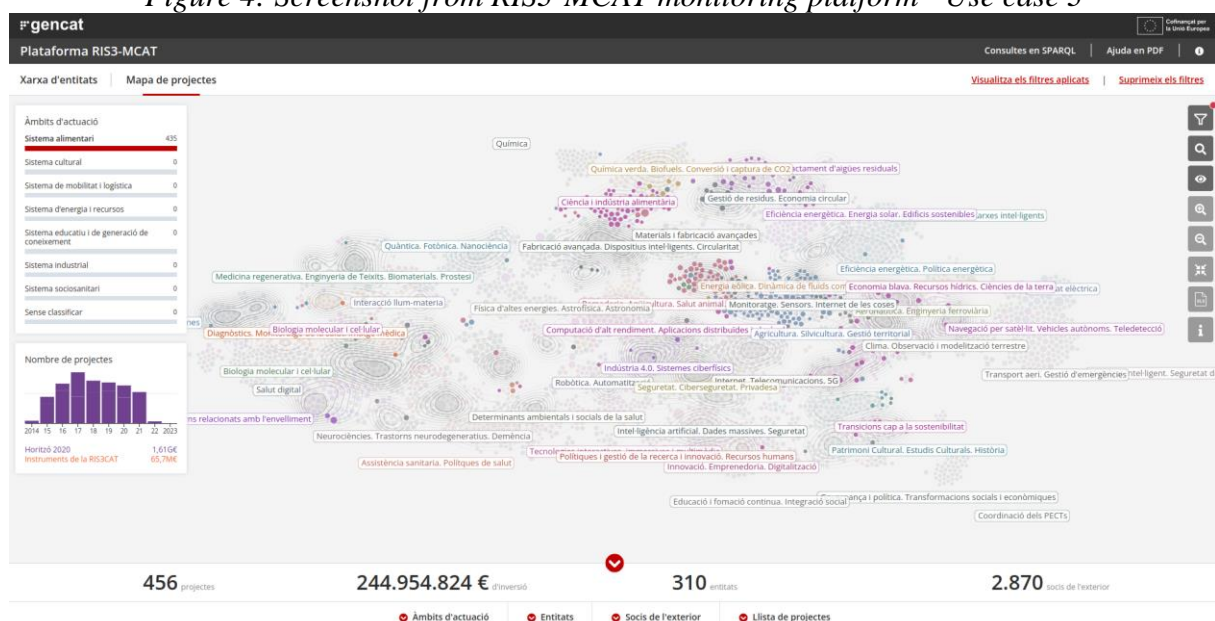
- **Use case 4. Collaboration promotion.** This use case allows the identification of current partners in other countries/regions (and their counterparts in Catalonia) by topic. The target user are mainly internationalization policy-makers.

Figure 5: Screenshot from RIS3-MCAT monitoring platform - Use case 4
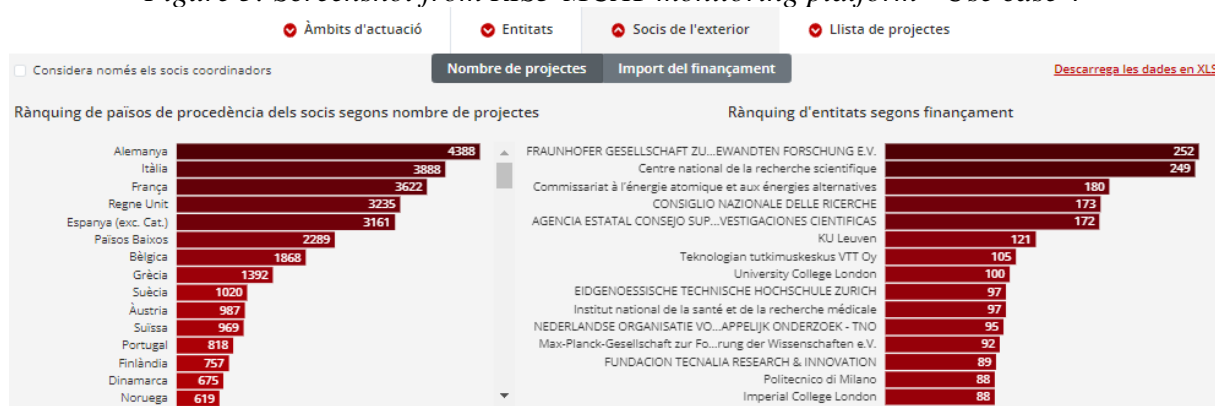
6. CONCLUSIONS AND FUTURE WORK

We introduce the new monitoring platform RIS3-MCAT, and addresses the questions and objective of monitoring, co-design and construction of monitoring platforms for a challenge-oriented smart specialisation with available technologies and open data. Putting the methodology, techniques, and knowledge at the service of questions of policymakers.. Furthermore, how to develop of interactive and exploratory visualisation tools providing an entry-point to complex and highly-intertwined data-sources.

Incorporating new projects and data sources is essential to keep the research and innovation (R&I) landscape up to date in the monitoring platform. The introduction of call-based funded R&I, other departments funded R&I projects, shared agendas, programme contracts, and direct public administration initiatives are excellent examples of expanding the scope of R&I projects for the future. Additionally, considering additional thematic niches and functional classifications, such as type of innovation/transformation in six TIPC dimensions (Geels, 2002; Geels, 2004), would help to capture a wider range of innovation efforts. Furthermore, there is the need to continue exploring additional ways to enhance visualizations, functionalities, and analytics. For instance, incorporating a geographic map, KPI analysis, European benchmarking, and interregional collaboration analysis can provide additional insights and perspectives on R&I projects.

Open science practices

This platform aims at promoting open science policies, allowing the exploration R&I activities in the region and encouraging collaboration. Our project itself is based and generates open data, and we have made intermediate reports about our codesign and lessons learned, publicly available. In our exploration and data integration, we have used open data sources and all data generated is available via SPARQL and CSV formats. Additionally, we have used open source technologies available on GitHub and the data of the platform is under a CC0 licence. Tye SDG vocabularies and VocTagger, partially developed in this context, are available on GitHub and Zenodo. Open science practices are crucial to advance research and innovation, but also public policies. For this reason, this article is intended to explain, formalise, and communicate the results, decision, and process of this research, so that other actors and regions can benefit in their own initiatives.

Acknowledgments

Competing interests

This article is authored by the key responsibles of RIS3MCAT at Generalitat de Catalunya alongside its private sector providers (SIRIS Academic). We believe we do not have competing interests.

References

Bigas, E., Duran, N., Fuster, E., Parra, C., Cortini, R., Massucci, F., Quinquillà, A., Fernández, T., Romagosa, M., & Cortijo, M. (2021). Monitoring smart specialisation with open data and semantic techniques. "RIS3CAT Monitoring" collection, number 16. http://catalunya2020.gencat.cat/web/.content/00_catalunya2020/Documents/angles/fitxers/monitoratge-ris3cat-dades-obertes-tecniques-semantiques-en.pdf

Calvanese, D., Cogrel, B., Komla-Ebri, S., Kontchakov, R., Lanti, D., Rezk, M., ... & Xiao, G. (2017). Ontop: Answering SPARQL queries over relational databases. Semantic Web, 8(3), 471-487.

Cohan, A., Feldman, S., Beltagy, I., Downey, D., & Weld, D. S. (2020). Specter: Document-level representation learning using citation-informed transformers. arXiv preprint arXiv:2004.07180.

Duran-Silva, Nicolau, Fuster, Enric, Massucci, Francesco Alessandro, & Quinquillà, Arnau. (2019). A controlled vocabulary defining the semantic perimeter of Sustainable Development Goals (1.2) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.3567769

European Commission. (2014). Research Innovation Strategies for Smart Specialisation. Cohesion Policy. https://ec.europa.eu/regional_policy/sources/docgener/informat/2014/smart_specialisation_en.pdf

European Commission. (2016). *Open innovation, open science, open to the world: a vision for Europe*, Publications Office. European Commission, Directorate-General for Research and Innovation.

European Commission. (2019). The European green deal. Eur. Comm., 53(9), 24.

European Commission (2021). Horizon Europe. https://op.europa.eu/en/publication-detail/-/publication/3c6ffd74-8ac3-11eb-b85c-01aa75ed71a1

Fuster, E., Marinelli, E., Plaud, S., Quinquilla, A., & Massucci, F. (2020). Open Data, Open Science and Open Innovation for Smart Specialisation monitoring, EUR 30089 EN, Publications Office of the European Union, Luxembourg, 2020, ISBN 978-92-76-10726-2, doi:10.2760/55098, JRC119687. https://publications.jrc.ec.europa.eu/repository/handle/JRC119687

Fuster, E., Massucci, F., & Matusiak, M. (2020). Identifying specialisation domains beyond taxonomies: mapping scientific and technological domains of specialisation via semantic analyses. In R. Capello, A. Kleibrink, & M. Matusiak (Eds.), Quantitative Methods for Place-Based In-novation Policy (pp. 195–234).

Geels, F. W. (2002). Technological transitions as evolutionary reconfiguration processes: A multi-level perspective and a case-study. Research Policy, 31(8–9), 1257–1274. https://doi.org/10.1016/S0048-7333(02)00062-8

Geels, F. W. (2004). From sectoral systems of innovation to socio-technical systems: Insights about dynamics and change from sociology and institutional theory. Research Policy, 33(6–7), 897–920. https://doi.org/10.1016/j.respol.2004.01.015

Generalitat de Catalunya. (2022). RIS3CAT 2030: Strategy for the Smart Specialisation of Catalonia 2030. https://fonseuropeus.gencat.cat/web/.content/ris3cat/documents/angles/ris3cat-2030-en.pdf

Gimenez, X., Mosca, A., Roda, F., Rondelli, B., & Rull, G. (2018). UNiCS: The open data platform for Research and Innovation?. In Proceedings of the Posters and Demos Track of the 14th International Conference on Semantic Systemsco-located with the 14th International Conference on Semantic Systems (SEMANTiCS 2018) (Vol. 2198). CEUR-WS.

United Nations (2019). Global indicator framework for the sustainable development goals and targets of the 2030 agenda for sustainable development. Developmental Science and Sustainable Development Goals for Children and Youth, 439.

Vila, D., Aranda, F. (2023). Argilla - Open-source framework for data-centric NLP (Version 1.2.0) [Computer software]. https://github.com/argilla-io/argilla

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., & Rush, A. M. (2020). Transformers: State-of-the-Art Natural Language Processing [Conference paper]. 38–45. https://www.aclweb.org/anthology/2020.emnlp-demos.6