REX Metabolomics Semantic Datalake

*D. Benaben, M. Boudet, C.Dupérier, **O. Filangi**, F. Giacomoni*

Objectifs du Data lake

- **Plateforme de gestion et d'intégration de données pour la métabolomique**
  - **Fédération de données à l'échelle du consortium et référentiel central unifié MetaboHUB**
  - Renforcer l'interopérabilité entre les données métabolomiques et les autres silots/ressources omiques
  - Garantir la traçabilité des données et des traitements
  - Automatiser l'enrichissement en métadonnées, des données inférées et la production de bases de connaissances MetaboHUB

Historique

- Avril 2021 - AAP SAPI DIPSO
  - CATI BARIC/EMPREINTE/PROSODIE
- Juin-Septembre 2021
  - réception machines (4 noeuds)
- Septembre 2021 – Administration d’un cluster Spark
  - 5 demi-journées
- Janvier/Fevrier 2022 – Programmation Fonctionnelle / Spark
  - 2 x 3 demi-journées
- Juin-Septembre 2022
  - Reception 4 noeuds

REX - Formations

- Monter en compétences à la fois sur les technologies d'administration du cluster et de développement applicatif
  - => pas évident car cela demande de s'investir pleinement
  - => nécessaire car les performances et l'implémentation s'obtiennent avec une bonne connaissance des technos et de l'infra
- Formation virtuelle plutôt efficace
  - Construction en demi-journée
  - Très accessibles car tout les outils pour les TP sont en ligne
    - Databricks
      - Ouverture compte : https://www.databricks.com/try-databricks#account
      - Utilisation https://community.cloud.databricks.com/login.html
  - les sessions sont enregistrées (CF le PDF *Cours_WebSem_Sansa* – dernière slide)
- => Des journées de formations auraient pu être réservées après une année d'utilisation du cluster

Production / Réalisations sur 18 mois

- Configuration/Deploiement du cluster SPARK / HADOOP
  - Expertise métier : **David Benaben / Christophe Dupérier**

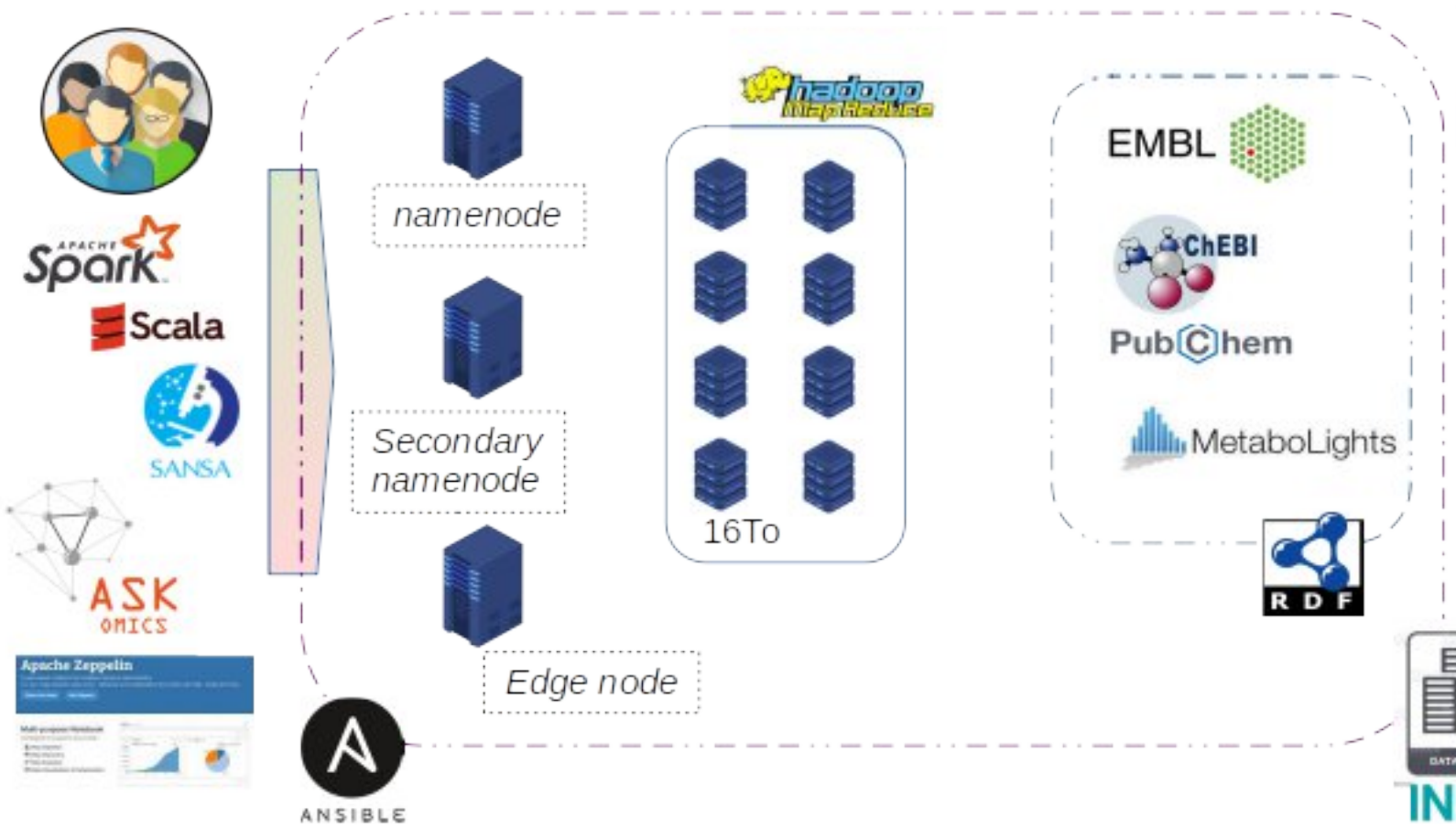
- Développement de template YAML pour GITLAB CI pour l'intégration des sources de donnée RDF

- Production/Analyse de données
  - Élévation sémantique de la base « Metabolights » (études métabolomique)
  - Portage de FORUM sur l'infrastructure Big Data
  - Étude de la cohérence des informations contenues dans plusieurs bases de métabolites
  - Expertise métier : **Olivier Filangi** / **Franck Giacomoni / Ghina Hajjar**

Metabolomic Semantic Datalake

Using RDF resources in a SPARK environment

Automation and refreshing of scientific data

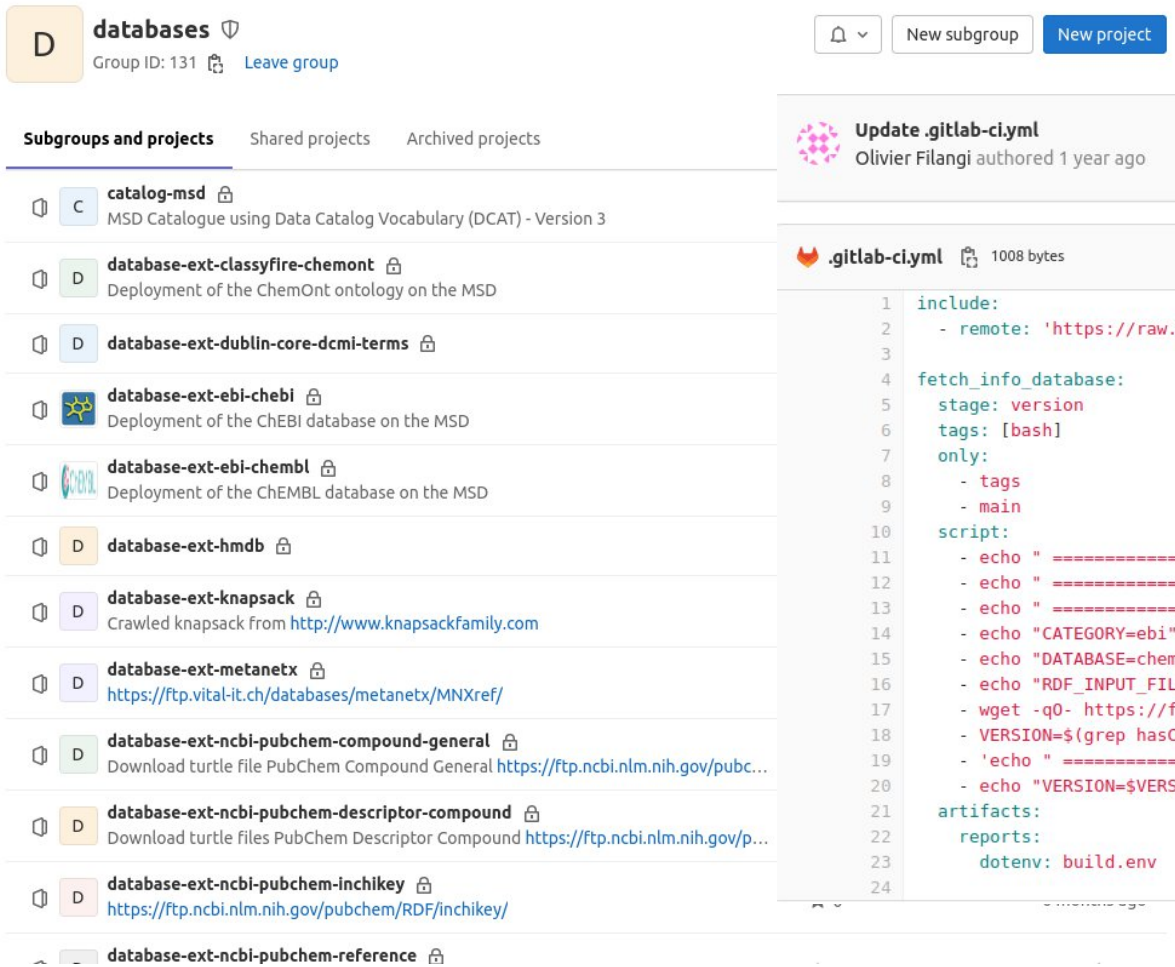


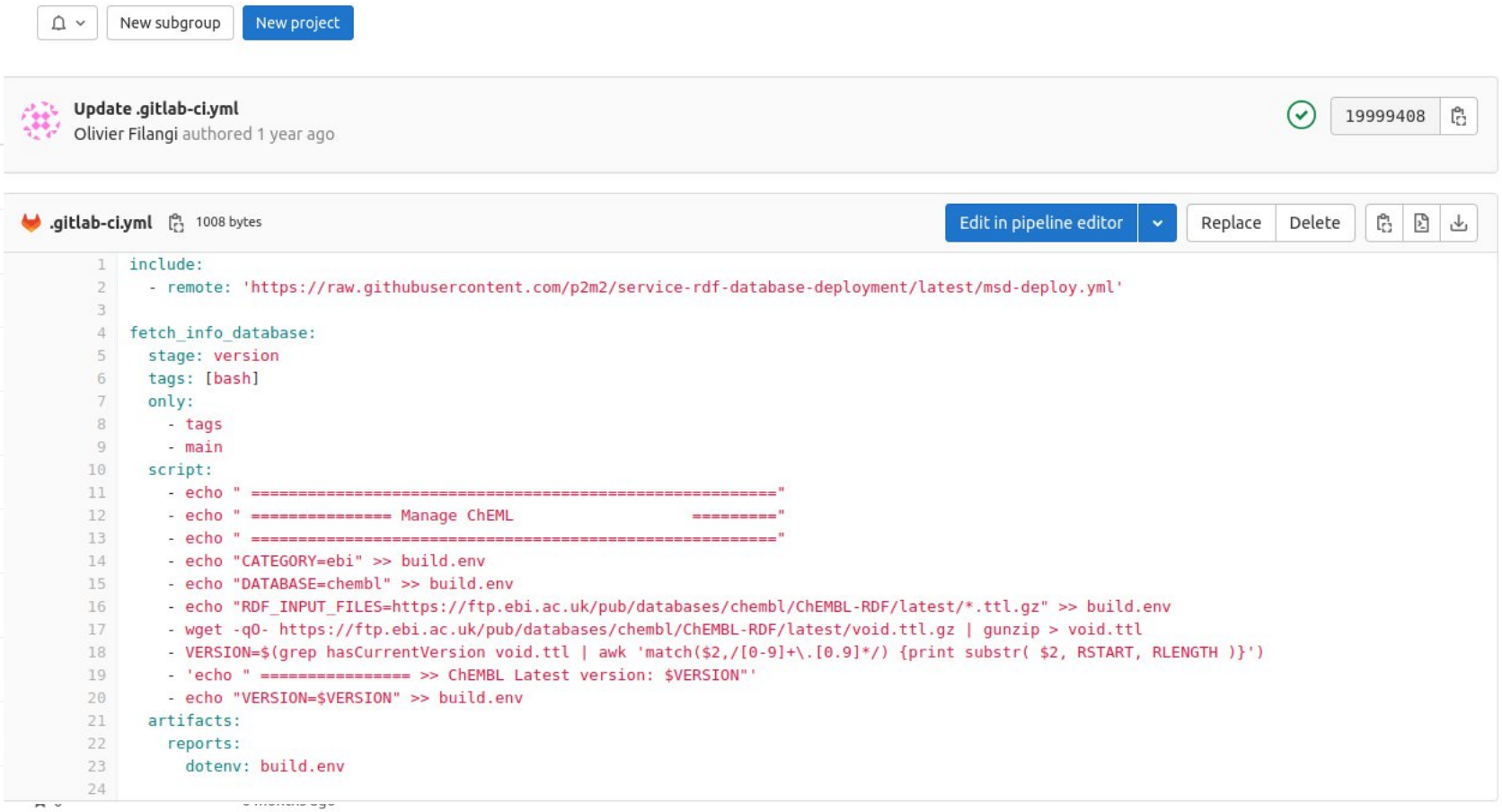
« Metabolomic Semantics DataLake » Project aims to build and deliver large-scale distributed Infrastructure for data processing and massive integration of semantic information on metabolomic data studies.
This system is strongly structured from semantic web technologies to maximize the reuse of existing ontologies and knowledge (NCBI, EBI) and finally manage the heterogeneous metabolomics and bioinformatics content of the MetaboHUB consortium.

Contact : empreinte-contact@groupes.renater.fr

Gestion des données scientifiques

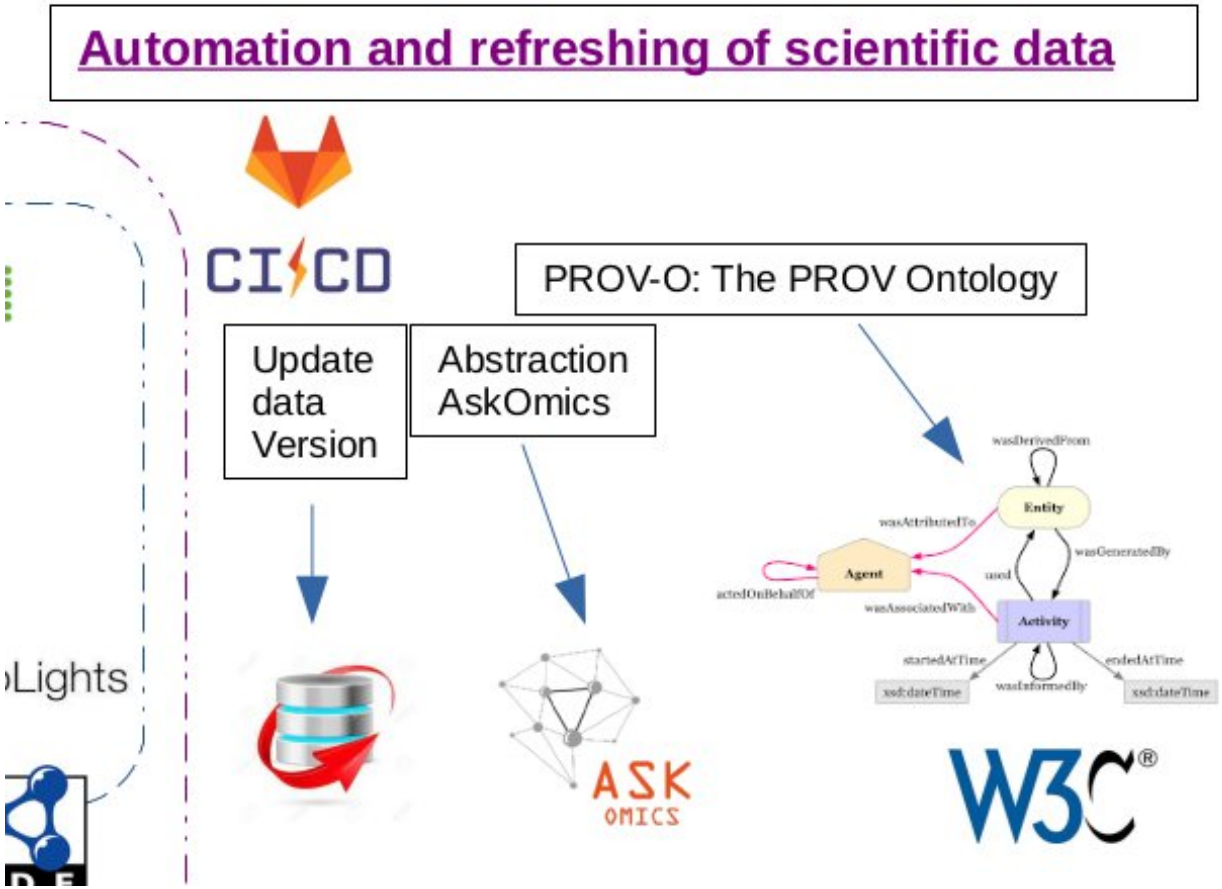
https://services.pfem.clermont.inrae.fr/gitlab/metabosemdatalake/databases





Projet : https://github.com/p2m2/service-rdf-database-deployment

Gestion des données scientifiques



Automatisation / versionning des bases via GITLAB CI avec un yaml de quelques lignes :
=> abstraction Askomics
=> information de provenance sur les données

TODO : ?passage à Airflow/NiFi ?

nifi

Apache Airflow

- Planifier et de surveiller des workflows (flux de travail)
- orchestration de pipelines de données complexes
- open source

Utilisation des ressources RDF

- SANSA
  - Difficile à utiliser. C'est un produit de recherche et la production des releases est dépend de la production des travaux de thèses en cours.
    - Documentation pas à jours
    - Pas de réponses aux messages envoyés sur la liste de diffusion...
  - Module R/W et Query . Fait vraiment l'affaire et distribue correctement les traitements (le temps de réponse est associé à la quantité de ressources réservées sur le cluster)
  - Module inférence, problématique sur l'inférence pour le graphe du projet FORUM. Ce graphe demande beaucoup de ressources qui n'est pas disponible sur le cluster.
  - TODO : Utilisation du ML
- Visualisation des ontologies / Concepts / Relations disponibles
  - TODO : AskOmics : https://askomics.org/
  - RH IE : Développement d'un catalogue des ressources

Utilisation/Gestion du cluster

- GITLAB / Ansible
  - Recettes spark/hadoop : https://github.com/andrewrothstein/ansible-spark
  - Nous n'avons pas eu encore le temps de tweaker le cluster en modifiant les paramètres spark/hadoop
  - Réplicat 4 => 2 dernièrement. Perte de performance 1/3
- Zeppelin (notebook)
  - installation/utilisation pas optimal
- Spark history server (http)
  - Bien pratique pour évaluer les problèmes de perf .
  - problèmes de latences entre l'arrêt d un job et l'affichage des informations d'exécution
- Yarn Hadoop UI (http)
  - Utiliser seulement pour la page principale (les logs des noeuds d'execution ne sont pas accessibles, préférence pour la commande yarn)

Utilisation/Gestion du cluster

- Spark-Shell / Scala / Sansa
  - Mise au point des méthodes
  - Permet de fouiller facilement les jeux de données RDF avec une requete SPARQL ou via DataFrame pour les autres de données
  - Evaluation de la quantité des ressources à attribuer à un job spark
- IntelliJ / Scala / SBT / Sansa / GITLAB CI/CD
  - Destiné à la production de contenu avec rafraichissement

TODO / Pistes d'amélioration

- **SAPI 2022 Metasaurus** : animation/conception d'une ontologie pour le consortium orientée métabolomique
- Finalisation de l'intégration **FORUM**
- **Catalogue** des ressources du datalake : IE 2023/2024
- Gestion / rafraichissement des ressources
- Visualisation des ressources dans **AskOmics**
- **Zeppelin** (notebook) opérationnel / promotion et utilisation du datalake par les chercheurs bioinformaticiens en ingénierie des connaissances

Questions ?

```
ofilangi@ara-unh-elrond:~$ hdfs dfs -ls /rdf/ebi/chembl
Found 3 items
drwxrwxr--   - ubuntu sparkuser          0 2021-12-02 16:23 /rdf/ebi/chembl/29.0
drwxrwxr--   - ubuntu sparkuser          0 2022-03-23 19:12 /rdf/ebi/chembl/30.0
drwxr-xr-x   - ubuntu sparkuser          0 2022-08-24 09:35 /rdf/ebi/chembl/31.0
```

Gestion des données scientifiques

résultats du yaml GITLAB CI
sur le cluster MSD

```
ofilangi@ara-unh-elrond:~$ hdfs dfs -cat /rdf/prov/ebi-chembl-31.0.ttl
@prefix : <http://www.metabohub.fr/msd#> .
@prefix owl: <http://www.w3.org/2002/07/owl#> .
@prefix rdfs: <http://www.w3.org/2000/01/rdf-schema#> .
@prefix xsd: <http://www.w3.org/2000/10/XMLSchema#> .
@prefix rdf: <http://www.w3.org/1999/02/22-rdf-syntax-ns#> .
@prefix prov: <http://www.w3.org/ns/prov#> .

<https://services.pfem.clermont.inrae.fr/gitlab/metabosemdatalake/databases/database-ext-ebi-chembl>
  a prov:Entity, <http://www.w3.org/ns/dcat#Dataset>;
  <http://purl.org/dc/terms/title> "chembl";
  <http://purl.org/dc/terms/description> "Category ebi / Database chembl";
  <http://purl.org/dc/terms/modified> "2022-08-24T07:28:14"^^<http://www.w3.org/2001/XMLSchema#dateTime>;
  prov:wasGeneratedBy "https://github.com/p2m2/service-rdf-database-deployment/";
  <http://www.w3.org/ns/dcat#Distribution> "https://services.pfem.clermont.inrae.fr/gitlab/metabosemdatalake/databases/

<https://services.pfem.clermont.inrae.fr/gitlab/metabosemdatalake/databases/database-ext-ebi-chembl/tags/31.0>
  a prov:Entity, <http://www.w3.org/ns/dcat#Distribution>;
  <http://purl.org/dc/terms/title> "31.0";
  <http://purl.org/dc/terms/modified> "2022-08-24T07:28:14"^^<http://www.w3.org/2001/XMLSchema#dateTime>;
  prov:wasGeneratedBy "https://services.pfem.clermont.inrae.fr/gitlab/metabosemdatalake/databases/database-ext-ebi-cher
  <http://www.w3.org/ns/dcat#accessURL> "hdfs://rdf/ebi/chembl/31.0" .

<https://services.pfem.clermont.inrae.fr/gitlab/metabosemdatalake/databases/database-ext-ebi-chembl/-/pipelines/39926>
  a prov:Activity;
  prov:used "https://services.pfem.clermont.inrae.fr/gitlab/metabosemdatalake/databases/database-ext-ebi-chembl";
  prov:startedAtTime "2022-08-24T07:28:14"^^<http://www.w3.org/2001/XMLSchema#dateTime>;
  prov:endedAtTime "2022-08-24T07:28:15"^^<http://www.w3.org/2001/XMLSchema#dateTime> .
```