

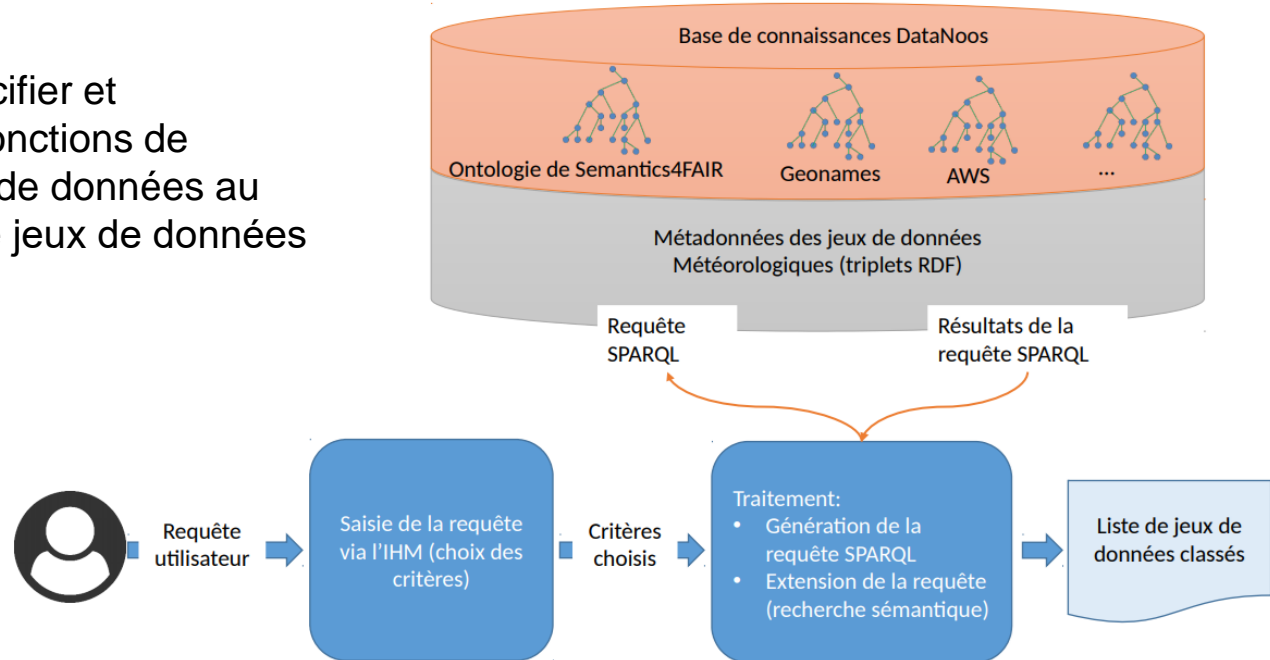
# Plan

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

# Présentation sujet



- Quelles sont les différentes manières de rechercher des jeux de données ?

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 manières de rechercher des jeux de données?

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

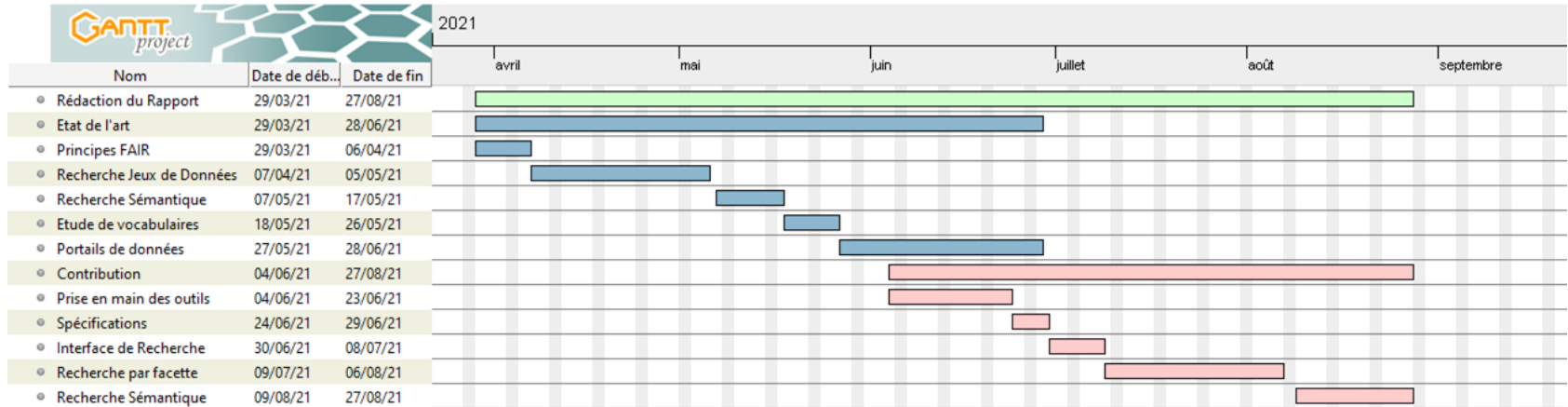


Diagramme de Gantt du stage

# 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 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	<b>dcat:Dataset</b>
Catalogue	Tous sauf AIR BREIZH	<b>dcat:Catalog</b>
Catégories/Thèmes	Tous les portails étudiés	<b>dcat:theme</b>
Localisation	Tous sauf Harvard Dataverse	<b>dct:spatial</b>
Mots-clés	Tous les portails étudiés	<b>dcat:keyword</b>
Format	Tous les portails étudiés	<b>dct:format</b>
Licence utilisée	Tous les portails étudiés	<b>dct:licence</b>
Éditeur	Tous les portails étudiés	<b>dct:publisher</b>
Langue	Tous sauf European Data Portal et Google Data Search	<b>dct:language</b>
Méthode d'obtention des données	DataOsu et Google Data Search	<b>prov:wasGeneratedBy</b>
DataDownload	Google Data Search	<b>dcat:Distribution</b>
Author name	dataOsu, Google Data Search et Harvard Dataverse	<b>dct:creator</b>
Contact	Tous sauf European Data Portal et Google Data Search si on considère que juste un lien n'est pas bon	<b>dcat:contactPoint</b>
Ressource créer	Tous sauf Google Data Search	<b>dct:issued</b>
Mise à jour	Tous les portails étudiés	<b>dct:modified</b>
Détails sur les accès	European Data Portal et dataOsu	<b>dct:rights</b>
Fréquence mise à jour	Tous sauf Google Data Search	<b>dct:accrualPeriodicity</b>
title	Tous les portails étudiés	<b>dct:title</b>
description	Tous les portails étudiés	<b>dct:description</b>
identifiant	Tous les portails étudiés	<b>dct:identifier</b>
temporal coverage	Harvard Dataverse, dataOsu et Google Data Search	<b>dct:temporal</b>
variableMeasured	Google Data	<b>dvq:QualityMeasurement</b>

# 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

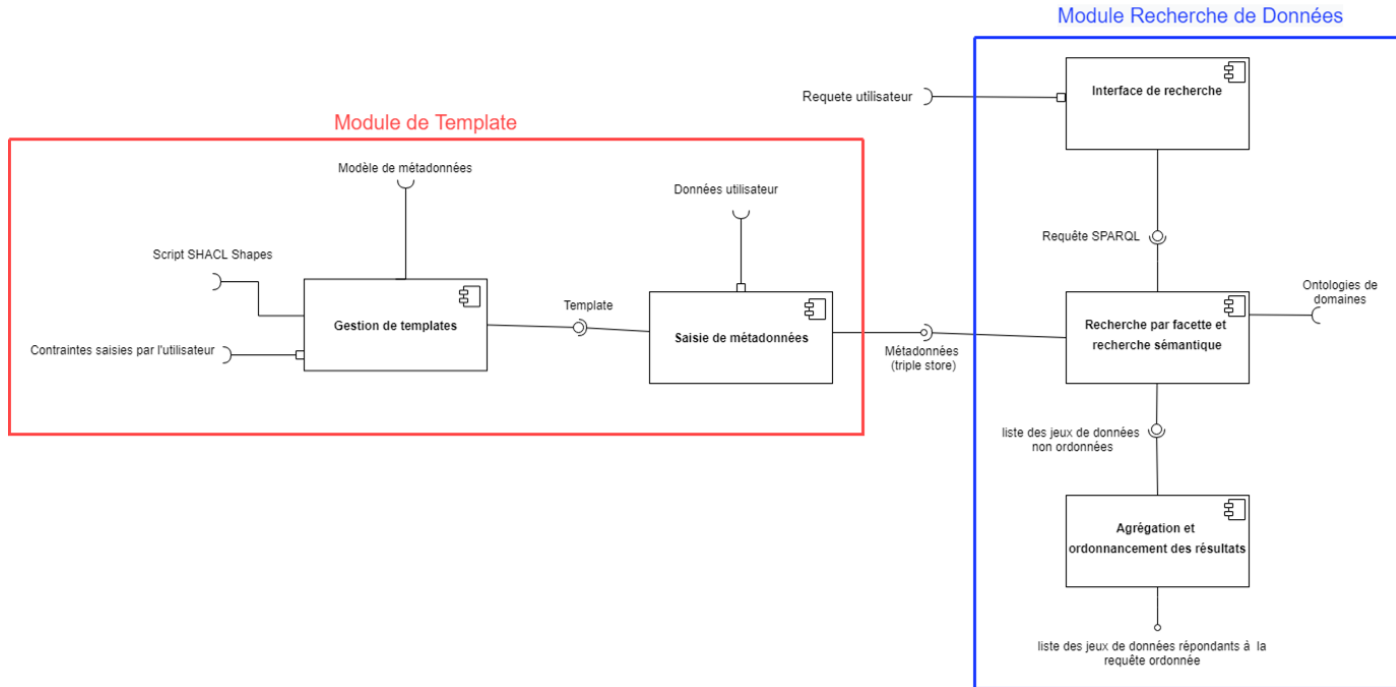
- La recherche sémantique est la plus-value qu'offre le projet Semantics4Fair par rapport aux autres portails de données. Les portails de données ne l'utilisent pas pour le moment.
- Elle permet d'étendre la recherche originale lancée par un utilisateur grâce à des structures de données comme des ontologies (elle exploite les relations hiérarchiques entre les concepts), 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 enregistrements 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


Datanoos

Mots clés...  Chercher

TRI: Date Création  Nombre de résultats par page : 10

TITRE
Description... Dataset 1
Dataset 2
Dataset 3
Dataset 4

**Paramètre atmosphérique**  
 OPTION 1  
 OPTION 2

**Couverture spatiale**  
  
OU : ENTREZ VILLE ...

**Couverture temporelle**  
DU : 20/05/2019  
AU : 25/05/2019

**Résolution spatiale**  
 OPTION 1  
 OPTION 2

**Résolution temporelle**  
 OPTION 1  
 OPTION 2

# Spécifications

## Cas d'utilisations

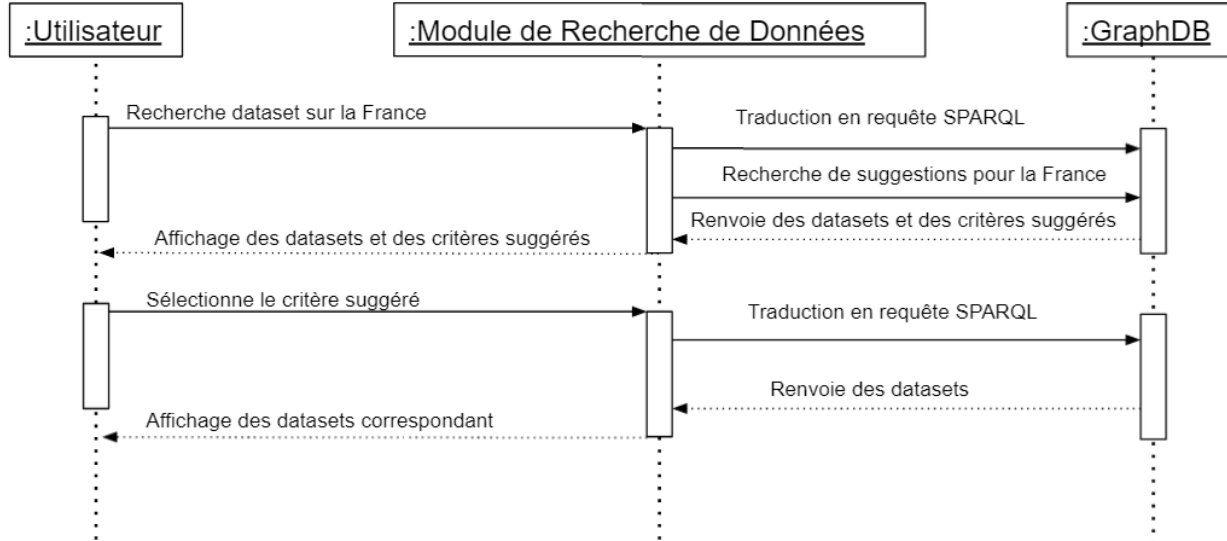


Diagramme de Séquence du cas d'utilisation : Recherche sémantique



# Implémentation

## Interface de recherche

### Search

Search

Keywords

Paramètres atmosphériques

Couverture spatiale   
Suggestion:

Couverture temporelle  
Date de début   
Date de fin

Résolution spatiale

Résolution temporelle

### LISTE DATASETS

« ‹ 1 2 3 › »

Title	Description
Aerodynamic data from Airbus	Aerodynamic data from Airbus 11-2019...
Synop for 11-2020 from MeteoFrance	Synop for 11-2020 from MeteoFrance, weather measures collected by over 50 stations...
Fiche climatologique des stations de Météo France	Fiche synthétique contenant les normales et les records d'une station. Il existe deux types de f...
51 Station météo Toulouse Lardenne	Station météo située à Lardenne....
Aéroports d'Occitanie	Cette couche géographique représente les 10 principaux aéroports de la région Occitanie....
Fichier adresse commune de Sainte Foy Saint Sulpice	Fichier adresse csv de la commune de Sainte Foy Saint Sulpice. Code Insee 42221...
Réseaux de transports en commun de la Métropole d'Aix-Marseille-Provence et des Bouches-du-Rhône	La Métropole d'Aix-Marseille-Provence ouvre ses données publiques "Transport" afin de favoriser l'ém...
36 Station météo Toulouse Purpan	Ce jeu de données est issu du capteur n° 36 situé sur le site de l'hôpital Purpan (Quartier Purpan à...

« ‹ 1 2 3 › »

# 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

## 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

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/
```

### Search

Keywords

Paramètres atmosphériques

Couverture spatiale

Toulouse

- France
- Toulouse
- Barcelona

Couverture temporelle

Date de début

Date de fin

Résolution spatiale

Résolution temporelle

#### LISTE DATASETS

« < 1 > »

Title	Description
test2	test numero2 toulouse...

« < 1 > »

# 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

The screenshot displays a Semantic Web editor interface with several panels:

- Individuals: air\_temperature:** A list of individuals including `air_pressure_at_convective_cloud_base`, `air_pressure_at_convective_cloud_top`, `air_pressure_at_freezing_level`, `air_pressure_at_sea_level`, `air_temperature` (highlighted), `air_temperature_anomaly`, and `air_temperature_at_cloud_top`.
- Annotations: air\_temperature:** A section for adding annotations to the selected individual.
- Description: air\_temperature:** A section for defining the property's types, currently showing two instances of `Temperature`.
- Property assertions: air\_temperature:** A section for defining property assertions, including `propertyType scalar`, `isPropertyOf air` (highlighted), and `generalQuantityKind temperature`.

At the bottom right, there is a footer with the text: "To use the reasoner click Reasoner > Start reasoner" and a checkbox for "Show Inferences".

# Implémentation

## Recherche sémantique

On peut donc suggérer des instances de type “Temperature”

The screenshot displays two panels from a software application. The left panel, titled "Class hierarchy: Temperature", shows a tree view of classes. The "Temperature" class is highlighted in blue. The right panel, titled "Annotations: Temperature", shows a list of properties associated with the "Temperature" class. The properties listed are: air\_temperature, air\_temperature\_anomaly, air\_temperature\_at\_cloud\_top, air\_temperature\_threshold, brightness\_temperature, canopy\_temperature, dew\_point\_depression, dynamic\_tropopause\_potential\_temperature, equivalent\_potential\_temperature, and equivalent temperature. Each property has a diamond icon and a set of control icons (question mark, at-sign, X, and a circle with a dot).

The screenshot displays two panels from a software application. The left panel, titled "Individuals: air", shows a list of individuals. The "air" individual is highlighted in blue. The right panel, titled "Annotations: air", shows a list of annotations for the "air" individual. The annotations listed are: rdfs:label and air. Below the annotations, there are sections for "Description: air" and "Property assertions: air". The "Description: air" section shows two instances of the "Medium" type. The "Property assertions: air" section shows a list of properties associated with the "air" individual. The properties listed are: hasProperty mass\_fraction\_of\_atomic\_chlorine\_in\_air, hasProperty isotropic\_shortwave\_radiance\_in\_air, hasProperty air\_temperature, hasProperty tendency\_of\_air\_temperature\_due\_to\_model\_physics, hasProperty mass\_concentration\_of\_beta\_pinene\_in\_air, hasProperty mass\_concentration\_of\_dinitrogen\_pentoxide\_in\_air, hasProperty mass\_concentration\_of\_nitrous\_acid\_in\_air, hasProperty mole\_concentration\_of\_nitrous\_oxide\_in\_air, and hasProperty surface\_roughness\_length\_for\_momentum\_in\_air. Each property has a blue bar and a set of control icons (question mark, at-sign, X, and a circle with a dot).

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 !!!