# RISIS final report. An overview of one decade of construction of a distributed European infrastructure on STI datasets and positioning indicators.

Philippe Laredo

LISIS
LABORATOIRE INTERDISCIPLINAIRE
Sciences Innovations Sociétés

RISIS


RESEARCH INFRASTRUCTURE FOR SCIENCE
AND INNOVATION POLICY STUDIES

# FINAL REPORT

## AN OVERVIEW OF ONE DECADE OF CONSTRUCTION OF A DISTRIBUTED EUROPEAN INFRASTRUCTURE ON STI DATASETS AND POSITIONING INDICATORS

Philippe Laredo
Université Gustave Eiffel
RISIS Project coordinator

Groupe Thématique Innovation Working Paper

Number 2 | March 2024

Required citation:

Laredo, Philippe (2024) RISIS final report. An overview of one decade of construction of a distributed European infrastructure on STI datasets and positioning indicators. Champs-sur-Marne, France: LISIS.

http://doi.org/10.5281/zenodo.13293265



RISIS

RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 824091

**Laboratoire Insterdisciplinaire Sciences Innovations et Sociétés (LISIS)**
UMR CNRS (No. 9003) - INRAE (No. 1326) - ESIEE Paris - UGE
Bâtiment Albert Camus | 2 allée Jean Renoir | 93160 Noisy-le-Grand

Postal address:
Université Gustave Eiffel
Cité Descartes
5, boulevard Descartes
Champs-sur-Marne
77454 MARNE-LA-VALLÉE Cedex 02
Telephone: +33 (0)1.60.95.71.89
Fax: +33 (0)1.60.95.72.38
Web: www.umr-lisis.fr
Email: direction@umr-lisis.fr

LISIS

LABORATOIRE INTERDISCIPLINAIRE
Sciences Innovations Sociétés

# FINAL REPORT

## AN OVERVIEW OF ONE DECADE OF CONSTRUCTION OF A DISTRIBUTED EUROPEAN INFRASTRUCTURE ON STI DATASETS AND POSITIONING INDICATORS

Philippe Laredo[1]

[1] Université Gustave Eiffel, LISIS, CNRS, INRAE, F 77454 Champs-sur-Marne, France

RISIS


RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

# Outline

## Introduction

RISIS is a European research infrastructure promoted by 20 European academic institutions, operated by researchers themselves and co-financed by two successive FP projects. Its objective is to support the development of positioning indicators for informing science technology and innovation policies. It combines different datasets answering FAIR principles and that are all harmonised thanks to a set of integrating services. Most datasets in the STI world being mainly textual, RISIS integrates open access semantic analysis and visualisation tools. Finally on key policy topics RISIS has developed interfacing tools that offer access to standardised positioning indicators.

RISIS is dedicated to 'publishable research', our specific way to inscribe itself in the open science movement. Users must follow the RISIS code of conduct to access our resources. RISIS services and some datasets are free access, but most datasets, for legal reasons, require that the corresponding projects are accepted by the overall RISIS access operator. A specific RISIS Core Facility enables users to access datasets and services and work online developing their own 'scenarios' for data integration, enrichment, processing and visualisation.

**Box 1: The academic institutions supporting or having supported RISIS across countries**

Austria: AIT, Joanneum

Czech Republic: TC CAS

France: Université Gustave Eiffel (UGE) – CNRS – INRAE

Germany: Fraunhofer, DZHW

Greece: ATHINA-OpenAire

Israel: Technion

Italy: CNR, Politecnico de Milano

The Netherlands: University of Leiden, Amsterdam Free University (RISIS1)

Norway: NIFU

Spain: CSIC

Switzerland: Univ della Svizzera Italiana (USI)

UK: University of Sussex – University of Sheffield – University of Manchester (RISIS1) and Strathclyde University

RISIS as an infrastructure is the outcome of nearly 20 years of research investment. The approach to RISIS (with the central notion of positioning indicators) was born within the PRIME Network of excellence (2004-2010). RISIS1 (2014-2018) was dedicated to demonstrating the feasibility of a research infrastructure: professionalise, harmonise and enrich relevant research datasets, start developing integrating services, opening them systematically for transnational access, start developing integrating services, test the interest of the research community. This being successfully done, RISIS2 (2019-2023) had the ambition of and in our view succeeded in turning the infrastructure fully operational and in extending its coverage and use. This explains why core participating institutions have decided to maintain RISIS in the long term and for this purpose are creating an international NGO under the Belgian law, RISIS association, to operate the infrastructure with a central feature based upon in-kind contributions.

## What is RISIS today after such investment?

RISIS gathers

- 14 core datasets accessible via a unique entry point and user-oriented platform, the RISIS Core Facility (RCF). They cover 4 main areas (see Box 2): (i) STI outputs with publications, patents, trademarks, publicly funded projects, and social innovation projects; (ii) firm innovation activities covering world largest firms, European venture capital backed start-up firms, and European fast growing mid-sized firms; (iii) public research datasets covering organisations (with a special focus on universities) and careers; (iv) policy learning with the SIPER repository of policy evaluations and the EFIL database for the analysis of policy portfolios.
- 3 core integrating services dealing with annotation at actor level (the ORGREG and FIRMREG reference DB also included in box 2), geospatial level (CORTEXT Geo) and thematic level (the RISIS SDG topic classifier and the RISIS SDG landscape explorer),
- 3 core platforms for semantic analysis (CORTEXT and GATE) and for visualisation (VOSVIEWER),
- a very active training programme based upon a monthly research seminar, advanced methods courses, and periodic data science summer schools,
- and active processes to connect with the broader STI community (in particular linked to the EUSPRI forum that binds the community), policy analysts and policy stakeholders.

### Box 2: RISIS datasets

| Dataset | Content | Coordinator |
|---|---|---|
| | **Output oriented datasets** | |
| CWTS Publication | Enhanced and vastly enriched copy of Web of Science (WoS) (standardized organization names, and other improvements or standardisations) | UL |
| RISIS Patent | Enriched and cleaned version of the PATSTAT database focused on priority patents, with standardized names and geolocalisation and other | UGE |
| ISI-TM* | The ISI-Trademark Data Collection (ISI-TM) provides detailed information on trademarks filed at the EUIPO and at the USPTO | Fraunhofer |
| EUPRO | Systematic and standardized information on R&D projects of different European and national R&D policy programmes | AIT |
| ESID* | ESID is a comprehensive source of information on social innovation projects and actors in Europe and beyond | Strathclyde |
| | **Firm oriented datasets** | |
| VICO | Geographical, industry and accounting information on start-ups that received at least one venture capital investment | POLIMI |
| CIB / CinnoB** | Database about largest R&D performers and their subsidiaries, providing patenting and other indicators | UGE |
| Cheetah | Geographical, industry and accounting information on mid-sized firms that experienced fast growth | POLIMI |
| FIRMREG** | Reference dataset providing unique identifiers for firms present in RISIS datasets with a set of aggregated core indicators | US-POLIMI |

| | | |
|---|---|---|
| | **Public research datasets** | |
| RISIS-ETER | Extension by additional indicators in terms of research activities of the European Tertiary Education Register | USI |
| DDC* | The Doctoral Degree and Career Dataset (DDC) is an experimental dissertation-centric database | NIFU |
| ORGREG** | Reference dataset providing unique identifiers for European public research organisations dealing with their dynamics (in particular mergers) with a set of aggregated core output indicators | USI-JOANNEUM |
| More*** | Integration of empirical studies on researcher's mobility in Europe | NIFU |
| Profile*** | Longitudinal study on the situation of doctoral candidates and their postdoctoral professional careers at German universities | DZHW |
| | **Policy learning datasets** | |
| SIPER | Unique repository and knowledge source of science and innovation policy evaluations worldwide | Fraunhofer |
| EFIL* | EFIL provides data for characterizing research funding instruments managed by selected European RFOs | CNR |
| JOREP*** | European trans-national joint R&D programmes, storing a basic set of descriptors on the programmes and agencies | CNR |

** new dataset developed during the RISIS2 project*
*** reference databases comprising coreset of positioning indicators*
**** databases accessible but no longer maintained in RISIS2*

**Box 3: RISIS services**

| Facility/ platform | Content | Coordinator |
|---|---|---|
| RCF | RISIS Core Facility enables integrated distant access, provides user workspace and offers processes for data integration, enrichment, treatment and visualisations (through scenarios) | UGE |
| CORTEXT Geo | Platforms for geocoding and geospatial allocation (especially for metropolitan areas) to large size corpuses | UGE |
| SDG classifier & landscape explorer | The classifier enables entity tagging of SDG in all types of datasets (including very large ones) and the SDG landscape explorer enables users to select their own definition of a SDG or a SDG topic | Sheffield & UL |
| CORTEXT Semantic Platform | The platform offers an integrated process for semantic analysis | UGE |
| GATE platform | GATE offers a large variety of tools and software for natural language processing | Sheffield |
| VOSviewer | VOSviewer is a visualisation tool for the analysis of publications | UL |
| RISIS VRE | The RISIS VRE is part of the D4Science developments for virtual research environments. It offers a privileged connection for the exploration of OpenAire | CNR |

The full final report, which is about 250 pages,[1] is organised in three parts:

Part 1 is an overall presentation of what RISIS has achieved and where it goes.

Part 2 gathers the reports done by each element (dataset and/or service) about what they are, how they have developed or deepened, and presenting exemplary cases of use.

Part 3 presents the WP of the project one by one with their activities and achievements.

This document presents only Part 1.

.

[1] The full report is available upon request.

# RISIS


RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

## PART I
## WHAT HAS THE SUPPORT OF THE EC FOR THE OPERATIONALISATION OF RISIS ENABLED?

## (RISIS2 PROJECT)

The objective of this second EC support was to build an 'advanced research community' manifested by a set of achievements that this first part of the report will review (part 3 will deal with them following the work package structure).

Before addressing them, we need to come back to the dynamics of the project itself, which was heavily hurt, as numerous research developments, by the COVID crisis. The crisis has had three impacts that have driven to a very important amendment signed in 2022.

First and foremost, it did not change the ambition nor the objectives of the project. It postponed them by one year, extending the project until the end of 2023.

Second it drove to move online far more rapidly than anticipated, with a strong impact on the very important training effort which had to be largely redesigned, quite successfully we shall see later.

Third the move online of training activities offered financial possibilities to address the core recommendation of the two very positive interim reviews that were made. They pushed us to be more pro-active in the move for providing a coreset of positioning indicators, while this third dimension was not initially part of the project, leaving these developments to users of our infrastructure. We shall detail in the review of our achievements, the results arrived at for this third pillar of RISIS activities after databases and services.

## User access as the core indicator of success

*(the final report on access, deliverable D8.5, provides more detailed information)*

The core performance indicator for the infrastructure is the number of project-based accesses asked for. Note that this refers to research projects or studies by researchers asking for access to raw micro-data of one or more RISIS datasets (see Box 2). The ambition was to receive 400 requests during the life of the RISIS2 project (which has been considered as a quite high number given the goal of 100 in RISIS1), we achieved **436 requests** mid-December 2023 out of which **94% were distant accesses**, far above our initial expectations. Similarly, one key indicator of the 'power of integration' provided by the infrastructure lied in the share of **requests for multiple datasets**. While we expected 33% the last year, this overall share has evolved from about one fifth (21%) until M24, to 31% until M36, further to 38,5% until M48, and to **43%** to M60. We have thus largely exceeded our objectives, demonstrating the validity of the approach developed for interoperability and integration built upon the RISIS positioning triangle of actors, geography and themes.

The process adopted has played a central role in this result. It combines a data request platform (part of the RCF, see below) on which all elements for each access request are stored, with a completely revised internal process compared to RISIS1. In view of the development of requests for multiple datasets, a central access operator was created that works in conjunction with the 'access managers' of each dataset. For the evaluation of requests, a 'project review board' has been created with an external president (for RISIS2 Rémi Barré) who operates the external review process and is in charge, when needed, of selecting the specialised external reviewers. This flow worked really well and enabled most requests to be handled within less than two weeks. As the RCF took longer than expected to become operational, most distant accesses took place, until 2023, by data transfer via e-mail or other file transfer mechanisms, even if this was far more time consuming for teams, and in particular for access managers.

The rate of accepted applications for the 5 years has been 83%. The main reasons for rejecting incoming proposals have been an insufficient description of how the requested data can support the proposed research, or even a scientifically questionable or commercially intended usage of RISIS data by users not covered by the Code of Conduct. A few proposals failed as the proposed projects lacked scientific novelty and originality.

## What datasets are used, by whom and for what?

### Datasets used.

Table 1 shows the attractiveness of output datasets. This attractiveness has been reinforced by the progressive addition of two datasets, and the deepening of the three pre-existing ones inherited from RISIS1 (see specific section below). They represent just over half the requests. Nearly one third of requests is linked to the unique feature of RISIS2, its 3 datasets on firms and especially the dataset on venture capital backed firms (VICO). The use of the two other dimensions is vastly underestimated as two of the core datasets are freely accessible: SIPER and mostly ETER, which is widely used as is demonstrated in the ETER report in part2. Also, the very recent opening of DDC for careers and EFIL for policy portfolios of funding agencies gives very important hopes of wider use in the coming years.

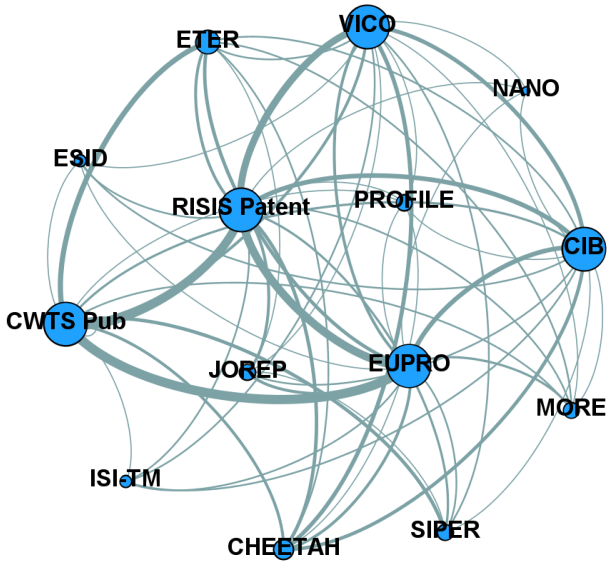
These features are further reinforced by the linkages between datasets provided by requests for multiple datasets. Figure 1 shows the very rich linkage structure that is dominated by 2 triangles: RISIS-Patent, CWTS publication and EUPRO on R&D activities and outputs; RISIS Patent, VICO and CIB on firms.

**Table 1: dataset requests**

| STI Output DB | | Firm innovation DB | | Public research | | Policy learning | |
|---|---|---|---|---|---|---|---|
| RISIS Patent | 65 | VICO | 88 | RISIS-ETER* | 28 | SIPER* | 15 |
| CWTS publication | 50 | CIB*** | 26 | PROFILE | 6 | JOREP | 9 |
| ISI-TM | 11 | CHEETAH | 15 | MORE | 13 | EFIL** | 3 |
| EUPRO | 62 | | | DDC** | 0 | | |
| ESID | 45 | | | | | | |
| | 233 | | 129 | | 47 | | 27 |
| | 53% | | 30% | | 11% | | 6% |

*Notes: *these datasets have mostly been accessed freely (without the need for a specific request, see part 2 reports); **these datasets have been opened too recently to be widely known yet; ***A new version of CIB will only be opened in 2024 while the CINNOB panel dataset has only been made available (as anticipated) at the end of the project and will enter the category of freely available datasets.*

**Figure 1: networks of datasets linked in multiple access requests**




*Notes: Node size corresponding to the number of co-occurrences in multiple access request with other datasets, links to the intensity of joint occurrences of datasets*

## Users of datasets

We had two clear objectives: supporting European researchers while remaining open to wider uses; and playing an important role in preparing the new generation. We can clearly see that we have achieved both.

Europe gathers 88% of requests coming for 20 EU countries (9 gather 10 requests and more) plus the UK (the 4th largest country for requests) and Switzerland. 16 non-European countries represent together 12% of access requests, the USA building the majority of such requests, followed by China.

Slightly more than 40% of users are senior researchers. This means that nearly ***60% of our users are the new generations of researchers***: 9% are at master level, 22% at PhD level and 26% are junior academics, most of them at the post-doctoral level. Our ambition to focus on the new generations and prepare them is clearly materialized by these numbers.

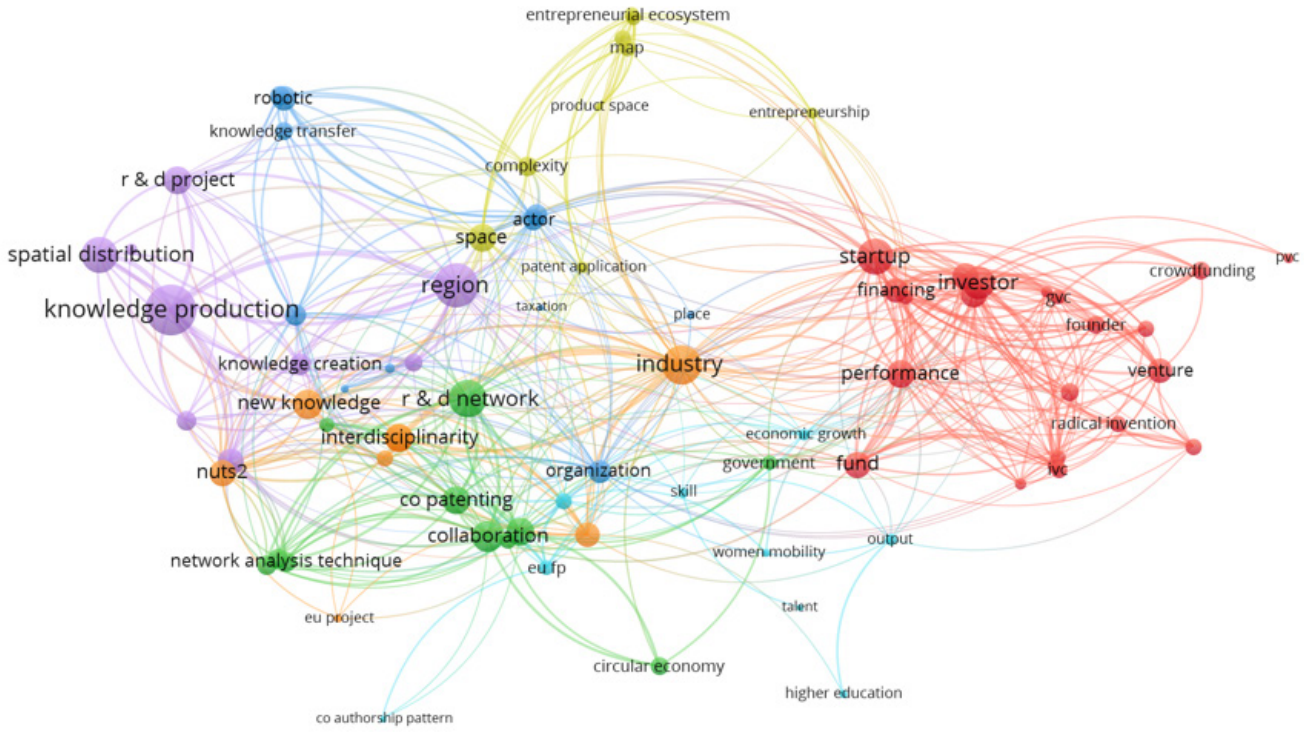
## What for?

One of our fears was that classical policy or scientometrics issues would dominate use. This is why we have conducted extensive studies to better follow usage. Figure 2 (word cloud) and figure 3 (co-occurrences of terms) demonstrate a wide thematic coverage that is well in line with the thematic objectives highlighted by the project: enlarging the geographical coverage (regional differentiation being a crucial dimension); a focus on actors, with an interest on different types of firms, their R&D activities and their performance, on public research (especially universities) and on collaborations and networks; a preoccupation about new forms of innovation (see the role of social innovation in requests).

**Figure 2: Main terms in access requests**


*(top 125 terms occurring in project descriptions, excluding stop words)*

**Figure 3: thematic network of terms occurring in RISIS access requests**




*(size of nodes correspond to the number of term occurrences)*

## Enlarging services offered

A core dimension of the project was to enlarge the services offered to the STI community. We have done so in six complementary dimensions.

- First, we have maintained and deepened our core datasets and we have complemented them with 4 new datasets on key topics of interest for STI policies.
- Second, we have further harmonized our core datasets. This has been based upon the development and extension of tools supporting our three foci of integration: actors, places and themes.
- Third, we have deepened and tailored software and platforms to help users integrate, treat and visualize their data as well as interact online and access open access resources.
- Fourth, we have developed a specific facility, the RCF that enables users not only to access the data they need but also to mobilise all the tools and services in an original integrated way.
- Fifth, we have provided extensive training in data sciences and in advanced quantitative methods.
- And sixth, thanks to the two amendments and the strong support of our reviewers, we have offered to the wider community of stakeholders, a first coreset of positioning indicators.

We review these 6 modalities through which we have enlarged services and generated impact in some more detail, highlighting the key achievements enabled by the RISIS2 project.

## Enlarging services (1): Enlarging the coverage and relevance of existing core datasets and creating four new datasets of interest.

We present the achievements at the end of RISIS2 project along the four core dimensions covered by RISIS: STI output datasets, firm level datasets, datasets on public research and human resources, datasets supporting policy-level learning.

Note: They combine activities about maintenance (WP5), deepening (WP9) and creation of new datasets (WP10).

### *STI output indicators*

RISIS2 now offers to researchers five databases that all are harmonized at the level of actors, geography and themes. It enables unique 'integrated' approaches of RDI activities and represents more than half of total requests. Below is a short presentation of the 3 core databases that have been integrated from the start of RISIS1 and extensively developed: CWTS publication database, RISIS patent database and EUPRO for funded projects. One ambition of RISIS2 was to complement them with two new databases that help to address the issue of non-technological innovation: Trademarks with ISI-TM and social innovation projects with ESID.

**CWTS publication database** is built on the core collection of the Web of Science. It has now a long history of validation and enrichment to serve for analyzing the dynamics of knowledge and to produce the well-known Leiden ranking. This has been greatly enhanced for European actors by the adoption of persistent identifiers (through ORGREG for universities and public research and through FIRMREG for large firms), by the geocoding of all affiliations and by the linking with the SDG classification. Beyond annual updates, RISIS has supported four major developments: (i) the bottom-up citation-based clustering of scientific activities (around 4000 clusters) provides longitudinal analyses of science dynamics at all levels (in particular organizations); (ii) CWTS has developed indicators of open science which are incorporated into ORGREG; (iii) A third development lies in the use of acknowledgements to better understand the role of funding organizations and project-based funding in the dynamics of knowledge production; (iv) CWTS has developed a new approach to analyse the role of science in society combining the publication DB with Crossref-Event data (see 2020 paper on the ABC method). To prepare for the future, support has been granted for the development of Open Citations and experimentations have been made for the use of open access resources (e.g. OpenAlex or Open Citations) to enable researchers to have a direct distant access (while for the time being their requirements have to be executed internally and only aggregated data can be circulated to them).

**The RISIS patent database (RPD)** derives from the EPO PATSTAT collection. It is designed for analysing the dynamics of technological knowledge creation. This explains why it focuses on priority patents, using extensions and families as markers of knowledge deployment. RISIS2 has enabled the development of a completely new version that covers the 2000-2020 period and contains over 20 million patents. RPD addresses one critical weakness of the PATSTAT collection for studying firm dynamics, that is the importance of 'artificial patents' (13% of the total, but often over 50% for large firms, see Laurens et al. 2017). It also integrates a specific development that enables to integrate for semantic analyses the definition of the 75000 IPC classes which gives a very rich background to the content of patents. It has developed specific indicators of knowledge deployment using the 5 main patenting offices (IP5). All patents are tagged along the 3 firm datasets (large firms, fast growing mid-sized firms and venture capital backed start-up firms) and for public research organisations (using ORGREG) and are geocoded (using the inventors addresses to locate knowledge production). This explains why it is widely used (65 requests during RISIS2).

**EUPRO** is a database on ongoing projects funded by public agencies in Europe. Projects are important markers of science in the making. They enable to study actors and their collaborative networks between public and private actors, between places (and in particular metropolitan areas) and to address issues of capabilities at lower levels than national (at regional or metropolitan level) for different domains and in particular now for different societal development goals. EUPRO was created for validating and enriching the successive databases produced by the EC on EC funded projects (from FP1 to Horizon Europe). It has been enlarged to EUREKA, JTI and COST actions, building now EC-PRO (130000 projects and 700000 participations). EUPRO has changed nature with now the incorporation of nationally funded projects (since

2010) from 14 EU countries (including all large ones) plus UK and Switzerland. Altogether NATPRO contains some 530000 projects and 790000 participations. This first major achievement goes along with an extended harmonization effort, for place (using CORTEXT Geo at the NUTS3 plus level), actors (based upon ORGREG and FIRMREG) and theme (using the SDG classifier). It turns EUPRO into a unique resource to study the role of projects in the spatial and thematic dynamics of knowledge production (with over 60 requests during RISIS2).[2]

**ISI-TM** is a new database of trademarks derived from the European Union Intellectual Property office (EUIPO) and from the USPTO. It has been developed during RISIS2. More and more trademarks are used to describe the diffusion of distinctive products or services. For products, it complements the technological description associated with patents, and it offers a unique access to new services, and business models innovation. Contrary to patents, the NICE classification remains relatively raw with 45 classes (and no subclasses). This is why one major development has been a new classification in five levels (with some 8600 subclasses at level 5) by applying a string-matching algorithm to the (partly standardized) text by which the applicant can describe his trademark. This enables articulation both with patents and KET, and with SDGs. Its geocoding enables regional and metropolitan analyses. An exemplary case of use, showing the growing interest by academics, deals with the use of trademarks as markers of start-up internationalisation (Guerini et al., 2023).

**ESID** is a new database on social innovation projects. It offers another complementary view on non-technological innovation, with often a different actor set, with an important role of co-creation and participation, being also part of firm responsible management and playing a large role, mostly at the local level, for sustainability transitions. This explains why, even though only open for the last two years, it has been already highly demanded by researchers. RISIS2 has enabled to transform the prototype developed in another EC project (Knowmak; later fully integrated in the RCF)) into a fully-fledged database gathering more than 12000 projects worldwide, using advanced machine learning and natural language processing techniques to collect information and characterise projects from public web sources (see 2022 descriptive paper in Nature Scientific Data).[3]

### *Firm level datasets*

RISIS contains 3 specific datasets dealing with firm innovation capabilities: VICO for venture capital backed start-up firms in Europe, CHEETAH for fast growing mid-sized firms in Europe and CIB for world largest firms. FIRMREG incorporates all firms from these 3 datasets allocating them persistent identifiers that enable to trace them in the different STI output databases. FIRMREG now incorporates core innovation-related indicators to foster enlarged econometric studies and a specific subset of 35 indicators has been created to characterize the innovation activities of world largest firms (CINNOB). The latter two have just been 'opened' on a free access basis and will represent a key element in the widening of RISIS user community in 2024.

**VICO.** Start-up firms are an important phenomenon of the 21st century and a key locus for STI policies over the last 30 years. However, despite previous attempts, researchers missed an 'extensive' dataset both longitudinally and in term of geographical coverage. This had been identified since the start of the RISIS1 project and was given priority in RISIS2. This proved a really relevant choice as VICO is by far the most used dataset within RISIS covering a wide range of aspects, e.g. firm distribution and agglomeration patterns; firm sectoral distribution, distribution of VC deals over countries, metropolitan areas or sectors, role of venture capital in growth (effects of their value-added activities), role of cross-border investments, importance of radical innovation, or internationalisation patterns. VICO focuses on venture capital backed start-up firms that, once supported, are followed over time in term of assets and financial results and of changes of status (exit, mergers and acquisitions, IPOs, other forms of financing, in particular crowdfunding). It incorporates a counter-factual to identify the key aspects of VICO firms. During RISIS2 VICO has been extended to cover all EU countries (plus the UK and Israel), all VC firms operating in Europe (with now a specific VC investor

[2] Scherngell, T., Barber, M., Zahradnik, G. et al. EUPRO - A reference database on project-based R&D collaboration networks. Sci Data 11, 291 (2024). https://doi.org/10.1038/s41597-024-03129-y

[3] Gök, A., Antai, R., Milošević, N. et al. Building the European Social Innovation Database with Natural Language Processing and Machine Learning. Sci Data 9, 697 (2022). https://doi.org/10.1038/s41597-022-01818-0

sub-database and a specific sub-database on all VC investments).

**CHEETAH**. Mid-sized firms have been poorly studied compared to other types while numerous studies have highlighted their importance in growth (and more specifically employment growth). This explains why a first version of a database was explored during RISIS1 focusing on fast growing firms (firms with a growth in their turnover or employment of 20% per year over a 3-year period). At the end of RISIS1 there were 45000 firms for 3 3-year periods. At the end of RISIS2 CHEETAH contains 130000 firms covering 10 3-year periods and 30 European countries. Information has been enlarged to address ownership and governance, to incorporate Mergers and Acquisitions, and to identify a specific subset of business models: platform-based firms. Key policy relevant results have been summarized in a policy brief that discusses the wide sectoral coverage, their uneven distribution throughout Europe and a growing relationship over time with urbanization. CHEETAH is unique worldwide in the landscape of firm datasets and represents a clear long-term asset for RISIS.

**CIB** (Corporate Invention Board) is a database identifying the world largest R&D performers and providing a group-level view of their technological activities (mobilizing their patent portfolio). The largest world firms combine on average around 500 subsidiaries and their composition rapidly evolves, which makes them difficult to follow over time. This explains why the approach selected is to work by batch looking retrospectively on their technological portfolio (see Laurens et al., 2015). CIB1 was built and open during RISIS1, CIB2 was opened in the second year of RISIS2 and updated end of 2022. It has gathered significant interest (more than 25 requests) focused on different sectors, different countries or continents, different technologies (pharma, semiconductors, space, automobile but also agri-food or construction) and on relations with country capabilities/policies, with firm strategies (role of M&A), with sustainability policies or with the role of green technologies. This has driven to focus on a fully open access panel database integrating a set of innovation-related indicators, see CINNOB below). The consequence was to postpone the opening of the third batch and the choice of a more selective approach to the selection of large firms, not only relying on the EC scoreboard, but also incorporating thresholds in term of minimal size (of turnover or assets). CIB3 is planned to open mid 2024.

**CINNOB** is a user-friendly repository providing robust indicators on the R&D activities of world largest firms. It is based on CIB2 firms incorporated into FIRMREG and is linked to their tagging in the different RISIS datasets: scientific data coming from CWTS publication database, technological data coming from RISIS Patent database, Trademark data coming from ISI-TM and participation in European R&D projects coming from EUPRO. With general information on firm size and finance, it provides a set of 35 indicators allowing the study of the firms' activities according to several dimensions: time, technologies, scientific domains, geographical location of the R&D activities, collaborations. The first cases of use - dealing with a comparison between pharma and electronics firms, or the role of scientific activities (as captured by co-publication with universities) in their inventive performance - were presented at the final RISIS Vienna Conference. It has been opened for wider use at the end of 2023 and will represent a key feature of the new developments proposed to users for 2024.

The other key open-access feature for 2024 is the full opening of **FIRMREG** in its two dimensions: the repository and the set of key indicators. As a register FIRMREG integrates firms from the different RISIS datasets, providing a unique firm identifier for more than 750,000 firms which represent over 600,000 parent-subsidiary linkages and 1.2 million name variations. FIRMREG has a front-end that is fully open, while the backend containing subsidiaries remains under the controlled access RISIS process. The completely new feature is a limited set of core indicators of R&D activities (numbers of publications, patents, trademarks and projects per firm) for all 750000 firms which represents a unique resource for all quantitative analyses that wish to incorporate R&D activities in their models.

### *Datasets on public research and human resources*

Two entry points have focused the attention of RISIS members regarding public research. The first one deals with organisations, with ORGREG covering all public research organisations and ETER focused on universities. Both are publicly available, even if the latter one is also accessible on request for researchers that wish to directly exchange with the operators of the dataset (28 requests received). The second one focuses on researcher careers with a core preoccupation on non-academic careers of PhD holders (over 70% of all PhDs

diplomas today). These are complemented by two external datasets that have been made accessible through the RISIS portal on PhD trajectories in Germany (PROFILE) and on the mobility of European Researchers (MORE).

**ORGREG** has a fundamental role in RISIS and beyond as it provides stable identifiers of public research organisations in Europe to be used in different datasets within and outside RISIS. Besides identifiers, it includes basic information such as foundation years and geography. The register documents history, linkages and geography and provides matching with publications, patents and EU-FP projects. It is unique in its coverage of demographic changes in European public research such as mergers ( Heller-Schuh et al., 2020). As of December 2023, OrgReg includes more than 7,500 unique entities. An important recent extension has been to include European University Alliances and the membership of HEIs in the alliances (Lambrechts et al., 2023).

An important feature of OrgReg is to provide matching with widely used identifiers in other settings, particularly the Research Organizations Registry identifiers (ROR; https://ror.org/), and, as of HEIs, identifiers from the World Higher Education Database, DEQAR, Erasmus Charter codes. For easy mapping with other registers and user data, the OpenRefine API service has been implemented in OrgReg. This service allows users to reconcile their own data with Orgreg data and import automatically OrgReg IDs in their datasets. This is therefore a core service to support the use of OrgReg for entity matching.

**ETER** is now a world reference about European Universities;[4] it has been and is directly supported by the EC in consortia that mobilise all European Statistical Offices. It includes descriptive information, such as legal status and foundation year, geographical information, students and graduates, personnel, HEI expenditures, research activities, as well as a set of pre-defined indicators characterising relevant dimensions of HEI activities, like the extent of subject specialisation, international mobility, or gender balance. Combined with ORGREG users can also incorporate STI output data. It has been extensively used and the specific report in part 2 analyses the more than 200 papers that have mobilised ETER discussing 5 main themes: HEI systems and types, European higher education and mobility, rankings and benchmarking, regional development and innovation and knowledge transfer. ETER also provides country reports that are mostly oriented towards policymakers.

**DDC** – Doctoral Degree and Career – dataset is a new experimental dataset that has just been opened for testing. It aims at analysing the labour market of trained PhDs as all analyses show that now a large majority of PhD holders leave public research directly or after a few years in post-doc positions. At the same time previous attempts (e.g. OECD CDH) have shown how difficult it was to address the issue. The initial approach was based on new developments in the analysis of CVs but was fully stopped as an indirect effect of GDPR that closed access to large CV databases. In this context DDC has taken a novel approach. It is dissertation-centric (as dissertations remain open public data): Information about the dissertation and the degree-holder is used to construct the population base for a given country each year. This is complemented by the publications retrieved in the CWTS publication database and connected to the corresponding university via ORGREG. This enables to arrange the dataset in an array of variables at granular level about when (the degree was issued), where (the degree granting institution and geolocation), who (the degree-recipient, gender) and what (topic mapping). The experimental dataset so far covers 2 years in 6 countries (nearly 100000 entries), and work done has demonstrated the widespread possibilities it offers, including clear identification of non-academic careers. But this would require extensive data development and depends on policy interest and will to invest into such developments.

The RISIS **MORE** database incorporates EC surveys made on researcher mobility. More specifically, it integrates MORE2 (2012, 10500 answers) and MORE3 (2016, 10400 answers) surveys. The owners of MORE 4 (2020) did not wish to provide the information, driving to no longer maintain RISIS-MORE, and this is mirrored in the progressive loss of interest by users for the dataset. The surveys are very rich (more than 100 questions covering 7 dimensions of research careers) and cover 31 European countries. Enriched by spatial and organisational data from other RISIS datasets (especially ETER), it has enabled rich analysis of career stages, e.g. differences in careers between countries, or the role of long-term mobility in accelerating careers (both at R3 and R4 stages).

**PROFILE** was a German database providing a longitudinal analysis of doctoral candidates and holders. It

[4] Lepori, B., Lambrechts, A.A., Wagner-Schuster, D. et al. The European Tertiary Education Register, the reference dataset on European Higher Education Institutions. Sci Data 10, 438 (2023). https://doi.org/10.1038/s41597-023-02353-2

covered all fields and a sample of institutions, with initial answers from 2009 to 2016. All in all it covered over 20000 PhD candidates and over 5000 PhD holders that were later surveyed. Connected with ETER and Leiden Ranking, it raised interest in particular about the role of graduate schools and about continuation in post-doc positions. However, interest faded when the survey was stopped in 2017 due to institutional changes.

### *Datasets supporting Policy Learning*

This fourth group is made of 2 core datasets: first the international repository of STI policy evaluations (SIPER) which is fully open access (even if it is also accessible on request for researchers that wish to directly exchange with the operators of the dataset, 15 requests received), and second the completely new European dataset of public R&D funding instruments (EFIL) only very recently open in its full format to users. The 2 datasets are connected through the joint use of ORGREG identifiers and of EFIL instrument codes in SIPER, enabling further articulation with EUPRO (and in particular NATPRO). Both datasets are complemented by a third dataset on European Transborder R&D programmes (JOREP) that remains accessible though it is no longer maintained as such.

**SIPER**, the Science and Innovation Policy Evaluation Repository, gathers evaluation reports worldwide (mostly EU and OECD countries) since 2000. Thanks to the effort done during RISIS2, more than 1220 reports are included (with 70% fully coded). These reports are fully available in pdf format (a unique feature not existing anywhere before RISIS2) and they are coded. The codings include information on the underlying policy measure and a qualitative assessment of each report.

Thus, interested researchers and policy makers can freely exploit the database choosing between two different approaches: a predefined search based on the criteria used for coding the reports and a free text search function. SIPER follows two complementary objectives. First, the database aims at facilitating policy learning, e.g. what works (and what does not) under which circumstances and why, the policy brief and the joint workshop with OECD focused on societal and environmental impact being a good example of such use. Second, it supports further academic research. Three main types of use have been identified: methodological aspects (e.g. the use of scientometrics), comparative analyses of policy evaluation practices (e.g. in Europe); presence and dynamics of thematic aspects (e.g. gender, system-oriented policy evaluations, societal & environmental impacts).

**EFIL**, The European dataset of public R&D funding instruments, is a completely new dataset proposed by RISIS2. Funding instruments are taken as the basic unit of any governance mode (Capano et al., 2019). EFIL proposes thus 3 sets of descriptors for each funding instrument: (i) attributes of the managing RFOs (e.g., domain, mission); (ii) attributes of the instruments (e.g., aim, mode of funding, composition of decision-making body, eligible beneficiary categories etc.); (iii) time-variant descriptors on budgetary information (funding amounts provided for each single reference year). It provides unique information about the use of given instruments in Europe and about RFO effective policy mixes. EFIL covers 10 European countries (8 EU plus UK and Switzerland), 55 RFOs and over 700 distinct funding instruments. It works by 'batch', the first one in 2017-18 and the second one in 2021, covering data from 2010 upwards. A unique second part of EFIL consists of a complete repository of all instrument documents that can be searched directly through textual analysis (this has in particular enabled to analyse the engagement of instruments in KETs and SDGs).

**JOREP 2.0** is a dataset created through a distinct EC project and deployed in RISIS1 to analyse a specific aspect of Europeanisation: transnational joint R&D programmes at European level. It focuses on two years (2013-14) with a retrospective view back to 2000. It contains 152 joint R&D programmes from EU-27 countries plus Israel, Norway, Switzerland, United Kingdom and Turkey. About 65% are European-level initiatives (e.g. ERA-Net, JPIs, JTIs), while the others include bilateral/multilateral programmes. It provides data on organisational, thematic and financial aspects of programmes. It has been used for spatial analyses of Europeanisation, for measuring the relative engagement of Research Funding agencies, or for their role in pushing interdisciplinarity.

## Enlarging services (2): Reference databases, tools and platforms supporting the harmonization and integration of datasets

*(this gathers work from WP6)*

From the start of the RISIS project, we have recognized that we were a distributed infrastructure, not only because teams are distributed geographically, but also and even more because we have never anticipated to integrate together the different datasets, each having its own 'raison d'être' in the landscape of STI policies and/or of characterising knowledge dynamics. What we considered important was that all datasets shared a number of attributes that foster their interoperability and enable their joint mobilization or dimension-based integration. This was the object of intense work in the first RISIS project where we built the RISIS triangle of actors, geography and themes. There have been extensive investments on the development of reference databases, tools and platforms enabling the tagging of RISIS datasets, but also well beyond since all these facilities are completely open access and are used by a wide range of actors far beyond RISIS participants.

### *Actor reference DBs*

This was probably the biggest shift in the way to look at knowledge dynamics: while it is standard to allocate innovations (but less often inventions) to firms, for a long time, science dynamics was looked at through the eyes of individual scientists forgetting the conditions under which they undertake their work. Markers of this shift were first the debates about the collective dimension of research activities and second the rise of questions and rankings about the excellence of organisations, and especially universities. We thus took advantage of the extensive work done about the characterization of universities to focus our attention on institutional actors: firms and within them groups, public research organisations and universities. It was a strong decision taken in RISIS1 to consider that online identification of actors (like ROR) would not satisfy our needs for enabling a robust integration of distinct databases, and to enter in the construction of reference databases providing unique and stable identifiers for organisations in Europe (following in a way what had been done for some time already in the US). ORGREG and FIRMREG propose such stable identifiers for European organisations. They have been presented in the previous section as datasets of their own. This is because the propagation of their identifiers within the STI output datasets has enabled to provide a coreset of organisation-level indicators about publications, patents, trademarks and projects.

### *Geospatial allocation*

The distribution of knowledge production has gone far beyond classical country-based analyses to consider both finer grained distribution processes (e.g. regional distribution in the EU) and agglomeration processes and the role of large metropolitan areas (with such notions as global cities). We thus considered it central to be in a position to address these geospatial issues (at a time, once more where open-access tools, such as google, became paying). UGE and its CORTEXT platform thus entered in a dual development of a geocoding service and a geospatial exploration tool.

Geocoding services allocate geographical coordinates to any given entity. These are complex processes because addresses vary in format (even within national borders), often associate non-geographic information, with repeated or alternative toponyms within and between countries.

The geospatial exploration tool addresses the issue of the relevant scale of analysis and thus mechanisms of aggregation. The tool offers an original way to solve this, by combining in one unique tool a large variety of well-established sources of shapes: national and regional at world level, urban boundaries (based upon OECD and extended at world level) and linked to this, rural boundaries; and EUROSTAT NUTS shapes that enable access to generic statistics (in particular NUT3+ recently developed for the analysis of metropolitan areas).

The first full release took place in 2019, with regular enrichments since. In 2021 the service handled more than 4500 user calculations, which represents a very important workload since both services are time- and resource-consuming. Both together they have concerned around 40 countries and 4 continents and requests coming from around 160 different research organisations.

### *Ontology-based semantic annotation and the identification of Societal Development Goals in STI outputs*

The move towards directed and/or transformative policies raised a complex question to analysts of knowledge dynamics as traditional categorisations in science and technology could not capture the knowledge base that could be mobilised for such problem-based issues. This required new textual analysis tools to look through each individual output, may it be a publication, a patent, a trademark or a project. This explains why the development of ontologies and tagging tools became more and more crucial for RISIS, even when capitalizing on efforts done more broadly. We thus concentrated far more effort than initially anticipated in developing an ability to identify relevant outputs corresponding to the 'Societal Development Goals' that now build the backbone of European and most national STI policies. We did this in two main directions.

First, we developed an ontology to link research outputs to SDGs, forming the basis for indicators to be developed (e.g. how many patents were produced relating to a particular SDG by which metropolitan area in which year). This is supported by the development by GATE of a specific tool – the RISIS SDG topic classifier – that is available on its own[5] and through the RCF. The tool is equipped to tag large amount of data (e.g. the 20 million patents of RPD) and of a dashboard that after selection of the relevant elements (e.g. nearly 500000 patents and 700000 patent class-SDG pairs) enables standard visualisations. The report (in part 2) focuses on SDG11 publications to demonstrate its use.

Second, to face the variety of interpretations about what an SDG is, we developed a platform in which a user can explore the relations between research outputs (at present only publications) and SDG from different perspectives, as different stakeholders can make different choices about how to link research output to a societal challenge or goal. This is important because there is a very large disagreement in how to map SDGs to data, as underlined by recent research investigating the mappings of SDGs carried out by commercial data providers such as Clarivate, Dimensions and Elsevier. The RISIS SDG landscape explorer[6] enables the user to select within relevant subclasses identified, most prominent topics, journals and policy documents (linked to OVERTON DB) the subset that corresponds best to a user's own understanding of what a given SDG or sub-SDG is about. The report (in part 2) focuses on a controversial SDG (SDG1 on no poverty) to illustrate its use.

### *Data quality assessment*[7]

Finally, a specific action was conducted for data quality assessment. It has been developed through a data quality package adapted to RISIS datasets and implemented in a specific workshop (2020). Its originality has generated 4 visible academic papers. A second step was to develop an advanced data quality and data preparation 'visual analytics environment' for users of RISIS datasets. It has been tested for the production of RISIS positioning indicators (ORGREG and RISIS Knowmak) and presented in one of the methods in action courses (2022).

[5] http://services.gate.ac.uk/knowmak/risis/filter-search/ and http://services.gate.ac.uk/knowmak/risis/faceted-search/
[6] https://public.tableau.com/app/profile/ed.noyons/viz/ExploreRISISSDGlandscape/RISISSDGdashboard
[7] This was an action inscribed in WP7 due to its important 'training' aspect.

## Enlarging services (3): The RISIS Core Facility

*(This builds the core of WP4)*

The RISIS Core Facility (RCF) has been built around three objectives: (i) enable users to access at a distance to the datasets extracts they require for their projects; (ii) offer users a private working space in which they find their project-based data, and at the same time enable them to access all the services that RISIS offers; (iii) provide them with ways in which to combine the tools and software they need for enriching, integrating, treating their data and visualising their results.

Objective 1 is made complex as most of our datasets are not freely accessible ('controlled access') and 'access managers' need to select what is useful for users to deploy their accepted projects. This requires a first level infrastructure where users register, accept RISIS charter of use and propose their projects. This infrastructure organises the de facto 'controlled access' made of two parts: First the agreement process (with the overall handling of requests, the approval of dataset managers and of the external review panel). Second the automatic extraction of the dataset extracts that will be made available to users in their workspace. This has required us to develop a parametrised datastore that enables the selection by dataset access managers of only what is needed from a dataset by a researcher.

Objective 2 has benefitted of all the work done by the CORTEXT platform about the interactions with users from authentication to the management of user workspaces via the monitoring system (that allows to help users that face repeated difficulties in their scripts).

Objective 3 is to enable users to mobilise and combine different tools and software for addressing their question, while not being computer specialists. The reader can understand the infinite possibilities when you combine dozens of datasets and multiple service.

This has driven the RISIS development team to build an architecture whereby the core is an "orchestrator" that builds and operates specific scenarios of use. It has to consider a large variety of possibilities of combinations both of datasets and of services. This has driven to two original software developments: RISIS Orchestration System (RISOR) and the RISIS Scenario Modelling Language (RSML). The RSML enables to develop specific scenarios of use (articulating access, integration, enrichment and eventually treatment and visualisations).

The core originality of the RSML is to render it possible for non-expect computer engineers with the objective that all scenarios developed are stored so that users can mobilise the ones that correspond to their type of use with limited parametrisation.

Even if it took more time than expected, we are proud to now have an operational RCF that fulfils these three objectives fully. The report in part 2 presents first the RCF as a user- oriented application, and then as a web-oriented architecture of services. It really enables to demonstrate the originality of this facility tailored to handle the specific issues of our domain.

## Enlarging services (4): Dataset analytical services

*(These are part of RISIS2 WP6)*

There are ample tools and platforms for statistical analysis (in particular using R). We thus considered that the issue was not creating new ones or even tailoring existing ones, but rather to showcase their use in our community and offer specific training activities, which explains a significant part of our training efforts (see below). This drove us to focus on three major services: NLP tools and semantic analysis, visualization, Virtual Research Environment and access to all OpenAire Data.

The starting point of our focus on natural language processing lies in the recognition that most of our datasets are mostly textual, and that we probably lose 90% of their analytical potential when leaving texts aside. However, at the time we conceived RISIS1, most available software for semantic analysis were private or small scale, and we considered it critical to offer at least one open-access large scale alternative. RISIS1 integrated the online CORTEXT platform, which very rapidly witnessed high levels of use (what we expected the last year of RISIS1 was achieved in one month!). This pushed us to widen the NLP services offered and we integrated in RISIS2 the GATE platform, a very large European-level resource for textual analysis tools.

A second issue we considered important lay in the ways to visualize results of textual analyses, so that it helps researchers and policymakers to mobilise them in their work. Both CORTEXT and GATE have shared their visualization methods and developed new ways (e.g. for CORTEXT Sashimi for the analysis of groups of documents and their relations, or the contingency matrix to visualize the joint distribution of two fields over the documents in a corpus). However, the core of the efforts has been in the wider dissemination of VOSviewer as a key visualization tool relevant for the STI field.

Below we present where we stand with the 3 key analytical resources we provide for the community: CORTEXT, GATE and VOSviewer platforms as well as the RISIS VRE.

### *GATE platform*

GATE toolkit for Natural Language Processing is one of the most widely used of its kind in the world, with approximately 300000 software downloads and over 5 million API requests on GATE cloud in the last 5 years. GATE is freely available under LGPL from Github and has a repository of over 150 text mining and NLP models and algorithms that can be organised in 'pipelines' for term recognition and tagging, morphological analysis, and other complex processing.

All GATE tools are available as ELG compatible services (European Language Grid) and this has been the support for the integration of GATE in the RCF by leveraging the API specification and docker-based tool packaging mechanisms from the ELG project. This choice was made to enable the RCF to integrate other tools (especially for translation) that are available through the ELG platform. One important aspect of GATE inscription in the RCF lies in the fact that its use requires far less computing abilities from researchers, thanks to the scenarios through which tools can be mobilised (see above).

A second important aspect is linked with the development of the new collaborative document annotation tool and platform, Teamware2, that enables far easier ways to administrate annotation projects (first version released in April 2023). It is available in the RCF as a standalone demonstrator, and a seamless service should be available in 2024.

Finally, GATE has developed a specific tagging tool, the RISIS SDG topic classifier (see above) which builds today a unique resource for the identification, characterisation and analysis of transformative and sustainable capabilities.

### *CORTEXT platform*

CortText addresses probably the most transformative element for research in SSH, the exponential growth of digital traces and the will of many researchers in our field to simultaneously consider the textual content

of most of our databases (e.g. publications) and the social relations that are constitutive of research activities. The ambition has been to offer to a researcher in our field and more globally in SSH (without programming capabilities nor access to intensive computing capabilities), the ability to treat large textual corpuses for answering his/her research question.

The choice has been to propose him/her an integrated process that covers the different stages: accepting multiple sources and formats of databases, covering data enrichment and curation, conducting data parsing & term extraction, combining semantic and social network analyses, offering multiple parametrisations for network analysis, enabling to analyse their evolutions over time and their spatial distribution, and providing multiple visualisations. A critical choice has been for the user at any stage to take the results in a generic format and to mobilise them in other tools (in particular for visualisations).

These developments have mobilized complementary competences: computer sciences but also scientometrics, social network analysis, automatic treatment of language, statistics, geographic cartography and spatial analysis, data visualization.

Recent developments deal, beyond the geospatial analysis (see above), with a new method for group level topic modeling (Sashimi, Addo A.H. et al. 2021) and a new visualization tool (the contingency matrix) to analyse the relations between 2 fields of data (e.g. country and WOS category).

CORTEXT use develops very rapidly with today more than 8000 registered users from more than 100 countries, even if French users still represent just under 50%. In 2023, we counted 1200 effective different users coming from 500 different institutions and 50 countries. They undertook over 45000 calculations.

Overall, since 2016 more than 1000 publications mention CORTEXT. This represents approximately 10% of users, highlighting the variety of uses. We have identified 4 main ones: quali-quantitative analysis of calibrated corpuses (e.g. scientometrics analyses, temporal analysis of knowledge production or systematic reviews), support to SSH surveys (at all stages including initial exploration to design the surveys), important use in teaching (CORTEXT as a pedagogical instrument, e.g. analysis of controversies), support to strategic analysis and expertise, especially focusing of characterizing past activities.

### *VOSviewer*

VOSviewer is a tool to visualize scientific literature and create interactive visualisations of scientometrics networks. For a long time, it was a successful standalone software and one key aspect of RISIS has been to support the release of VOSviewer Online (2021) that fosters interactive exploration and facilitates sharing with colleagues and users. Within 2 years it has been used nearly half a million times and is overall mentioned in more than 19000 articles. It can be accessed through the VOSviewer platform (https://app.vosviewer.com) and through the RCF.

### *Virtual Research Environment*

The RISIS VRE is a web portal that has 2 objectives: (i) give access to the OpenAIRE datasets (via API and web Use Interfaces); (ii) offer a VRE lab for collaborative work. It is empowered and operated by D4 Science. Its identity and access management process are connected with the authentication services of the RCF to enable a seamless access for RISIS users. For the first aspect, an Open VRE data catalogue is available through web user interfaces enabling researchers to navigate and operate their own selection. At present the catalogue contains 350000 references (including over 220000 datasets and 93000 publications). The Lab VRE proposes a computational environment for the collaborative exploration and validation of datasets. It provides RISIS users with 2 main services: the RStudio (for execution of real-time statistical analyses using R) and Jupyter Lab Data analytics (to manage shared notebooks that support coding and computing).

Though available on its own and in the process of being integrated in the RCF, it has witnessed limited use during the project, probably because it requires a high level of computational capabilities (this was in particular the case when it was considered for the 'methods in action' training effort, see below).

## Enlarging services (5): Fostering community capabilities in use of datasets and advanced analytical methods

*(see recapitulative analysis of WP2 and WP7 activities: D2.7 and D7.4)*

In RISIS1 we developed a conventional training approach based mostly on short training courses. It proved successful in boosting the use of datasets, but we also discovered that many researchers faced difficulties in the use of advanced analytical methods. We thus enlarged the traditional approach to methods. We started deployment in 2019 with a set of dataset-oriented short courses and a first batch of 6 analytical methods.[8] By stopping all face-to-face activities, the COVID crisis obliged us to completely revise our approach and to completely reorganize training activities.

First, **tutorials** were extensively developed that drastically reduced the need for dataset-oriented short courses. All materials linked to training have been made available through the RISIS Zenodo, and there has been more than 3800 downloads over the project (while we anticipated 1000).

Second, we created a **monthly online research seminar** based upon existing papers or presentations having used both our datasets and advanced analytical methods. It helped researchers (at whatever stage) to grasp the intricacies of dataset mobilization and treatment. This seminar has had a tremendous success with 30 sessions and nearly 1600 participants. Furthermore, they are transformed into videos (more than 2500 visualisations) and all the relevant material is then accessible (nearly 1600 views and over 600 downloads). On average in 2023 one research seminar was followed by 50 people, generated 250 views, around 90 downloads and around 4000 impressions. In brief, for us the seminar has proven the lasting core solution to our training objectives.

Third, this has driven us to focus **face-to-face training activities** on two mains aspects: (a) summer or winter schools on data science (as an overall introduction, mostly for PhD students) following the very important success of the first one delivered in 2019; and (b) targeted schools combining RISIS datasets and new approaches (e.g. semantic analysis or machine learning). Altogether we have conducted 16 training courses with 600 participants.

And fourth, we developed a new interaction format labelled as '**RISIS methods in action'**, where 'complex' methods adopted for RISIS usage cases are explained, tested and discussed with interested researchers. It is based on a three-step process: (i) a half-day presentation session of the method based on an effective case of use in RISIS. (ii) Individual work by participants (to reproduce elements of the usage case presented) and a second half-day session one week later where participants present their work and their questions as a source for a collective discussion on the use of the method. (iii) a follow-up session 2-3 months later to discuss with participants their mobilisation of the method (and provide eventually further guidance). The overall setting of the RISIS Methods in Action courses therefore emphasizes practical learning of using statistical methods for the specific problems of STI communities, and their application to RISIS datasets. 7 sessions have been organised with 275 participants (with very positive evaluations about the format).

[8] 6 methods were addressed in trainings that gathered nearly 200 participants and generated over 700 downloads of the documentation. It was also successful for organisers in generating 8 academic presentations and 5 articles.

### Enlarging services (6): Making a first set of open-access positioning indicators

*(see recapitulative analysis of WP7 activities: D7.5)*

One strong recommendation of the first interim review of the project was for RISIS to be more proactive and not leave to users the sole production of new positioning indicators. This drove to three complementary actions to give users (both academics and policy analysts and makers) free access to a core set of positioning indicators: the integration of Knowmak into an enlarged RISIS_KNOWMAK interface; the development of public research indicators associated with ORGREG and the development of firm-based research and innovation indicators with both FIRMREG and an enlarged specific subset for large firms with CINNOB.

We have already presented the efforts made and the new services provided to users for the last two aspects: public research (ORGREG) and firms (with both FIRMREG and CINNOB, see point 1 above).

We thus focus here on the first aspect **RISIS Knowmak**. Knowmak is an interface focused on European regional and metropolitan areas connecting them to key STI indicators and enabling thematic and actor-based analyses. De facto it offers an online interface on core indicators derived from RISIS databases distributed by geographical area and by key themes of interest (Key enabling technologies KET and now also Societal Development Goals SDG). It was initially developed in a separate EC project gathering 7 RISIS partners. The first amendment redirected resources so that the Knowmak central database built from RISIS datasets could be maintained and updated and the Knowmak interface fully integrated in the overall RISIS infrastructure.

Three main developments have taken place since: (i) a more reasonable geographical breakdown using NUTS3+ was implemented to provide a better view of metropolitan areas (see CORTEXT geo); (ii) a complete SDG identification was enabled by the work done by the GATE classifier (see above); and (iii) the set of indicators developed in the initial DB (based mostly on publications, patents and projects) has been extended for Higher Education (number of different types of students using ETER), on innovative activities (using ISI-trademark, Cheetah and VICO), on international co-publications (using Leiden publication DB) and on industry-public research interactions (using EUPRO and the share of industry in FP projects).

The last version of RISIS-KNOWMAK has been put online in Autumn 2023 and demonstrated at the RISIS Vienna conference (November 2023).

## RISIS interactions with its user communities

*(see recapitulative analysis of WP2 activities: D2.7)*

***Whom is the infrastructure useful to? And for what purposes? These are two central questions all European supported research infrastructures ask themselves, but answers are far from simple.***

As a research infrastructure, our first audience is our core community, in our case researchers that study STI dynamics and policies. There we had and have two core objectives: be useful to our colleagues and extending the use of advanced quantitative approaches within the community, with a focus on the younger generation (PhDs and post-docs). On both we were quite ambitious at the end of RISIS1 when we demonstrated its feasibility and usefulness. The second RISIS project thus encompassed high level objectives that we have nearly all exceeded by far: the number of requests for projects (more than 400 over the period, out of which 60% of younger researchers), and the visibility of our training activities: 1600 direct participants to our monthly research seminar (a number nearly tripled if one considers the visualisations of the recorded sessions on You Tube), the some 900 participants to our dedicated training sessions, the importance of downloads of our tutorials and training materials (more than 3800 over the project) and more widely the extensive use of our Zenodo archive that contained end of 2023 185 documents and generated over 21000 downloads). All participate to demonstrate how RISIS has been relevant to an increasing share of our community, especially the younger generation.

STI indicators are not only useful to our core community, they are also useful to many neighboring communities especially in economics, management and geography, with academics not interested in producing indicators rather in mobilizing existing indicators in their models. We have thus developed open access readymade 'datasets', 'panels' and platforms where academics can find indicators of firm STI activities, of public research organisations (and in particular universities) and of place-based and problem-based capabilities (though the completely redesigned Knowmak interface). All these are too new to already have measures of their take-over. But they constitute a critical resource for transforming RISIS in a long-term sustainable European research infrastructure.

Infrastructures are costly ventures and, like many other public investments, funders expect that they have also an impact beyond their direct and indirect research communities. It is often more complex to demonstrate this for SSH research infrastructures, as most of the ways through which effects are measured are focused on STEM type interactions with economic actors and their money counterparts (e.g. patents and royalties). How to consider impact, when the corresponding stakeholder communities are mostly 'public', in our case policymakers and policy analysts. Our approach to measuring their interest in our 'productions', beyond participating to huge periodic conferences (e.g. ESOF), was to invite them to discuss the policy briefs we produced (14 over the period) and that showcase key RISIS results that emerge from the treatment of our datasets (remember that one specific characteristics of RISIS is that it is an infrastructure made by researchers themselves). De facto, Policymakers Sessions have evolved into forums for engaging in collaborative discussions and thoughtful reflections on the resources of RISIS. These sessions serve as dynamic platforms for fostering dialogue with a participative approach, addressing significant issues in Science, Technology, and Innovation (STI) policies, and establishing a prominent institutional communication channel with policymakers. One noteworthy highlight includes the 13th RISIS Policymakers Session, a crucial event focused on monitoring and analyzing research careers to guide policymaking in the European Research Area. The meeting which took place in Brussels on 8th June 2023, did not only generate rich and open discussions between the 97 participants on a key topic for the next EU framework programme, it also had a far wider circulation thanks to the views and impressions (respectively 546 and 3450 in the following weeks) on its You Tube version. You Tube has become for us a critical channel with more than 60 videos. The last 5 policymakers' sessions have gathered nearly 500 participants, 27500 views on You Tube, and generated 730000 impressions, highlighting

the level of dissemination obtained through this specific social media.[9]

All these impacts are only possible when researchers and stakeholders are kept 'informed' about what RISIS does, produces, disseminates and puts in discussion. This is a hard task to keep an active and lively website or to organize awareness and dissemination events.[10] But we soon discovered that our users still expected a periodic newsletter. We have produced 9 newsletters[11] and have now 2000 subscribers (following GDPR this requests a pro-active engagement). The newsletter is also a way to interact with our policy community. We have incorporated in recent editions interviews from policymakers. This was for instance the case in 2022 with a specific interview of Maryia Gabriel, at the time European Commissioner for innovation, research, culture, education and youth who insisted on the key role of open research infrastructures, or with Ute Gunsenheimer, the secretary general of the EOSC association discussing its role in nurturing Open Science. Followers of our newsletter and social media accounts include major organisations dealing with infrastructures (like ESFRI or EOSC Pillar), with actor representation (like Science Europe or the EUA) and with science policy analysis (like Research Fortnight).

[9] We are of course present on other social media: Twitter/X (nearly 800 twits, more than 1000 followers), Facebook (over 500 followers) and LinkedIn (750 contacts, 200 followers).

[10] As part of awareness raising activities, we have participated in 12 conferences where we organised specialised panels, mostly at EUSPRI and STI annual conferences, but also for 2022-2023, ICRI and ESOF conferences. Each time these sessions gathered on average 50 researchers.

[11] See Newsletter page (https://www.risis2.eu/newsletter/

## RISIS future

We have documented the achievements of this decade of activity. It has turned a vision into an operational research infrastructure that serves our community in taking advantage of the digital turn (both in exploration and validation of new concepts); and it offers new resources to neighbouring research communities. We also consider that we now have demonstrated our 'impacts' on our key stakeholders, policymakers and their policy analysts.

We have done this as a project enabled and supported by two successive EC projects. And our central question, these last three years, has been to inscribe it in the long-term landscape of European research.

Our first attempt was to consider ESFRI as the central direction to follow. Whatever the quality of our proposal, we learned that in an intergovernmental process, we were too small for each country even if globally we represent both a sizable investment and capacity. We clearly are not the sole one in this position at European level.

This drove us to enter in a completely different path, which is well illustrated by associations created by research organisations to handle long-term projects, such as in forestry. We thus explored the interest of our organisations to enter in such a scheme. The global answer was very positive and pushed to design four main principles for such an association. (i) It recognises the variety of levels of involvement with the notion of full and associated members; (ii) it will raise limited annual fees only focused on the shared activities of the infrastructure (with a restrictive list); (iii) the core input of members will be in-kind, based upon mutual engagements between the members thanks to an agreed 'plan of activities'. This third dimension is crucial as it leaves all IPR to members; (iv) it acts as a coordinating place and represents the members in the promotion of the infrastructure, especially at European level.

Once agreed upon, the choice was made, as other European research infrastructures, to become an AISBL under the Belgian law, with standard governance processes. The early 2023 GB put a trio (Benedetto Lepori, Emanuela Reale and Matthias Weber) in charge of operationalising the principles. A new Governing Board held April 17 decided to implement their proposal. The statutes have been now established and January 2024 has witnessed the starting of its signature. Meanwhile the members have agreed to maintain the functioning of the infrastructure as it stands end of 2023 (for data access as for the use of services).

LISIS LABORATOIRE INTERDISCIPLINAIRE Sciences Innovations Sociétés

# RISIS


RESEARCH INFRASTRUCTURE FOR SCIENCE AND INNOVATION POLICY STUDIES

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement N° 824091