Le portail Data Inra et les services associés

Fanny Dedet[1], Anne-Sophie Martel[2], Sylvie Cocaud[3], Esther Dzalé-Yeumo[4],

Résumé. Dans un contexte de « science ouverte » (open science) promu par la Commission européenne[5] et par l'Etat (Plan national pour la science ouverte)[6], l'Inra intègre la révolution numérique dans ses orientations stratégiques[7] et s'emploie à mettre à disposition des scientifiques des outils et des services en faveur de la gestion et du partage des données. L'ensemble de ces outils et services sont décrits et accessibles depuis le site web Datapartage[8]. Les enjeux de la bonne gestion et du partage des données sont scientifiques, stratégiques, juridiques, déontologiques et techniques. Ouvert en mars 2018, le portail Data Inra[9] est un des outils mis à disposition par l'établissement pour accompagner les scientifiques dans la gestion et le partage des données. Il vise à faciliter la gestion et le partage des données de l'établissement dans le respect de la loi pour une République Numérique et des principes FAIR (Findable, Accessible, Interoperable, Reusable). Le portail Data Inra n'a pas vocation à centraliser la gestion de toutes les données produites par l'établissement qui, par leur hétérogénéité nécessitent différents outils et règles de gestion. Il a été conçu pour s'intégrer avec les autres systèmes d'information de l'établissement afin de leur permettre de partager facilement des données et de favoriser/amplifier leur visibilité, leur accès pérenne et leur citation.

Mots clés : portail Data Inra, entrepôt de données institutionnel, API DOI, API REST, partage de données.

Introduction

Comme rappelé sur le site du Plan national pour la science ouverte[10], la diffusion sans entrave des données de la recherche contribue à « faire sortir la recherche financée sur fonds publics du cadre confiné des bases de données fermées » et réduit la duplication des efforts pour collecter, créer, transférer et réutiliser du matériel scientifique. Une telle diffusion des données scientifiques suppose que celles-ci soient gérées et partagées d'une manière qui favorise leur découverte, leur accès pérenne et leur réutilisation. Par ailleurs, la citation des données et la reconnaissance des acteurs qui contribuent à les collecter, créer, gérer et partager constituent un enjeu majeur. Le portail Data Inra offre aux scientifiques de l'établissement et à leurs partenaires un outil et un accompagnement pour gérer et partager des données scientifiques (liées ou non à des publications scientifiques) dans le respect des principes FAIR (Findable, Accessible, Interoperable, Reusable), tout en favorisant leur citation et la reconnaissance

[1] *UAR DSI-UA Direction du Système d'Information-Unité d'Appui, Inra, Paris, France*
fanny.dedet@inra.fr

[2] *UMR EPIA Epidémiologie des maladies animales et zoonotiques, Inra, VetAgro Sup, 63122, Saint-Genès-Champanelle, France*
anne-sophie.martel@inra.fr

[3] *UAR SDAR Grand Est-Nancy/Colmar Services déconcentrés d'appui à la recherche Grand Est-Nancy/Colmar, Inra, Champenoux, France*
sylvie.cocaud@inra.fr

[4] *UAR DIST Délégation Information scientifique et technique, Inra, Versailles, France*
Email : esther.dzale-yeumo@inra.fr

[5] https://ec.europa.eu/digital-single-market/
[6] https://bit.ly/2MU8m9n
[7] http://2025.inra.fr/openscience/
[8] https://datapartage.inra.fr
[9] https://data.inra.fr
[10] https://bit.ly/2MU8m9n

de leurs auteurs et contributeurs. Il permet également de répondre aux obligations réglementaires françaises en matière d'ouverture des données.

Le portail Data Inra repose principalement sur la technologie Dataverse, logiciel open source, développé depuis une dizaine d'années par l'Institut de science sociale quantitative (IQSS) de Harvard et de plus en plus adopté par la communauté scientifique.

Dans la suite de l'article, nous présentons le portail Data Inra, ses fonctions et principaux cas d'usage, l'organisation des données, son architecture et son positionnement dans l'écosystème des systèmes d'information internes et externes gérant des données, puis concluons en indiquant quelques perspectives.

Fonctions et principaux cas d'usage du portail Data Inra

Le portail Data Inra propose deux fonctions principales. La première est une **fonction entrepôt** pour stocker, organiser et documenter les données dès leur collecte. Cette fonction offre plusieurs possibilités aux scientifiques, aux responsables de collectifs ou de projets en fonction des cas d'usage. La deuxième fonction offerte par Data Inra est une **fonction annuaire** qui favorise la découverte, l'accès et la citation des données dès leur publication. Le portail Data Inra permet de maîtriser le moment et les conditions de diffusion des données. Une fois les données rendues publiques, Data Inra contribue à favoriser leur découverte, leur accès fiable et pérenne ainsi que leur citation via un DOI (Digital Object Identifier). Le portail Data Inra est accessible et utilisable (dans ses fonctions entrepôt et annuaire) par les humains via une interface Web et par les agents logiciels via des API (Application Programming Interface). La **Figure 1** et la **Figure 2** ci-dessous présentent les principaux cas d'usage du portail Data Inra et son positionnement dans son environnement.

Figure 1. *Principaux cas d'usages du portail Data Inra.*

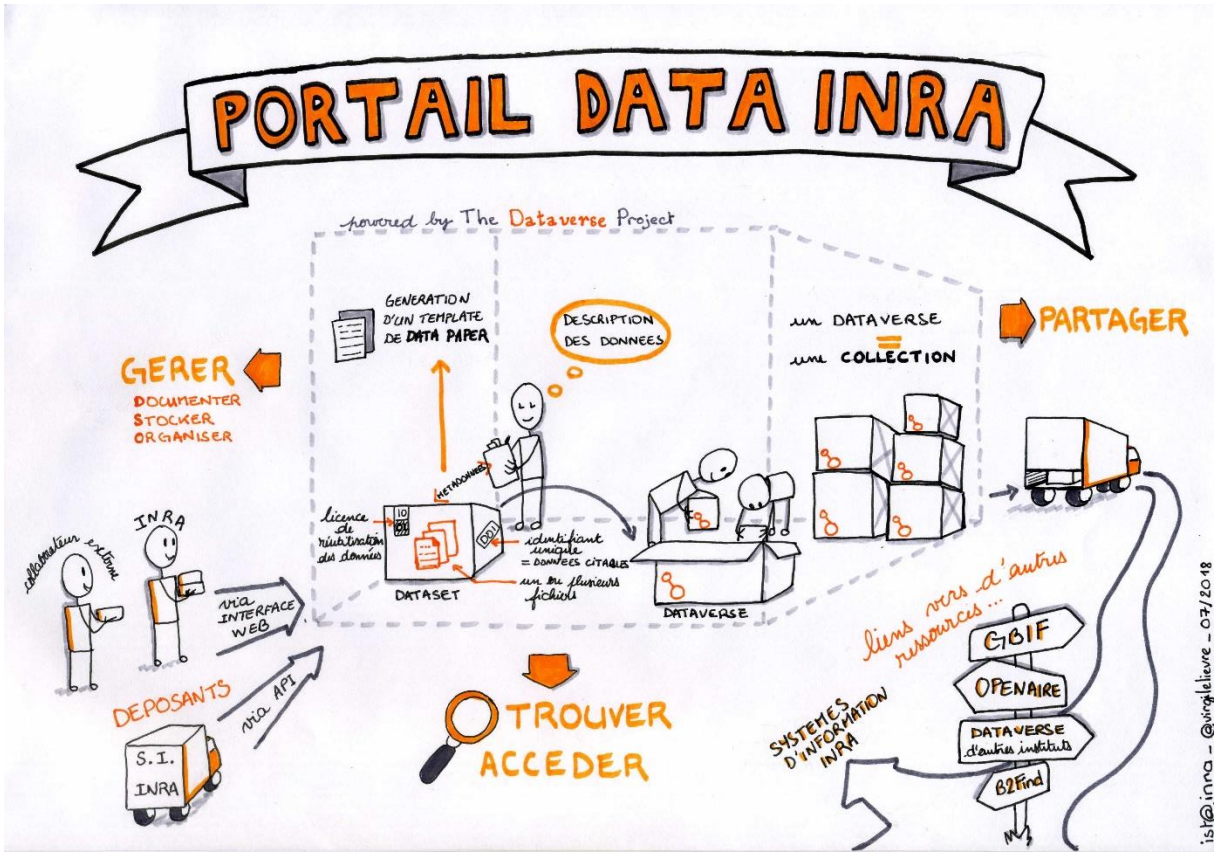

Figure 2. *Présentation générale de Data Inra et de son environnement.*

Fonction entrepôt : organisation des données dans Data Inra

Un **dataset** ou jeu de données peut être composé d'un ou plusieurs fichiers de données ; il est décrit par un ensemble cohérent de métadonnées, est identifiable par un DOI et citable comme un objet unique. Les fichiers de données peuvent être déposés dans le portail ou dans un autre entrepôt (dans ce dernier cas, le déposant doit fournir un lien permettant d'accéder aux données). Chaque jeu de données est obligatoirement accompagné de termes d'usage. En l'absence d'indications contraires, nous recommandons l'utilisation de la licence ouverte Etalab promue par le décret d'application de la loi pour une République numérique. Enfin, un jeu de données peut faire l'objet de plusieurs versions reflétant les changements dans les métadonnées et/ou les données.

Un **dataverse** est une collection contenant des datasets ou d'autres dataverses. Un dataverse peut être créé pour organiser et gérer les données d'un projet spécifique ou d'un collectif (Département, Unité, etc.). À travers la création d'un dataverse dédié, le projet ou le collectif concerné peut spécifier les métadonnées nécessaires et utiles pour décrire et comprendre les données qui s'y trouvent. Le dataverse dédié permet également au collectif d'avoir une meilleure maîtrise de la qualité des métadonnées et des données publiées, de spécifier un workflow de publication, et de gérer les droits relatifs à la collection. Chaque dataverse est géré par un administrateur désigné par le projet ou le collectif concerné. Dans le cadre d'un projet de recherche impliquant des partenaires externes, l'administrateur du dataverse dédié au projet peut autoriser les partenaires externes à y déposer, documenter, modifier et publier des données.

Fonction entrepôt : déposer, documenter, mettre à jour et publier des données dans Data Inra

Tout agent Inra ou agent travaillant dans une Unité Inra peut se connecter à Data Inra grâce à ses identifiants de connexion LDAP pour déposer, décrire, mettre à jour et publier des données. Il peut déposer les données dans une des collections (dataverses thématiques) qui se trouvent directement sous de la racine du portail (Experimental - Observation - Simulation Dataverse, Omics Dataverse, Surveys & Texts Dataverse, Genetic Resources Dataverse). Lors du dépôt, l'utilisateur doit fournir un minimum de métadonnées (titre, auteurs, description) permettant notamment de générer une citation pour le jeu de données concerné. Cependant, il est recommandé de renseigner un maximum de métadonnées par la suite pour favoriser davantage la découverte et la bonne réutilisation des données.

Le portail Data Inra attribue un DOI au jeu de données (voir l'encart « Focus sur le DOI ») et génère la citation associée dès le dépôt. Ce DOI est activé dès la publication du jeu de données et ne change pas en cas de modification mineure de version. Tout changement dans les métadonnées ou les données d'un dataset entraîne la création d'une nouvelle version, les changements majeurs entraînant en plus une mise à jour de la citation associée. Un jeu de données publié ne peut pas être supprimé, mais peut être « dé-publié » : il n'apparaît plus dans les résultats de recherche et le DOI associé mène vers une page qui contient uniquement la citation associée et la raison pour laquelle le dataset n'est plus accessible. Le processus de dépôt et de publication des données dans Data Inra est résumé dans la **Figure 3** ci-dessous.

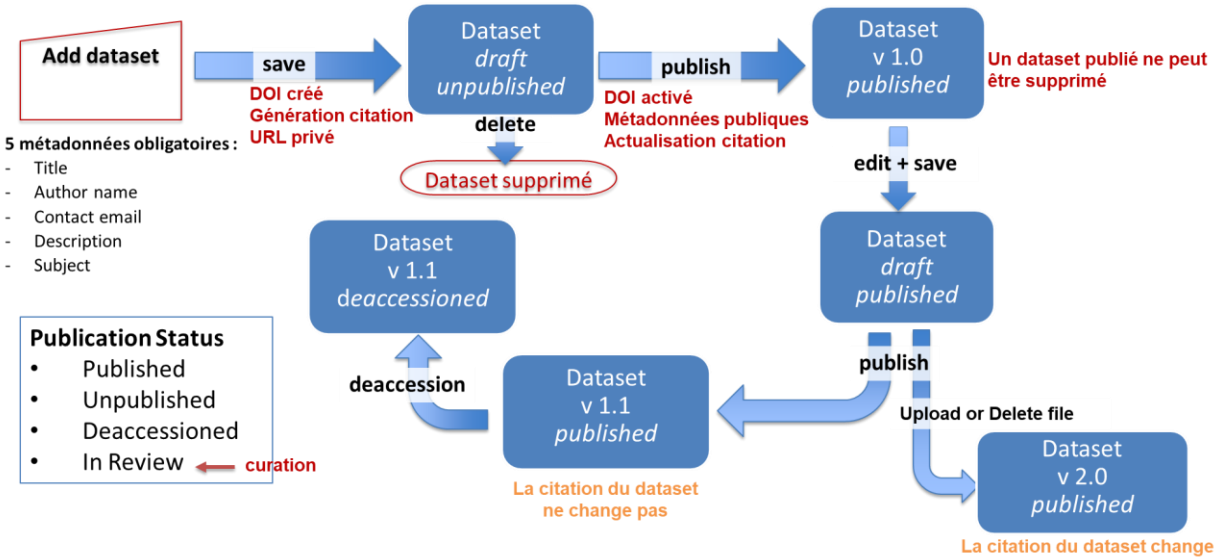

Figure 3. *Processus de dépôt et de publication d'un jeu de données dans Data Inra.*

L'utilisation de la fonction entrepôt est particulièrement recommandée dans les situations suivantes (voir **Figure 4** ci-dessous).

- Il n'existe pas d'entrepôt thématique recommandé dans le domaine concerné par le jeu de données : en effet, l'Inra recommande aux scientifiques de privilégier l'utilisation des entrepôts thématiques recommandés dans leurs domaines.
- Le jeu de données sous-tend un article scientifique et l'éditeur n'impose pas un entrepôt particulier

Par ailleurs, si le jeu de données contient des données sensibles nécessitant un niveau de sécurisation particulier, il est recommandé de contacter les administrateurs de Data Inra avant tout dépôt.

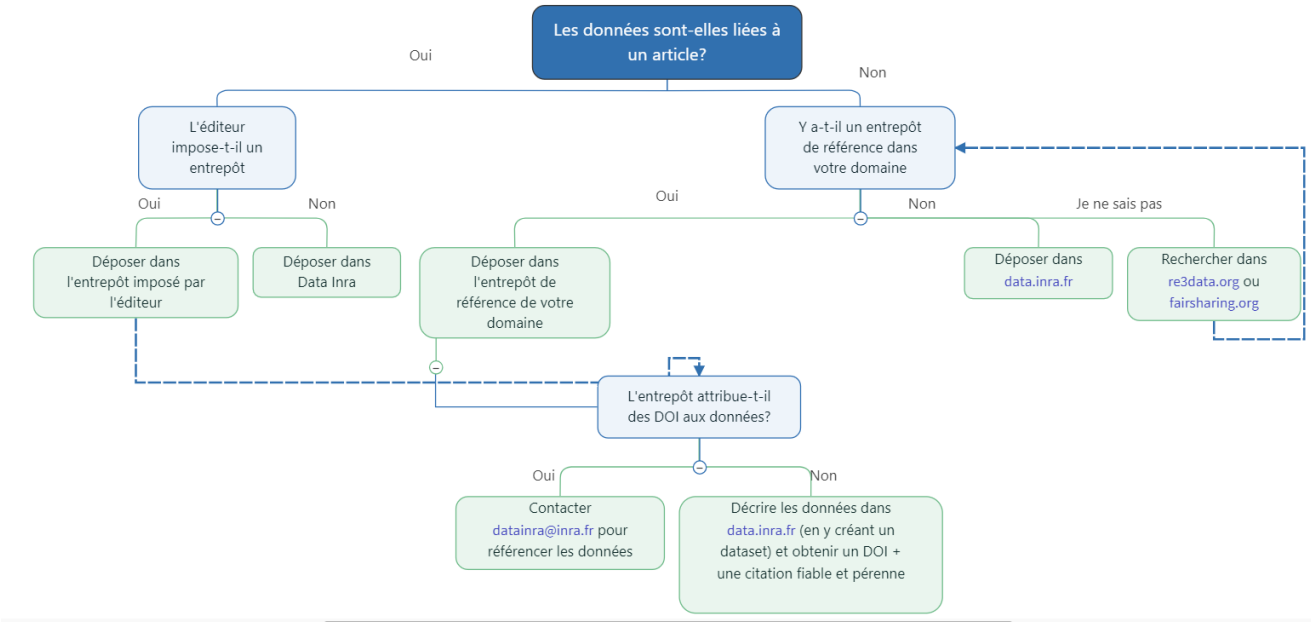

Figure 4. *Choix d'un entrepôt.*

Fonction annuaire : décrire, identifier, rechercher, accéder, citer

La fonction annuaire de Data Inra permet de décrire et d'identifier les données de manière unique et pérenne afin de favoriser la recherche, la découverte, l'accès et la citation des données de l'Inra dès leur publication, qu'elles soient déposées et gérées au sein de Data Inra ou ailleurs (voir **Figure 4** ci-dessus).

Décrire les données : Le portail Data Inra permet de décrire les données suivant des métadonnées compatibles avec des standards internationaux tels que Dublin Core DCMI[11], DDI Lite[12], DDI 2.5 Codebook[13], DataCite 3.1, ISA-Tab[14], iso 19115[15] et la norme INSPIRE Spatial data set[16]. Certains champs sont adossés à des vocabulaires contrôlés tels que ISO 3166-1[17] pour le champ Country, OBI Ontology et NCBI Taxonomy pour les organismes. Les métadonnées sont utiles pour la recherche, la découverte, et la bonne réutilisation des données, d'où l'importance d'en renseigner un maximum avec des informations de qualité (pertinentes, exactes, précises, etc.). De plus, le portail Data Inra permet de générer pour un ou plusieurs jeux de données un modèle de data paper pré-rempli avec les métadonnées de ce(s) jeu(x) de données, facilitant ainsi la création de ce nouveau type de publication.

Identifier et citer les données : Data Inra attribue un DOI à chaque jeu de données et génère, dès sa création, une citation associée (que le déposant y dépose les fichiers associés ou indique un lien d'accès vers un autre entrepôt). Le DOI et la citation associée peuvent être utilisés par les auteurs pour citer le jeu de données dès cet instant. Le DOI est activé dès la publication du jeu de données. Une citation fiable et pérenne des données permet une meilleure reconnaissance des personnes ou entités qui ont contribué à les collecter, créer, gérer, et partager. Dans le cas où le jeu de données est déposé dans un entrepôt (autre que Data Inra) qui attribue des DOI, nous

[11] http://dublincore.org/documents/dcmi-terms/
[12] http://www.ddialliance.org/specification/ddi2.1/lite/index.html
[13] http://www.ddialliance.org/
[14] http://isa-tools.org/format/specification/
[15] https://www.iso.org/fr/standard/53798.html
[16] https://inspire.ec.europa.eu/data-specifications/2892
[17] http://en.wikipedia.org/wiki/ISO_3166-1

recommandons aux auteurs de mentionner l'INRA dans les contributeurs et de se rapprocher des administrateurs de Data Inra pour envisager le moissonnage des métadonnées associées en évitant de créer une deuxième DOI pour ce jeu de données,

Rechercher et découvrir les données : tout internaute peut rechercher des données et, selon ses droits, télécharger les fichiers associés. Seules les données publiées peuvent être recherchées par tous les internautes, les données non encore publiées peuvent l'être uniquement par ceux qui ont des droits suffisants. Pour certains types de données (données tabulées, données géospatiales), le portail Data Inra propose des outils complémentaires pour les explorer (TwoRavens, WorldMap, et bientôt Data Explorer).

Le cas particulier des logiciels

Conformément à la Loi pour une République numérique du 7 octobre 2016, les codes sources sont des documents administratifs communicables et réutilisables, sous réserve des exceptions légales. L'Inra préconise que la « *production de codes-sources puisse être visible, partagée, collaborative et de qualité* » et recommande pour cela l'usage d'une forge logicielle institutionnelle, en l'occurrence SourceSup (https://sourcesup.cru.fr/). Pour que les logiciels soient citables de manière fiable et pérenne et plus faciles à trouver, il est utile de les décrire dans le portail Data Inra en indiquant le lien vers les codes dans SourceSup (voir **Figure 4**).

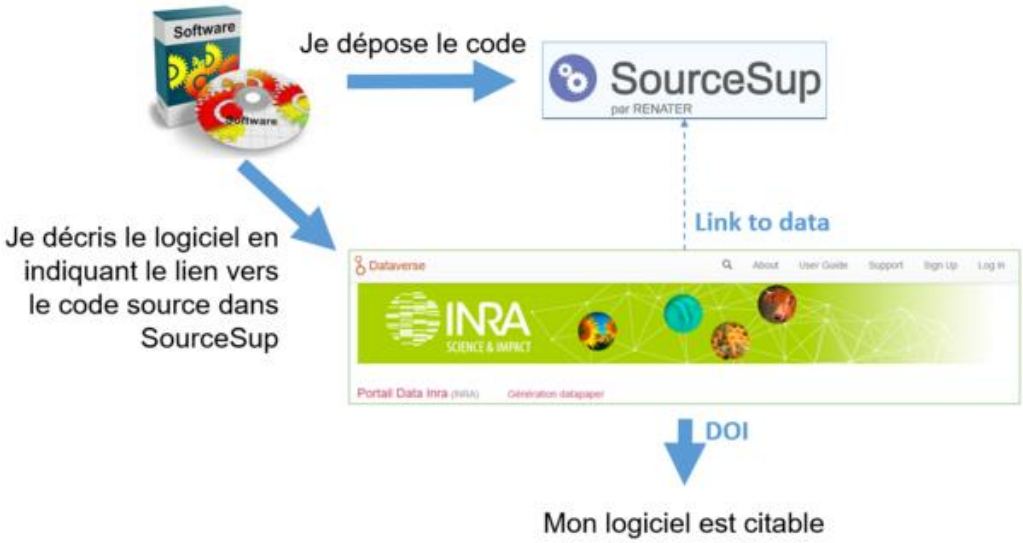

Figure 4. *SourceSup et Data Inra pour le partage de logiciels FAIR.*

Architecture et positionnement de Data Inra dans son environnement

Le portail Data Inra repose principalement sur l'utilisation du logiciel Dataverse et de DataCite pour les DOI, auxquels s'ajoutent progressivement des outils pour explorer ou analyser les données déposées dans le portail (exemple : Worldmap). Les composants techniques de Dataverse et leur déploiement actuel dans le cadre précis du portail Data Inra sont décrits dans la **Figure 5**.

Afin de permettre aux systèmes d'information (SI) qui gèrent des données d'utiliser Data Inra pour les partager et publier, et éviter ainsi de développer chacun une surcouche applicative, nous proposons deux solutions techniques. La première consiste à faire appel à une API et la deuxième consiste à exposer les métadonnées du système d'information via le protocole OAI-PMH pour que Data Inra les moissonne.

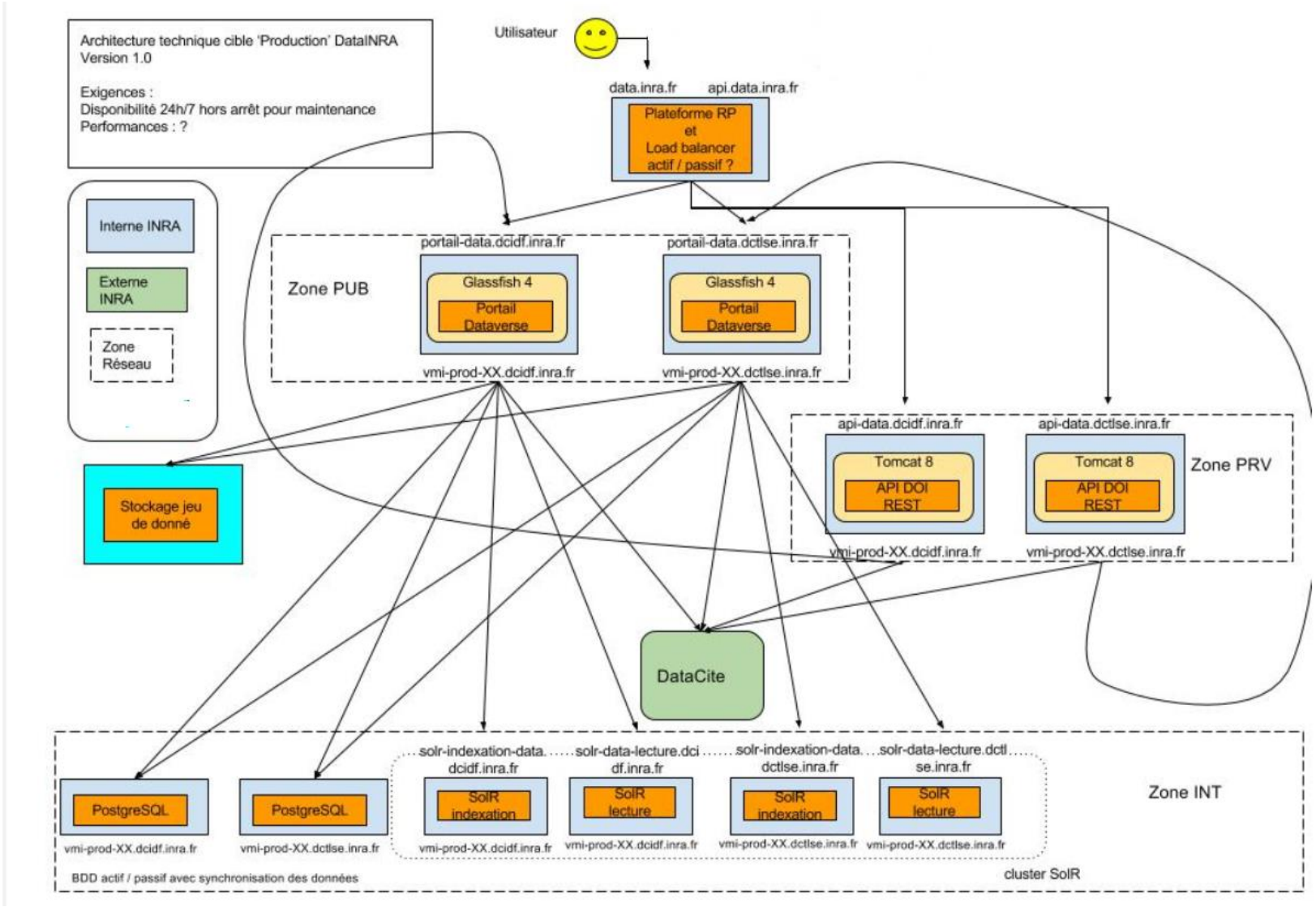

Figure 5. *Architecture de production Data Inra.*

Deux API sont disponibles : **l'API DOI**[18] qui permet au SI de demander un DOI pour un jeu de données qu'il veut publier. Cette API prend en entrée un fichier de métadonnées au standard DataCite[19] et renvoie un DOI. Les métadonnées sont utilisées pour créer une landing page dans le portail Data Inra vers laquelle pointe le DOI. La landing page contient le lien pour accéder aux données dans le SI concerné. Data Inra assure la pérennité de la landing page ainsi que celle du lien entre le DOI et la landing page. La landing page permet aux utilisateurs d'avoir des informations pour évaluer l'intérêt des données, en connaître les conditions d'utilisation et y accéder. Les données continuent d'être gérées par le SI d'origine. Cette première API est indiquée pour des SI qui souhaitent utiliser Data Inra pour publier des données sans les y déposer. Si au contraire le SI souhaite déposer les données dans Data Inra, il peut utiliser **l'API Sword de Dataverse**[20] qui permet l'envoi de données avec les métadonnées. **A terme, l'API DOI permettra de traiter les deux cas d'usage.** L'utilisation de l'API DOI nécessite un login et un mot de passe à demander auprès de doi@inra.fr. L'utilisation de l'API Sword de Dataverse nécessite un token qui peut être obtenu par tout utilisateur connecté à Data Inra qui dispose des droits suffisants.
L'environnement de Data Inra comprend aussi des entrepôts thématiques nationaux et internationaux. Il est rappelé que l'Inra recommande le dépôt des données dans les entrepôts thématiques à chaque fois que cela est possible. Si les données sont déposées et gérées dans un entrepôt thématique qui ne génère pas de DOI, elles peuvent en complément être décrites dans Data Inra en indiquant le lien vers l'entrepôt utilisé (voir **Figure 4** ci-dessus). Cela permet de bénéficier d'un DOI Inra pour citer et référencer les données concernées. Si l'entrepôt concerné octroie des DOI, il est recommandé de mentionner l'Inra dans les contributeurs et de se rapprocher des administrateurs de Data Inra pour le moissonnage des métadonnées associées.
Enfin, l'équipe de Data Inra travaille à ce que le portail et les données qu'il contient soient largement référencés dans les annuaires d'entrepôts et de données utilisés par la communauté scientifique tels que FAIRSharing, Re3Data, B2Find, OpenAire et DataCite. Il est d'ores et déjà référencé dans les annuaires d'entrepôts re3data.org[21] et FAIRshairing.org[22].

Conclusion et perspectives

Le portail Data Inra offre un service complet, non seulement pour gérer et partager les données, mais plus largement pour permettre leur publication en les rendant citables, documentées, et en leur garantissant une accessibilité pérenne afin, *in fine,* de faciliter leur découverte et leur réutilisation conformément à la démarche FAIR[23]. Il permet ainsi une meilleure lisibilité des données partagées par l'établissement mais aussi des données des collectifs (Département, Unité, etc.) ou des projets qui choisissent de disposer d'une collection (dataverse) dédiée dans le portail : il est possible de lister les données partagées par le collectif concerné. S'inscrivant dans la dynamique générale de l'Open Science, le portail Data Inra intégrera progressivement des outils pour permettre d'explorer et d'analyser les données qui y sont déposées. Enfin, Data Inra permet à toute personne travaillant à l'Inra de disposer d'un outil pour ouvrir les données issues de la recherche financée sur fonds publics (en application de la loi pour une République numérique), ou de rendre accessibles et citables des données soutenant une publication comme le demandent de plus en plus de revues scientifiques.

Le portail Data Inra continuera d'évoluer pour intégrer les développements de Dataverse, poursuivre son adaptation aux besoins spécifiques de l'Inra et élargir ses interactions avec d'autres systèmes d'informations Inra ou externes.

---

[18] https://www6.inra.fr/datapartage/Gerer/DOI/L-API-DOI-Inra et https://catalogue-api.intranet.inra.fr/store/apis/info?name=Datapartage-DOI-REST&version=v1&provider=admin
[19] http://schema.datacite.org/

[20] http://guides.dataverse.org/en/latest/api/sword.html
[21] REgistry of REsearch Data REposirtories : https://www.re3data.org/
[22] https://fairsharing.org/
[23] https://doi.org/10.1038/sdata.2016.18

Les évolutions de Data Inra à court et moyen terme concernent notamment le dépôt et le stockage de gros volumes de données (plus de 2 giga octets), l'intégration d'outils complémentaires pour explorer ou analyser les données, et l'amélioration de l'expérience utilisateur. En parallèle des développements informatiques, plusieurs actions d'information, de formation et d'accompagnement sont envisagées pour les utilisateurs. Enfin, le portail Data Inra vise à se conformer davantage aux bonnes pratiques et normes en matière de gestion et de partage des données (principes FAIR, certification) tout en continuant son ouverture vers l'extérieur.

Focus sur les DOI

Les produits de la recherche qui peuvent bénéficier d'un DOI sont divers : une publication[24], un jeu de données[25], un workflow scientifique[26], une ontologie[27], un logiciel[28] ou un objet physique, comme par exemple une accession génétique[29]…

Le DOI permet de mettre à disposition d'un lien pérenne vers une ressource identifiée, ce qui permet d'augmenter sa visibilité, de faciliter sa découverte, de fiabiliser sa citation et d'améliorer sa traçabilité.

Le DOI offre également la possibilité de relier les ressources entre elles, par exemple : les publications et les données sur lesquelles elles s'appuient, plusieurs versions d'un même jeu de données, ou encore un jeu de données avec ses sous-ensembles. Enfin, le DOI permet d'accéder à la version précise du jeu de données utilisée pour obtenir un résultat de recherche, favorisant ainsi la transparence et la reproductibilité de la science. La fiabilité et la pérennité de l'accès aux ressources sont garanties par l'engagement des Centres de données et organismes qui gèrent le DOI.

Le DOI est un type particulier de Persistent IDentifier (PID). D'autres types de PID existent tel que Handle pour des données ou ORCID[30] pour l'identification des personnes.

Dans l'attente d'un système d'identification global pour les structures (voir les travaux du groupe de travail orgID[31]), le DOI est parfois utilisé pour identifier des structures de recherche. Le projet THOR[32] vise à intégrer différents types de PID (DOI, ORCID, ORGID) pour que les contributeurs, leurs rôles, les organisations auxquelles ils sont affiliés et les ressources qu'ils produisent soient décrits de manière cohérente à travers les différents systèmes d'informations.

[24] Exemple : https://dx.doi.org/10.1001/jama.2016.3608
[25] Exemple : https://doi.org/10.15454/EORHL8
[26] Exemple : https://doi.org/10.15454/1.4811287270056958e12
[27] Exemple : https://doi.org/10.15454/1.469094224299365E12
[28] Exemple : https://doi.org/10.18130/V3/DQQZOG
[29] Exemple : https://doi.org/10.15454/1.493806880218088E12
https://doi.org/10.15454/TRBJTB
[30] https://orcid.org/
[31] https://orcid.org/content/organization-identifier-working-group
[32] https://project-thor.eu/thor-plans/