

Scripting pour l'agrégation d'annotations de données biologiques

Vincent ROCHER Mathieu GACHET

Université de Rouen

27 Février 2015

Sommaire

1 Déroulement du pipeline

- Principe
- Mind Map

2 Connexion aux différentes databases

3 Déroulement du script

Déroulement du pipeline
Connexion aux différentes databases
Déroulement du script

Principe
Mind Map

Objectif

On cherche à créer un pipeline de scripts permettant l'annotation de gènes à partir de différentes bases de données.

Objectif

On cherche à créer un pipeline de scripts permettant l'annotation de gènes à partir de différentes bases de données.

Déroulement

- ① Fichier d'entrée : Fichier tabulé en format texte
Gene symbol Organism

Objectif

On cherche à créer un pipeline de scripts permettant l'annotation de gènes à partir de différentes bases de données.

Déroulement

- ① Fichier d'entrée : Fichier tabulé en format texte
Gene symbol Organism
- ② Récupération des données via un programme principal (main.pl).

Objectif

On cherche à créer un pipeline de scripts permettant l'annotation de gènes à partir de différentes bases de données.

Déroulement

- ① Fichier d'entrée : Fichier tabulé en format texte
Gene symbol Organism
- ② Récupération des données via un programme principal (main.pl).
- ③ Annotation progressive via différents modules perl qui seront appelés par le programme principal.

Objectif

On cherche à créer un pipeline de scripts permettant l'annotation de gènes à partir de différentes bases de données.

Déroulement

- ① Fichier d'entrée : Fichier tabulé en format texte
Gene symbol Organism
- ② Récupération des données via un programme principal (main.pl).
- ③ Annotation progressive via différents modules perl qui seront appelés par le programme principal.
- ④ Sortie : Génération d'un fichier Json qui sera interprété en javascript par la page html.

Déroulement du pipeline

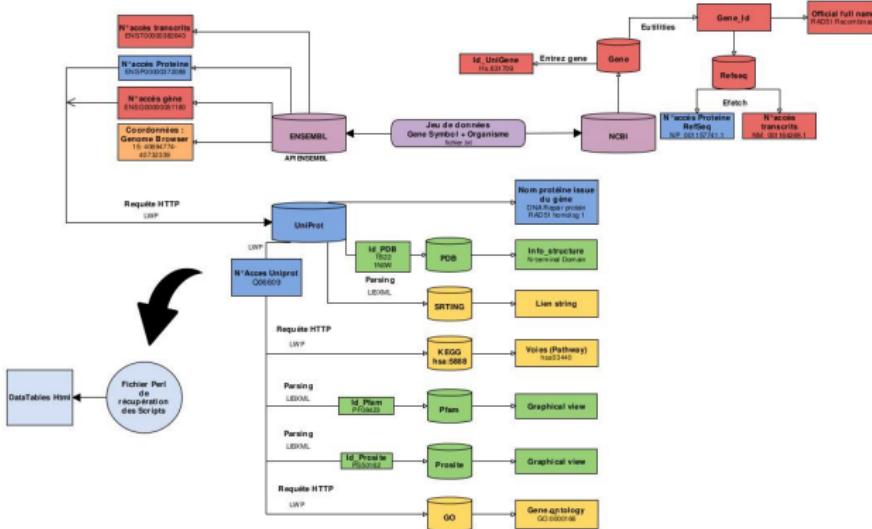
Connexion aux différentes databases

Déroulement du script

Principe

Mind Map

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



Sommaire

1 Déroulement du pipeline

2 Connexion aux différentes databases

- NCBI
- EnsEMBL
- Uniprot
- PDB
- Pfam
- Prosite
- KEGG
- GO :Ontology

Déroulement du pipeline
Connexion aux différentes databases
Déroulement du script

NCBI
EnsEMBL
Uniprot
PDB
Pfam
Prosite
KEGG
GO :Ontology

Objectif

Récupérer les annotations à partir des données présentes sur le NCBI
Output : Nom complet ,ID du gène, des transcrits, des protéines, et
UniGene

Objectif

Récupérer les annotations à partir des données présentes sur le NCBI
Output : Nom complet ,ID du gène, des transcrits, des protéines, et
UniGene

Déroulement

- ① On récupère les ids Gene via easearch

Objectif

Récupérer les annotations à partir des données présentes sur le NCBI
Output : Nom complet ,ID du gène, des transcrits, des protéines, et UniGene

Déroulement

- ① On récupère les ids Gene via easearch
- ② On envoi les ids et on récupère un format gene table via efetch qui contient les données

Objectif

Récupérer les annotations à partir des données présentes sur le NCBI
Output : Nom complet ,ID du gène, des transcrits, des protéines, et UniGene

Déroulement

- ① On récupère les ids Gene via easearch
- ② On envoi les ids et on récupère un format gene table via efetch qui contient les données

Modules utilisés

- ① EUtilities pour Gene

Objectif

Récupérer les annotations à partir des données présentes sur le NCBI
Output : Nom complet ,ID du gène, des transcrits, des protéines, et UniGene

Déroulement

- ① On récupère les ids Gene via easearch
- ② On envoi les ids et on récupère un format gene table via efetch qui contient les données

Modules utilisés

- ① EUtilities pour Gene
- ② DB : :EntrezGene pour UniGene

Déroulement du pipeline
Connexion aux différentes databases
Déroulement du script

NCBI
EnsEMBL
Uniprot
PDB
Pfam
Prosite
KEGG
GO :Ontology

Objectif

Récupérer les annotations à partir des données présentes sur EnsEMBL
Output : ID du gène, des transcrits (et du canonique), des protéines et sa localisation sur le chromosome.

Objectif

Récupérer les annotations à partir des données présentes sur EnsEMBL
Output : ID du gène, des transcrits (et du canonique), des protéines et sa localisation sur le chromosome.

Déroulement

- ① On fournit le Gene Symbol et l'organisme à EnsEMBL

Objectif

Récupérer les annotations à partir des données présentes sur EnsEMBL
Output : ID du gène, des transcrits (et du canonique), des protéines et sa localisation sur le chromosome.

Déroulement

- ① On fournit le Gene Symbol et l'organisme à EnsEMBL
- ② On récupère les données sous la forme d'un objet.

Objectif

Récupérer les annotations à partir des données présentes sur EnsEMBL
Output : ID du gène, des transcrits (et du canonique), des protéines et sa localisation sur le chromosome.

Déroulement

- ① On fournit le Gene Symbol et l'organisme à EnsEMBL
- ② On récupère les données sous la forme d'un objet.

Objectif

Récupérer les annotations à partir des données présentes sur EnsEMBL
Output : ID du gène, des transcrits (et du canonique), des protéines et sa localisation sur le chromosome.

Déroulement

- ① On fournit le Gene Symbol et l'organisme à EnsEMBL
- ② On récupère les données sous la forme d'un objet.

Modules utilisés

- ① API EnsEMBL

Déroulement du pipeline
Connexion aux différentes databases
Déroulement du script

NCBI
EnsEMBL
Uniprot
PDB
Pfam
Prosite
KEGG
GO :Ontology

Objectif

Récupérer les annotations à partir des données présentes sur Uniprot
Output : ID uniprot, nom complet, ID PDB et lien STRING.

Objectif

Récupérer les annotations à partir des données présentes sur Uniprot
Output : ID uniprot, nom complet, ID PDB et lien STRING.

Déroulement

- ① On fournit l'id EnsEMBL du gène et de la protéine canonique

Objectif

Récupérer les annotations à partir des données présentes sur Uniprot
Output : ID uniprot, nom complet, ID PDB et lien STRING.

Déroulement

- ① On fournit l'id EnsEMBL du gène et de la protéine canonique
- ② On récupère l'id uniprot de la protéine canonique d'EnsEMBL

Objectif

Récupérer les annotations à partir des données présentes sur Uniprot
Output : ID uniprot, nom complet, ID PDB et lien STRING.

Déroulement

- ① On fournit l'id EnsEMBL du gène et de la protéine canonique
- ② On récupère l'id uniprot de la protéine canonique d'EnsEMBL
- ③ On construit le lien xml et on parse pour récupérer les données

Objectif

Récupérer les annotations à partir des données présentes sur Uniprot
Output : ID uniprot, nom complet, ID PDB et lien STRING.

Déroulement

- ① On fournit l'id EnsEMBL du gène et de la protéine canonique
- ② On récupère l'id uniprot de la protéine canonique d'EnsEMBL
- ③ On construit le lien xml et on parse pour récupérer les données

Modules utilisés

- ① LWP : :UserAgent pour récupérer l'id Uniprot

Objectif

Récupérer les annotations à partir des données présentes sur Uniprot
Output : ID uniprot, nom complet, ID PDB et lien STRING.

Déroulement

- ① On fournit l'id EnsEMBL du gène et de la protéine canonique
- ② On récupère l'id uniprot de la protéine canonique d'EnsEMBL
- ③ On construit le lien xml et on parse pour récupérer les données

Modules utilisés

- ① LWP : :UserAgent pour récupérer l'id Uniprot
- ② XML : :LibXML pour récupérer le reste des informations

Déroulement du pipeline
Connexion aux différentes databases
Déroulement du script

NCBI
EnsEMBL
Uniprot
PDB
Pfam
Prosite
KEGG
GO :Ontology

Objectif

Récupérer les annotations à partir des données présentes sur PDB
Output : Informations structurales

Objectif

Récupérer les annotations à partir des données présentes sur PDB
Output : Informations structurales

Déroulement

- ➊ On fournit l'id PDB obtenu via uniprot (peut y en avoir plusieurs)

Objectif

Récupérer les annotations à partir des données présentes sur PDB
Output : Informations structurales

Déroulement

- ① On fournit l'id PDB obtenu via uniprot (peut y en avoir plusieurs)
- ② On construit le lien xml et on parse pour récupérer les données

Objectif

Récupérer les annotations à partir des données présentes sur PDB
Output : Informations structurales

Déroulement

- ① On fournit l'id PDB obtenu via uniprot (peut y en avoir plusieurs)
- ② On construit le lien xml et on parse pour récupérer les données

Modules utilisés

- ① XML : :LibXML pour récupérer les informations

Déroulement du pipeline
Connexion aux différentes databases
Déroulement du script

NCBI
EnsEMBL
Uniprot
PDB
Pfam
Prosite
KEGG
GO :Ontology

Objectif

Récupérer les annotations à partir des données présentes sur Pfam
Output : ID Pfam + ID du domaine

Objectif

Récupérer les annotations à partir des données présentes sur Pfam
Output : ID Pfam + ID du domaine

Déroulement

- ① On fournit l'ID Uniprot obtenu

Objectif

Récupérer les annotations à partir des données présentes sur Pfam
Output : ID Pfam + ID du domaine

Déroulement

- ① On fournit l'ID Uniprot obtenu
- ② On construit le lien xml et on parse pour récupérer les données

Objectif

Récupérer les annotations à partir des données présentes sur Pfam
Output : ID Pfam + ID du domaine

Déroulement

- ① On fournit l'ID Uniprot obtenu
- ② On construit le lien xml et on parse pour récupérer les données

Modules utilisés

- ① XML : :LibXML pour récupérer les informations

Objectif

Récupérer les annotations à partir des données présentes sur Prosite

Output : ID Prosite + Informations structurales qui permettent de

construire un dessin



Objectif

Récupérer les annotations à partir des données présentes sur Prosite
Output : ID Prosite + Informations structurales qui permettent de construire un dessin



Déroulement

- ① On fournit l'ID Uniprot obtenu

Objectif

Récupérer les annotations à partir des données présentes sur Prosite
Output : ID Prosite + Informations structurales qui permettent de construire un dessin



Déroulement

- ① On fournit l'ID Uniprot obtenu
- ② On construit le lien xml et on parse pour récupérer les données

Objectif

Récupérer les annotations à partir des données présentes sur Prosite
Output : ID Prosite + Informations structurales qui permettent de construire un dessin



Déroulement

- ① On fournit l'ID Uniprot obtenu
- ② On construit le lien xml et on parse pour récupérer les données

Modules utilisés

- ① XML : :LibXML pour récupérer les informations

Objectif

Récupérer les annotations à partir des données présentes sur KEGG

Output : On récupère le numéro d'accès du gène ainsi que celui du pathway (lien vers image)

Objectif

Récupérer les annotations à partir des données présentes sur KEGG

Output : On récupère le numéro d'accès du gène ainsi que celui du pathway (lien vers image)

Déroulement

- ① On fournit l'ID Uniprot obtenu

Objectif

Récupérer les annotations à partir des données présentes sur KEGG

Output : On récupère le numéro d'accès du gène ainsi que celui du pathway (lien vers image)

Déroulement

- ➊ On fournit l'ID Uniprot obtenu
- ➋ On construit la requête HTTP à partir de l'ID Uniprot, on récupère l'information directement dans une variable

Objectif

Récupérer les annotations à partir des données présentes sur KEGG

Output : On récupère le numéro d'accès du gène ainsi que celui du pathway (lien vers image)

Déroulement

- ① On fournit l'ID Uniprot obtenu
- ② On construit la requête HTTP à partir de l'ID Uniprot, on récupère l'information directement dans une variable

Modules utilisés

- ① LWP : :UserAgent pour construire la requête HTTP

Objectif

Récupérer les annotations à partir des données présentes sur Quick :GO
Output : On récupère le numéro d'accès Quick :GO, sa description partielle ainsi que la catégorie à laquelle il appartient

Objectif

Récupérer les annotations à partir des données présentes sur Quick :GO

Output : On récupère le numéro d'accès Quick :GO, sa description partielle ainsi que la catégorie à laquelle il appartient

Déroulement

- ① On fournit l'ID Uniprot obtenu

Objectif

Récupérer les annotations à partir des données présentes sur Quick :GO

Output : On récupère le numéro d'accès Quick :GO, sa description partielle ainsi que la catégorie à laquelle il appartient

Déroulement

- ① On fournit l'ID Uniprot obtenu
- ② On construit la requête HTTP à partir de l'ID Uniprot, on récupère un fichier tabulé qui contient les informations

Objectif

Récupérer les annotations à partir des données présentes sur Quick :GO

Output : On récupère le numéro d'accès Quick :GO, sa description partielle ainsi que la catégorie à laquelle il appartient

Déroulement

- ① On fournit l'ID Uniprot obtenu
- ② On construit la requête HTTP à partir de l'ID Uniprot, on récupère un fichier tabulé qui contient les informations

Modules utilisés

- ① LWP : :UserAgent pour construire la requête HTTP

Déroulement du pipeline
Connexion aux différentes databases
Déroulement du script

Explication
Démonstration
Conclusion

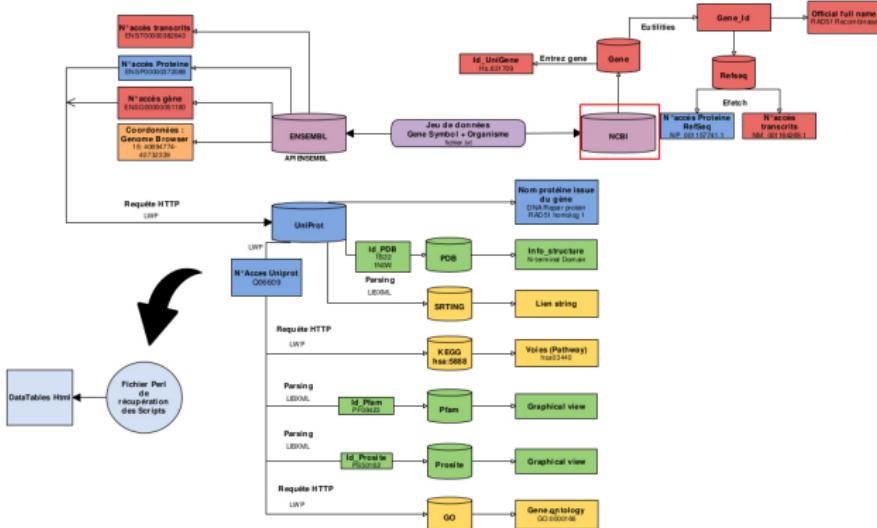
Sommaire

- 1 Déroulement du pipeline
- 2 Connexion aux différentes databases
- 3 Déroulement du script
 - Explication
 - Démonstration
 - Conclusion

Déroulement du pipeline Connexion aux différentes databases Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



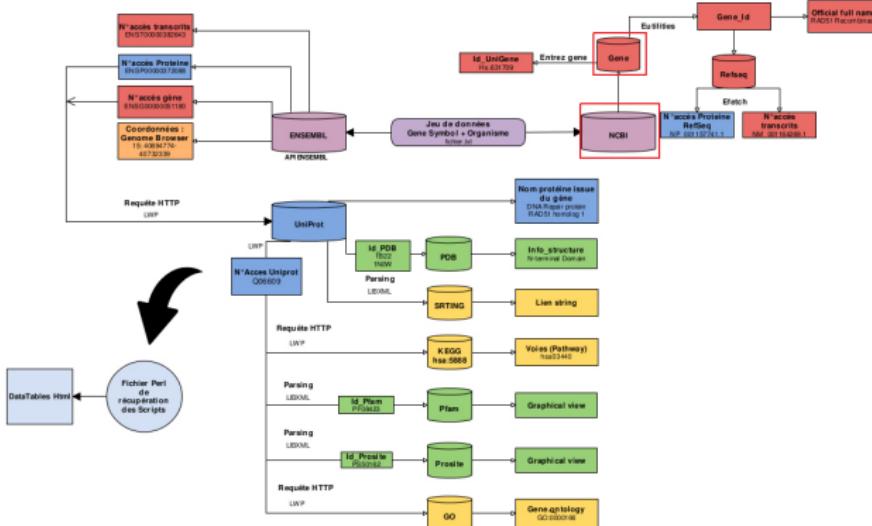
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

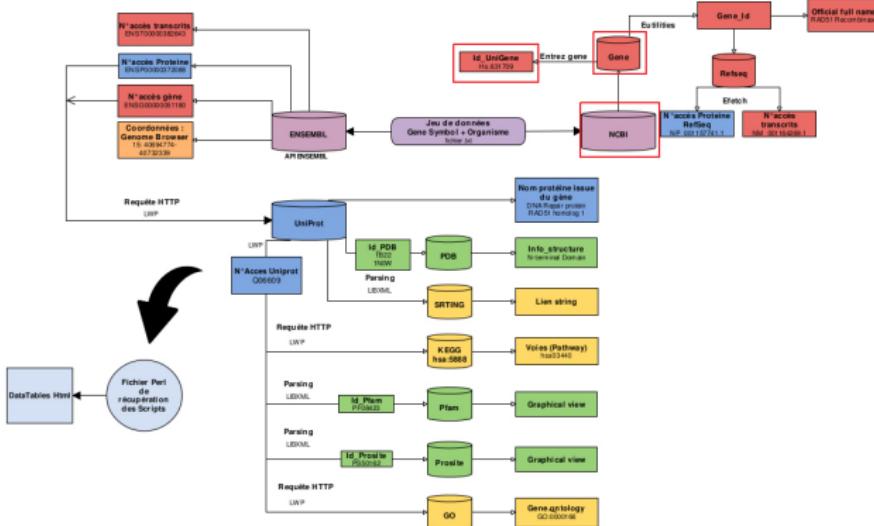
Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



Déroulement du pipeline Connexion aux différentes databases Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



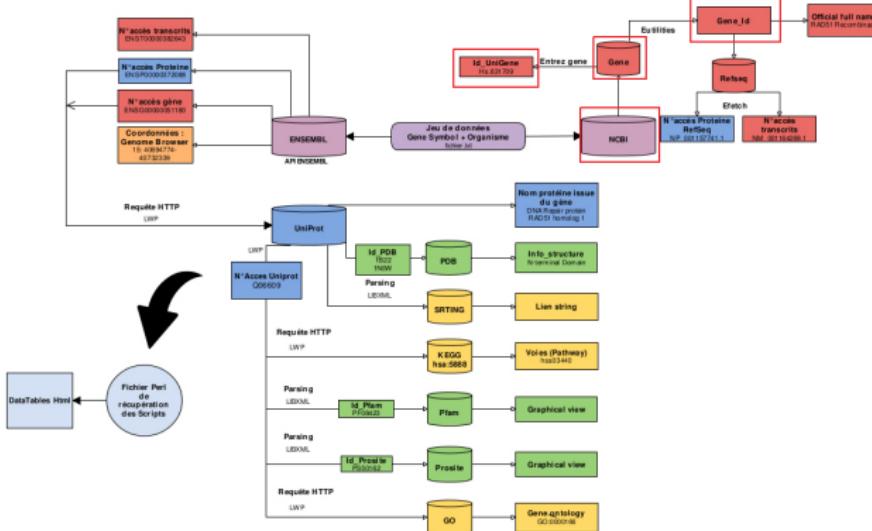
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



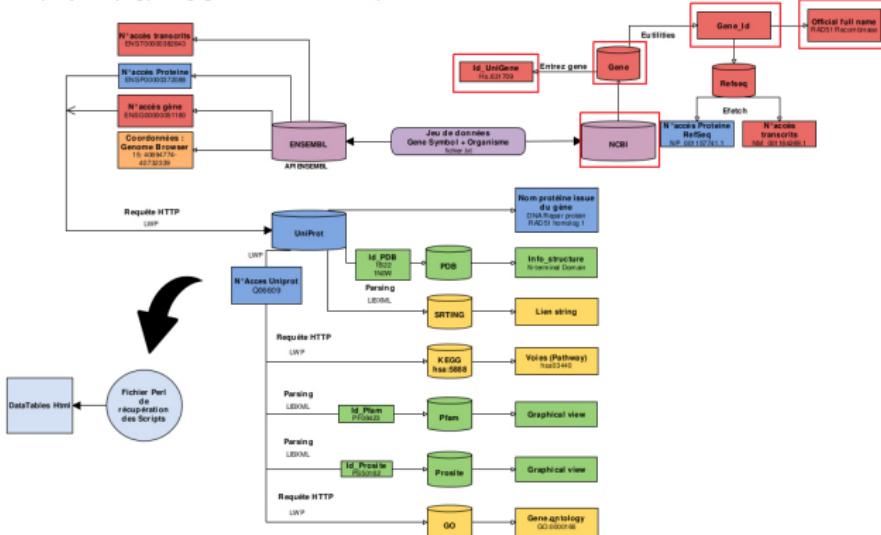
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

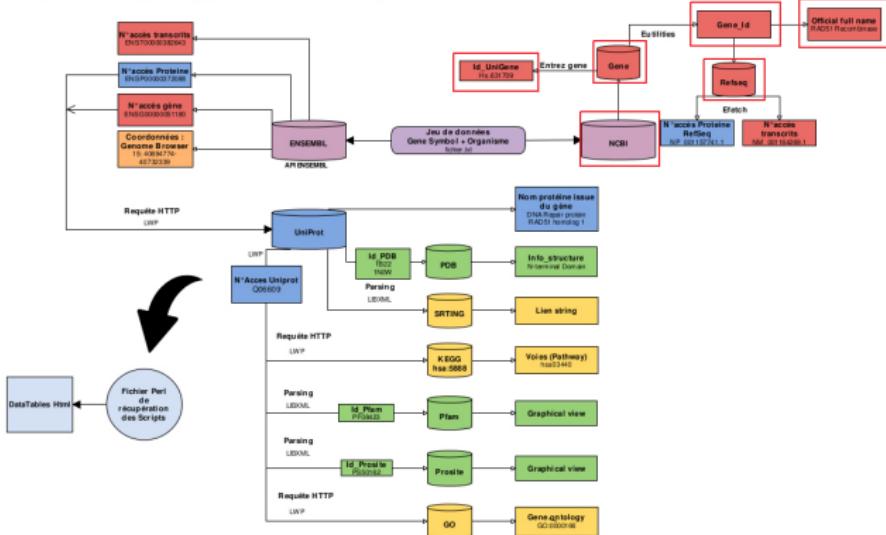
Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



Déroulement du pipeline Connexion aux différentes databases Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



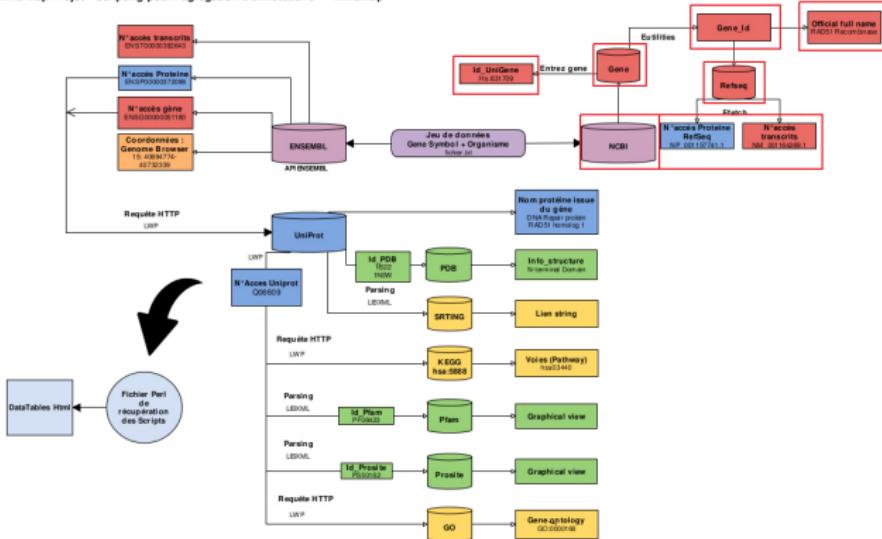
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



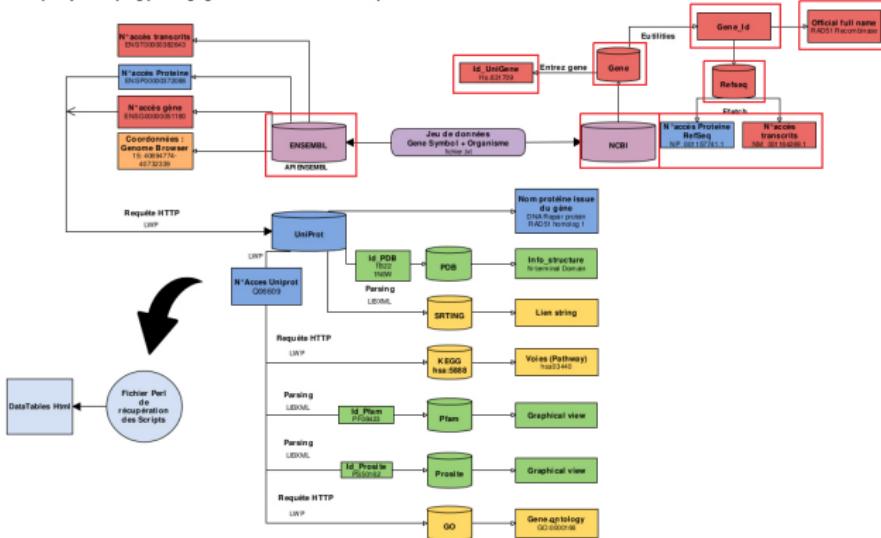
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

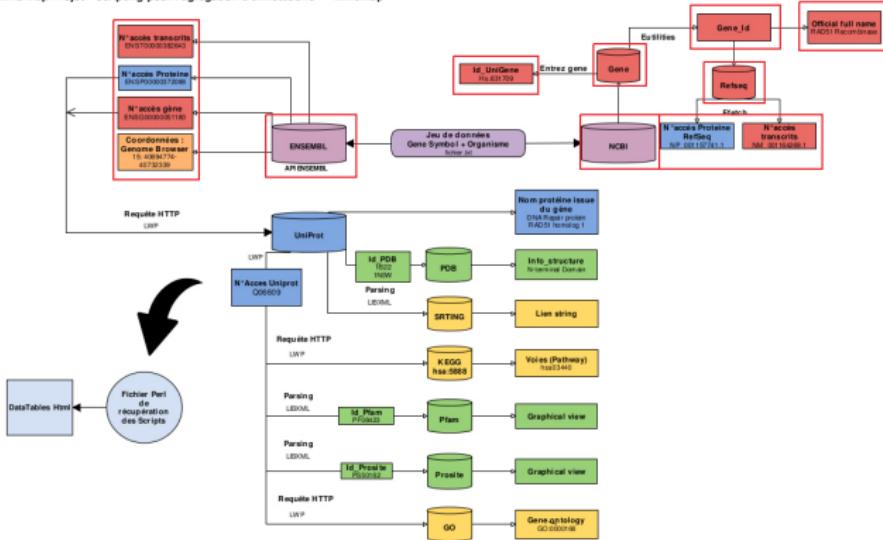
Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



Déroulement du pipeline Connexion aux différentes databases Déroulement du script

Explication Démonstration Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



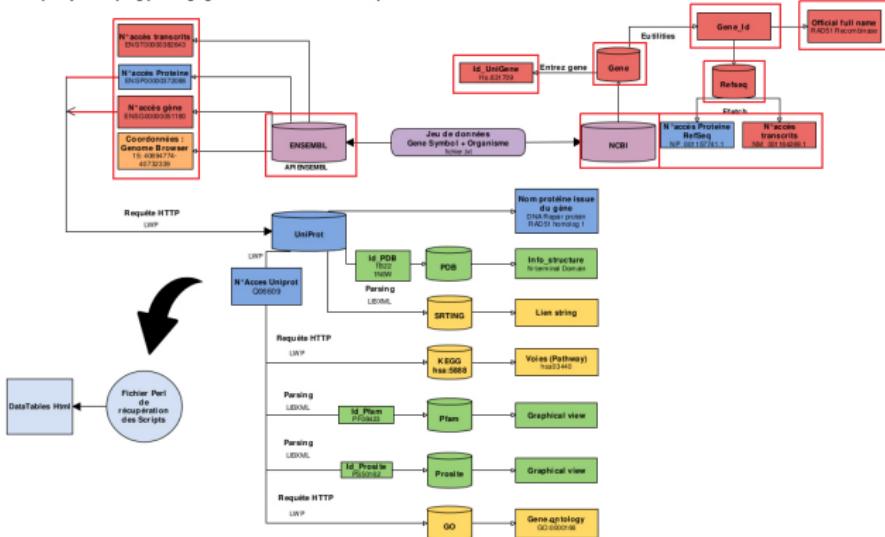
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



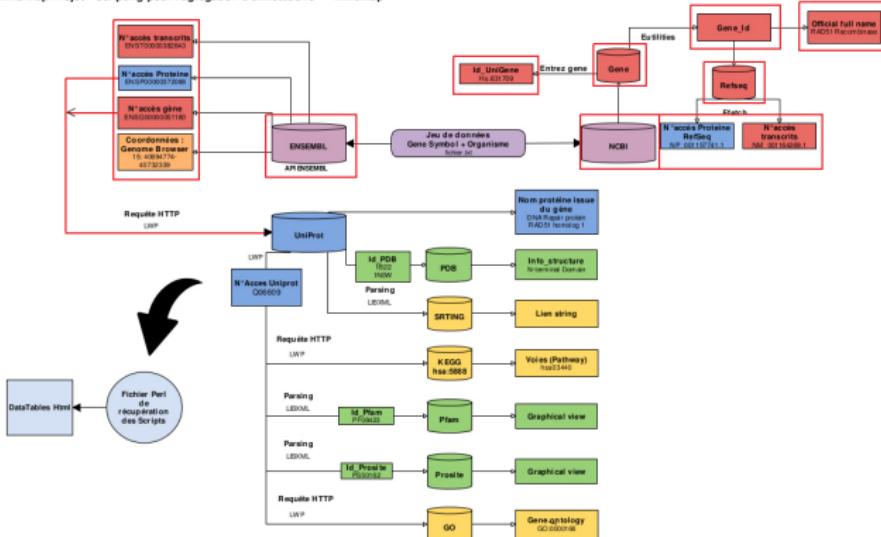
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



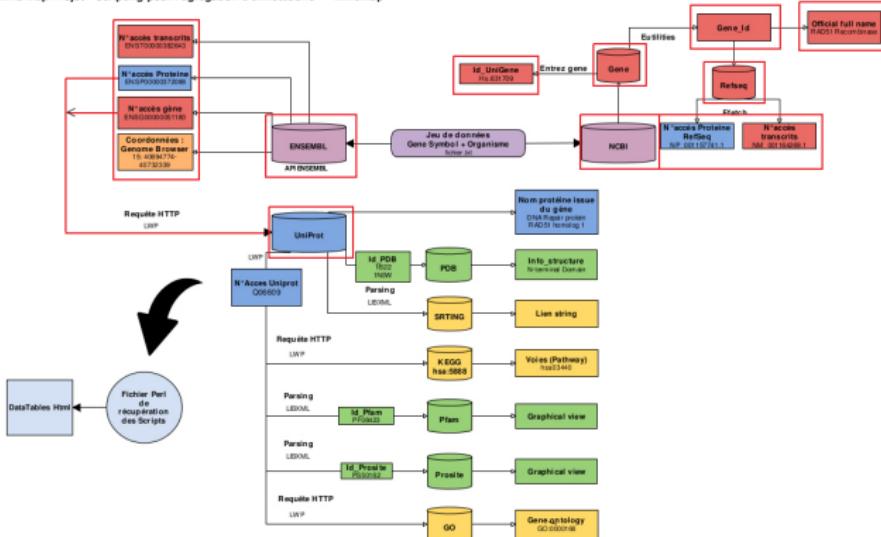
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



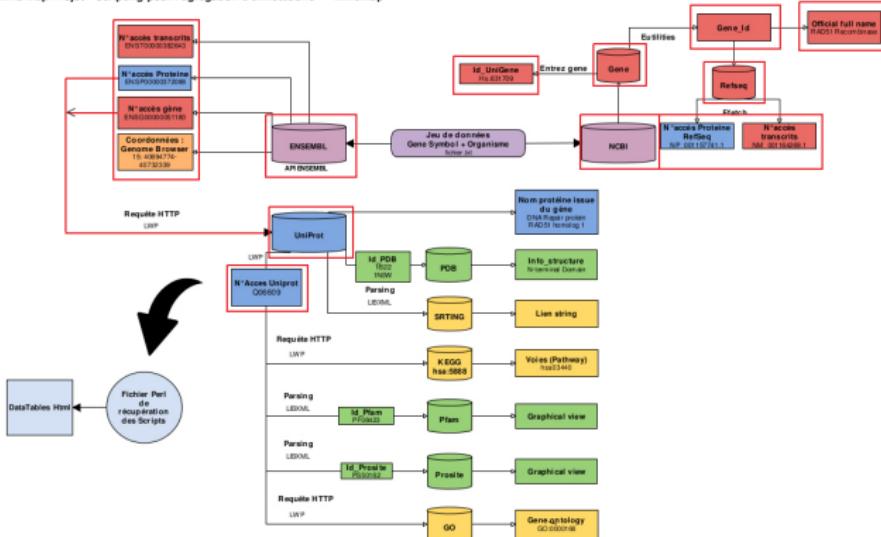
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



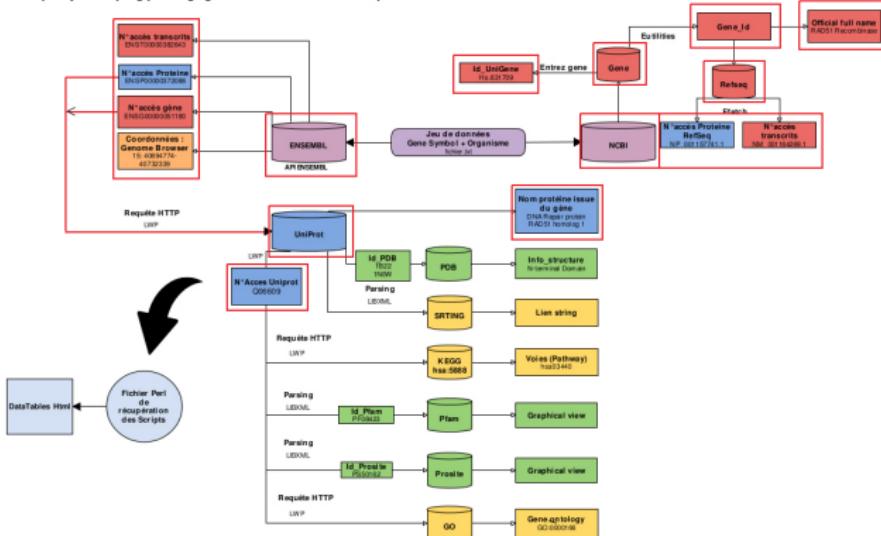
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



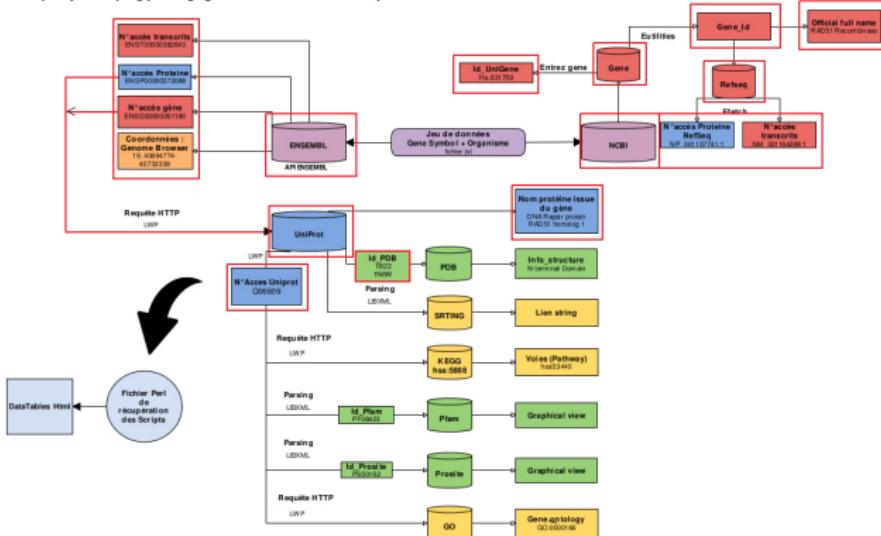
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

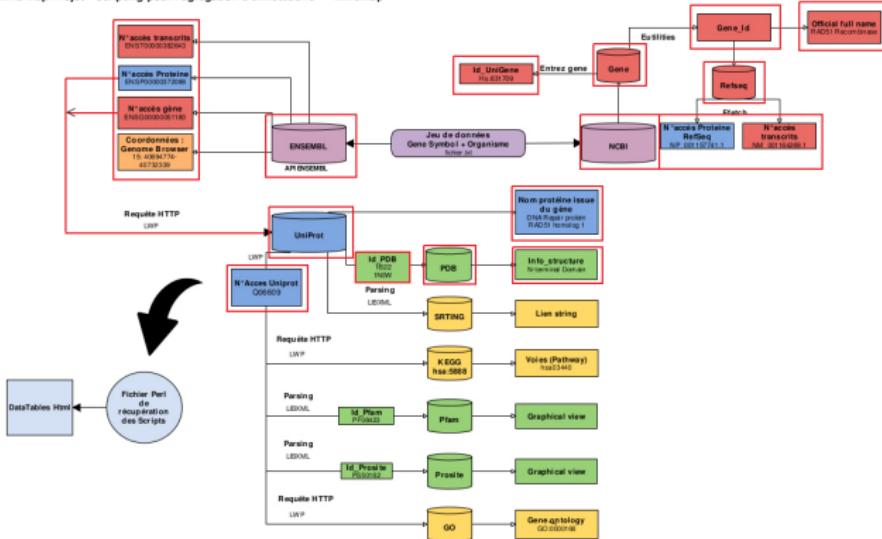
Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



Déroulement du pipeline Connexion aux différentes databases Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



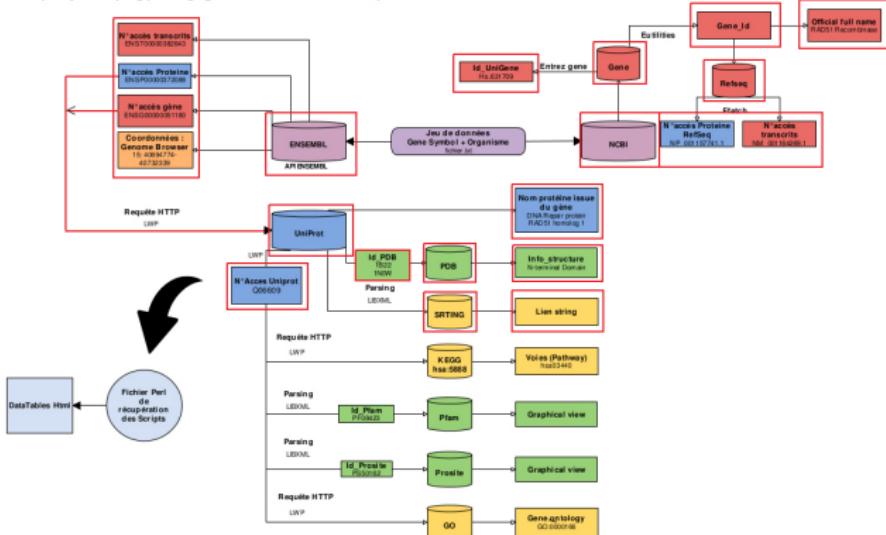
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



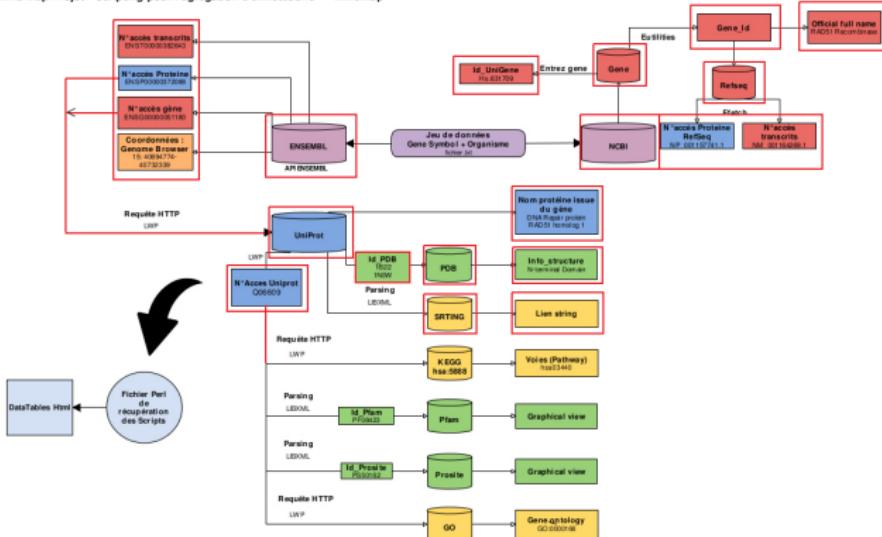
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

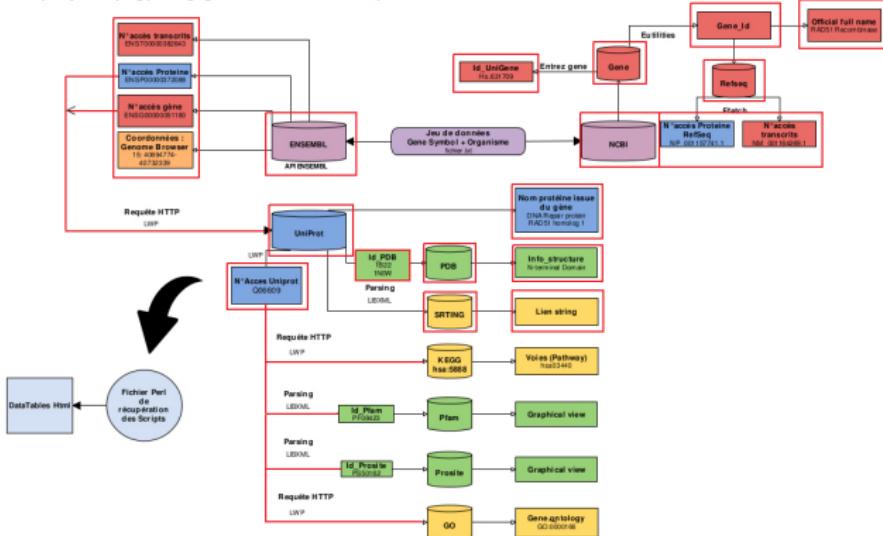
Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



Déroulement du pipeline Connexion aux différentes databases Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



Déroulement du pipeline

Connexion aux différentes databases

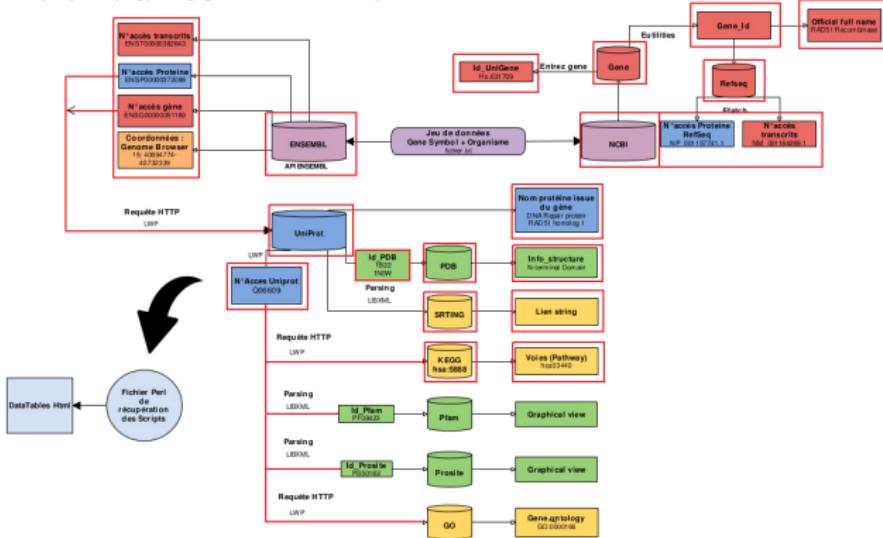
Déroulement du script

Explication

Démonstration

Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



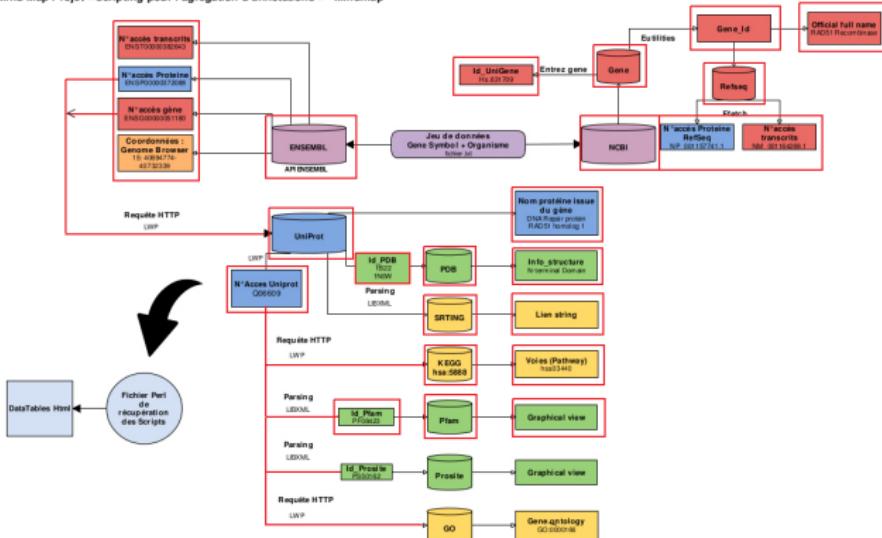
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



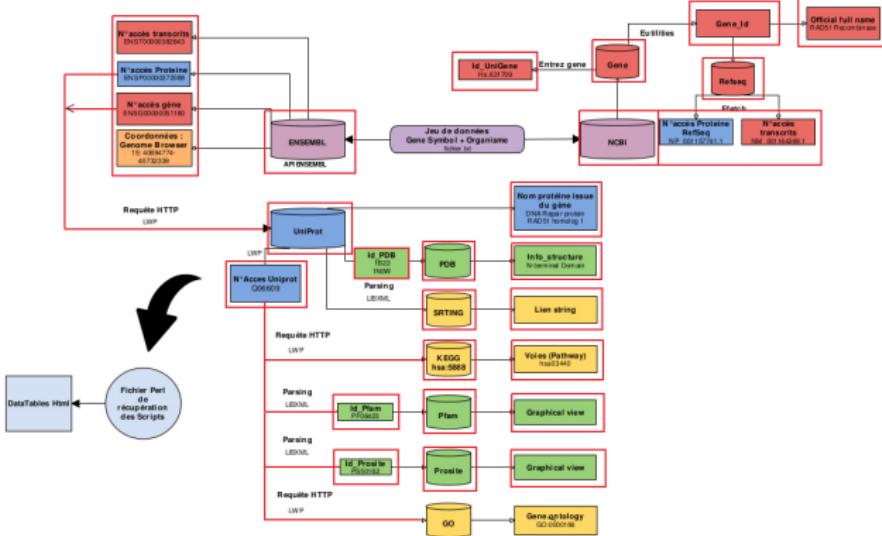
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



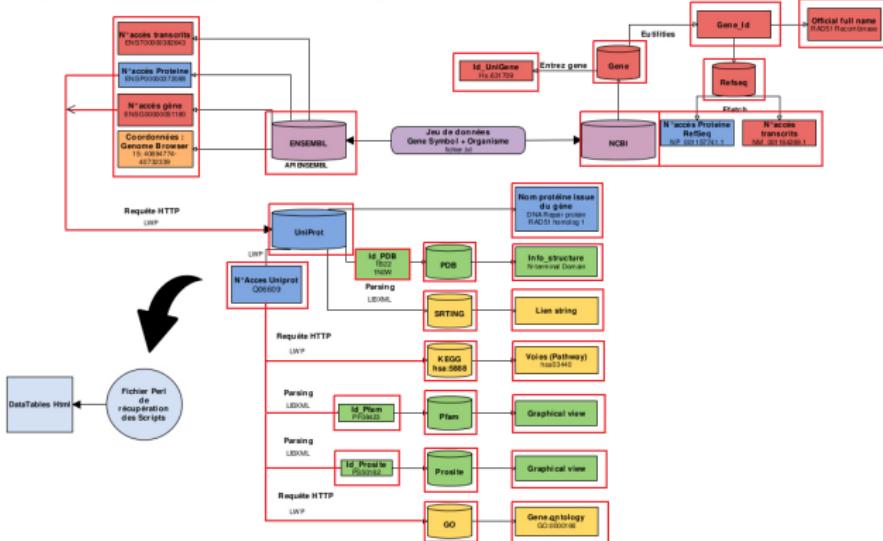
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



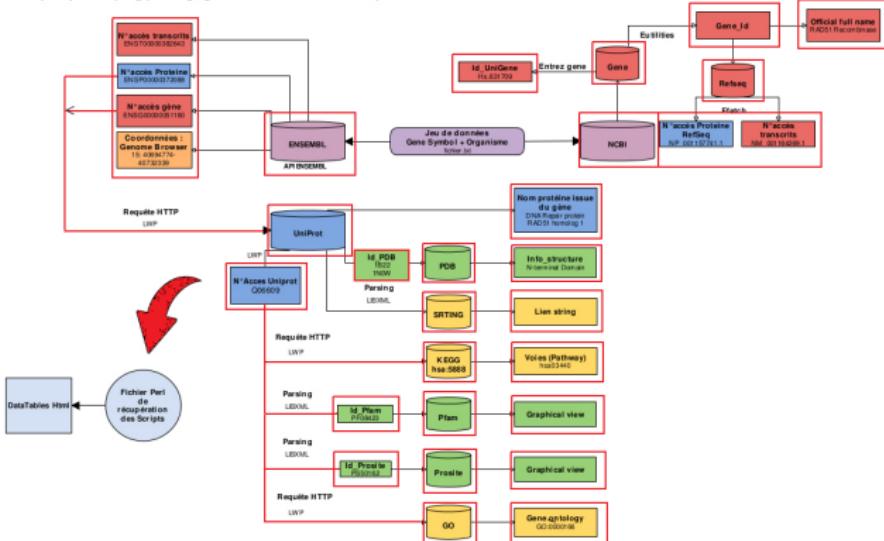
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



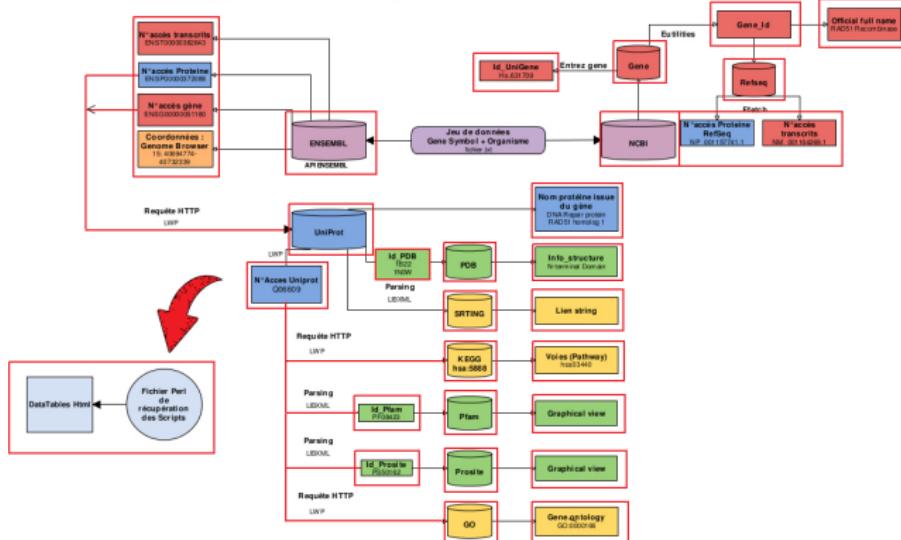
Déroulement du pipeline

Connexion aux différentes databases

Déroulement du script

Explication
Démonstration
Conclusion

Mind Map Projet « scripting pour l'agrégation d'annotations » - Mindmap



Déroulement du pipeline
Connexion aux différentes databases
Déroulement du script

Explication
Démonstration
Conclusion

Démonstration

Les annotations obtenues sont désormais disponibles sous la forme d'un tableau html dynamique

Conclusion

- ➊ Cette connexion multiple permet d'obtenir les informations de façon relativement fiable car on les obtient depuis les bases de données sources.

Conclusion

- ① Cette connexion multiple permet d'obtenir les informations de façon relativement fiable car on les obtient depuis les bases de données sources.
- ② Cependant ces connexions multiples rendent le programme très lent

Amélioration

- ① Il existe des programmes qui permettent de récupérer les annotations sur plusieurs bases de données, sans avoir besoin de réaliser ces connexions multiples.

Conclusion

- ① Cette connexion multiple permet d'obtenir les informations de façon relativement fiable car on les obtient depuis les bases de données sources.
- ② Cependant ces connexions multiples rendent le programme très lent

Amélioration

- ① Il existe des programmes qui permettent de récupérer les annotations sur plusieurs bases de données, sans avoir besoin de réaliser ces connexions multiples.
- ② L'abondance d'informations rend la génération du résultat très compliqué.