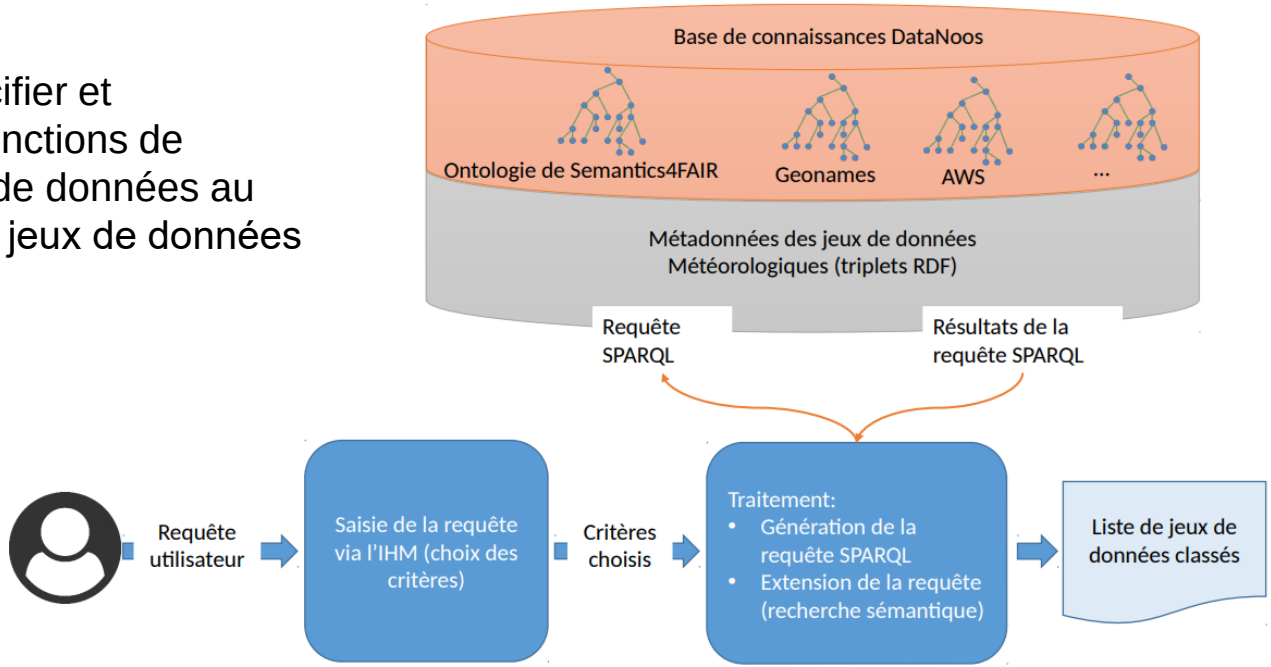
Plan

- Présentation du sujet
- Méthodologie
- Etat de l’art
- Spécifications
- Implémentation
- Conclusion

Présentation sujet

- But du stage : spécifier et implémenter des fonctions de recherche de jeux de données au sein d'un portail de jeux de données



Présentation sujet

- Quelles sont les différentes manière de rechercher des jeux de donnée?

- Comment améliorer la recherche de jeux de données à l’aide d’ontologies ?

Présentation sujet

- Quelles sont les différentes manière de rechercher des jeux de donnée?

Étudier différents portails de données et articles sur le sujet

- Comment améliorer la recherche de jeux de données à l’aide d’ontologies ?

Présentation sujet

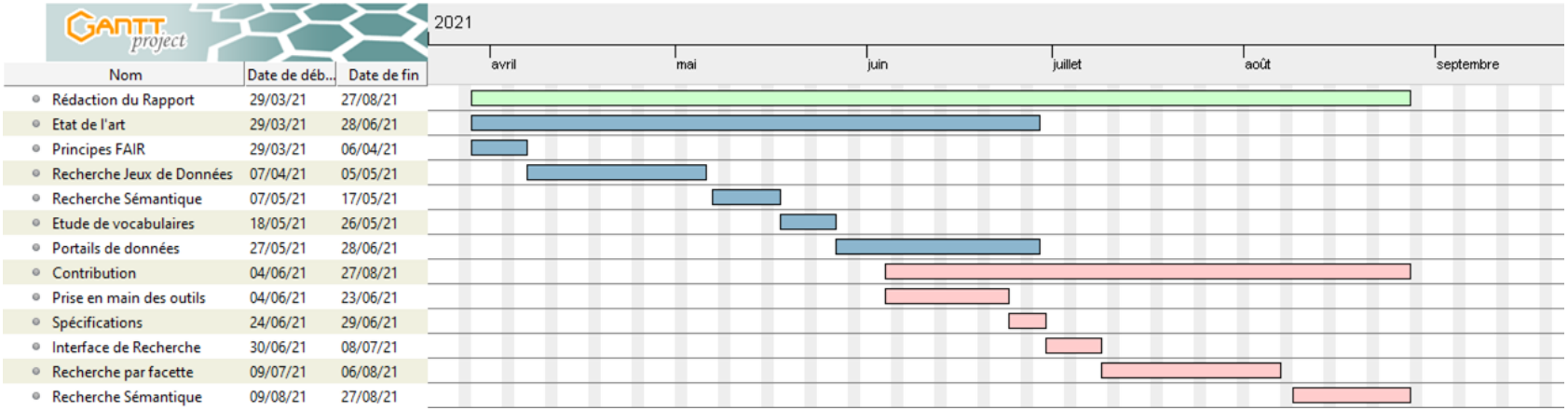
- Quelles sont les différentes manière de rechercher des jeux de donnée?

Étudier différents portails de données et articles sur le sujet

- Comment améliorer la recherche de jeux de données à l'aide d'ontologies ?

Implémenter des fonctions, interfaces et patrons de requêtes SPARQL

Méthodologie



Nom	Date de déb...	Date de fin
Rédaction du Rapport	29/03/21	27/08/21
Etat de l'art	29/03/21	28/06/21
Principes FAIR	29/03/21	06/04/21
Recherche Jeux de Données	07/04/21	05/05/21
Recherche Sémantique	07/05/21	17/05/21
Etude de vocabulaires	18/05/21	26/05/21
Portails de données	27/05/21	28/06/21
Contribution	04/06/21	27/08/21
Prise en main des outils	04/06/21	23/06/21
Spécifications	24/06/21	29/06/21
Interface de Recherche	30/06/21	08/07/21
Recherche par facette	09/07/21	06/08/21
Recherche Sémantique	09/08/21	27/08/21

Diagramme de Gantt du stage

Etat de l’art
Etude de vocabulaires / bases de connaissance

- Le modèle Semantics4FAIR comprend des vocabulaires ontologiques spécifiques:
  - GeoDCAT-AP version 2.0: basé sur DCAT, qui gère les métadonnées de datasets
  - CSVW NameSpace: décrit des métadonnées pour données tabulaires
  - RDF Data Cube: permet de représenter les jeux de données multidimensionnels
- Il est aussi lié à des bases de connaissances:
  - AWS: représente des connaissances lié au domaine de la météorologie (capteurs météorologiques, paramètres atmosphériques, etc…)
  - GeoNames: contient plus de 25M de noms géographiques. On peut accéder à des données telles que la latitude, la longitude, la subdivision administrative, le code postal, etc...

Etat de l’art
Portails de données

- Etude sur différents portails de données existants dans le but de s'informer sur les critères/métadonnées, les méthodes de recherche utilisées et leur interfaces:
    - European Data Portal
    - DataOsu
    - Google Dataset Search
    - Harvard Dataverse
    - DRIIHM
    - AIR BREIZH

Etat de l'art
Portails de données

- Utilisation de DCAT dans le modèle Semantics4FAIR donc on associe les critères de recherche aux aux entités DCAT qui seront nos métadonnées

- Voir les critères importants que se démarques des autres

Critère de recherche	**Portail(s) implémentant ce critère**	**DCAT entity**
Dataset	Tous les portails étudiés	**dcat:Dataset**
Catalogue	Tous sauf AIR BREIZH	**dcat:Catalog**
Catégories/Thèmes	Tous les portails étudiés	**dcat:theme**
Localisation	Tous sauf Harvard Dataverse	**dct:spatial**
Mots-clés	Tous les portails étudiés	**dcat:keyword**
Format	Tous les portails étudiés	dct:format
Licence utilisée	Tous les portails étudiés	**dct:licence**
Éditeur	Tous les portails étudiés	**dct:publisher**
Langue	Tous sauf European Data Portal et Google Data Search	**dct:language**
Méthode d'obtention des données	DataOsu et Google Data Search	**prov:wasGeneratedBy**
DataDownload	Google Data Search	**dcat:Distribution**
Author name	dataOsu, Google Data Search et Harvard Dataverse	**dct:creator**
Contact	Tous sauf European Data Portal et Google Data Search si on considère que juste un lien n'est pas bon	**dcat:contactPoint**
Ressource créer	Tous sauf Google Data Search	**dct:issued**
Mise à jour	Tous les portails étudiés	**dct:mofified**
Détails sur les accèes	European Data Portal et dataOsu	**dct:rights**
Fréquence mise à jour	Tous sauf Google Data Search	**dct:accrualPeriodicity**
title	Tous les portails étudiés	**dct:title**
description	Tous les portails étudiés	**dct:description**
iddentifiant	Tous les portails étudiés	dct:identifier
temporal coverage	Harvard Dataverse, dataOsu et Google Data Search	dct:temporal
variableMeasured	Google Data	dvq:QualityMeasurement

Etat de l’art
Recherche de jeux de données

- La recherche de jeux de données commence par une requête sur la barre de recherche ou bien un filtrage sur des catégories pré-identifiées [1]. On va par la suite associer ces mots-clés et catégories aux métadonnées des jeux.

- La recherche web classique n’est pas très efficace pour la recherche de jeux de données car comme les jeux de données sont structurés, ils ne peuvent pas être facilement mis en correspondance avec approches de texte non structurées [2].

- C’est une des raisons pour laquelle les portails de données utilisent plutôt la recherche par facette (l’utilisateur peut filtrer les jeux de données en fonction d’un ou plusieurs critères).

Etat de l’art
Recherche sémantique

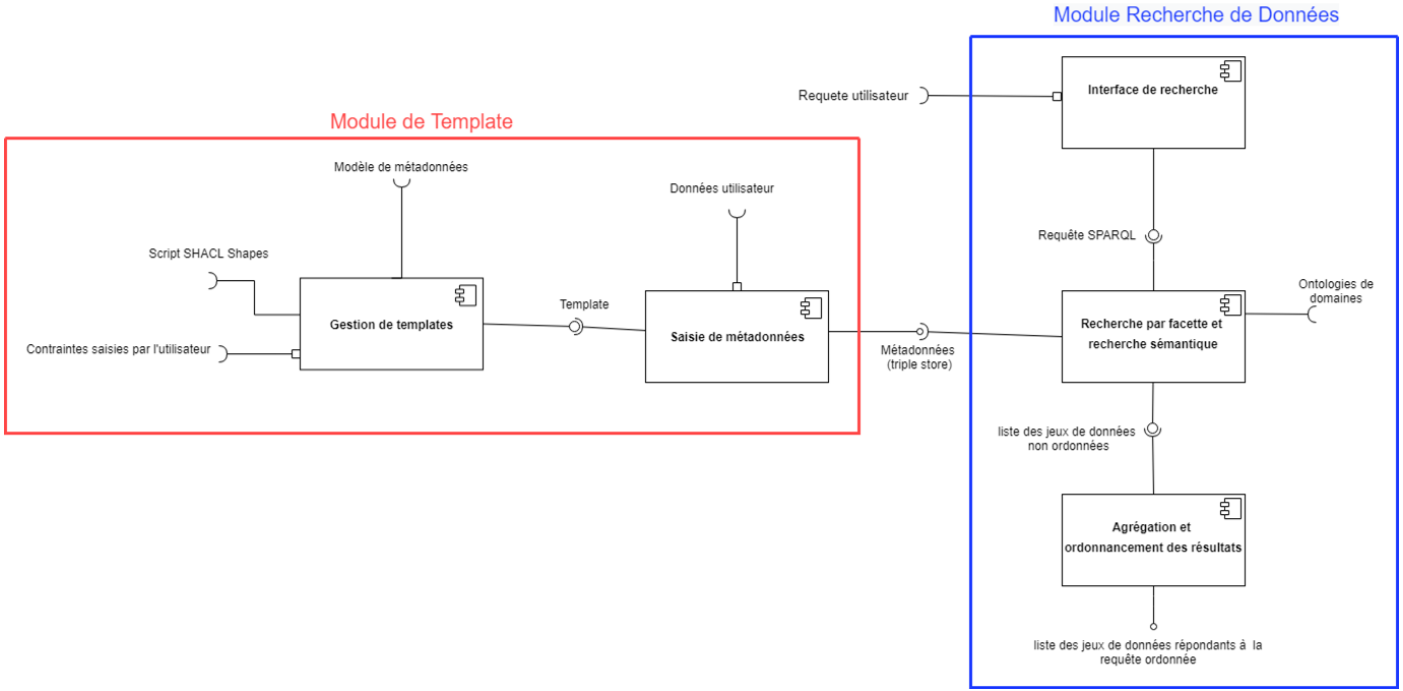
- La recherche sémantique est la plus-value qu’offre le projet Semantics4Fair par rapport aux autres portails de données. Les portails de données ne l’utilise pas pour le moment.

- Elle permet d’étendre la recherche originale lancée par un utilisateur grâce à des structures de données comme des ontologies (elle exploite les relations hiérarchiques entre les concept) , si il n’y a pas assez de résultats ou que l’utilisateur n’est pas satisfait, on lui offre des suggestions [3].

Contribution

- Spécifications
  - Diagrammes
  - Etat de l’art / Choix d’implémentation
  - Maquette
- Implémentation
  - Interface de recherche
  - Recherche par facette
  - Recherche sémantique

Spécifications
Diagramme des composants

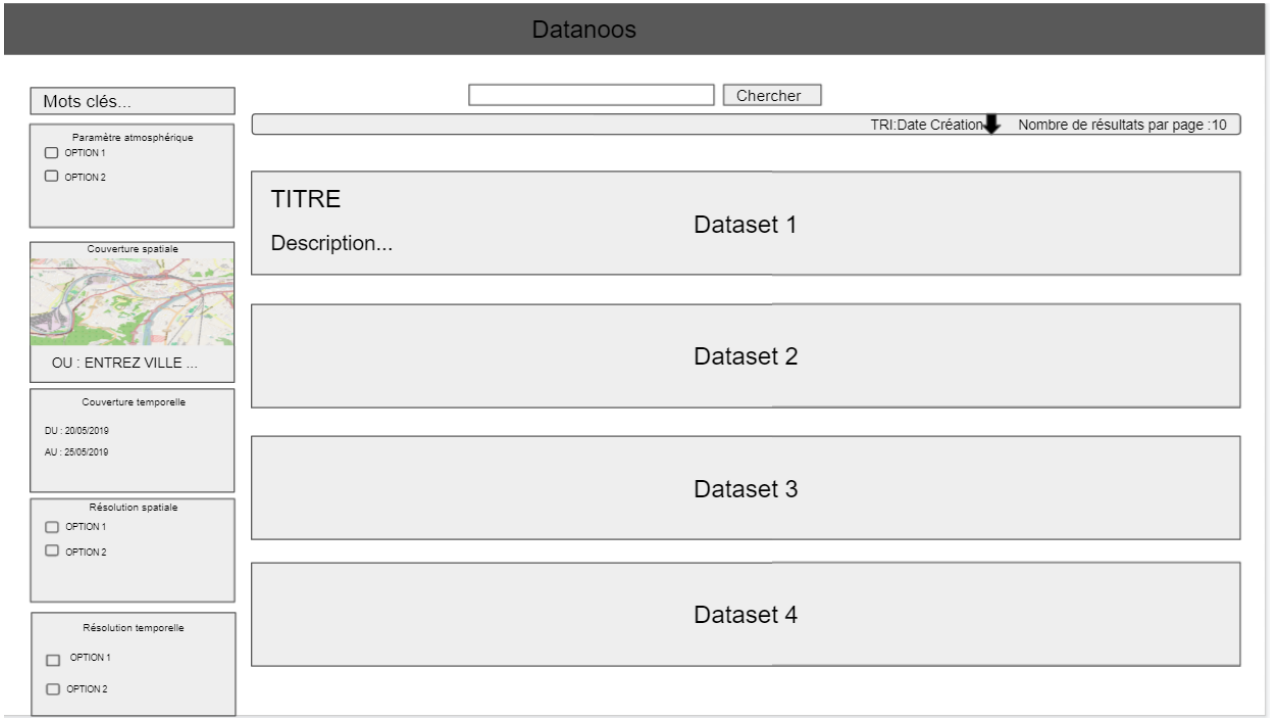


Spécifications
Choix des critères

- Couverture spatiale : la région géographique du jeu de données
- Couverture temporelle : intervalle de temps que couvre le jeu de données
- Résolution spatiale : distance en mètre entre les items du jeu de données
- Résolution temporelle : intervalle de temps entre les enregistrement d’un jeu de données
- Mots-clés : décrit le jeu de données
- Paramètre atmosphérique : critère spécifique au domaine car on traite des jeux de données météorologiques et il est utile pour la recherche de ce type de données

Spécifications

Maquette Interface de Recherche



Spécifications
Cas d'utilisations

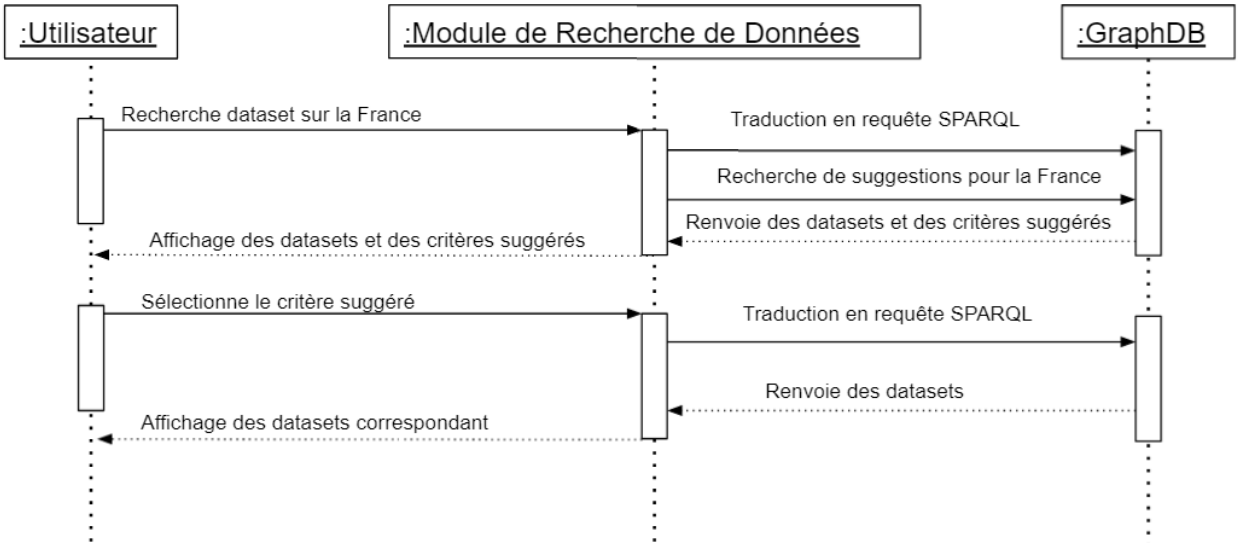
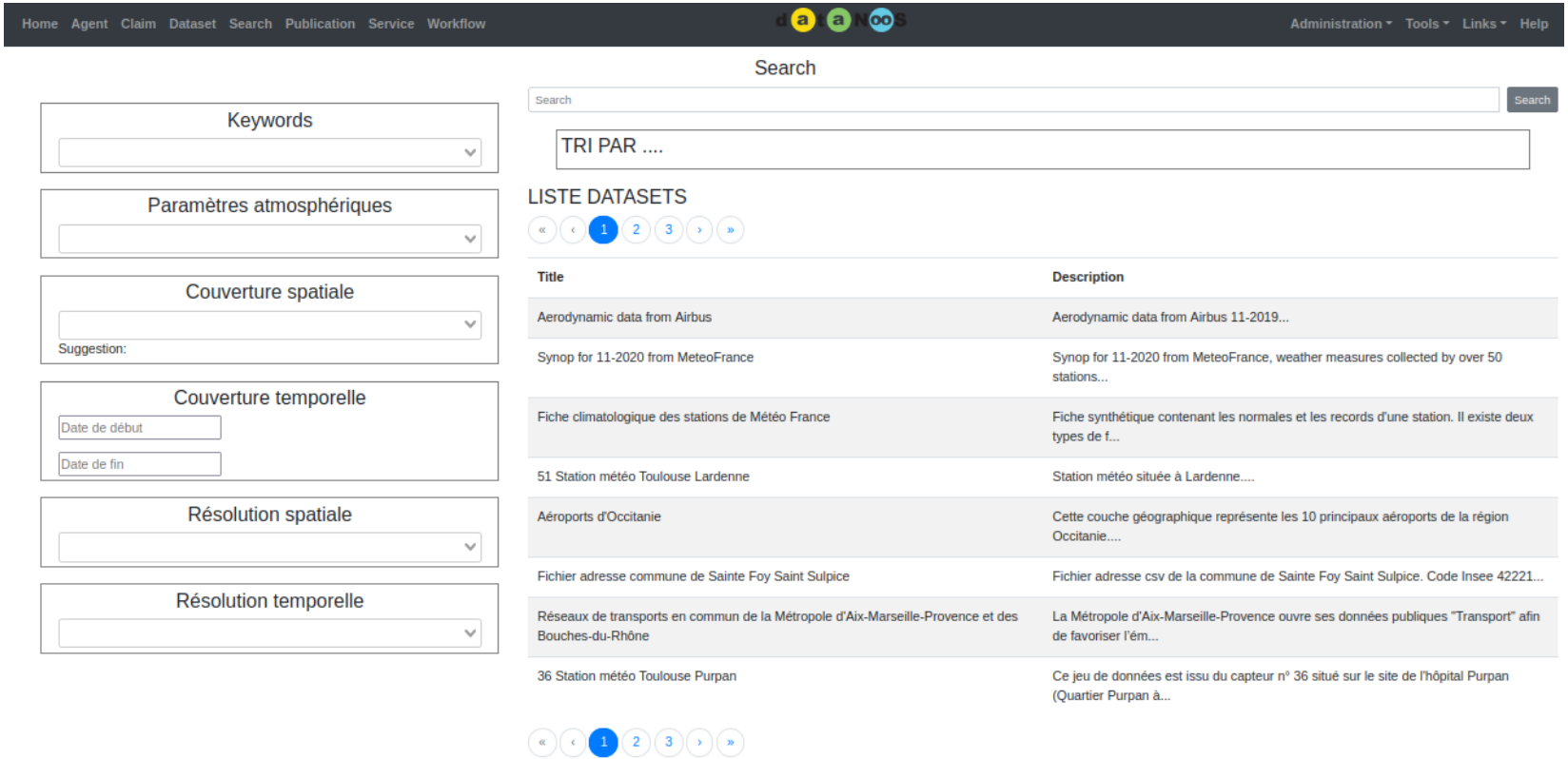


Diagramme de Séquence du cas d'utilisation : Recherche sémantique

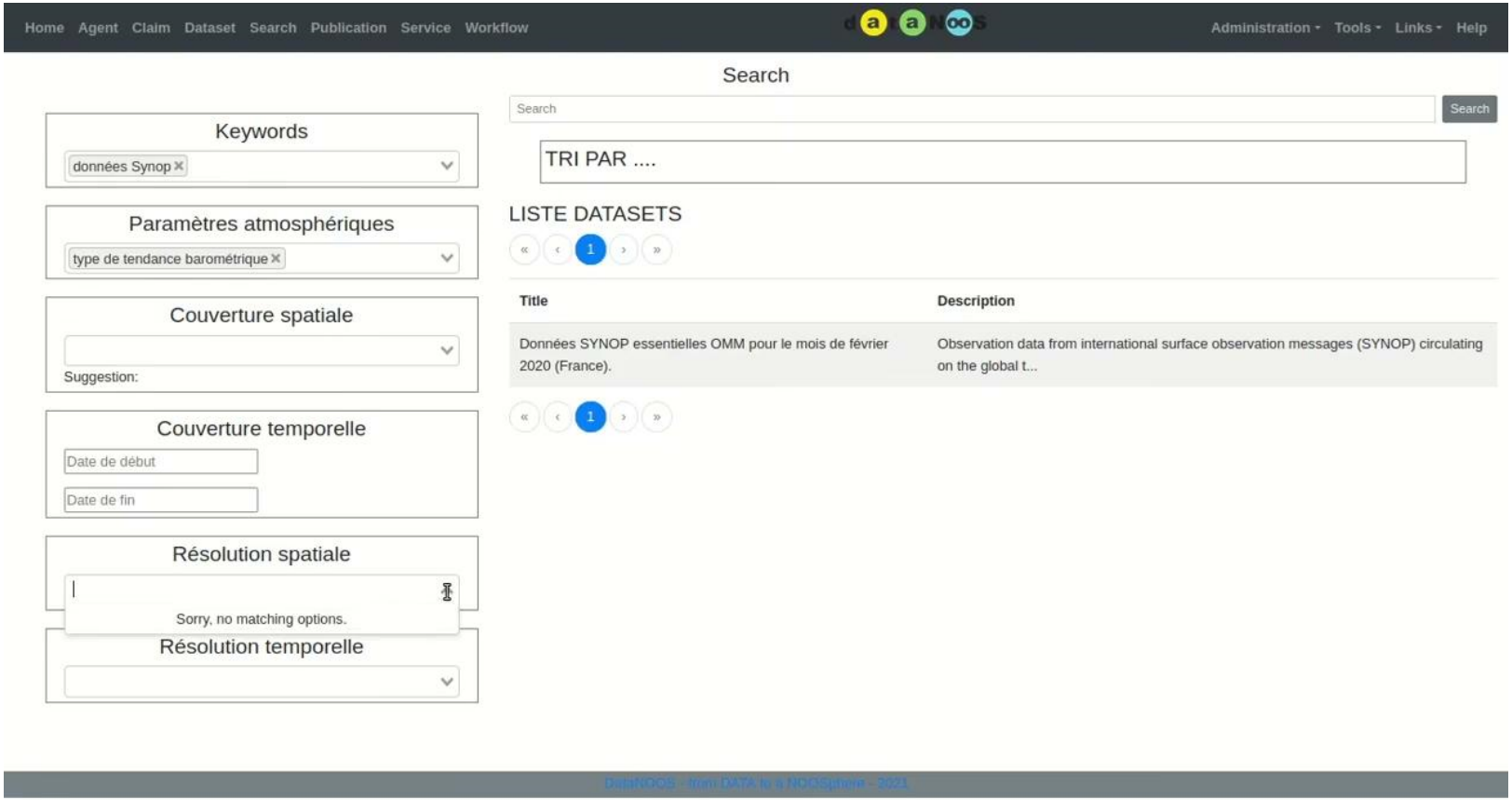
Implémentation

Interface de recherche



Implémentation
Recherche par facette

- Recherche multi critères avec utilisation de l’opérateur logique “ou” pour les éléments d’un même critère
- Opérateur logique “et” entre les différents critères
- Mise à jour des autres listes de critères lorsque l’on sélectionne un critère

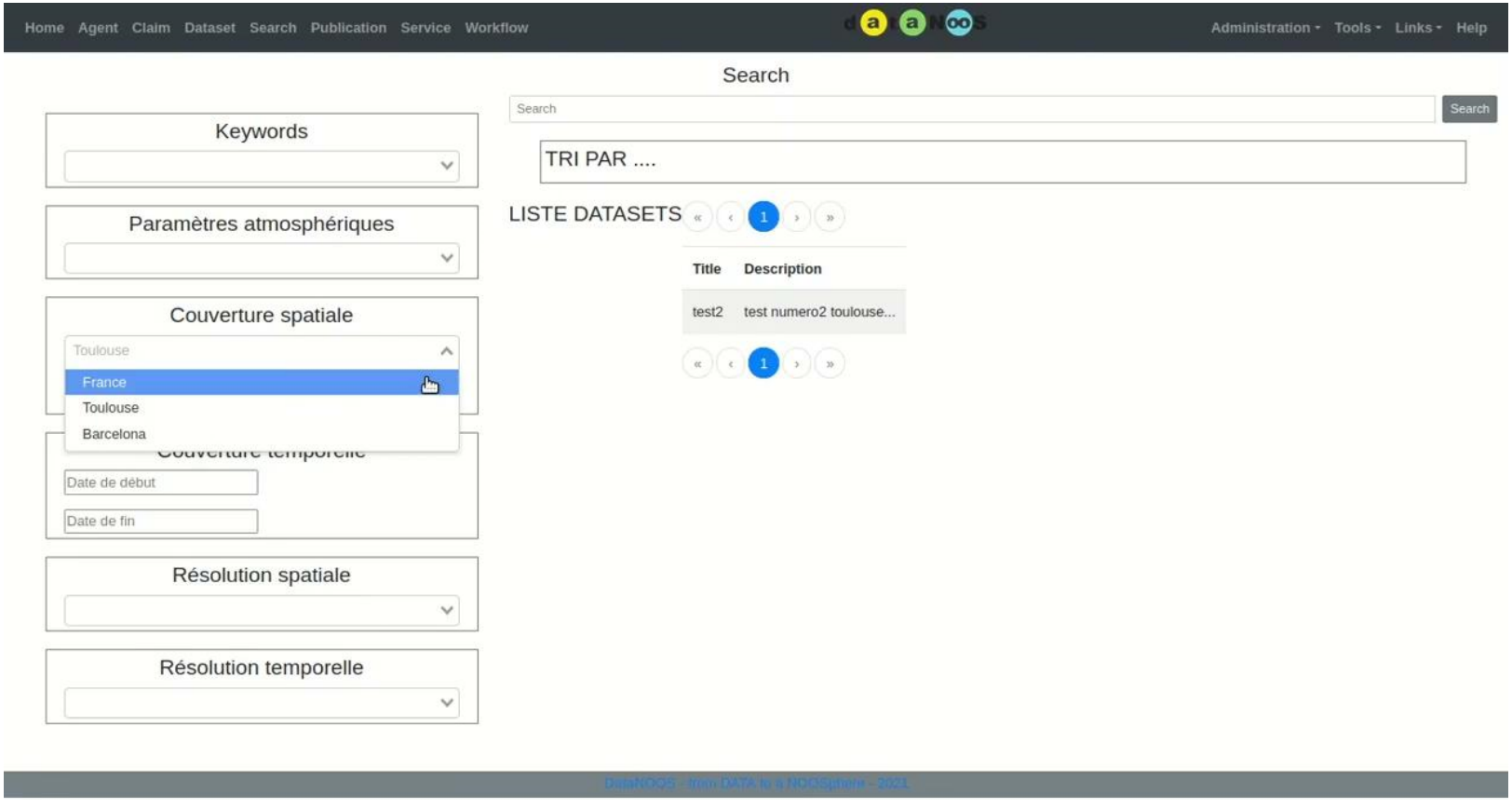

Home Agent Claim Dataset Search Publication Service Workflow
Administration Tools Links Help
Search
Keywords
données Synop
Paramètres atmosphériques
type de tendance barométrique
Couverture spatiale
Suggestion:
Couverture temporelle
Date de début
Date de fin
Résolution spatiale
Sorry, no matching options.
Résolution temporelle
Search
Search
TRI PAR ....
LISTE DATASETS
Title
Description
Données SYNOP essentielles OMM pour le mois de février 2020 (France).
Observation data from international surface observation messages (SYNOP) circulating on the global t...

Implémentation
Recherche sémantique

- Critère couverture spatiale
    - Utiliser les relations hiérarchiques entre les zones
    - Vocabulaire contrôlé Geonames donc utilisation d’une API pour récupérer la hiérarchie

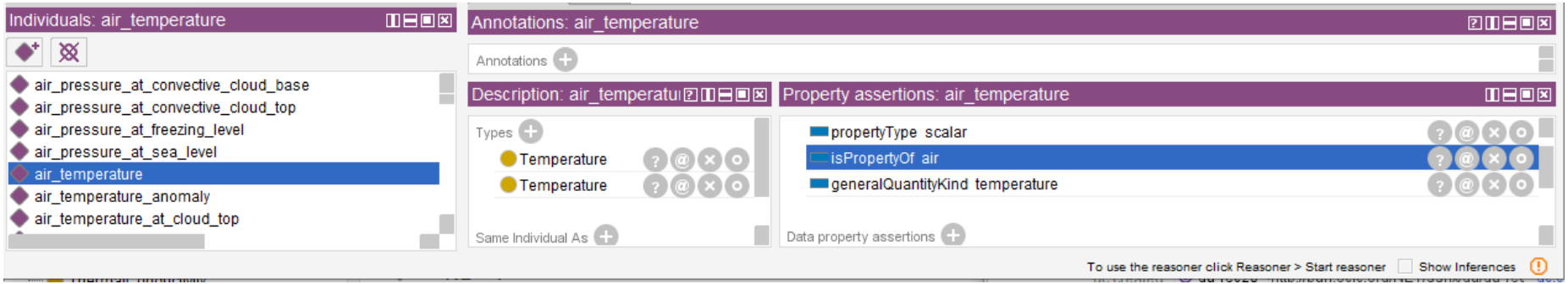
Hiérarchie de zone pour la ville de Toulouse

```
Earth/Europe/France/Occitanie/Upper Garonne/Arrondissement de Toulouse/Toulouse/Toulouse/
```


Home Agent Claim Dataset Search Publication Service Workflow
Administration Tools Links Help
Search
Search
Search
Keywords
TRI PAR ....
Paramètres atmosphériques
LISTE DATASETS
Couverture spatiale
Title
Description
test2
test numero2 toulouse...
Toulouse
France
Toulouse
Barcelona
Date de début
Date de fin
Résolution spatiale
Résolution temporelle

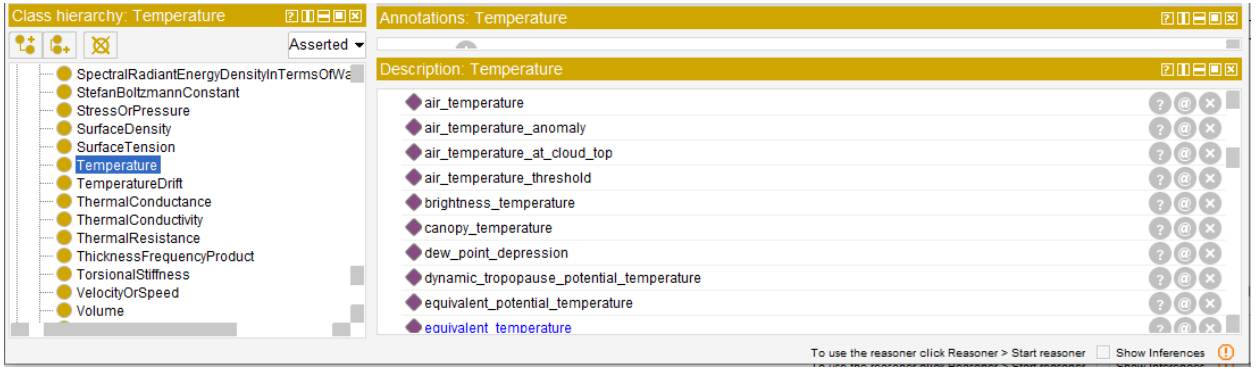
Implémentation
Recherche sémantique

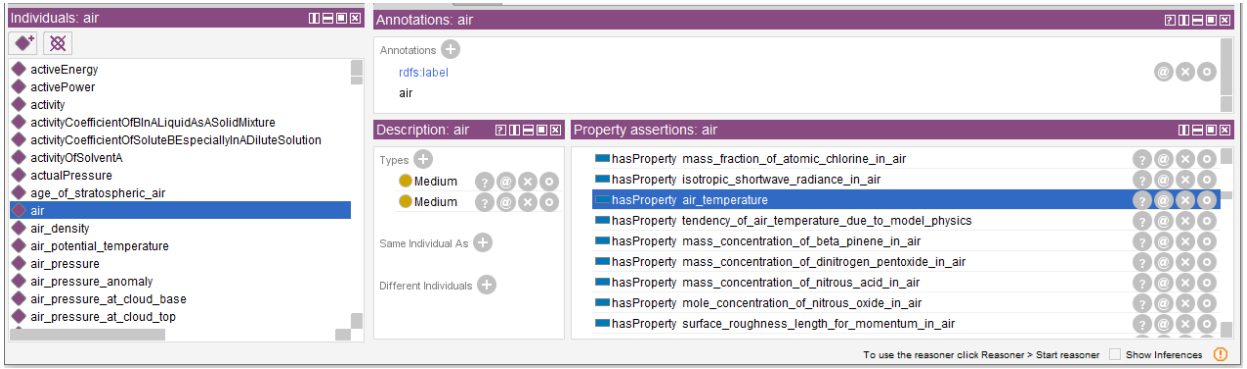
- Critère paramètre atmosphérique
    - Alignements des paramètres atmosphériques entre nos jeux de données et la base de connaissance AWS
    - On peut donc utiliser les relations hiérarchiques d'AWS pour suggérer d'autres paramètres atmosphérique



Implémentation
Recherche sémantique

On peut donc suggérer des instances de type “Temperature”





Ou bien suggérer des paramètres atmosphériques qui possèdent la relation “isPropertyOf” de l’entité “air”

Conclusion

- Bilan des résultats
  - Interface de recherche implémentée
  - Le prototype de recherche sémantique de jeux de données est opérationnel
  - Difficultés pour évaluer la recherche car on ne dispose que d'une vingtaine de jeux de données

- Perspectives d'évolutions
  - Implémenter la recherche sémantique sur le critère paramètre atmosphérique sur le prototype
  - Implémenter le trie des jeux de données en fonction de leur qualité de métadonnées
  - Implémenter la recherche de la barre de recherche

- Bilan personnel
  - Enrichissant car j'ai pu approfondir le domaine du Web sémantique que j'ai découvert cette année
  - Tiré à parti des enseignements du Master DC (TIR, Ontologies, RI, MCO)
  - Expérience acquise sur ces 5 mois de stage
  - Conforté dans mon orientation professionnelle

Références

[1] Adriane Chapman et al. “Dataset search: a survey”. In: arXiv:1901.00735 [cs] (Jan. 3, 2019). arXiv: 1901.00735. url: http://arxiv.org/abs/1901.00735 (visited on 08/23/2021).

[2] Emilia Kacprzak et al. “Characterising dataset search—An analysis of search logs and data requests”. In: Journal of Web Semantics 55 (Mar. 1, 2019), pp. 37–55. issn: 1570-8268. doi: 10.1016/j.websem.2018.11.003. url: https://www.sciencedirect.com/science/article/pii/S1570826818300556 (visited on 08/23/2021).

[3] Dario Bonino et al. “Ontology driven semantic search”. In: Sigir Forum 1 (June 1, 2004)

Merci de m’avoir écouté
Vous pouvez poser vos questions !!!