



**MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE**

*Liberté
Égalité
Fraternité*

scanR - Explore public data on French research and innovation

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

euroCRIS
27 novembre 2024

Eric Jeangirard

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

The web application scanR is a **research portal**

designed to help understand the French research and innovation ecosystem

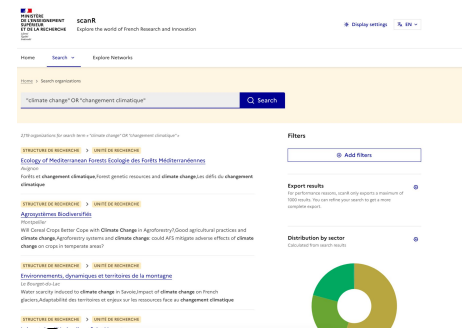
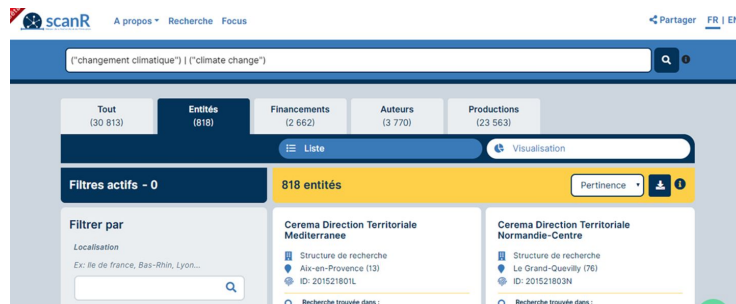
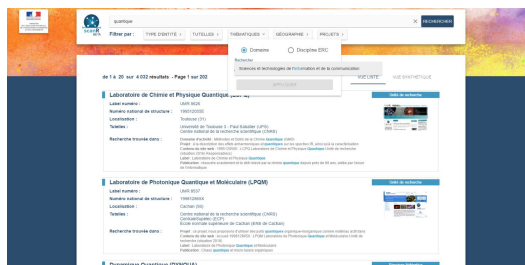
It gathers open information around 5 main objects: **structures, authors, public funding, publications and patents**

Launched in 2016, scanR adopts an open approach: method, data, source code, API
⇒ **the services and data provided can be used in other contexts by many other players in the ecosystem.**

The screenshot shows the scanR web application interface. At the top left is the logo of the Ministry of Higher Education and Research, with the text 'MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR ET DE LA RECHERCHE' and the motto 'Liberté Égalité Fraternité'. To the right of the logo is the text 'scanR' and 'Explore the world of French Research and Innovation'. Further right are 'Display settings' and a language selector 'EN'. Below the header is a navigation bar with 'Home', 'Search', and 'Explore Networks'. The main content area features a large yellow and white background with the text 'Explore the world of French Research and Innovation with scanR'. Below this is a section titled 'Explore public data on French research and innovation...' with five icons and labels: 'Organizations', 'Authors', 'Funding', 'Publications', and 'Patents'. Each icon has a 'Search' button below it.

The vision behind scanR remains the same but the service keeps evolving

Since 2016, the vision behind scanR remains the same offering an open service to provide a 360° view of research and innovation in France



2016

2018

2020

2022





2024

A new data pipeline to process all the French publications is implemented for the French Open Science Monitor - and will be used for scanR

Complete reinternationalization of the project: data, backend, frontend

scanR the French research portal

This tool has several facets:

-  the **data** itself, which aggregates, cross-references and standardizes numerous sources
-  the **User Interface** (website) which facilitates data exploration via a search engine and dataviz
-  fast, high-performance **infrastructure** providing technical APIs for a variety of use cases
-  the **community of users** that also provide feedback

Producing rich and linked metadata is the hidden part of the iceberg



PID are key for each of the 5 types of objects are indexed institutions/labs, authors, publications, patents, funding



Many global sources are collected / harvested.

Web scraping, PDF mining, Large dumps, APIs etc ...

Merging information (deduplication, enrichment of complementary information between sources)



The links between the objects (e.g authors to publications and affiliations) generally needs to be computed (AI, heuristics ...). Errors can be made.



Other enrichment can bring added values

Language detection, topics, classification, open access/licence, dataset and software use ...

Again some AI used, and again errors can be made.

The User interface is the tip of the iceberg



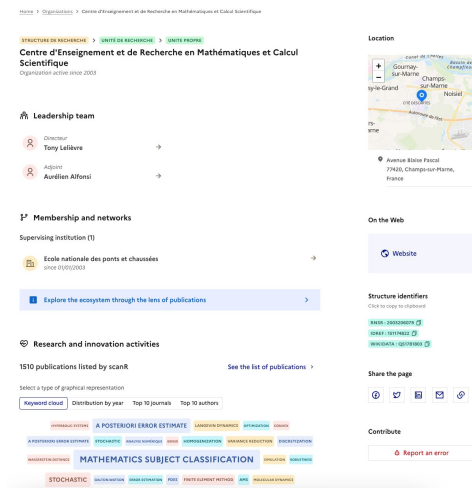
The first use case for scanR (from 2016) is the search engine
⇒ search interface to find an object
⇒ full description page of the object



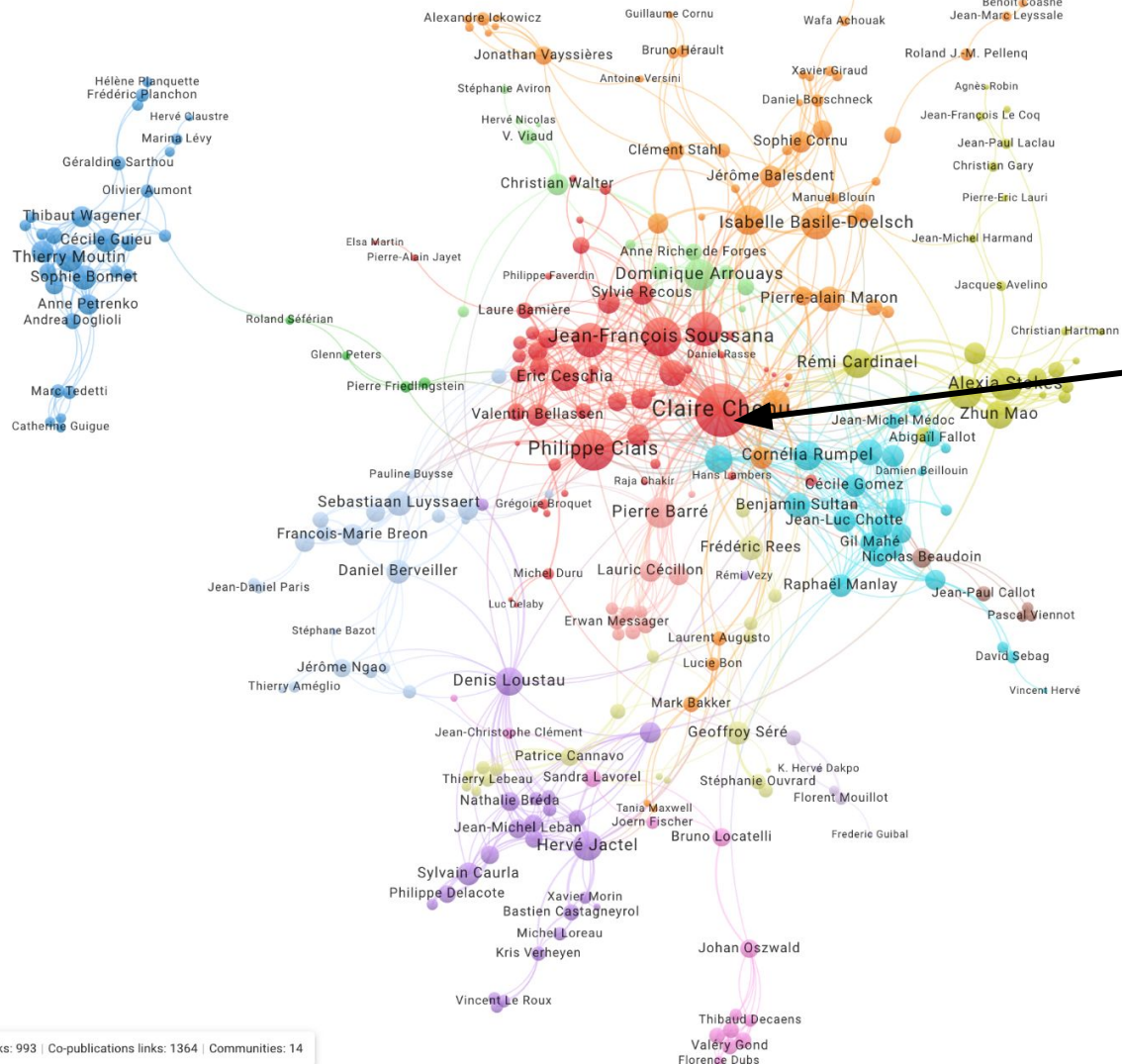
A new use case: mapping research communities
to give more insights than simple flat statistics (top authors,
top keywords ...)?

The idea is to give insights about the underlying structure of
the search results:

- detecting “groups” or clusters
- describing those groups
- analysing how those groups relate to each other



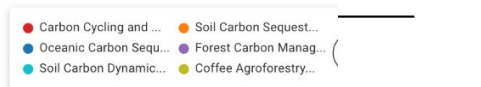
Query: carbon AND sequestration



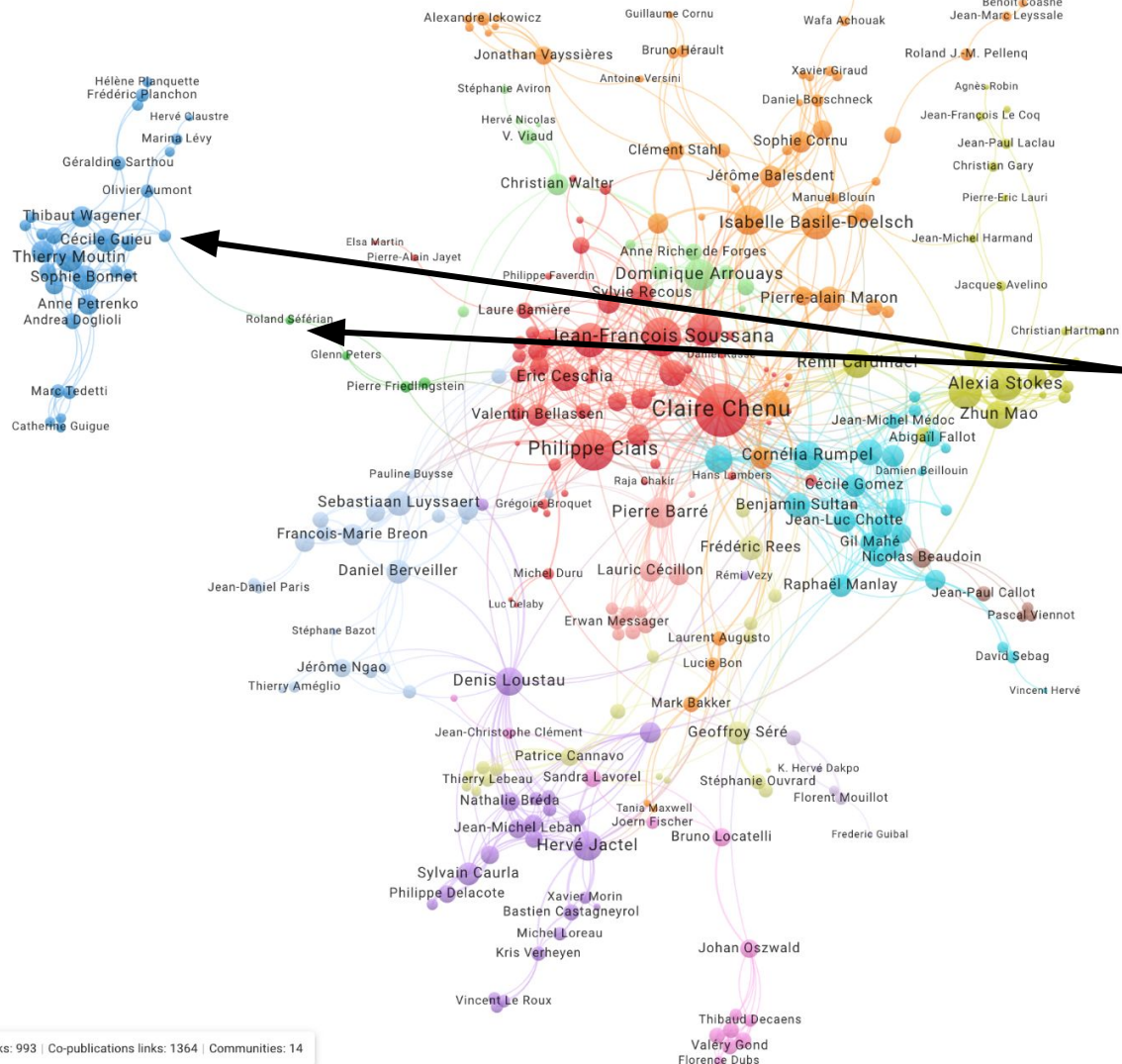
Each node is an author of at least 1 publication in the corpus

The size of the node is relative to the number of publications (in the corpus)

VosViewer (from CWT5) is used to visualize the network

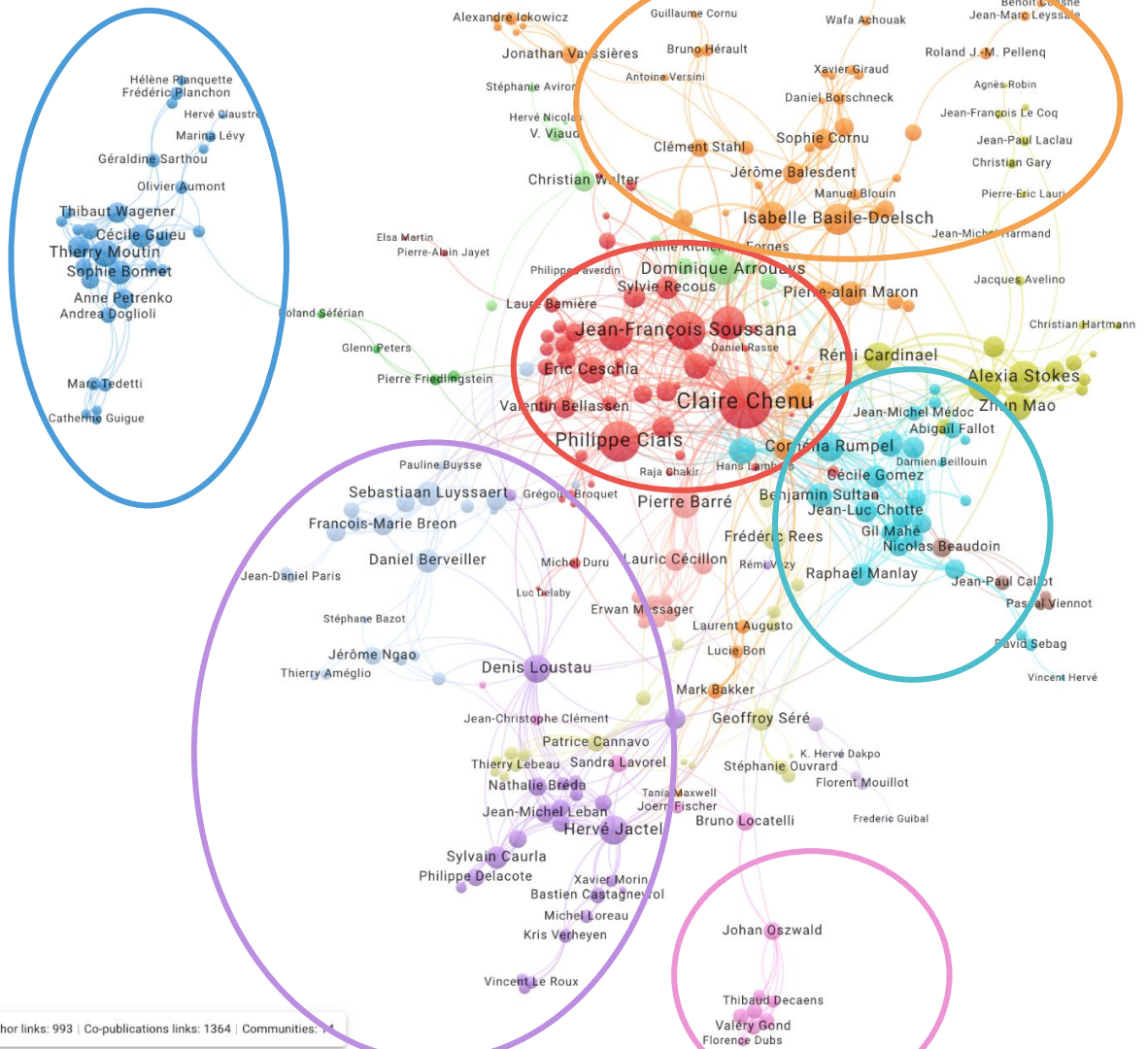


Query: carbon AND sequestration



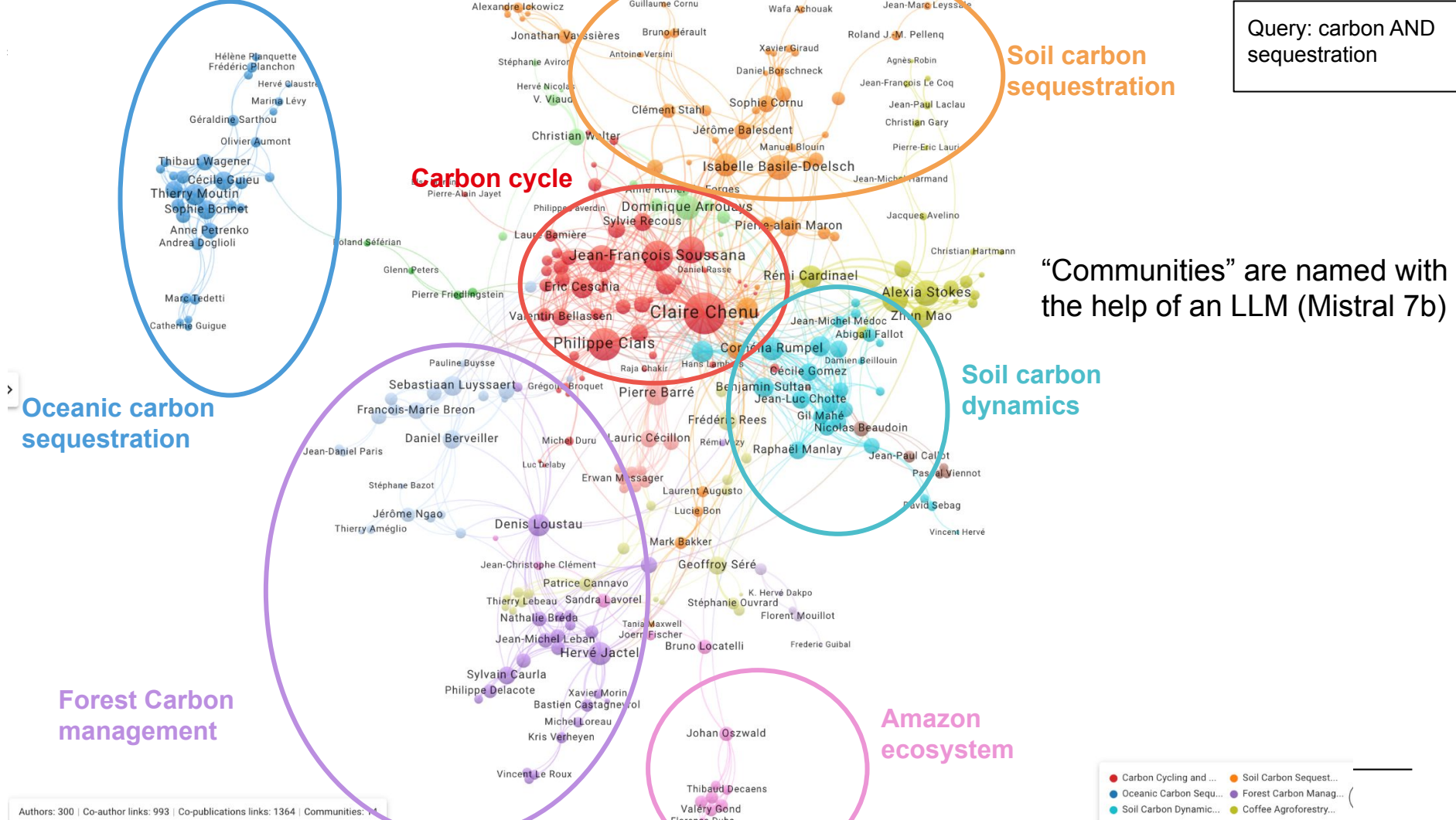
Two nodes are linked if they co-appear in the publication (co-author)

Query: carbon AND sequestration



“Communities” are detected with the Louvain algorithm (modularity optimization)

Query: carbon AND sequestration



Soil carbon sequestration

Carbon cycle

Oceanic carbon sequestration

Forest Carbon management

Soil carbon dynamics

“Communities” are named with the help of an LLM (Mistral 7b)

Amazon ecosystem

Authors: 300 | Co-author links: 993 | Communities: 14

Hide communities* information ↑

Using the Louvain method. Community labels are then calculated by generative AI from the community's corpus of publications.

Authors communities (14)

54 AUTHORS | 121 PUBLICATIONS | OPEN ACCESS: 71.1% | LAST PUBLICATION: 2024 | 3,144 CITATIONS (2023-2024)
CITATION SCORE: 26.0

Carbon Cycling and Soil Sequestration

Claire Chenu, Philippe Ciais, Jean-François Soussana, George R. R. Martin, Bertrand Guenet, Eric Ceschia, Peter H. Free-Smith, Sylvie Recoux, Valentin Bellassen, Nicolas Vuichard, ...

#Carbon, #Carbon Sequestration, #Soil, #Soil Organic Carbon, #Climate Change, #Climate Change Mitigation, #Soil Carbon Sequestration

50 AUTHORS | 55 PUBLICATIONS | OPEN ACCESS: 69.1% | LAST PUBLICATION: 2023 | 777 CITATIONS (2023-2024)
CITATION SCORE: 14.3

Soil Carbon Sequestration

Sabbele Basile-Doelsch, Sébastien Fontaine, Delphine Derrien, Pierre-alain Maron, Nicolas Fanin, Jérôme Balesdent, Sophie Cornu, Catherine Picon-Cochard, Christine Hatté, Laetitia Bernard, ...

#Soil, #Carbon Sequestration, #Soil Organic Matter, #Carbon, #Ecosystem Services, #Priming Effect, #Sequestration, #Amazonian, #Andosol

30 AUTHORS | 25 PUBLICATIONS | OPEN ACCESS: 92.0% | LAST PUBLICATION: 2022 | 439 CITATIONS (2023-2024)
CITATION SCORE: 17.6

Oceanic Carbon Sequestration

Thierry Moutin, Sophie Bonnet, Cécile Guieu, Sandra Helias Nunige, Thibaut Wagener, Anne Petrenko, Pascale Bouruet-Lubertot, Karine Leblanc, Nathalie Leblond, Karine Desboeufs, ...

#ACI, #Carbon, #Carbon Sequestration, #South Pacific Ocean, #Trichodesmium, #Crocospaera, #Crocospaera-vastronli, #Dinitrogen-fixation, #Dissolved Organic Nitrogen, #High-precision

25 AUTHORS | 30 PUBLICATIONS | OPEN ACCESS: 63.3% | LAST PUBLICATION: 2022 | 491 CITATIONS (2023-2024)
CITATION SCORE: 16.4

Forest Carbon Sequestration

Hervé Jactel, Denis Loustau, Sylvain Cauria, Laurent Saint-André, Jean-Michel Leban, Nathalie Bréda, Antoneilo Lobianco, Jean-François Dhôte, Eric Rigolot, Philippe Delacote, ...

#Carbon Sequestration, #Forest Sector, #Biodiversity, #Economie Forestiere, #Ecosystem Services, #Biomass Energy, #Bio-economic Model, #Carbon Storage

25 AUTHORS | 38 PUBLICATIONS | OPEN ACCESS: 76.3% | LAST PUBLICATION: 2024 | 877 CITATIONS (2023-2024)
CITATION SCORE: 23.1

Soil Carbon Management

Cornélia Rumpel, Nicolas Viovy, Julien Demenois, Benjamin Sultan, Raphaël Manlay, Jean-Luc Chotte, Cécile Gomez, Eric Blanchart, Maud Loireau, Tiphaine Chevallier, ...

#Soil Organic Carbon, #Carbon Sequestration, #Climate Change, #Climate Change Mitigation, #Carbon, #Grassland, #Soil Carbon Sequestration, #Ecosystem

22 AUTHORS | 39 PUBLICATIONS | OPEN ACCESS: 48.7% | LAST PUBLICATION: 2023 | 515 CITATIONS (2023-2024)

Export Results

JSON
Export results in JSON format

XLSX
Export results in XLSX format

Number of authors

Number of authors per community

- (54) Carbon Cycling and Soil Sequestration
- (50) Soil Carbon Sequestration
- (30) Oceanic Carbon Sequestration
- (25) Forest Carbon Sequestration
- (25) Soil Carbon Management
- (22) Coffee Agroforestry Carbon Sequestration
- (22) Forest Carbon Dynamics
- (20) Technosol Carbon Management
- (13) Amazonian Ecosystem Carbon Services
- (13) Soil Organic Carbon Sequestration

Number of publications

Number of publications per community

- (121) Carbon Cycling and Soil Sequestration
- (55) Soil Carbon Sequestration
- (25) Oceanic Carbon Sequestration
- (30) Forest Carbon Sequestration
- (38) Soil Carbon Management
- (39) Coffee Agroforestry Carbon Sequestration
- (25) Forest Carbon Dynamics
- (19) Technosol Carbon Management
- (8) Amazonian Ecosystem Carbon Services
- (34) Soil Organic Carbon Sequestration

Number of publications per year

Number of publications per community

Number of citations per year

Number of citations per community

Query: carbon AND sequestration

“Communities” are described with a few metrics:
main nodes, keywords, number of publications, number of citations ...

All the results can be extracted to Excel for further analysis / merge with other datasource etc...

How does it work at scale?

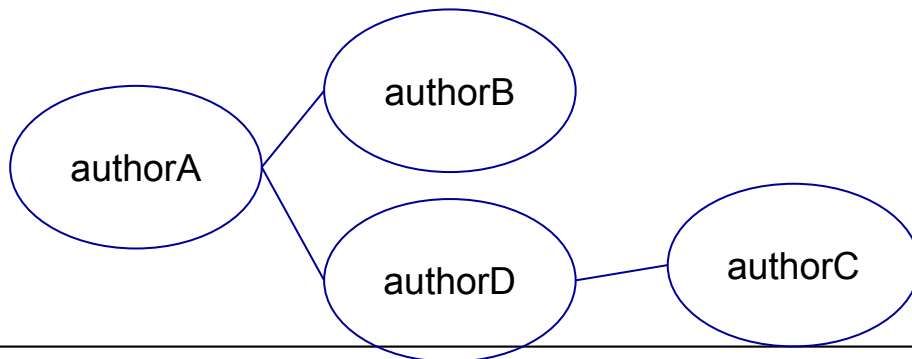
Network analysis on very large corpus tend to be very computationally consuming as the number of nodes and edges increase.

As a web application, scanR has to keep the user waiting time reduced to a few seconds maximum.

scanR does not build the network not from the corpus itself: it would need to loop through all the publications one by one

Instead, it builds the network from the list of the top pairs of entities, with an Elasticsearch aggregation

e.g Top co-authors
authorA - authorB
authorA - authorD
authorC - authorD

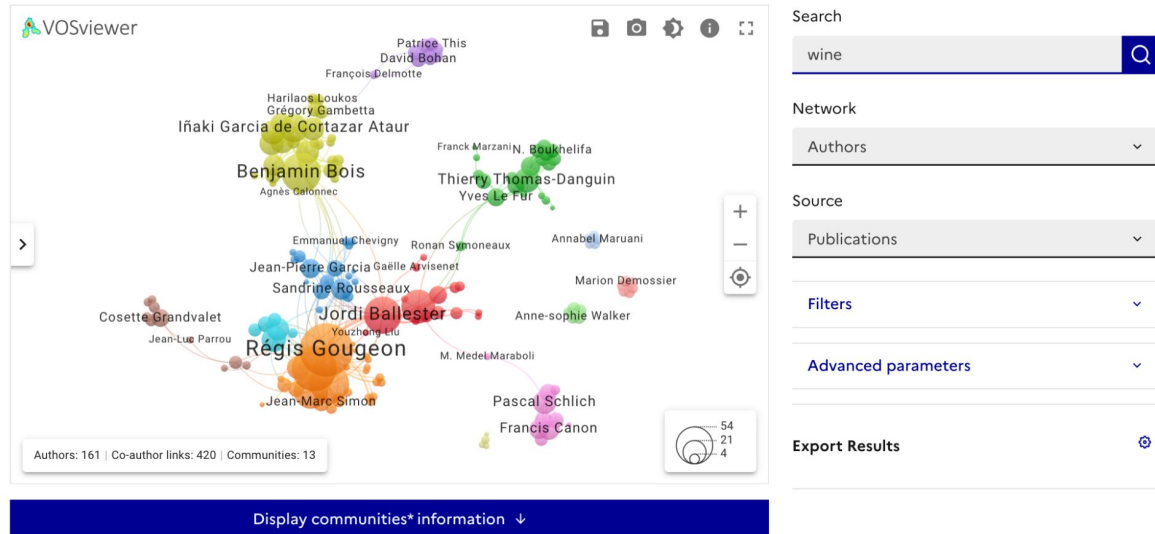


From a national to a local tool

The same tool can be deployed at an institutional level: only the publications from the institution are analyzed (example with uB - university of Burgundy)

Visualization of authors networks through a publications analysis of uB

i Estimated on the basis of co-signatures / co-appearances in the corpus of publications corresponding to the research carried out [ⓘ](#)



*Using the Louvain method. Community labels are then calculated by generative AI from the community's corpus of publications.

Infrastructure on cloud gives the flexibility we need

Public **cloud infrastructure** (OVH) hosts:

- the data processing pipelines
- the elasticsearch indices
- and the UI application

Several clusters part of a **managed Kubernetes service**

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

We **use open source** (like the VosViewer!) and **produce open source**: all our source repositories are public on github

User feedback is a valuable source of information

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

A dedicated webapp for the feedback management has been implemented.

Some feedback have pattern, like correcting affiliations. We develop a dedicated app, the works-magnet to collect this information.

<https://works-magnet.dataesr.ovh>

Contribution par objets

Rechercher par nom ou ID

Résultats: 1-20 de 1668

Rechercher dans les messages Désactivé

Contributeurs

- PERSONNES** **NOUVEAU**
DENIS Recu 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
- PERSONNES** **NOUVEAU**
Jean
Borjour, Je viens de démler un cas d'homonymie pour Ekate
- PERSONNES** **NOUVEAU**
Bert
Borjour, Cet article "Ontoquer" n'est pas co rédigé par "Barbi
- PUBLICATIONS** **TRAITÉ** **ACCÈS THÈSE**
Thierf
Depuis des jours ... le lien de téléchargement ne fonctionne

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

Works-magnet to improve affiliations metadata

Affiliation name, ROR of your institution  *

Press ENTER to search for several terms / expressions. If several, an OR operator is used.

02dga6j42 X

École Supérieure des Sciences Économiques et Commerciales X

ESSEC X

ESSEC Business School X

Institut Economique X

 [Get children from ROR](#)

ROR to exclude: exclude affiliation strings already mapped to a specific ROR in OpenAlex

You can focus on recall issues in OpenAlex (missing ROR). This way, only affiliation strings that are NOT matched in OpenAlex to this specific ROR will be retrieved. If several ROR to exclude, separate them by space.

You can search

- by string
- by ROR id









You can add ROR children

You can exclude ROR id


















Use cases

- I want to make sure all the publications from ESSEC have the ROR <https://ror.org/02dga6j42>
- I want to make sure all the publication with the ROR <https://ror.org/02dga6j42> are from the ESSEC institution

Error detection within the works-magnet

OpenAlex Raw affiliation	ROR computed by OpenAlex ↑↓	Click to improve / edit RORs	Modified by user? ↑↓	Works ↓↕
<input type="checkbox"/>				
<input type="checkbox"/> essec business school department of accounting and management control [source: OpenAlex]	ROR https://ror.org/02dga6j42 (École Supérieure des Sciences Économiques et Commerciales - FR)	• 02dga6j42		<ul style="list-style-type: none"> • W4386014955  • 10.2139/ssrn.45460..  • W4366396136  • 10.2139/ssrn.44116.. 
<input checked="" type="checkbox"/> essec higher school of economic and commercial sciences university of douala [source: OpenAlex]	ROR https://ror.org/02dga6j42 (École Supérieure des Sciences Économiques et Commerciales - FR) ROR https://ror.org/02zr5jr81 (University of Douala - CM)	• 02dga6j42 • 02zr5jr81		<ul style="list-style-type: none"> • W4385801793  • 10.47191/ijmra/v6-..  • W438577206  • 10.5281/zenodo.823.. 
<input checked="" type="checkbox"/> essec higher school of economic and commercial sciences university of douala [source: OpenAlex]	ROR https://ror.org/02dga6j42 (École Supérieure des Sciences Économiques et Commerciales - FR) ROR https://ror.org/02zr5jr81 (University of Douala - CM)	• 02dga6j42 • 02zr5jr81		<ul style="list-style-type: none"> • W4385801793  • 10.47191/ijmra/v6-..  • W438577206  • 10.5281/zenodo.823.. 

Error detection within the works-magnet

OpenAlex Raw affiliation	ROR computed by OpenAlex ↑↓	Click to improve / edit RORs	Modified by user? ↑↓	Works ↓↕
<input type="checkbox"/>				
<input type="checkbox"/> essec business school department of accounting and management control [source: OpenAlex]	ROR https://ror.org/02dga6j42 (École Supérieure des Sciences Économiques et Commerciales - FR)	• 02dga6j42 		<ul style="list-style-type: none"> • W4386014955  • 10.2139/ssrn.45460..  • W4366396136  • 10.2139/ssrn.44116.. 
<input checked="" type="checkbox"/> essec higher school of economic and commercial sciences university of douala [source: OpenAlex]	<div style="border: 2px solid red; padding: 2px;"> ROR https://ror.org/02dga6j42 (École Supérieure des Sciences Économiques et Commerciales - FR) </div> ROR https://ror.org/02zr5jr81 (University of Douala - CM) 	• 02dga6j42 • 02zr5jr81 		<ul style="list-style-type: none"> • W4385801793  • 10.47191/ijmra/v6-...  • W438577206  • 10.5281/zenodo.823.. 
<input checked="" type="checkbox"/> essec higher school of economic and commercial sciences university of douala [source: OpenAlex]	<div style="border: 2px solid red; padding: 2px;"> ROR https://ror.org/02dga6j42 (École Supérieure des Sciences Économiques et Commerciales - FR) </div> ROR https://ror.org/02zr5jr81 (University of Douala - CM) 	• 02dga6j42 • 02zr5jr81 		<ul style="list-style-type: none"> • W4385801793  • 10.47191/ijmra/v6-...  • W438577206  • 10.5281/zenodo.823.. 

7000+ feedback for now

Monthly ingestion by OpenAlex



dataesri commented on Oct 16

Member ...

Correction needed for raw affiliation Audencia Business School (Nantes), département Communication et culture, Rn'B Lab, GRIPIC-CELSA (EA 1498)
raw_affiliation_name: Audencia Business School (Nantes), département Communication et culture, Rn'B Lab, GRIPIC-CELSA (EA 1498)
new_rors: 000axn811;04p8das24
previous_rors: 000axn811
works_examples: W2738809070
contact: 39367f06034f5b6bb78787d12a24da65:da3c62f43fe20098eb0a11527b @ sorbonne-universite.fr



dataesri commented 2 weeks ago

Member Author ...

This issue was accepted and ingested by the OpenAlex team on 2024-11-01. The new affiliations should be visible within the next 7 days.



scanR usage

- around 50,000 monthly visits
- API used by research institutions and companies
 - finding partners
 - expert finding
 - identifying risk of conflict of interests
 - marketing targeting
 - ...

Key take aways (1 / 3)

a) No use of proprietary sources brings benefits!

- building a national research portal with a top down approach (no CRIS systems) without any proprietary data is possible (IA, cloud computing ...)
- work on the data quality
- that allows to openly share the results (indicators, code, data) but also services can be built on top of the resulting data and API

Key take aways (2 / 3)

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 ML models constructions are themselves iteratives processes
- Exploring hidden things is complicated, and the development of PID policy or implementation enables a more accurate description

Key take aways (3 / 3)

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, ORCID and OpenAlex 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.

Contact & questions

eric.jeangirard [at] recherche.gouv.fr