Archive

Structurer et préserver données et métadonnées

Ludovic LEGRAND, Ludovic COTTRET, Emmanuel COURCELLE, Erika SALLET, Sébastien CARRERE et Jérôme GOUZY
Laboratoire des Interactions Plantes Micro-organismes (LIPM), INRA-CNRS UMR 441/2594, F-31320 Castanet Tolosan, France

Contexte

Aujourd'hui, la pérennité des données brutes issues des sciences de la vie est assurée de façon très inégale puisque placée sous la responsabilité des laboratoires/instituts qui les produisent. Elle est pourtant indispensable pour assurer sur le long terme la traçabilité et l'exploitation de ces données aussi bien avec les méthodes et outils actuels que futurs. La diversité des données et des formats, l'évolution rapide des technologies à haut débit ainsi que la quantité croissante de données nécessite la mise en place d'un système à la fois **robuste et générique** qui permettra d'exploiter les données plusieurs années après leur production.

Pour répondre à ce besoin critique, nous proposons Archive, un système d'information dont les objectifs sont multiples :

- **conservation** pérenne des données expérimentales brutes
- **flexibilité** pour gérer tout type de données expérimentales
- **partage sécurisé** des données
- facilité d'**accès aux données** par une interface web et par un accès programmatique

Fonctionnalités

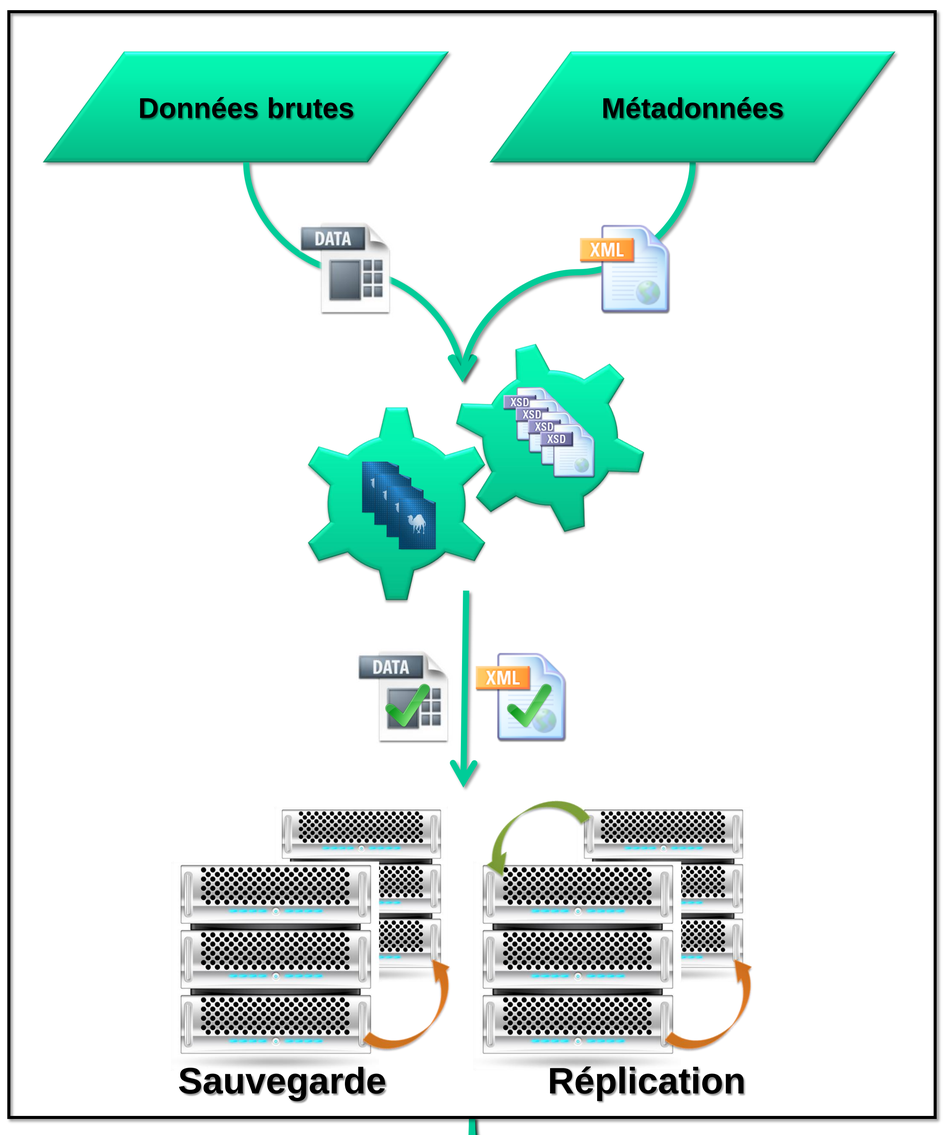
Le **transfert les données** peut se faire par HTTPS lors de la saisie des métadonnées ou en amont par SFTP pour les volumes de données supérieurs à 2 Go.

Après la saisie ou lors d'une édition, une **phase de validation asynchrone** contrôle l'ensemble des informations :

- **validation du fichier XML** de métadonnées par un schéma XSD
- **validation des données** par un script spécifique au type de données
- contrôle de l'**intégrité des fichiers** après copie dans l'Archive.

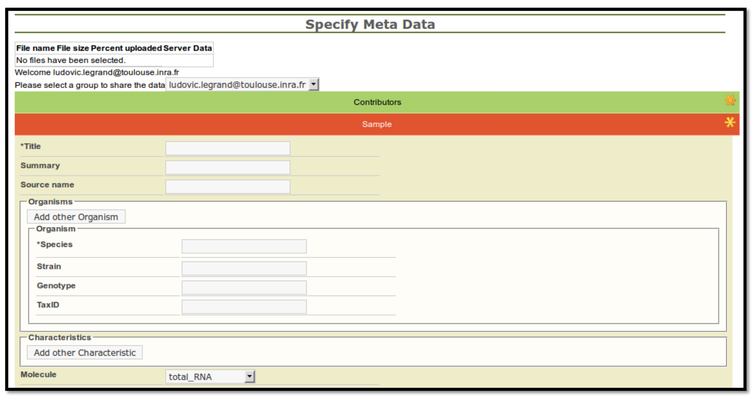
Le **stockage des données** sur le long terme nécessite la mise en place de mécanismes permettant d'assurer la conservation et/ou une reprise d'activité rapide en cas d'incident.

Archive supporte le stockage sur un **système de fichiers** classique, mais aussi sur une **grille de données** iRODS[3] permettant leur réplication entre plusieurs sites.



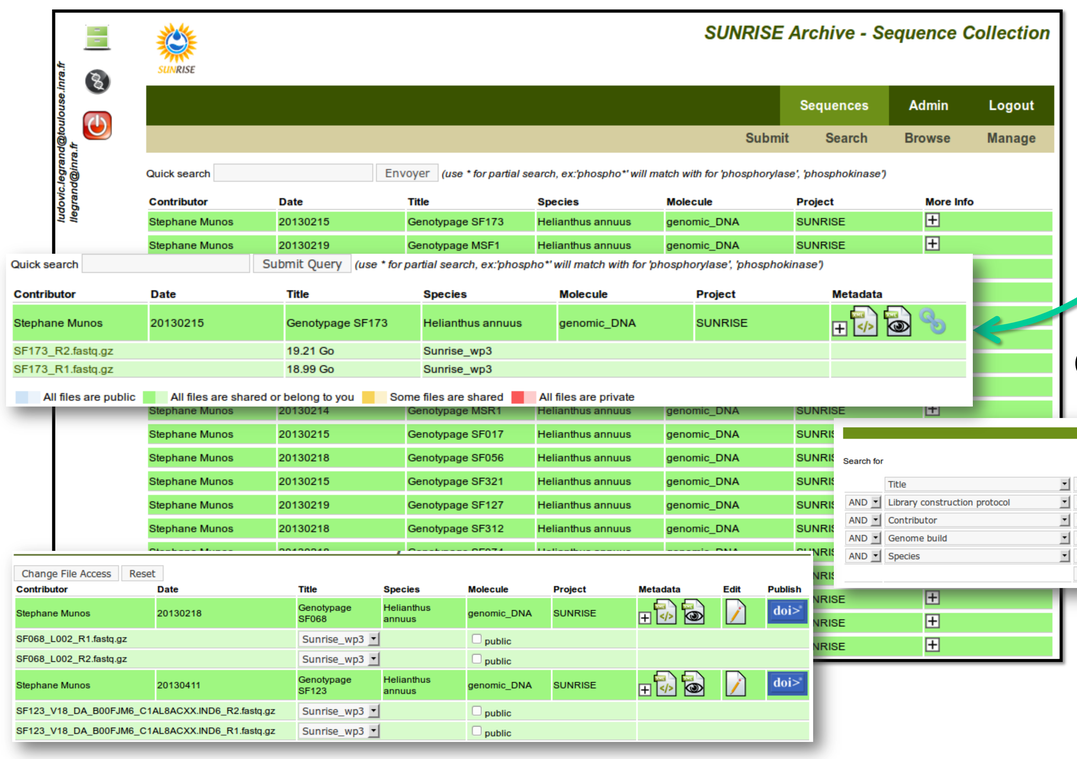
La **saisie et l'édition des métadonnées** se font *via* un formulaire web généré à partir d'un fichier XML de description **spécifique à chaque type de données**.
Les contraintes formulées dans le XML (cardinalités, énumérations…) sont utilisées pour valider les champs du formulaire.

Les métadonnées sont conservées dans un fichier XML similaire au XML de description.

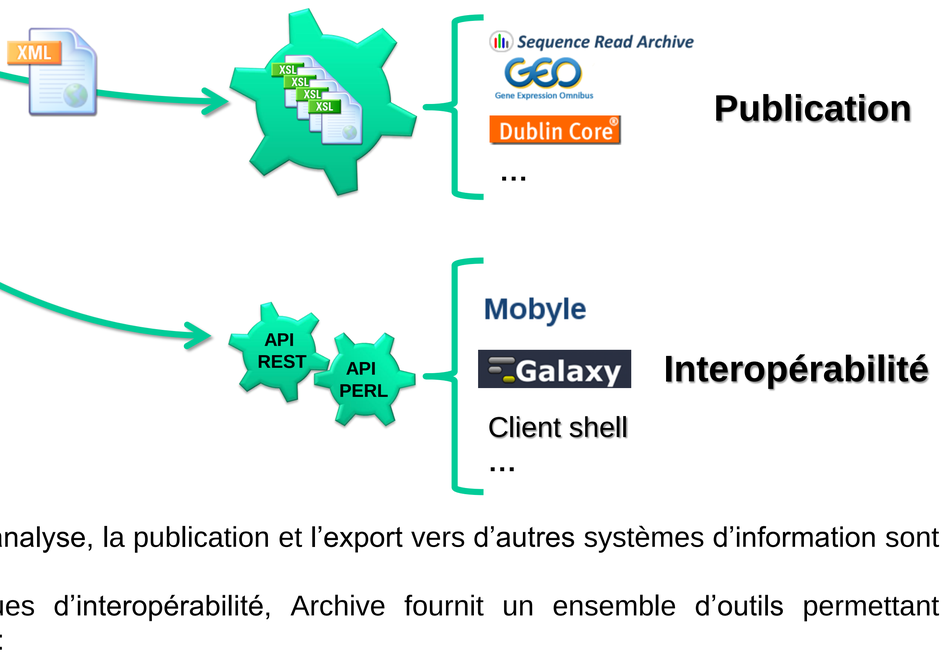


La **sécurisation** de l'accès aux données est assurée par trois composantes :

- la **Fédération d'identité Education-Recherche**[1] et le logiciel Shibboleth[2] assure l'authentification des utilisateurs
- la **base de données d'Archive** permet au propriétaire des données de contrôler l'accès et le partage des données privées par un système de groupe
- des **protocoles sécurisés** pour tous les échanges



Consultation



L'**interface web** permet de visualiser les métadonnées, de télécharger les données et d'effectuer des recherches sur les métadonnées grâce au moteur d'indexation Lucene[4].

Les responsables des données peuvent éditer les métadonnées, gérer les droits d'accès ou d'édition et publier les données.

Dans le cycle de vie des données, l'analyse, la publication et l'export vers d'autres systèmes d'information sont incontournables.
Pour répondre à ces problématiques d'interopérabilité, Archive fournit un ensemble d'outils permettant d'exploiter et de publier les données :

- **lien permanent** pour chaque entrée
- l'export des métadonnées vers le format **Dublin Core** pour la mise en place d'un **Digital Object Identifier**
- la transformation **XLST** des métadonnées vers d'autres formats comme SRA ou GEO pour les séquences
- une **API REST** et une **API Perl** permettent de développer des programmes pour interroger et récupérer les entrées dans d'autres systèmes comme Galaxy, Mobyle ou une arborescence Unix. L'API Perl permet de faire des requêtes sur un **réseau d'instances d'Archive**.

Mise en place

Le Laboratoire des Interactions Plantes-Microorganismes utilise Archive en production depuis février 2013 pour les données de séquençage. A l'heure actuelle, nous gérons 3 To de données ce qui représente 550 fichiers et 44 espèces.

Perspectives

Intégration de **nouveaux types de données**
- phénotypique
- métabolomique

Réplication entre plusieurs sites

Intégration avec des **moteurs de workflow** via l'API
- Galaxy
- Mobyle

Références

1. Fédération d'identité Education-Recherche. https://services.renater.fr/federation/index
2. Shibboleth. http://shibboleth.net/
3. iRODS. https://www.irods.org/index.php
4. Lucene. http://lucene.apache.org/core/

Code source disponible sous licence CeCILL à cette adresse :
http://lipm-svn.toulouse.inra.fr/svn/inra_archive