REGIONE DEL VENETO

InnovatiVE
SMART SPECIALISATION STRATEGY
VENETO

S3 Monitoring-System Veneto region 2021-2027

Technical documentation

innoveneto.org

Index

Introduction

S3 Veneto introduction

The Smart Specialisation Strategy (S3) is the tool that, since 2014, the EU Regions and member States must adopt in order to identify objectives, priorities, actions capable of maximising the effects of investments in research and innovation, by focusing resources on the specialisation areas of each territory.

The S3, a conditionality linked to the approval of the ROP-ERDF, is a cross-sectoral strategy, an integrated set of tools and actions to strengthen the capacity of the regional system to attract resources from national and European programmes for the support of research and innovation.

The Strategy for the 2021-2027 programming period takes into consideration the results of the previous seven years, the evolution of technology and production systems, the regional research ecosystem, but in particular takes into account the new global challenges, as highlighted by the objectives of the new EU Cohesion Policy 2021-2027 and Agenda 2030.

The S3 2021-2027 of Veneto region is characterised by the transition from a single block made of 4 ambiti of specialisation, as it has been in the 2014-2020 S3, to 3 blocks (ambiti-driver-missioni), more articulated, better fitted to represent the bottom-up emerging needs and more in line with the current transformations, such as the PNRR orientations and the actual social and economical complexity.

The Veneto S3 2021-2027 structure applies a matricial logic, which intersects vertical elements (*ambiti*) with transversal elements (driver) integrating them in goal dimensions (*missioni*). The Ambiti passed from being 4 in the 2014-2020 S3 to 6 in the actual one and they are divided according to 52 Traiettorie (specific trajectories of research and innovation relevant in the regional ecosystem).

Motivation of this platform

To answer the needs of accounting, transparency and public information regarding the S3 and its application, Veneto Innovazione, in collaboration with Siris Academic, has published an online platform that provides institutions, companies and citizens with a wealth of 'open data' on research and innovation policies and the projects they have funded and implemented. The platform, divided into pages with different purposes, allows consultation, by search filters, of various indicators such as the number of projects funded, millions of euro in grants received from the actors of the regional ecosystem.

The platform has been developed with the involvement of all representatives of the innovation ecosystem and has taken into account the results of the previous S3,

covering the 2021-2027 programming period, the evolution of technologies, and markets, the general objectives of regional policies and programmes, and global challenges. It is accessible in open data mode and uses semantic artificial intelligence technologies for the automatic classification of non-classified projects. Each of the projects can be classified, in fact, by *ambito* and *missione*, giving relevance to the interdisciplinary and thematic proximity of the various interventions. Therefore, a tool designed and developed around users and their possible cognitive demands.

Overview of the interactive platform

The interactive platform is a dashboard that allows monitoring the results, within the research and innovation field, of the policies in place in Veneto, whether they are regional, or European. It enables the visualisation, with data and graphs, of the distribution of projects and funding across the territory based on various categorizations and filters.

This dashboard is divided into sections. The header and footer are common elements across all sections of the web platform. They gather key information and contacts related to the Veneto Region and Veneto Innovazione. The footer also includes general contact information for the Veneto Region.

The “Presentazione” page introduces the S3 and its functions and it briefly explains the scope and objectives of the platform itself. It shows the numbers of projects, public funding and actors involved in total and divided by regional and European. It is explained here the structure of Veneto’s S3 for the period 2021-2027 and it is shown a first specialisation of the projects and funding in its categories.

The “Monitoraggio” page shows the numbers of projects, public funding and actors involved through regional calls for projects. This webpage has many dynamic graphical/analytical visualisations that automatically adapt to the imposition of filters. This page is divided into sections, the first section shows some S3 indicators calculated using these data. The second section shows the specialisation of projects and funding according to the S3 structure (ambiti-driver-missioni-traiettorie). The following section shows the distribution of projects, fundings and actors involved according to their geographical location (by province). The last section shows the data divided according to the type of actors participating in the projects/receiving the fundings.

The “Specializzazione” page shows the numbers of projects, public funding and actors involved through regional and European calls for projects. As the “Monitoraggio” page this one has many dynamic graphical/analytical visualisations that automatically adapt to the imposition of filters and it is divided into sections. The first section shows the numbers divided by origin of the calls (regional or European). The second section shows the specialisation of projects and funding according to the S3 structure (ambiti-missioni). The last two sections match the last two of the “Monitoraggio” page, with the only difference of the data taken into account.

The “Dati Aperti” page has been conceived as a query page for the data integrated into the platform. In the first part of the page, the SPARQL endpoint is explained, and it

is possible to download the selectable data in excel form. The second part of the page is dedicated to the SPARQL endpoint itself. Here, a query example is also provided to make requests for information to the platform.

The “Documenti” section presents the documents inherent to the platform itself. All the documents presented in this webpage are downloadable in pdf format either by clicking the document icon or the adjacent “Scarica” button.

Purpose of the technical documentation

This document serves as a comprehensive guide to the S3 Monitoring Platform of Veneto Innovazione and Regione del Veneto, the monitoring dashboard designed to monitor and explore data regarding the R&I policies applied in Veneto and the projects developed through these policies. The purpose of this technical document is to provide stakeholders, administrators, and users with detailed insights into the features, functionalities, and configuration options of the dashboard.

The key objectives of this document are as follows:

- Provide a Comprehensive Overview: Offer a detailed overview of the S3 Monitoring Platform of Veneto Innovazione and Regione del Veneto, outlining its purpose and intended audience.
- Explain Data Sources: Provide an explanation of the data that are used in the dashboard, their origin and their characteristics.
- Show the Infrastructure architecture: Highlight the technical configurations, structure, techniques and tools used to make and update the dashboard.

Co-design process

From the outset, the work carried out by Veneto Innovazione, Veneto Region and Siris Academic has been strongly oriented towards listening to each other with a view to fruitful and effective work. The idea was to arrive, together, to a product rich in the experience of all realities, bearers of different competences.

The co-design and implementation process of the monitoring platform consisted of several stages:

- shared analysis of lessons from the previous programming period;
- definition of the targets;
- definition of system functionalities;
- definition of the graphics and design of the platform webpages;
- testing and continuous improvement of the platform and of the products developed.

Along this path, Siris Academic, Veneto Innovazione and Veneto Region have

constantly discussed strategic and operational issues, bringing proposals and solutions into the discussion.

Data sources

The data sources integrated are outlined below, indicating whether the source will be supplied by the customer or will be retrieved directly from SIRIS (data gathering), and the update frequency is also proposed. The contact person for data curation is also indicated, in charge of the disambiguation of the names of the organisations and validation of the completeness of the information.

Type of date	Source	Data Gathering	Data Curation	Time-period
Research and Innovation European projects (Horizon Europe)	Comunitary informative system on research and innovation (Community Research and Development Information Service - CORDIS) https://cordis.europa.eu/	SIRIS	SIRIS	2021-2024
Projects funded in coherence of the research and innovation Interreg Europe programs	https://keep.eu/	SIRIS	SIRIS	2021-2024
Research and innovation projects funded by Veneto Region	Veneto Region internal data collecting system	Veneto Region	Veneto Region	2021-2024

In the following section are outlined the technical information regarding how the data are collected from the data sources and stored in the relational database.

CORDIS

Data and metadata related to R&D projects and related organisations which have received funding by the European Commission under the HEurope framework program. The data are accessible through Open Data licence, and updated monthly, provided in the format of CSV, XML and Linked Open Data on the CORDIS website. The CORDIS records are collected 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, which are not always correct in the original datasets.

Interreg

Data and metadata related to Inter-regional European projects which have received funding from the European Commission as part of the EU's Cohesion Policy. The data is accessible through Open Data licence, updated monthly, and downloadable in the format of an XLSX file on the Keep.eu website.

Regional funded projects from 2021 to 2027

The regional funded projects are collected manually—from the call and data management organisations—and the related information have been organised according to a common record layout. The relevant information regarding the regional funded projects is available in the document "documento_analisi_monitoraggio_S3".

Infrastructure architecture

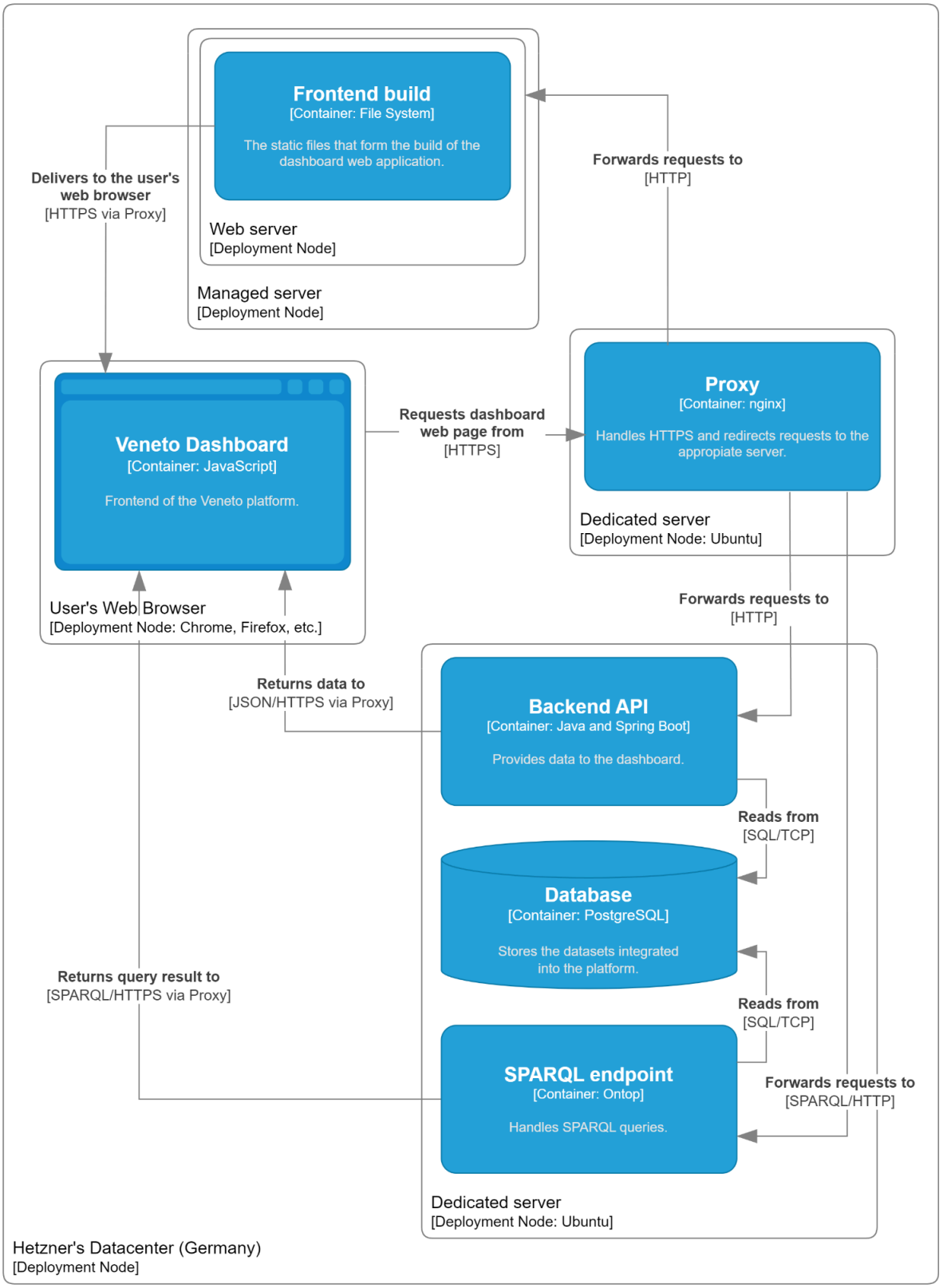
Infrastructure summary

The infrastructure of the platform is depicted in the figure below using the C4 notation. It shows the different containers (i.e. applications and datastores) that compose the platform, as well as the deployment nodes (i.e. the servers) where these are located.

The platform has a single entry point from the Internet in the form of a reverse proxy (deployed on a dedicated server[1]), which redirects the users requests to the proper container/deployment node. A managed web server delivers the frontend files to the user's browser, while a second dedicated server processes the data requests made by the frontend by means of an API and a SPARQL endpoint, which in turn read data from a database that is deployed on the same server.

The servers (both dedicated and managed) are hosted by the German provider Hetzner in one of their data centres.

[1] By dedicated server we mean a machine that comes with just the operative system and a SSH access that allows authorised users access to install whatever software is needed on the server. On the other hand, a managed web server comes with the necessary software already installed and configured, and allows authorised users access only for uploading files to the corresponding web site.


Frontend build
[Container: File System]
The static files that form the build of the dashboard web application.
Web server
[Deployment Node]
Managed server
[Deployment Node]
Delivers to the user's web browser
[HTTPS via Proxy]
Forwards requests to
[HTTP]
Veneto Dashboard
[Container: JavaScript]
Frontend of the Veneto platform.
User's Web Browser
[Deployment Node: Chrome, Firefox, etc.]
Requests dashboard web page from
[HTTPS]
Proxy
[Container: nginx]
Handles HTTPS and redirects requests to the appropiate server.
Dedicated server
[Deployment Node: Ubuntu]
Forwards requests to
[HTTP]
Returns data to
[JSON/HTTPS via Proxy]
Backend API
[Container: Java and Spring Boot]
Provides data to the dashboard.
Reads from
[SQL/TCP]
Database
[Container: PostgreSQL]
Stores the datasets integrated into the platform.
Returns query result to
[SPARQL/HTTPS via Proxy]
Reads from
[SQL/TCP]
SPARQL endpoint
[Container: Ontop]
Handles SPARQL queries.
Forwards requests to
[SPARQL/HTTP]
Dedicated server
[Deployment Node: Ubuntu]
Hetzner's Datacenter (Germany)
[Deployment Node]

Proxy

The proxy is the entrance to the platform. It servers all requests directed towards the platform's domain name, and redirects the requests to the appropriate server/container (i.e. web server, API, etc.)

It handles the security of the web connections by implementing the HTTPS protocol. To this effect, it holds a certificate for the "venetoinnovazione.unics.cloud" domain name that is issued by the Let's Encrypt certification authority.

It is hosted on a dedicated server that has Ubuntu as operating system and nginx as proxy software.

Web server

The web server holds the static HTML, CSS, Javascript and image files that form the frontend web application of the platform. Its job is to deliver these files to the user's web browser.

It is on a managed server (also provided by Hetzner) specific for hosting web sites, to which files are uploaded via secure FTP.

Backend API

The API provides to the frontend the data that feeds the different visualisations. This data is retrieved by the API from the database via SQL queries, and returned to the frontend application in JSON format.

It is implemented in Java 17, using the latest version of the Spring Boot framework.

Ontop-powered SPARQL endpoint

Ontop is a Virtual Knowledge Graph system that allows users to interrogate a relational database through an ontology.

An ontology is a conceptual model of a domain of interest. Knowledge is represented in an ontology in the form of concepts and properties. Some properties represent relationships between concepts (object properties), while other properties serve to characterise concepts by associating them with data values (numbers, texts, dates, etc.). Data that is structured according to an ontology follows the format defined by the Resource Description Framework (RDF) standard.

The users of an Ontop-powered information system see the data structured according

to the ontology, and query the data also according to the ontology. The language used to write queries to the system is SPARQL, a standard from the W3C. Ontop receives queries written in this language and translates them into queries that are executable by the underlying relational database, that is, the SQL language. Once in SQL, the query is executed by the relational database and provides a result. This result is translated by Ontop into the format of a SPARQL query result and returned to the user.

Ontop serves thus to hide the complexities of the underlying database from the users. Databases are designed mostly with performance in mind, and tend to be not particularly user friendly. Besides, the structure of a database tends to change over time as needs for new types of queries arise that require a restructuration of the data to perform efficiently, or new datasets are introduced that overlap with existing ones.

In this way, the SPARQL endpoint offered by Ontop complies with the requirements of a Linked Open Data (LOD) system, which is the recommended way of publishing open data on the web.

In the context of the Veneto S3 Monitoring system, a SPARQL console that allows access to the Ontop-powered endpoint is available in the "Dati Aperti" section of the dashboard.

Database

The data integrated into the S3 Monitoring platform is stored in a PostgreSQL relational database, which is accessible to both the API and the Ontop-powered SPARQL endpoint.

All the details about the database can be found in the “Database Structure” technical document.

Data Integration and Enrichment

Data disambiguation

Most of the process of cleaning and transforming raw data (data curation) is concentrated in the aspects of disambiguation, harmonisation and geolocation. This is because, given the various practices and standards that exist when creating and managing data, very often there are cases where an entity is identified in different ways. When it comes to research and innovation project data, the main entities to identify are:

- The projects, identifying a single piece of content in the data source for each project. If there is note an identifier with these capabilities, it will be automatically generated based on the title.

- Organisations, which need to be identified not only in an isolated data source, but also across sources. This makes unique identification of organisations generally more costly than individual project identification. In one data source—and more often when working with multiple data sources as in this case—each actor may appear with different names (either because they appear in different languages, or because their legal name is used instead of the most commonly used one). It is therefore necessary to uniquely identify the different nomenclatures present in the various data sources.

The unique identification of the actors is a fundamental task, since their duplication would lead to erroneous results (for example when quantifying the funding received by an actor). This procedure consists of the following two stages:

- Institutional name disambiguation, i.e. identifying different nomenclatures that refer to the same organisation in the same data source. This process is done using the OpenRefine tool, a free software tool equipped with various algorithms to identify entities with similar names. OpenRefine is one of the most used tools for processing and dealing with different nomenclatures coming from various data sources. Over the years it has become a point of reference for this and other tasks related to Data Science activities.

- Name alignment, i.e. identifying the same organisation with "disambiguated" names in different data sources. In this case, the tool used will still be OpenRefine, which includes reconciliation capabilities, used to identify an institution in an external metadata repository, such as DBPedia, Wikidata, GRID, ROR, etc. Identification in external metadata repositories is done via the institution name, which once found makes it possible to assign it a unique external identifier (for example, ROR ID, the unique internal identifier used by the ROR metadata repository). This external identifier helps align your organisation across data sources, since the same entity across multiple sources is assigned the same external identifier. Using metadata repositories also allows the retrieval of details such as common names, acronyms, external identifiers, official websites, etc.

This process of disambiguation and alignment of names is validated with domain experts with whom SIRIS regularly collaborates, in order to incorporate the context of the data into the identification process. Validation is done with an internal tool created by SIRIS Academic.

Once the organisation has been properly disambiguated in each source and aligned across the various sources, the corresponding typology is assigned. In the case of Horizon Europe, CORDIS provides this information for each institution, which takes on five possible values:

Acronimo	Etichetta	Descrizione
HES	Education institution	Secondary and higher education institutions

REC	Research centre	Research institutions (excluding educational institutions)
PRC	Enterprise	Private companies
PUB	Public Administration	Public administrations (excluding those of research and education)
OTH	Other	Other entities, such as non-profit organisations and philanthropic institutions

Within the platform, the categories are slightly modified from the ones provided by CORDIS. While "Enterprise", "Public Administration" and "other" remain the same, "Education institution" is splitted between "University" and "Formation institution". "University" is merged with "Research centre" and "Formation institutions" form a category of its own. So the platform categories are:

- Enterprise - Imprese
- Public administration - Ente pubblico
- Research centre/University - Organizzazione di ricerca/Università
- Formation Institution - Enti di formazione (compresi ITS)
- Other - Altro

The procedure for identifying the types of entities is a manual process, which can be simplified, for example, by using the abbreviations of the types of commercial companies (S.r.L, S.p.A., etc.) to identify the companies.

Data classification

The projects which are not previously classified, meaning all the projects that come from sources different from the regional database, are processed by an automatic classification system. The classification system is based on machine learning models that allows it to extrapolate meaningful information from textual data. It classifies projects according to the Veneto region S3 2021-2027 structure, assigning them Ambiti and Missioni.

All the details about the data classification can be found in the "Classification document" technical document.

Particular cases

Both in the case of projects from the regional database (previously classified) and of projects from the other data sources (classified with the automatic classifier), there have been cases of projects classified either with all the 6 Ambiti or with no Ambito at all. When this occurred, the projects have been singularly and manually evaluated and categorised in:

- Projects non-coherent to the S3 of the Veneto region, in which case they have been excluded from the database of the platform;
- Projects coherent to the S3 of the Veneto region but not specifically coherent to any of the Ambiti, in which case they have been kept in the database of the platform and assigned with a tag “Ambiti emergenti” and possibly a Missione, in case of non regional projects, and either a Missione, a Driver or both, in case of regional projects.