

scanR - Explore public data on French research and innovation

Providing an open portal and service to the higher education and research community in France

<https://scanr.enseignementsup-recherche.gouv.fr/>

Proposal for EuroCRIS - Paris November 2024

Authors: Eric Jeangirard (<https://orcid.org/0000-0002-3767-7125>)

Contact: eric.jeangirard@recherche.gouv.fr

Affiliation: French Ministry of Higher Education and Research, Paris, France

(<https://ror.org/03sjk9a61>)

scanR is a web application designed to help understand the French research and innovation ecosystem by concentrating as much open information as possible around 5 main objects (structures, authors, public funding from calls for projects, research publications, and patents). Since 2016, scanR has adopted an open approach (method, data, source code, API) that allows data, beyond the application, to be used in other contexts by many other players in the ecosystem.

1. Objectives

scanR aims to provide a 360-degree view of research and innovation in France, based solely on open data. As far as possible, the scanR service is developed and maintained by in-house teams, with a view to keeping costs under control and ensuring greater adaptability to change.

scanR aims to index several types of objects:

- entities
- publications, PhD dissertations, patents
- funding
- authors

and link them together. Each object is identified by a national or international PID, which scanR then seeks to link (for example, author identifiers are linked to a publication).

Unfortunately, the sources available to scanR do not generally contain these links between identifiers. scanR takes the approach of calculating these links, using heuristics or machine learning techniques, thus accepting potential errors.

In this way, scanR should not be seen as a monitoring or evaluation tool, but as an exploration tool, making it easy to search within a single tool.

This tool has several facets:

- the **data** itself, which aggregates, cross-references and standardizes numerous sources
- the **website**, which facilitates data exploration via a search engine and dataviz
- fast, high-performance **infrastructure** providing technical APIs for a variety of use cases

2. Methodology: data, website and infrastructure

a) Data: concentrate open data sources and fill in the gaps

scanR does not collect data from institutional CRIS (there are almost none in France). On the contrary, scanR uses massive (and open) sources and seeks to extract what concerns the French research and innovation ecosystem. A major step is to use crossref to identify publications with a French affiliation. Affiliation strings are crawled from the web landing pages and an algorithm detects the affiliation countries (L'Hôte, 2021). This original methodology was validated by an external benchmark comparing different bases of French publications. (Chaignon, 2022)

For entities and structures, several international (RoR in particular) and national registers are consolidated to provide a picture of the public and private research and innovation landscape.

For author disambiguation, ORCID proves useful, but still largely insufficient for the entire corpus. The national idref.fr register, maintained by the French Bibliographic Agency of Higher Education (ABES), is useful for identifying authors linked to the French research ecosystem.

Metadata generally needs to be included for links between publications, structures, and authors. Filling the gaps is done using automated matching processes, in an attempt to trace the scientific production of a laboratory or researcher.

The identification of funding is still very much linked to the willingness of funders to display transparent open data (even though transparency on the beneficiaries of public subsidies is, in theory, mandatory). European and ANR (main French funder) projects are the best described for the moment, but more and more funding agencies are getting organized to expose funding data (as well as calls on an emerging common platform <https://www.appelsprojetsrecherche.fr/>)

b) User interface and visualisations (website)

The website features three main functions

- 1. a search engine, to explore a theme, for example, and list all laboratories or authors related to that theme

Home > Search organizations

"climate change" Search

1,574 organizations for search term « "climate change" »

STRUCTURE DE RECHERCHE > UNITÉ DE RECHERCHE

[Interactions Hôtes-Pathogènes-Environnements](#)

Perpignan

Biophysical modelling to investigate the effects of **climate change** on marine population dispersal and connectivity, Rapid adaptive responses to **climate change** in corals

STRUCTURE DE RECHERCHE > UNITÉ DE RECHERCHE

[Ecology of Mediterranean Forests Ecologie des Forêts Méditerranéennes](#)

Avignon

Climate change impacts on forest fires : INRA contribution, Forest genetic resources and **climate change**, Agroclimatic conditions in Europe under **climate change**

STRUCTURE DE RECHERCHE > UNITÉ DE RECHERCHE

Filters

Add filters

Export results

For performance reasons, scanR only exports a maximum of 1000 results. You can refine your search to get a more complete export.

Distribution by sector

Calculated from search results

- 2. a specific page dedicated to each entity / author

STRUCTURE DE RECHERCHE > UNITÉ DE RECHERCHE > UNITE MIXTE

ractionnement des agro-ressources et environnement

active since 1994

The main research mission of FARE is to generate and promote knowledge in the areas of degradation and fractionation of lignocellulosic biomass, thereby contributing to major bioeconomy-related issues. The fields of application that pertain to the scope of its mission are maintaining soil fertility by controlling biogeochemical cycles (C, N, P and S), producing and using fibres and polymers for materials, and producing molecules and building blocks or chemistry and energy via biotechnology.

Leadership team

Directeur BERNARD KUREK	→	Directeur GABRIEL PAES	→
Adjoint CAROLINE REMOND		Adjoint Gwenaëlle LASHERMES	→

Membership and networks

Supervising institution (2)

University of Reims Champagne-Ardenne since 01/01/2004	→
Institut national de recherche pour l'agriculture, l'alimentation et l'environnement since 01/01/2020	→

Location

2 Esplanade Roland Garros
51100, Reims, France

On the Web

Website

Structure identifiers

Click to copy to clipboard

RNSR : 199417885W

WIKIDATA : Q30262421

LABEL_NUMERO : UMR 0614

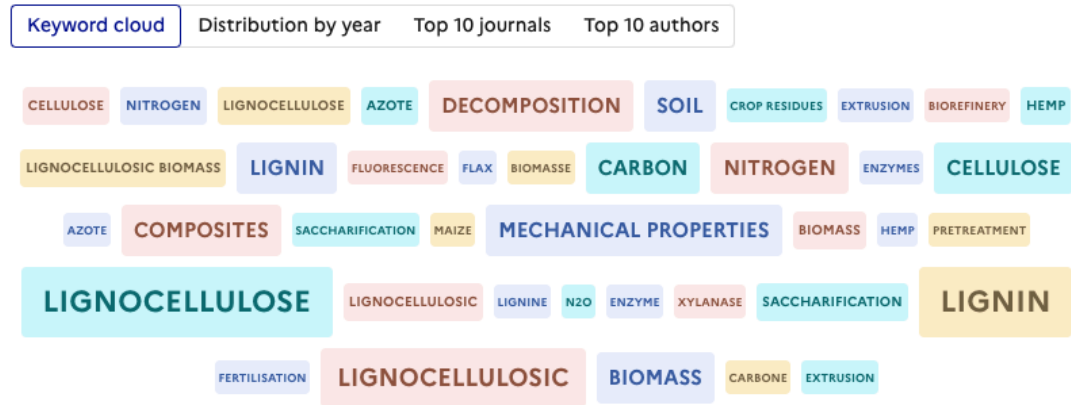
LABEL_NUMERO : UMR 614

Research and innovation activities

766 publications listed by scanR

[See the list of publications >](#)

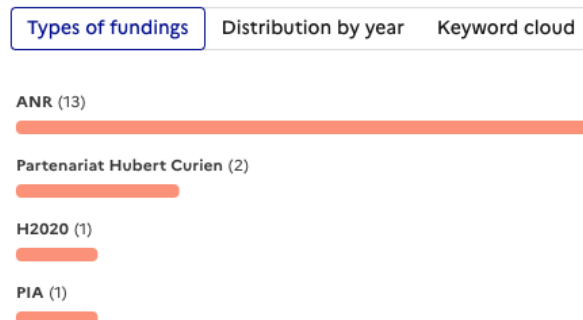
Select a type of graphical representation



17 funding listed by scanR

[See the list of fundings >](#)

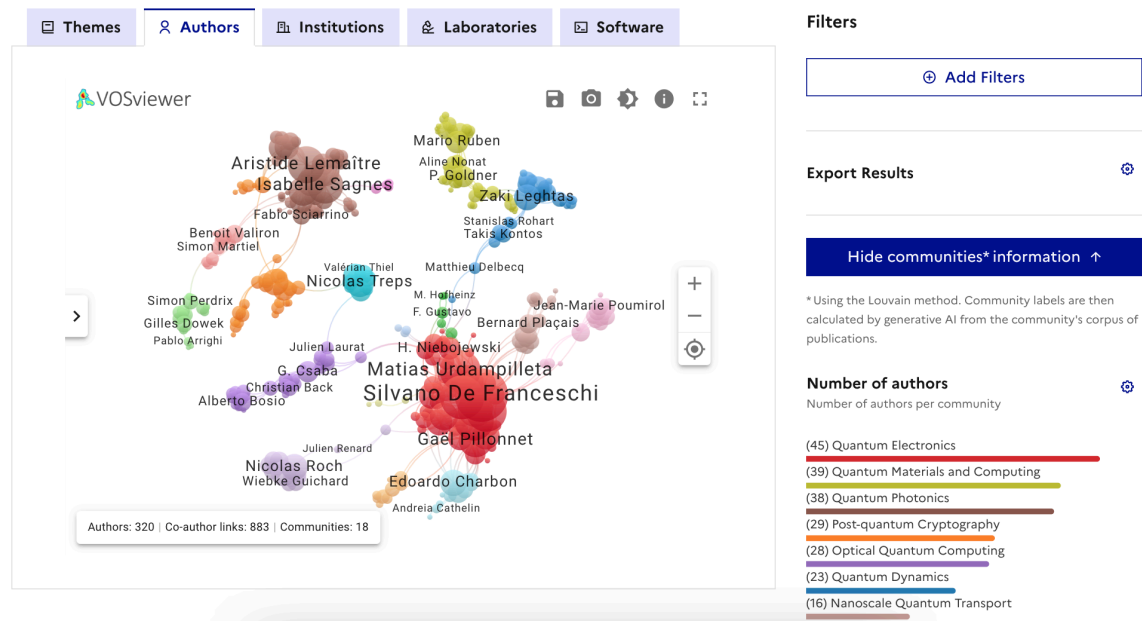
Select a type of graphical representation



- 3. a network analysis, based on a search query, leveraging all the links in the underlying data

Visualization of networks through a publications analysis

Estimated on the basis of co-signatures / co-appearances in the corpus of publications corresponding to the research carried out ②



Network clusters are computed on the fly and an LLM is used to label dynamically those clusters.

c) Technical infrastructure and API

The data processing pipelines, the elasticsearch indices, and the UI application all run on a public cloud infrastructure (OVH), on several clusters part of a managed Kubernetes service.

Using a cloud provider gives us access to flexible resources and easier deployment of microservices (docker).

d) Feedback

Users are encouraged to send feedback to improve and correct data.

A dedicated webapp for the feedback management has been implemented.

Contribution par objets

Rechercher par nom ou ID

Q Rechercher

Résultats: 1-20 de 1668

Retour à la page 1

Dernière page

Plus récentes

Toutes les contributions

☐ Rechercher dans les messages

Contributeurs

PERSONNES NOUVEAU

DENIS 12/09/2024

Bonjour, Je souhaite rectifier une confusion sur ma page Scan

PERSONNES NOUVEAU

Jean 11/09/2024

Bonjour, Je viens de démêler un cas d'homonymie pour Ekate

PERSONNES NOUVEAU

Bert 09/09/2024

Bonjour, Cet article "Ontoquer" n'est pas co rédigé par 'Barbi

PUBLICATIONS TRAITÉ ACCÈS THÈSE

Thierr /2024

Depuis des jours ... le lien de téléchargement ne fonctionne

NOUVEAU PERSONNES

DENIS 66e2b0fd73a066ee347d7b72) Reçu le 12/09/2024

Aucune réponse apportée à ce message pour l'instant

ID de l'objet concerné: Nom:

Email:

Organisation non renseignée Fonction: Maître de Conférence

Sur scanR Sur dataEsr

Bonjour, Je souhaite rectifier une confusion sur ma page ScanR : dans la catégorie "Participations à des jurys de thèse", les 4 résultats qui apparaissent

3. Usage

On the website, scanR registers around 50,000 monthly visits.

New API keys are requested every month, by new research institutions and companies building services on top of scanR data. About 50 API keys have been generated between April and September 2024.

Use cases are very different from one another, for example

- help in setting up expert panels (avoiding conflicts of interest)
- feed for local RIS
- building collaborative public-private projects
- building research consortium
- ...
- and even marketing targeting

4. Key take-aways

a) No use of proprietary sources brings benefits!

- It is possible at the national level to build a national research portal with a top down approach (no CRIS systems) without any proprietary data, taking advantage of the progress in machine learning and cloud computing.

- That requires more work on the data quality, that has to be checked with local knowledge of the Higher Education and Research ecosystem
- That allows to openly share the results (indicators, code, data) but also services can be built on top of the resulting data and API

b) An iterative process is needed to improve and extend the results

- scanR has been built step by step, layer after layer since 2016
- The machine learning models constructions are themselves iteratives processes as they improve with more feedbacks
- Exploring hidden things is complicated, and the development of PID policy or implementation enables a more accurate description

c) Collaborations at different scales are key

scanR harvests, creates and provides open data and open services. This fosters collaborations at many scales.

For example, large institutions use scanR services in their own infrastructure, which incites large actors (funders for example) to make their data open so that it can be disseminated, creating a virtuous circle around data enrichment and circulation.

scanR also uses existing large open infrastructures, especially for PID, like Crossref, RoR, and ORCID, but also French infrastructures like idref and HAL. scanR uses shared data, but also code, like the online VOSviewer developed by the CWTS.

Smaller collaborations like user feedback are also key to building trust and higher quality metadata.

References

- Lauranne Chaignon, Daniel Egret; Identifying scientific publications countrywide and measuring their open access: The case of the French Open Science Barometer (BSO). *Quantitative Science Studies* 2022; 3 (1): 18–36. doi: https://doi.org/10.1162/qss_a_00179
- Anne L'Hôte, Eric Jeangirard. Using Elasticsearch for entity recognition in affiliation disambiguation. 2021. <<http://hal.science/hal-03365806>>