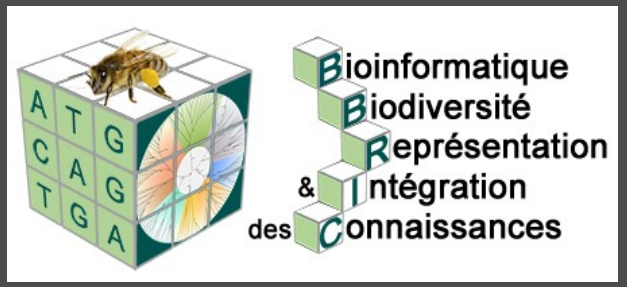


Mise en place de serveurs Galaxy dans le cadre du réseau CATI BBRIC

{Sebastien.Carrere, Ludovic.Legrand,Jerome.Gouzy}@toulouse.inra.fr
{Fabrice.Legeai,Anthony.Bretaudeau}@rennes.inra.fr

CATI BBRIC

- 35 bioinformaticiens
- 15 unités
- 10 sites

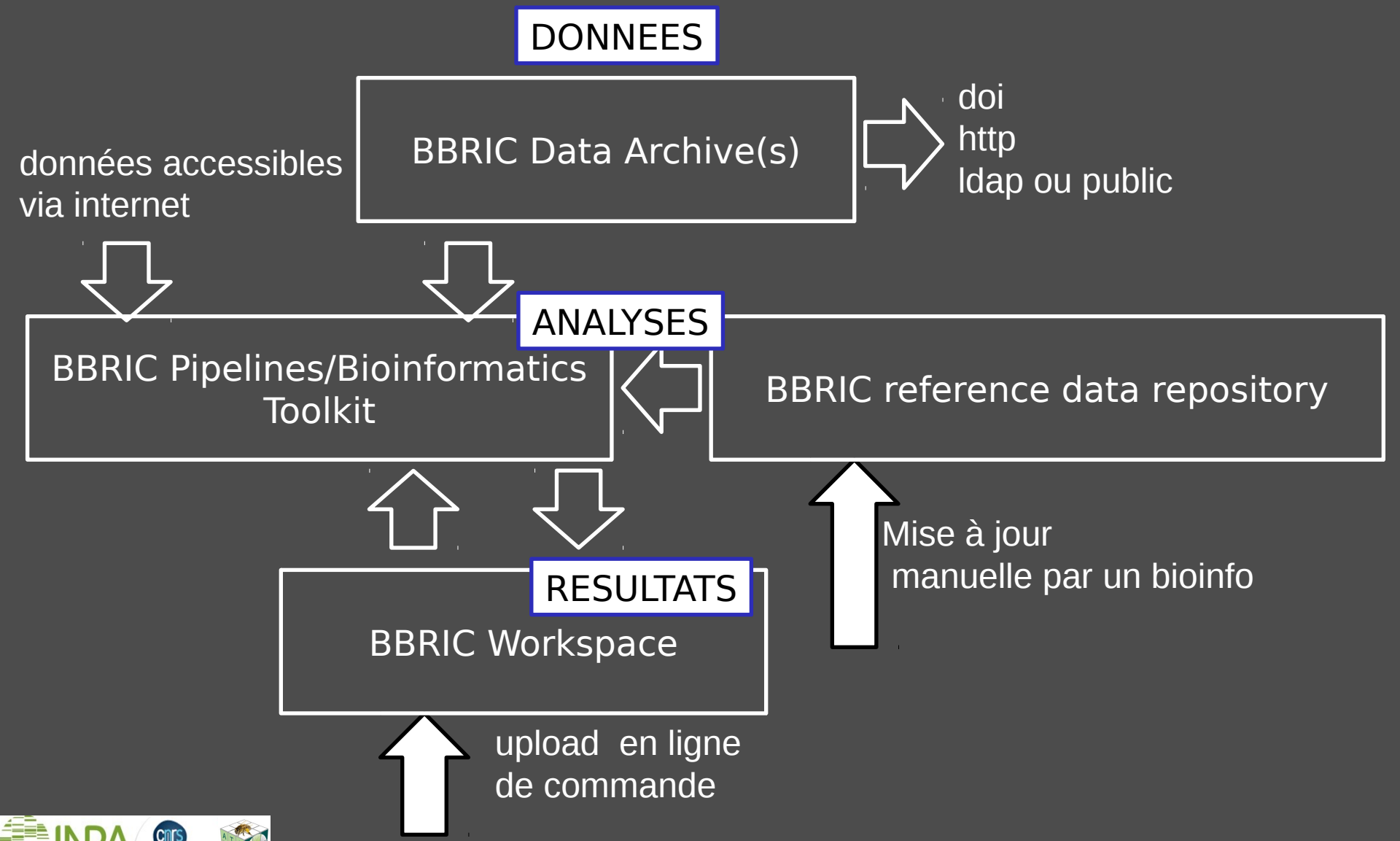
Modèles d'intérêt

- Plantes
- Bactéries
- Insectes
- Nématodes
- Champignons
- Oomycetes
- Virus

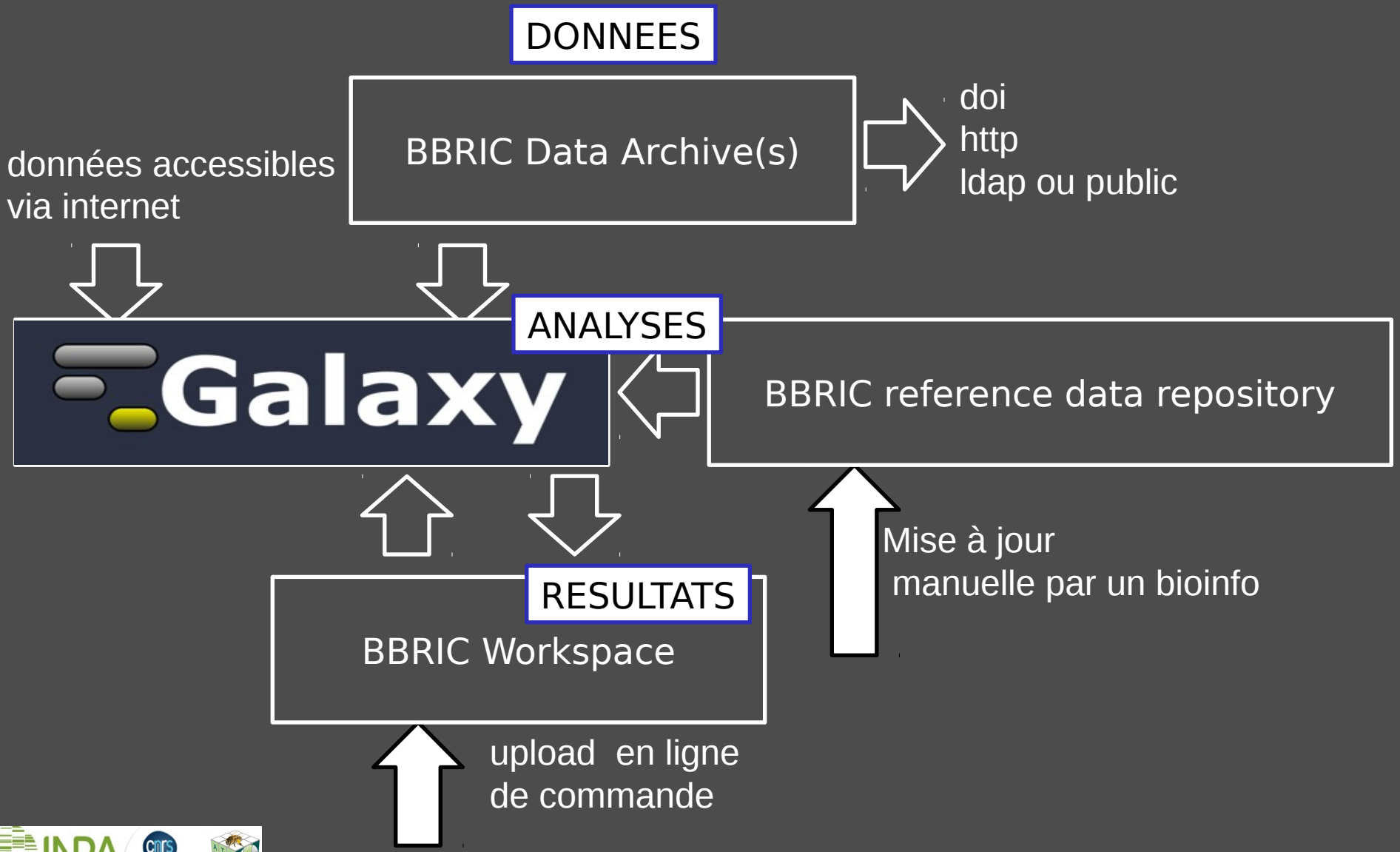
Domaines d'applications

- Assemblage (genomes/transcriptomes)
- Annotation structurale des génomes (gènes codants pour des protéines / ncRNA)
- Annotation fonctionnelle (genomes/transcriptomes)
- Analyse de l'expression (RNAseq, puces)
- Détection et analyse du polymorphisme
- Metagenomique
- Epigénomique
- Modélisation des réseaux métaboliques et de régulation.
- Gestion de collections (bactéries, insectes, etc.)
- Systématique: identification des espèces de groupes d'espèces d'intérêt
- Phylogénie, évolution
- Génomique des populations

Architecture bioinformatique BBRIC
4 composants interopérables



Architecture bioinformatique BBRIC
4 composants interopérables



Galaxy @ BBRIC
S'adapter au réseau

- Des instances au plus proche des données
  - @Rennes
  - Accès contrôlé
    - annuaire LDAP
  - @Toulouse
  - Accès contrôlé
    - Fédération Education Recherche (Shibboleth)
  - @Ailleurs ?
- Un toolshed pour partager les outils
- Un système pour maintenir les données de référence

Galaxy @ BBRIC

Pré-requis avant un passage en production

- Utilisation de fichiers compressés
  - fastq.gz
- Interopérabilité avec les instances d'Archive
  - Les utilisateurs doivent pouvoir analyser leurs données
- Interopérabilité avec les instances Workspace
  - Les utilisateurs doivent pouvoir être autonomes dans le post-processing des résultats intermédiaires

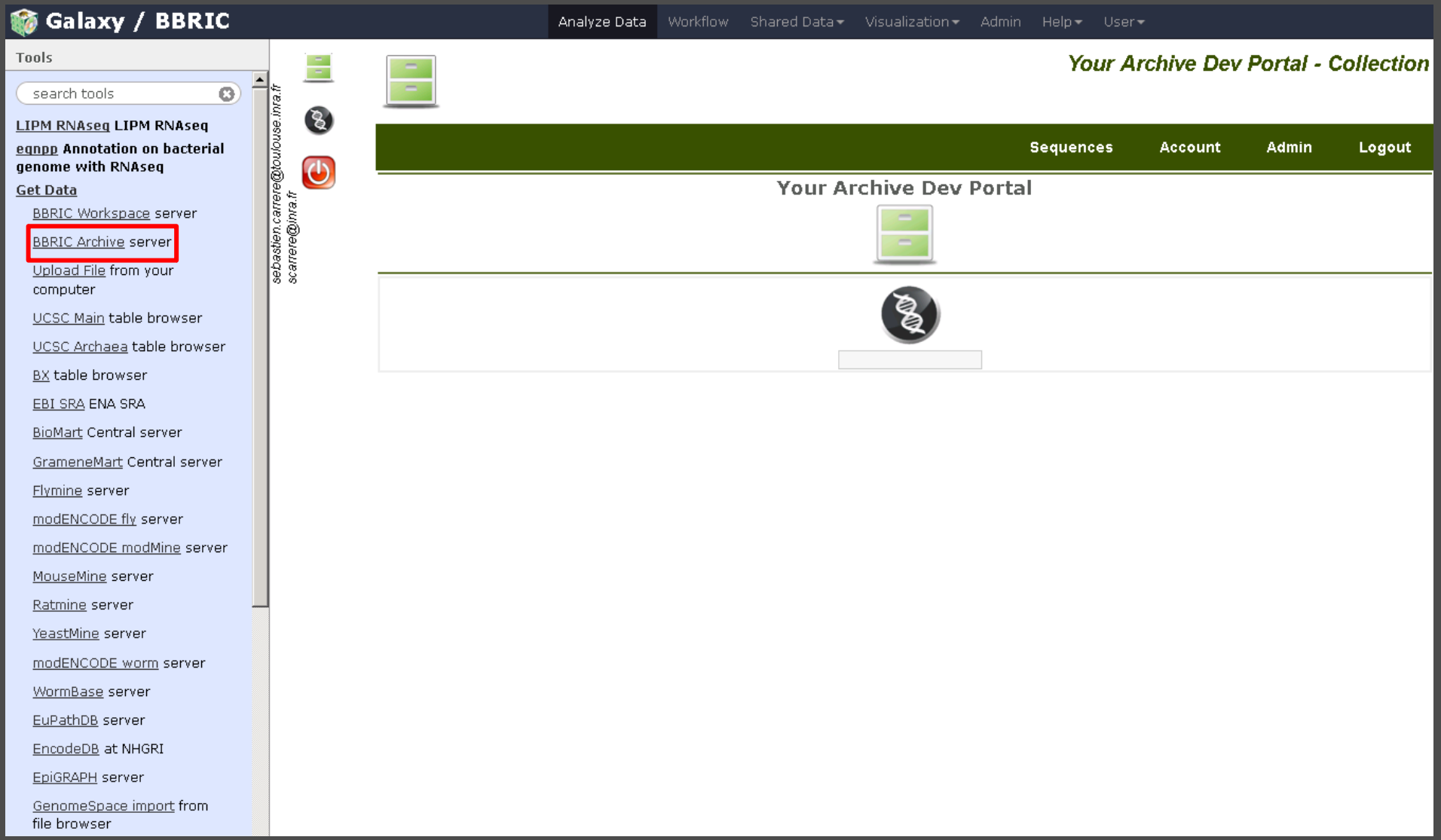
Utilisation de fichiers compressés

- De nombreux outils le permettent, pourquoi s'en priver ?
- Pourquoi Galaxy décompresse automatiquement ?
- Extension des datatypes
  - datatypes_conf.xml
  - lib/galaxy/datatypes/data.py
- Modification du code de l'upload
  - tools/data_source/upload.py
- Modification des fiches des outils pour ajout du nouveau type accepté

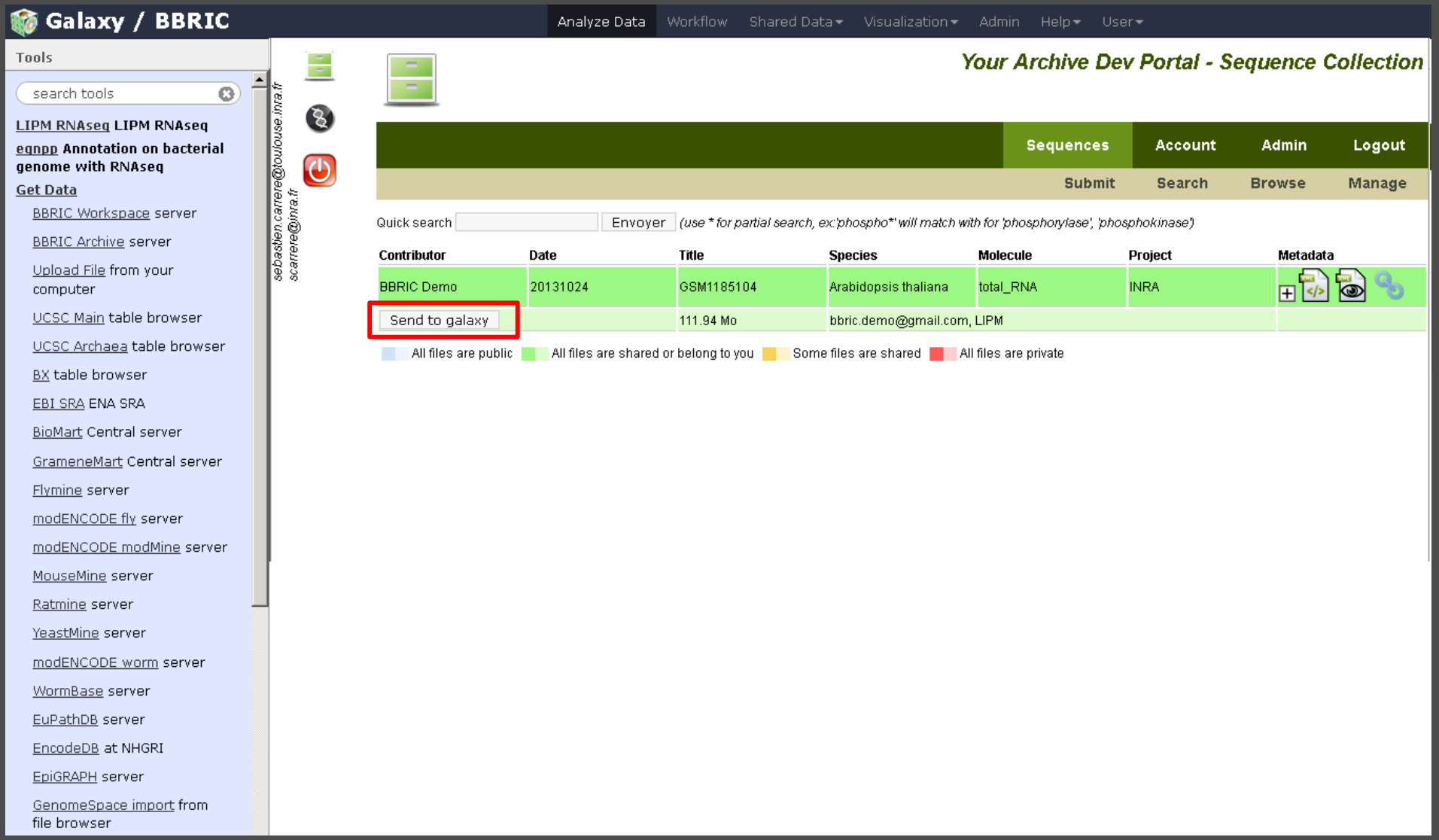
Interopérabilité avec Archive
Qu'est-ce que l'Archive ?

- Un système de conservation des données brutes
  - générique
  - sans limite de taille de fichier
- Un système ouvert
  - accès programmatique (API, client)
  - conversion vers les formats standards (SRA, GEO pour les séquences)
  - publication des données/métadonnées (DOI, lien permanent)
- Un accès sécurisé et authentifié
  - droits d'accès par fichier
  - authentification par la fédération Education Recherche
- Un outil accessible aux producteurs de données (les biologistes)
  - interface web (consultation, gestion)
  - moteur de recherche

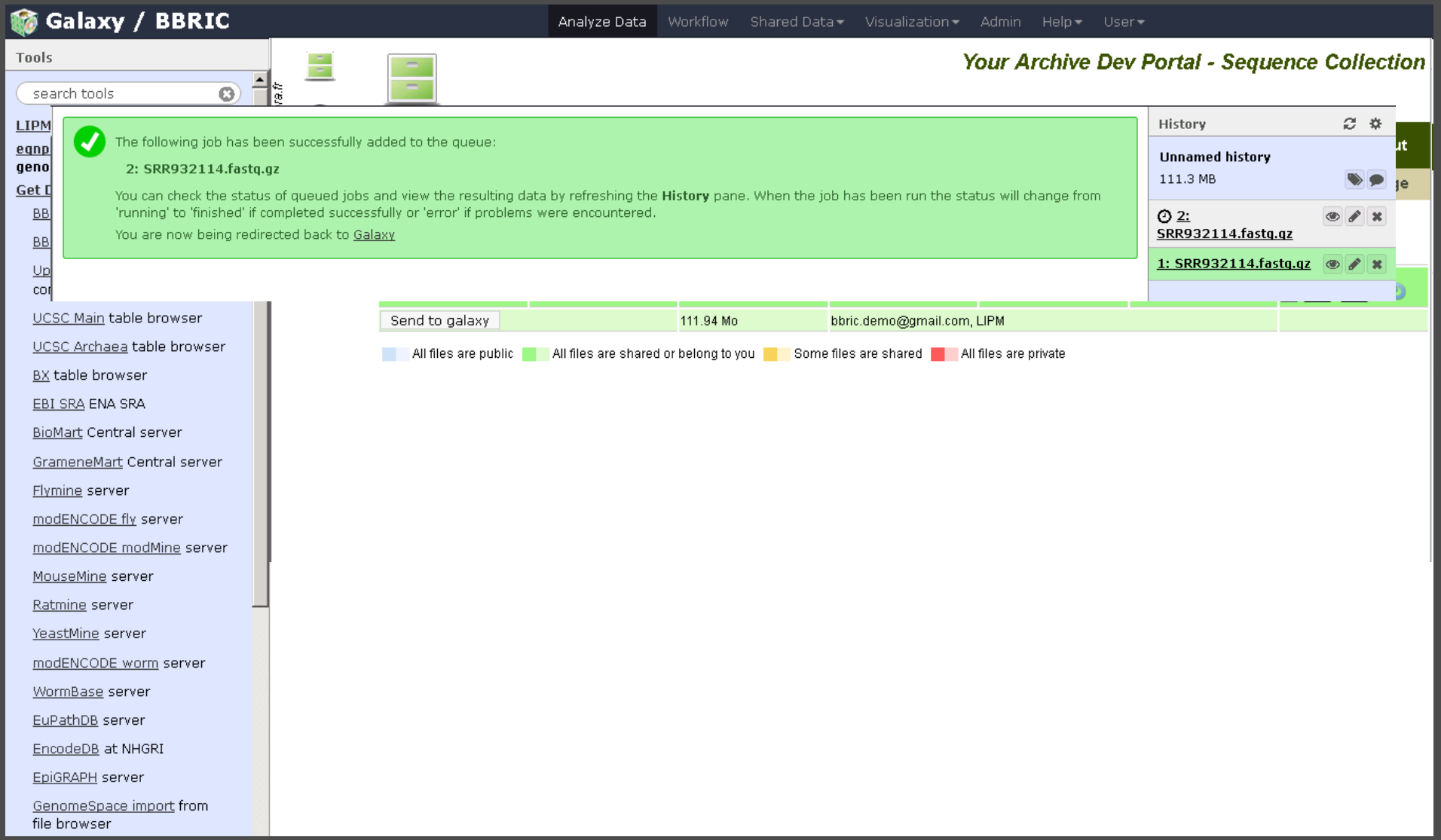
Interopérabilité avec Archive



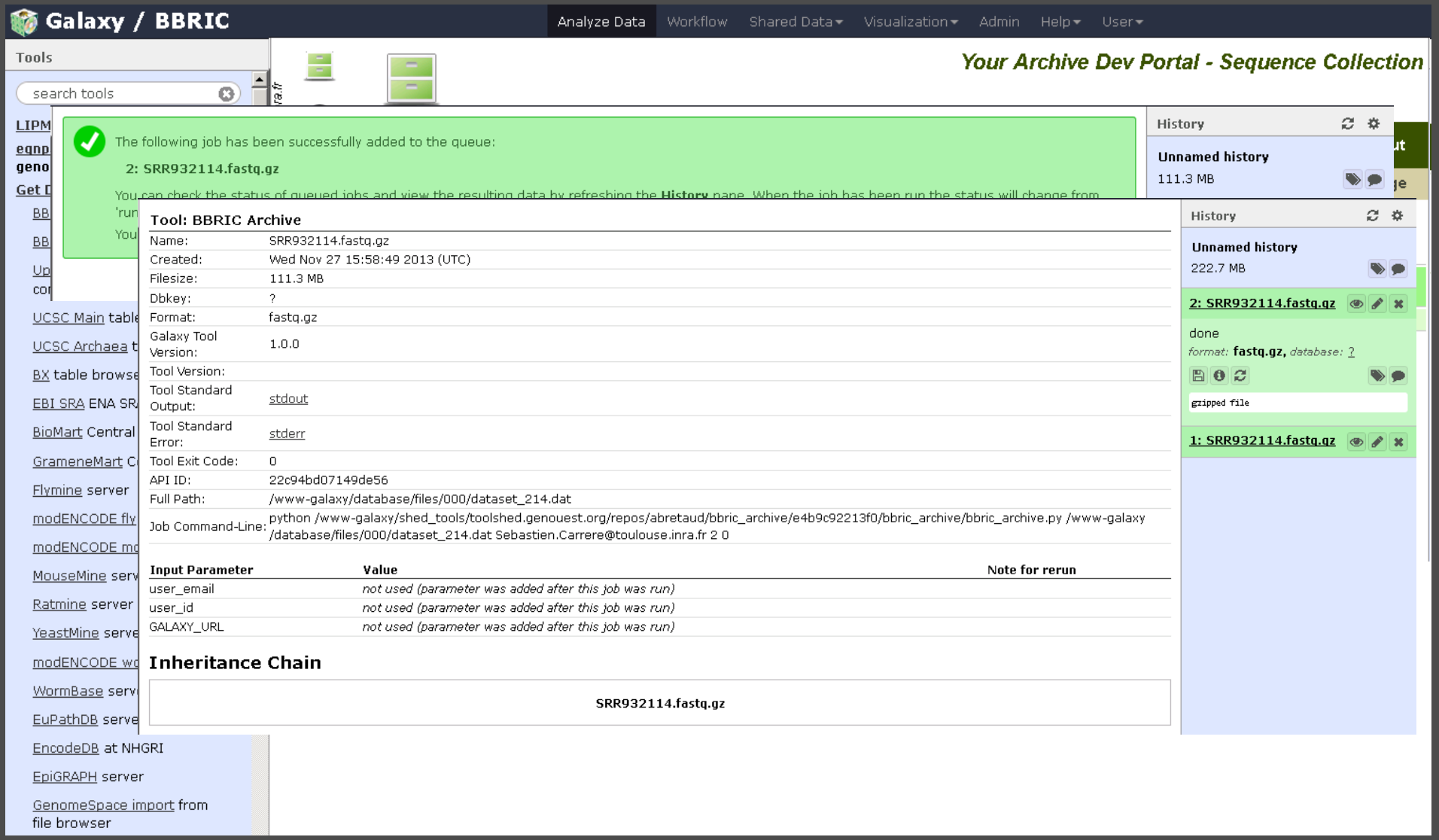
Interopérabilité avec Archive



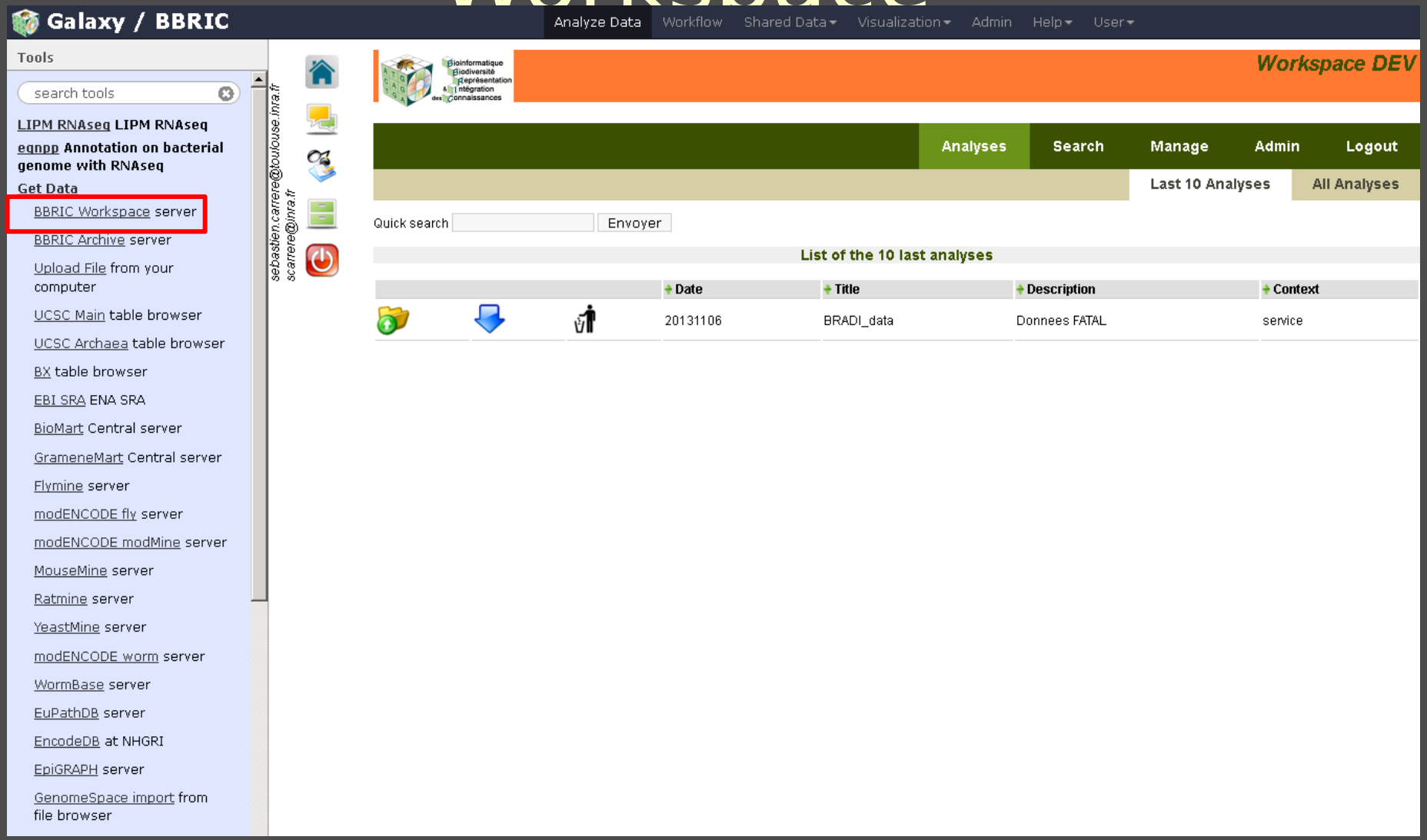
Interopérabilité avec Archive



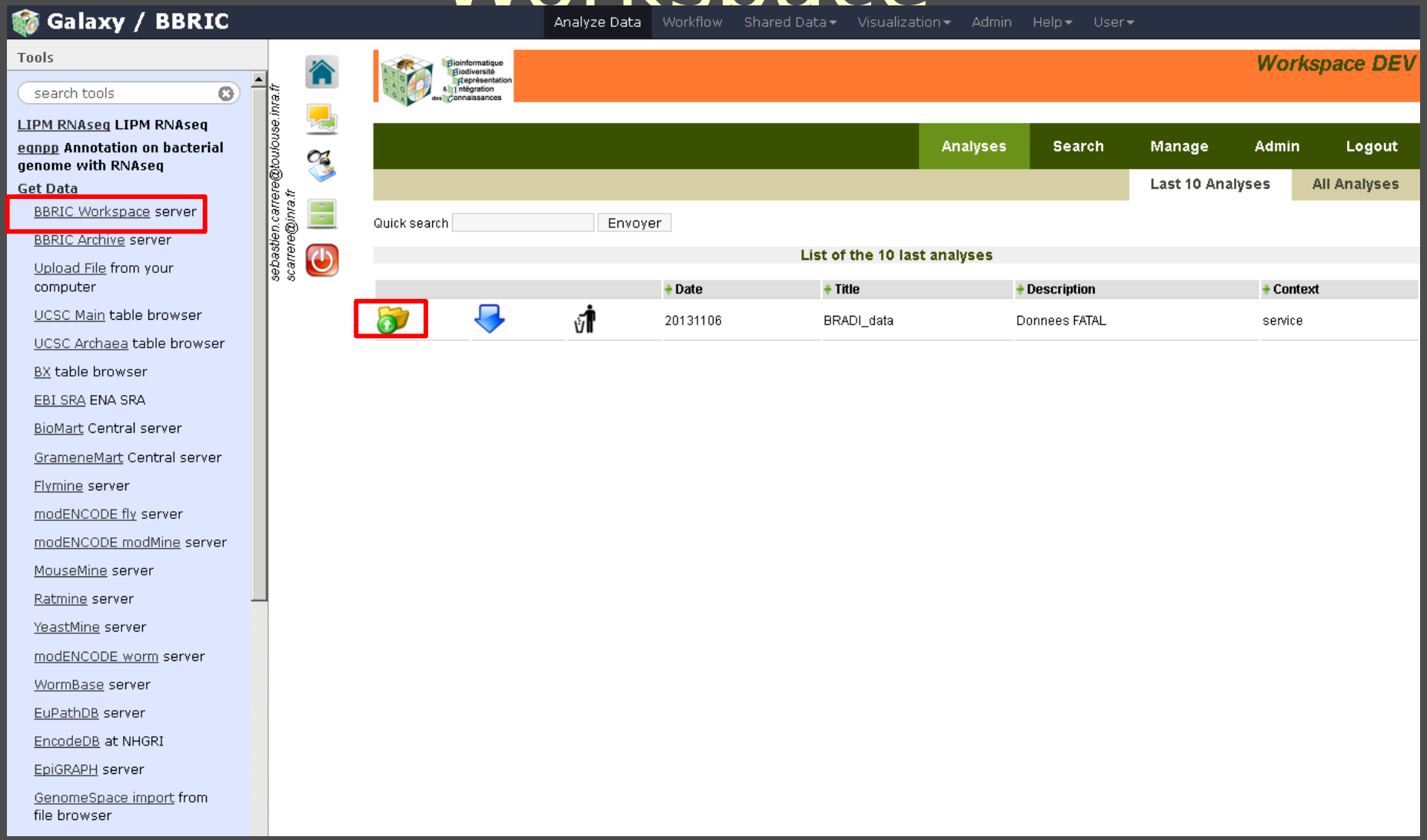
Interopérabilité avec Archive



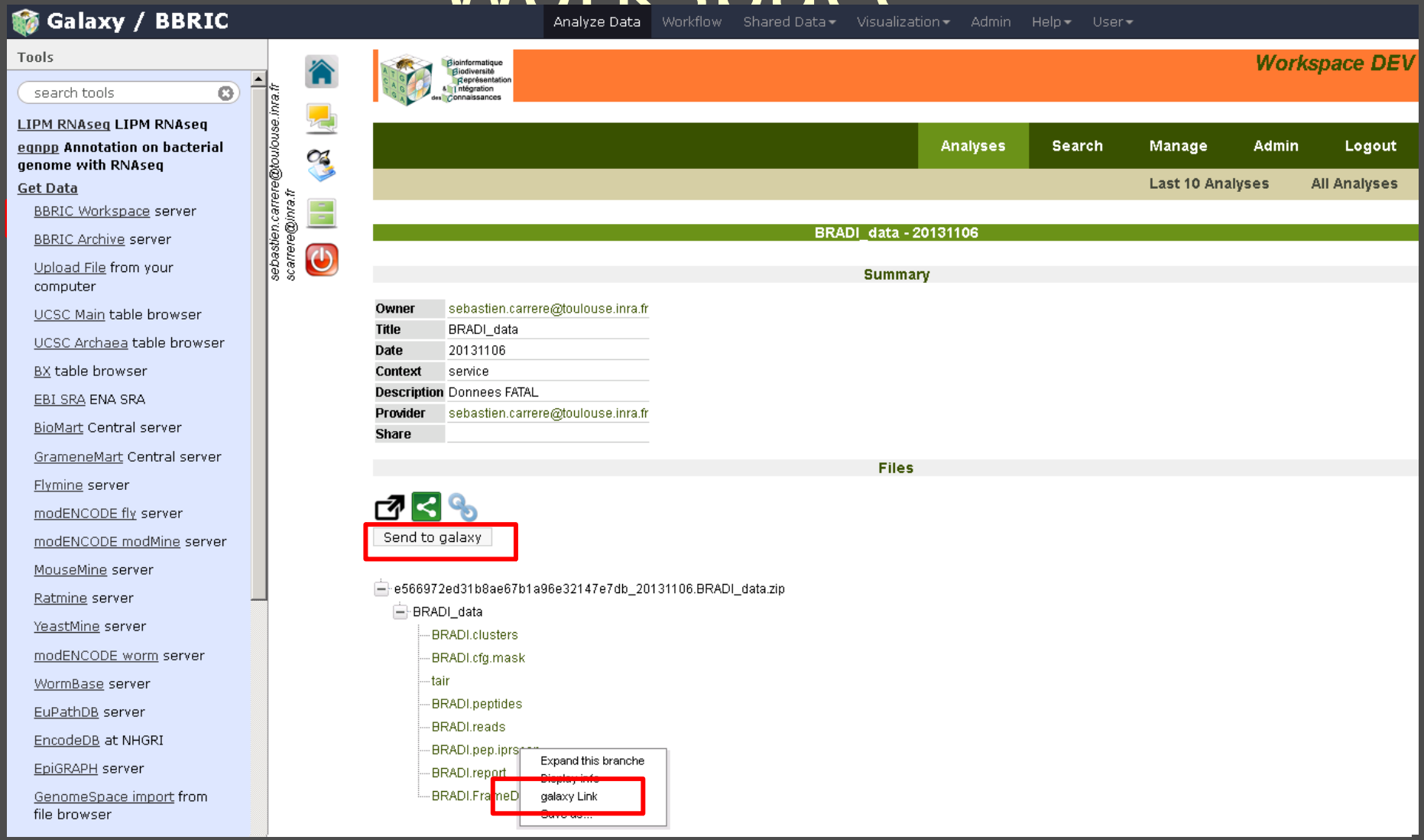
Interopérabilité avec Workspace



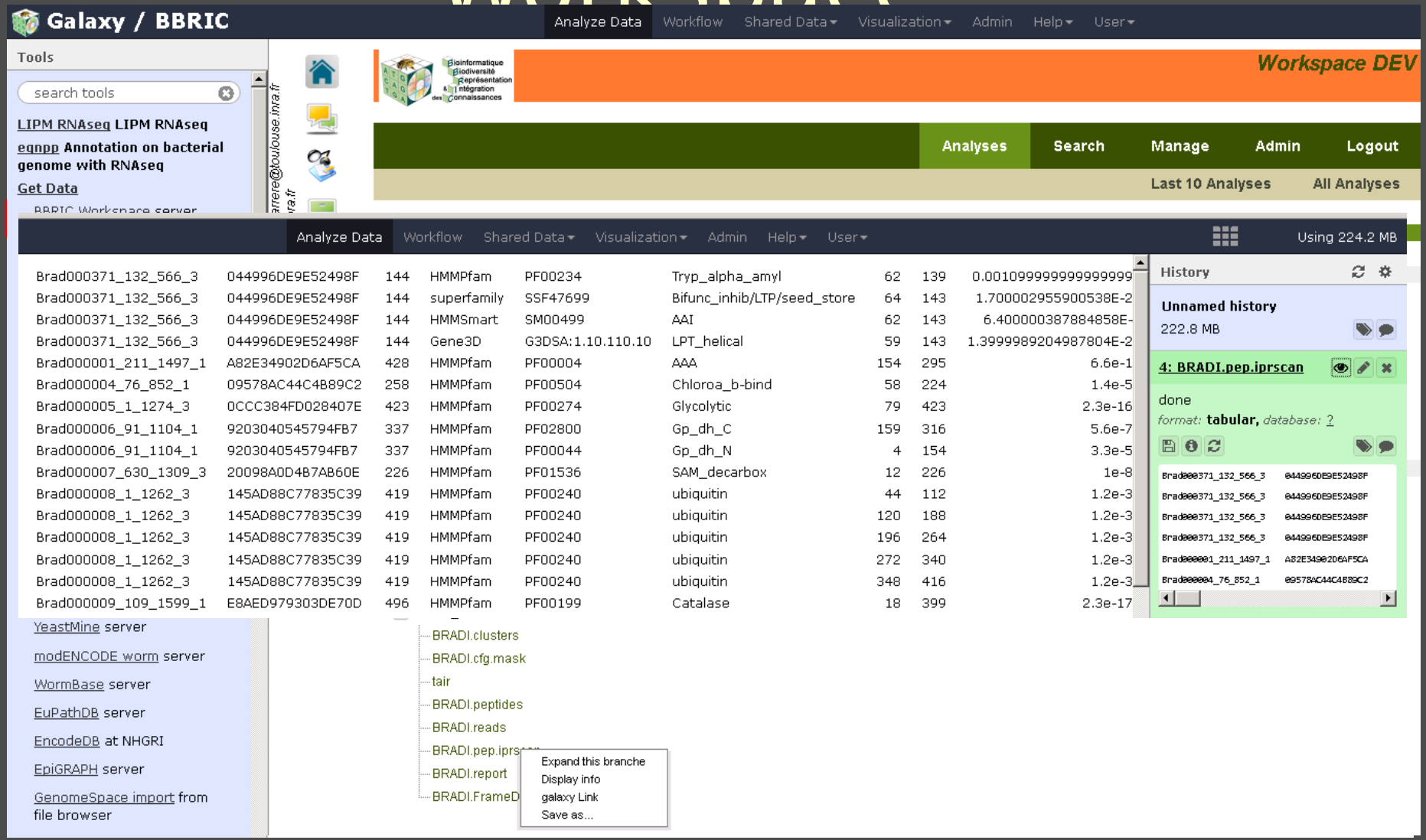
Interopérabilité avec Workspace



Interopérabilité avec Workspace



Interopérabilité avec Workspace



Atelier 29-30/10/2013

- Quinzaine de bioinformaticiens du CATI BBRIC
- 1/3 connaissait les aspects techniques de Galaxy

- Etape 1: déployer un outil simple dans Galaxy from scratch
  - Démystifier Galaxy
- Etape 2: utiliser Appli.pm
  - Se faciliter la vie
- Etape 3: Déployer un outil complexe (pipeline)
  - Se rendre utile

Atelier 29-30/10/2013

- Etape 1: déployer un outil simple dans Galaxy from scratch
  - Predotar, tmhmm, signalpHMM
  - Bilan des difficultés rencontrées
  - Gestion dynamique des formats de sortie
    - Paramètres conditionnels
    - Itérateurs / collection
    - Sélection des *datatypes*
    - Gestion des paths
    - Débogage

Mais tout le monde peut y arriver et tout le monde y est arrivé !

Atelier 29-30/10/2013

- Etape 2: Appli.pm

- Qu’est-ce ?
  - Un module Perl de la « lipmutils » permettant de décrire un programme.
  - Développé pour déployer des services Mobyle & BioMOBY étendu à Galaxy

- Pourquoi faire?
  - décrire une et une seule fois son programme
  - générer autant d’usage que d’outils (CLI, Mobyle, Galaxy, JSON, **next ?**)

- Sous quelles contraintes?
  - Perl
  - Connaitre les datatypes des interfaces (mobyle ≠ biomoby ≠ galaxy ≠ GetOptLong ≠ EDAM)

Atelier 29-30/10/2013

- Etape 2: Appli.pm

- Quel cas d'utilisation ?
  - un programme à ajouter dans Galaxy
  - ce programme a une interface en ligne de commande
  - j'écris un wrapper Perl utilisant Appli pour générer son usage Galaxy

- Résultat
  - Permet d’obtenir rapidement un outil fonctionnel
  - Permet de générer facilement une trame du XML Galaxy
    - ➢ *Gère les collections, mais pas les paramètres conditionnels*

Atelier 29-30/10/2013

- Etape3: Déployer un outil complexe (pipeline)
- EuGène-P
  - Pipeline d'annotation de génomes procaryotes
    - Intègre des données de similarité (Blastx)
    - Intègre des données de transcriptomique (RNA-seq)
    - Intègre des résultats de prédicteurs *ab-initio* (Prodigal, tRNAScan-SE,…)
  - Le wrapper devait
    - prendre en entrée X banques de RNA-seq, un génome nu
    - générer l'environnement (fichiers de conf, arborescence pour l'exécution du pipeline
    - produire un fichier d'annotation au format GFF3

OK pour la majorité des binômes, quelque soit le langage et la méthode utilisée

Bilan de l'atelier

\+ Montée en compétence commune

\+ Prise en main de l'outil

\- Crash des *handlers*

En cause: l'écriture de caractères spéciaux sur stderr !

\- Difficultés logs

Comment trouver les commandes exécutées ?

~ Les types de données, les ontologies

~ Nécessité de redémarrer régulièrement

Problèmes de cache

Bonnes pratiques

- Toolshed
  - Partager les outils
- Appli.pm (ou autre ?)
  - Si demain, l'outil à la mode n'est plus Galaxy
- Déployer des outils de haut niveau
  - Utilisables
  - Utiles
- Fichiers compressés
  - Obligatoire dans un environnement de production

Merci