Partage des données de la recherche : état de l'art, étude d'opportunité IST

Windpouire Esther Dzale Yeumo, Dominique L'Hostis, Sylvie S. Cocaud, Pascal Aventurier, Diane Le Henaff, Veronique Batifol, Virginie Lelievre, Caroline Dandurand

Etat de l’art – Etude d’opportunité / IST

24/01/2014

État de l'art, étude d'opportunité : Data citation index, données liées aux publications, stratégies éditeurs, place de ProdInra…

Rappel de la commande

Périmètre

- Data journals : type d'articles, reviewing, stockage des données
- Mise en œuvre des DOI
- Dépôt centralisé des données publiées : Prodinra2.0 ou autre alternative
- Citation des données : conseils aux auteurs
- Data citation index : Compréhension du fonctionnement, conseils à donner
- Repérage des bonnes ou mauvaises pratiques des éditeurs
- Évaluation des nouveaux entrepôts de données externes (Dryad, Pangaea ….)

s de fonctionnement

Étude pour produire analyse des points du périmètre voire étude d'opportunité

Binômes à identifier en charge de chaque item

Échanges avec les axes données et méthodes-outils

Livrables - Jalons

- Chaque item fait l'objet d'un livrable. L'ensemble sera présenté au séminaire du 2 et 3 décembre 2013

Diapo n°47 / Bilan du Séminaire de lancement avril 2013

Plan du Wiki

PARTAGER SES DONNÉES

→ **Déposer ses données**

- Qu'est ce qu'un entrepôt de données
- Typologie des entrepôts
- Certification des entrepôts
- Choisir son entrepôt
- Place de Prodinra

→ **Décrire ses données**

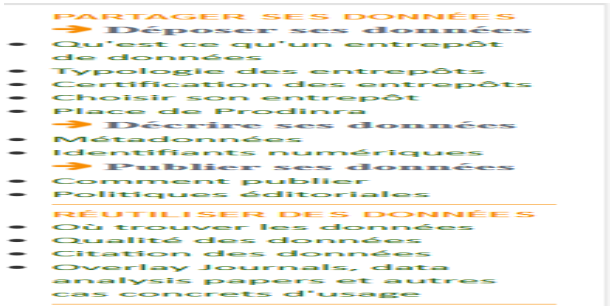
- Métadonnées
- Identifiants numériques

→ **Publier ses données**

- Comment publier
- Politiques éditoriales

RÉUTILISER DES DONNÉES

- Où trouver les données
- Qualité des données
- Citation des données
- Overlay Journals, data analysis papers et autres cas concrets d'usage



Accès (avec authentification LDAP)
http://wiki.inra.fr/wiki/donneesrechercheist/

Dans ce diaporama :
Un clic sur XWIKI permet d'aller à la page spécifique correspondante du wiki

Rendre visible ~ partager les données Inra

Disponibilité

7 critères ~ qualités à atteindre pour faciliter le partage et la réutilisation des données (*)

Visibilité

Interprétabilité

Dans le contexte Inra, quelles questions, quelles recommandations ?

Réutilisabilité

2 points de vue :

- celui du chercheur,
- celui de l'institut

Citabilité

Curation

Préservation

(*) Reilly, S., Schallier, W., Schrimpf, S., Smit, E., & Wilkinson, M. (2011). *Report on Integration of data and publications. 87 p.*
http://www.stm-assoc.org/2011_12_5_ODE_Report_On_Integration_of_Data_and_Publications.pdf

_1

Rendre visible ~ partager ses données

Producteur/Utilisateur de données

Où déposer ?

Dans quel entrepôt ?
Sur quels critères choisir un entrepôt ?
Quelles caractéristiques prendre en compte ?
Puis-je déposer dans un entrepôt Inra ? dans Prodinra ?

Comment décrire et citer ses données ?

Avec quelles métadonnées?
Quel identifiant utiliser ?, Comment l'obtenir ?, Qui peut m'aider ?
Bonnes pratiques, contraintes : comment les connaitre ?

Où et comment publier ses données?

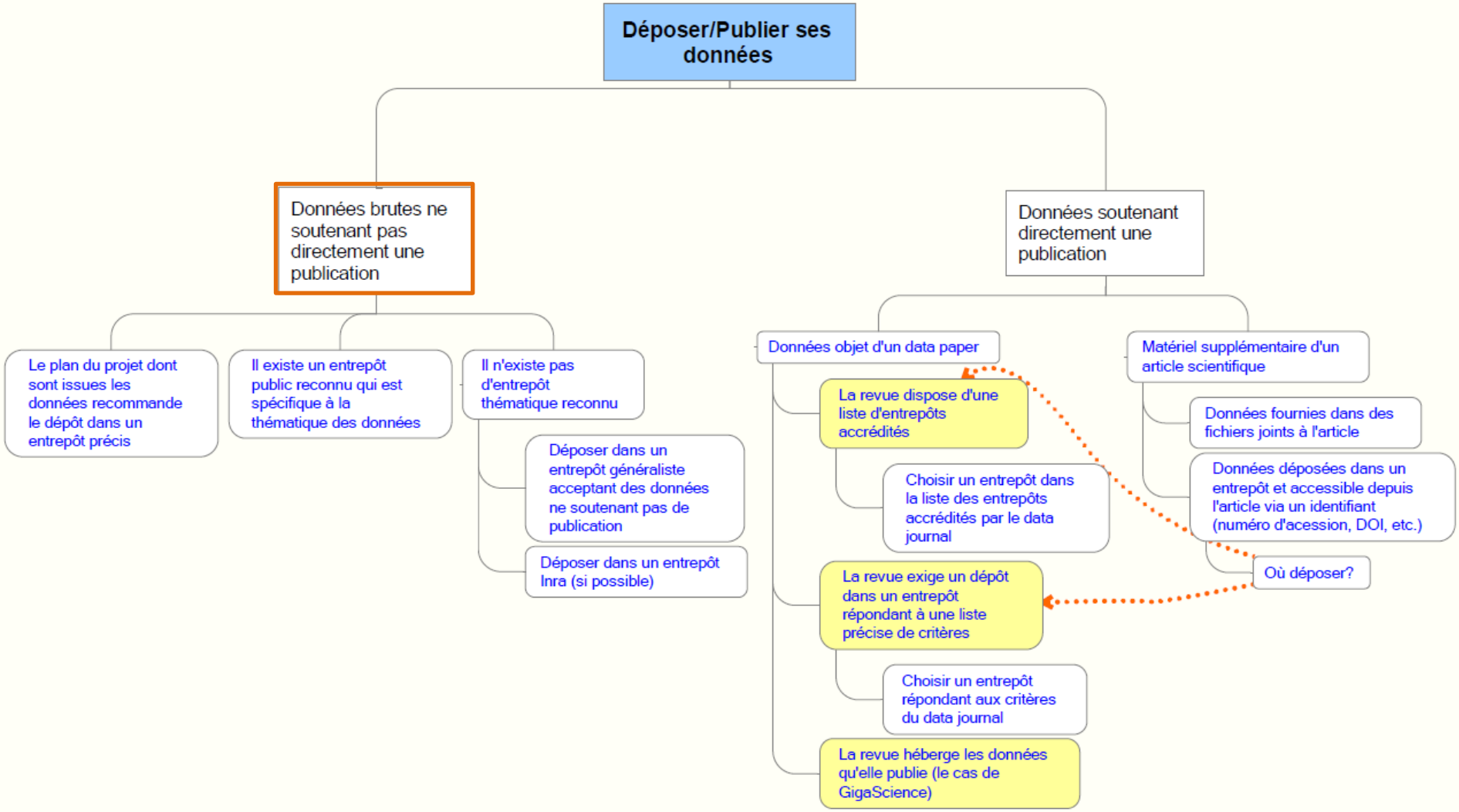
Quelles voies utiliser ?
Publier dans un Data Paper, quels avantages ?
Quels coûts ?

Où et comment trouver, réutiliser et citer des données ?

Où puis-trouver des données ?
Comment citer des données ?
Puis-je protéger mes droits ?
Comment suivre l'utilisation de mes données ?

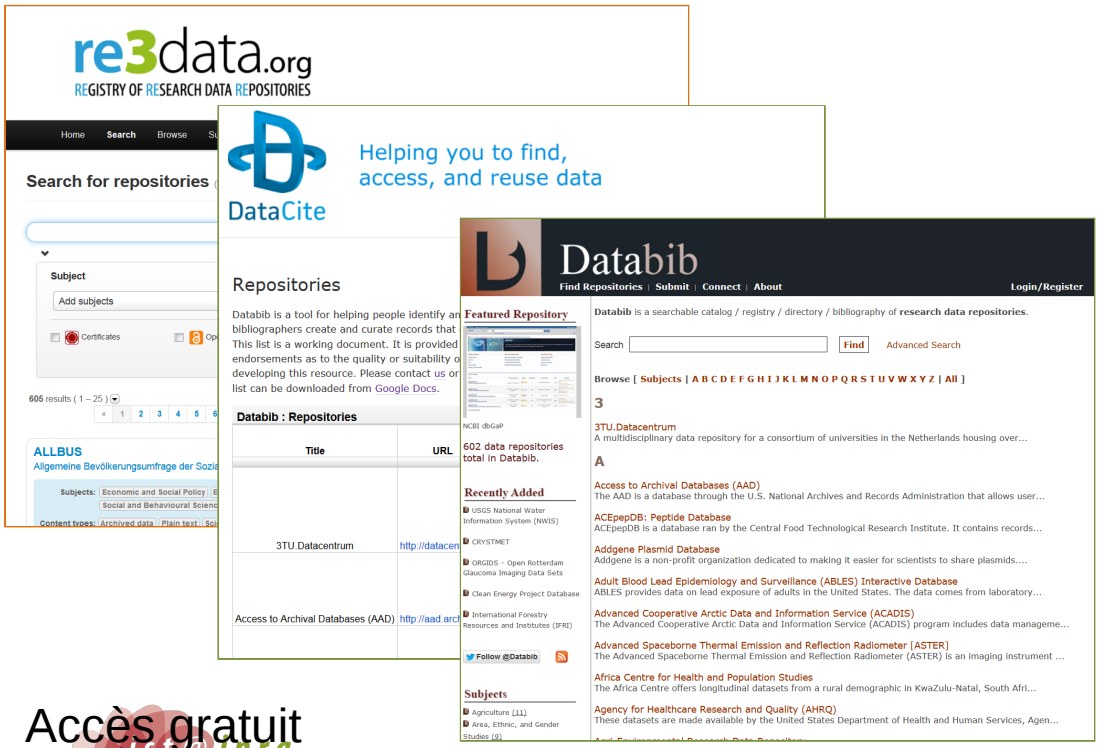
Où déposer ses données ?

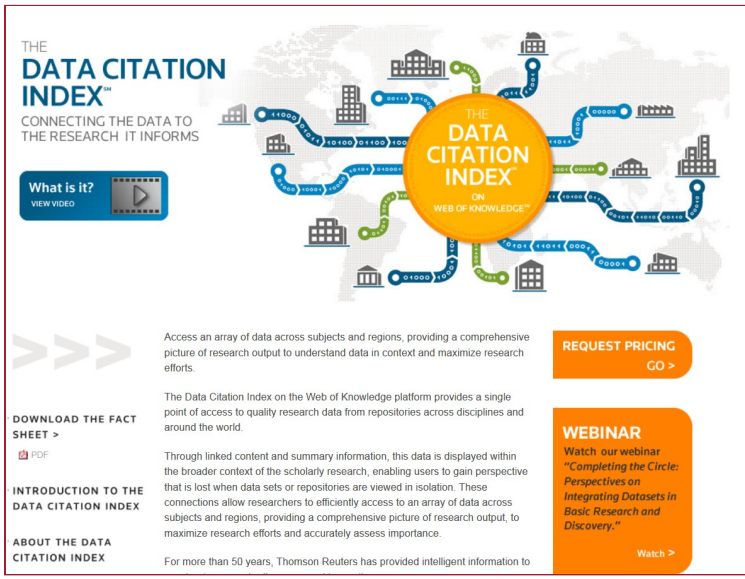
Où déposer ses données ?



Comment trouver un entrepôt ?

- Multiplication et hétérogénéité des entrepôts
- Conseil : utiliser des annuaires ou répertoires d'entrepôts





Accès gratuit

Accès payant

Comment choisir un entrepôt ?

- **Selon les recommandations d'un financeur, éditeur, de son organisme de rattachement**
- **les types d'entrepôts et leurs caractéristiques**
  - Discipline, modèle économique, type d'identification, licence, partenariat éditeurs, certification

re3data.org
REGISTRY OF RESEARCH DATA REPOSITORIES

Disciplinaire / Propriétaire de l'entrepôt	Institution publique	Organisation à but non lucratif	Organisation à but lucratif
Thématique	PANGAEA GenBank Knowledge Network for Biocomplexity (KNB)		
Pluridisciplinaire	Zenodo 3TU.Datacentrum	Dryad Datahub	Figshare

Exemple d'entrepôt

Propriétaire	Dryad
Thématique	Pluridisciplinaire (Biology and Biochemistry; Ecology and Environment; Health and Medicine)
Modèle économique	gratuit pour les chercheurs si le dataset <10GB, coût si >10GB. Sponsorship proposé aux institutions. Voir http://datadryad.org/pages/pricing
Formats de données supportés	Tout format (texte, tableurs, vidéos, photographies, code, y compris des archives compressées de fichiers multiples)
Plateforme utilisée	DSpace
Identifiant attribué au jeu de données	DOI
Licence des données publiées	CC0
Sécurité, persistance, préservation	Partenariat avec CLOCKSS qui garanti l'accès aux données indéfiniment
Accessibilité, réutilisabilité des données	Oui.
Compatibilité OAI-PMH pour l'ouverture des données	Oui
Partenariat avec des éditeurs	Liste des éditeurs membres de Dryad : http://datadryad.org/pages/membershipOverview#members Integrated journals table : http://datadryad.org/pages/integratedJournals
Liens externes	Données liées à la fois vers et depuis la publication correspondante. Lorsque c'est approprié, liées également vers et depuis des entrepôts spécialisés (e.g. GenBank).
Contenu de l'entrepôt	Uniquement des données associées à des publications scientifiques
Gestion des versions des fichiers	oui
Gestion/Curation des données	Curation : verification de l'intégrité des fichiers, verification de la complétude et de la qualité des métadonnées, conversion des fichiers dans des formats adaptés à la preservation. S'intègre au workflow de soumission du manuscript aux journaux partenaires. Accès reserve aux relecteurs Durant la revue par les pairs

Dryad Membership: Members

- American Association for the Advancement of Science *
- American Society of Nauralists *
- The American Genetic Association *
- British Ecological Society *
- BMJ Publishing Group, Ltd. *
- The Biological Journal of the Linnean Society (Linnean Society of London) *
- BioMed Central *
- Ecology Letters *
- Ecological Society of America *
- Elementa: Science of the Anthropocene
- European Society for Evolutionary Biology *
- Evolutionary Applications *
- The Genetics Society *
- German National Libary of Medicine
- HighWire - starting January 1, 2014
- Molecular Ecology *
- Molecular Ecology Resources *
- Molecular Phylogenetics and Evolution *
- Oikos *
- Oxford University Press *
- The Paleontological Society *
- Pensoft Publishers *
- PLOS *
- The Royal Society
- Society for Molecular Biology and Evolution *
- Society for the Study of Evolution *
- Society of Systematic Biologists *
- United States Fish and Wildlife Service *
- Wiley-Blackwell *

European Framework for Audit and C of Digital Repositories

L’europe se donne un cadre pour l'audit et la certification des entrepôts numériques, à trois niveaux :

certification de base

2014 DSA 2015

- accordée au entrepôts ayant obtenu le Data Seal of Approval (DSA) : Accréditation attribuée aux entrepôts numériques ayant mis en place des procédures d’assurance qualité garantissant un accès aisé et à long terme aux données stockées. Demande par procédure d’auto-évaluation

certification "étendue"

certification "formelle" réalisée par des

Comment décrire ses données ?

Décrire/documenter ses données

Fonctions des métadonnées

- recherche, découverte, localisation, accès, contextualisation, validation, citation.

Choix d'un standard de métadonnées

- standards communs tels que DataCite et le Dublin Core, ou spécifiques à un domaine scientifique (voir site du Digital Curation Center).

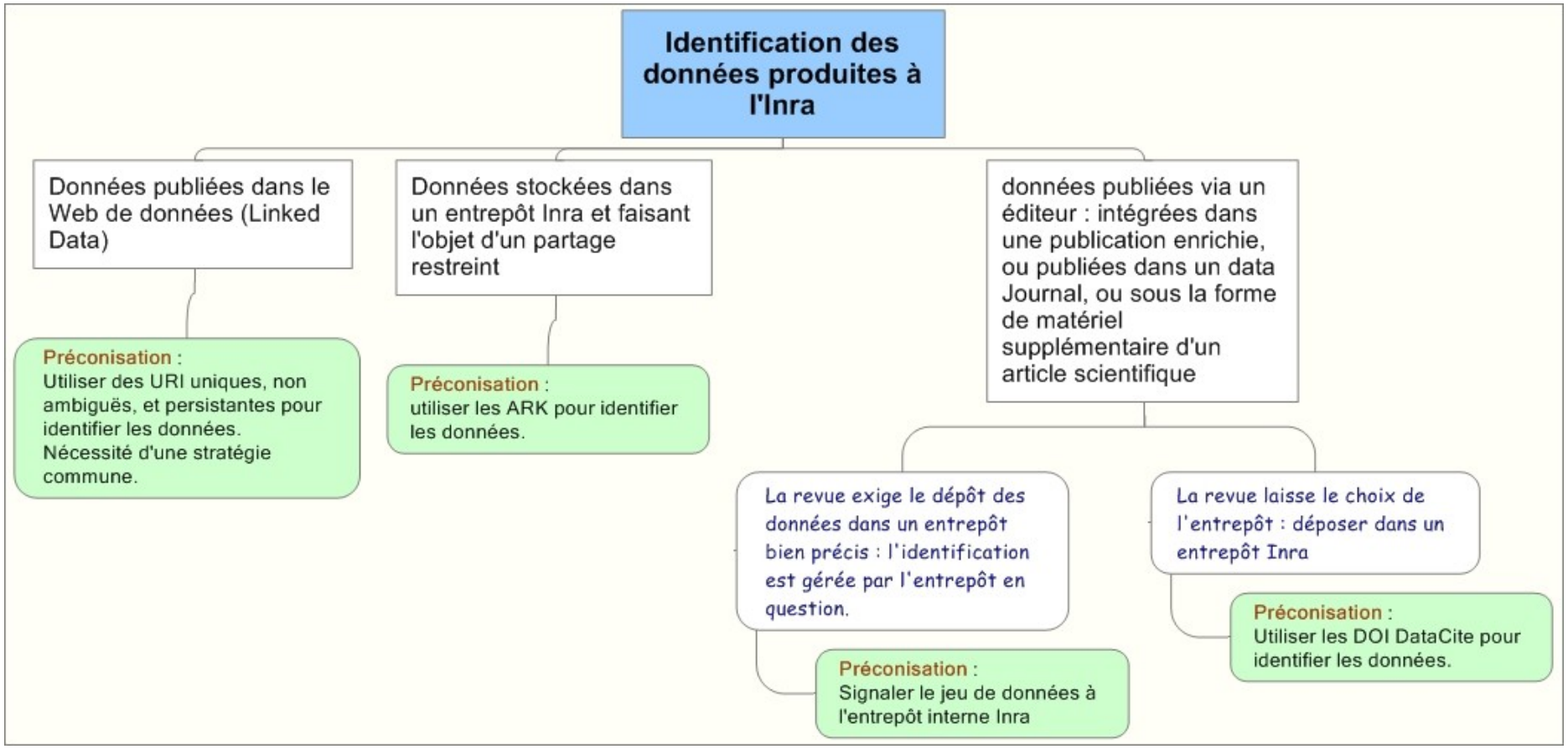
Identifiants

- plusieurs identifiants sont compatibles avec les données de la recherche

Nommage des fichiers de données

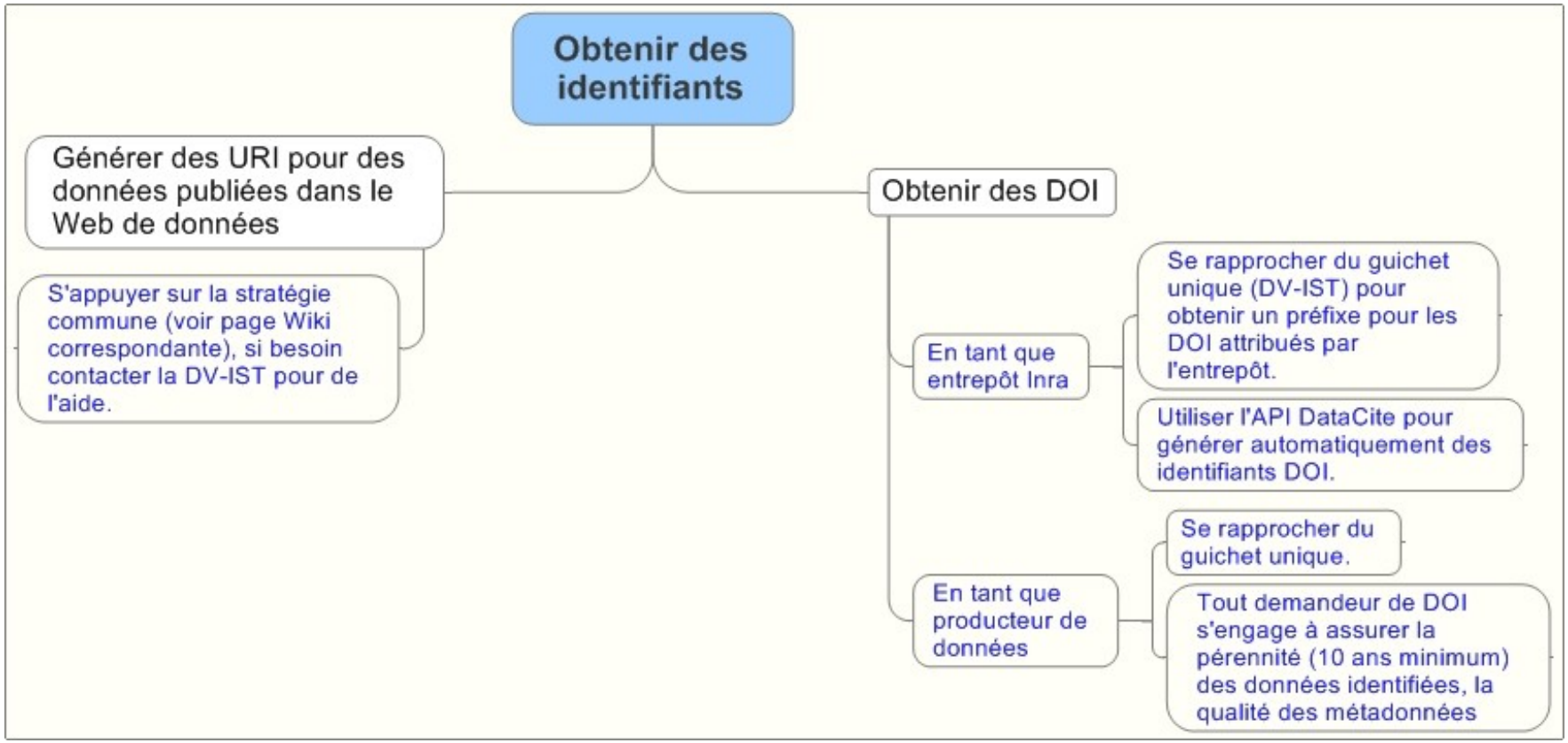
- dans le cas où ils doivent permettre de rechercher et accéder aux données, ils doivent faire l'objet de conventions garantissant unicité et persistance

Identifiants numériques préconisés



DOI : Digital Object Identifier
ARK : Archival Resource Key
URI : Uniform Ressource Identifier

Comment obtenir des identifiants ?

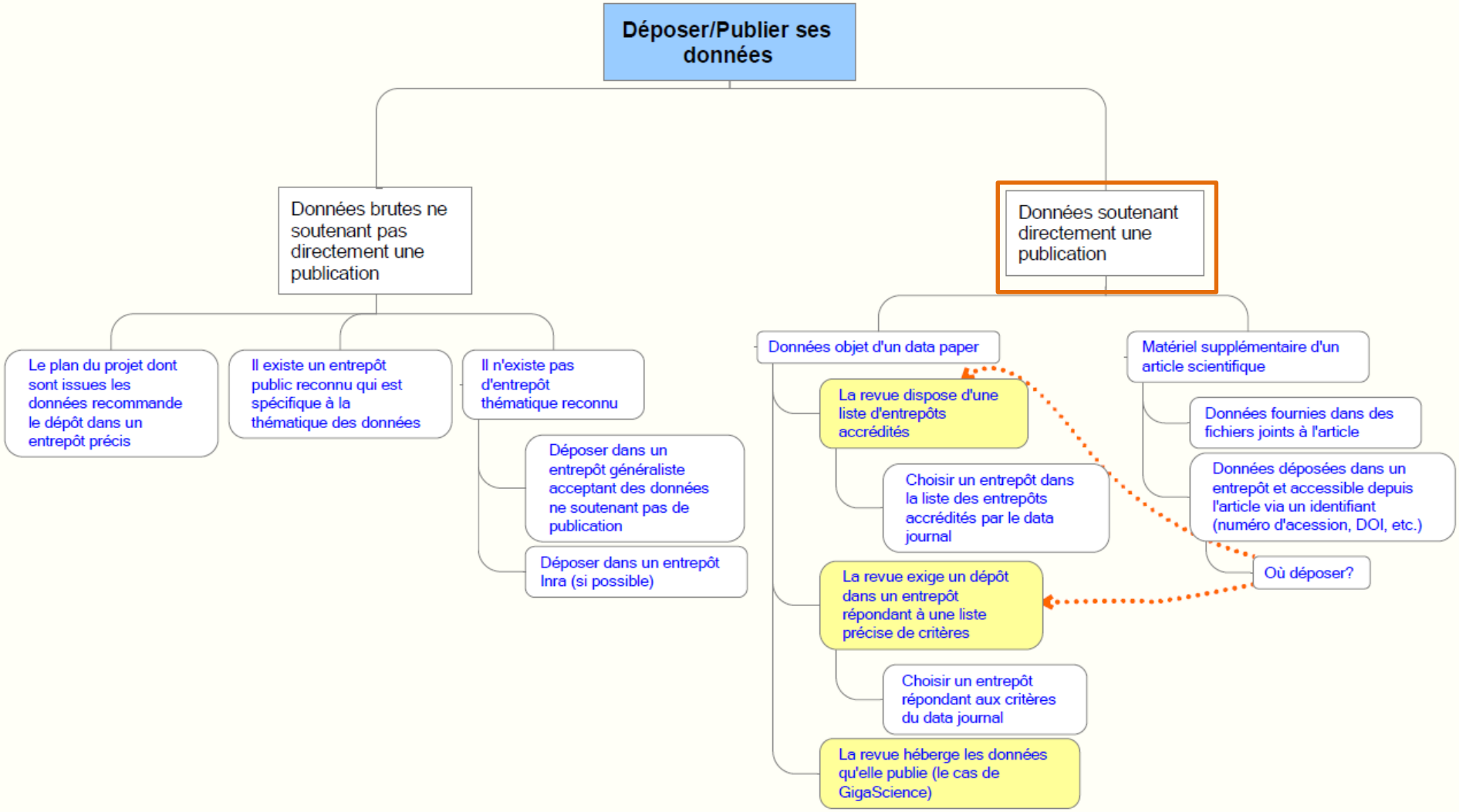


Comment publier ses données ?

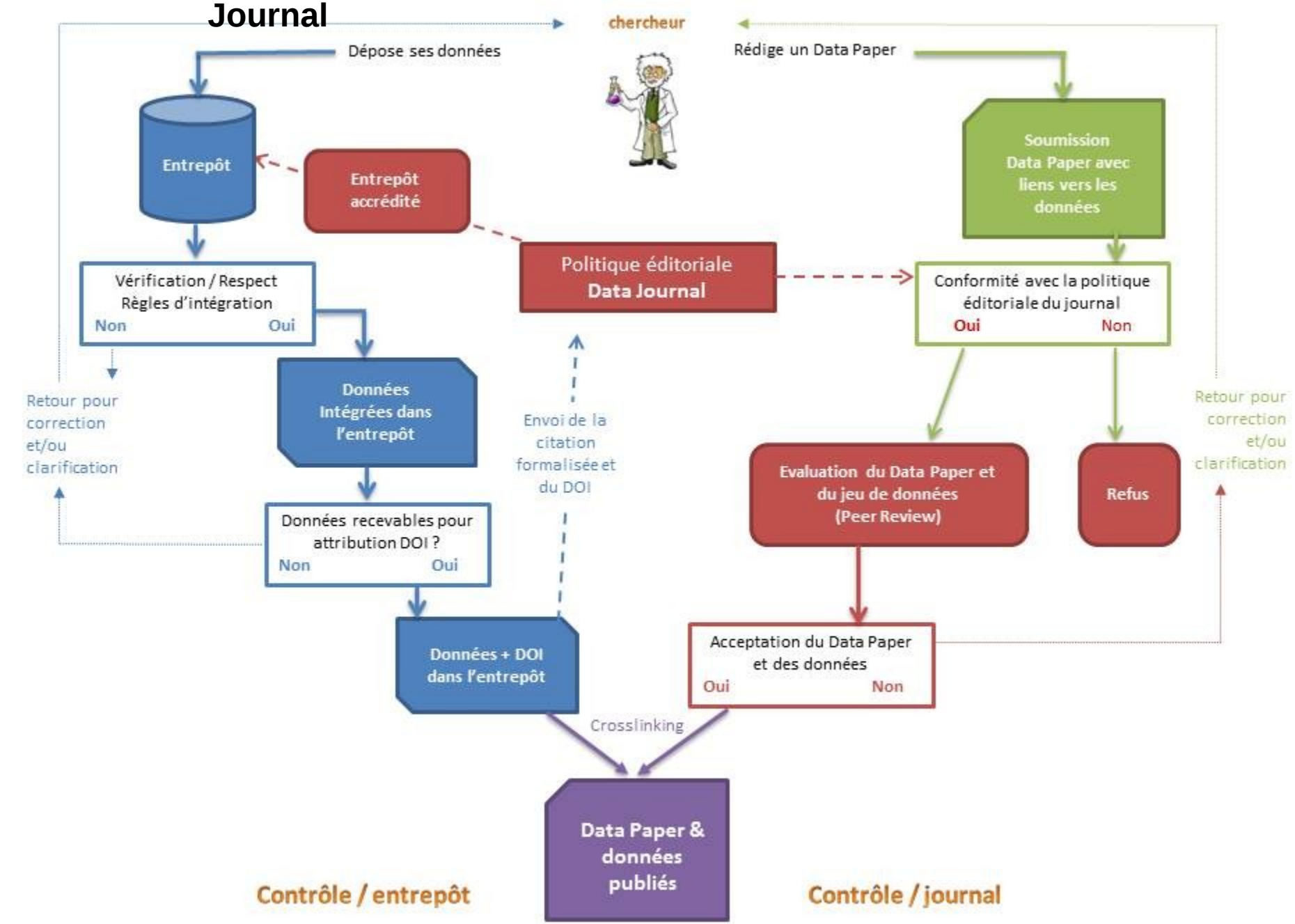
Plusieurs stratégies de publication …

- Publier ses données dans un entrepôt
- Fournir ses données sous la forme de **matériel supplémentaire** à la publication
- Publier ses données dans un **Data Paper** : publication scientifique spécifique décrivant les données, publié soit dans une revue «classique» soit dans un Data Journal qui peut fournir ou préconiser des entrepôts de confiance**.**
  NB : Se renseigner sur le modèle économique
- Publier dans le **web des données** (linked data)

Comment publier ses données ?



Processus de publication d'un Data Paper dans un Data Journal



Valeur ajoutée de la liaison données & publication

- Permet de retrouver plus facilement les données (pointées par des liens),
- Facilite l'interprétation des données et donc leur réutilisation en fournissant une explication sur les traitements et les résultats obtenus,
- Permet à l'auteur d'avoir une reconnaissance immédiate et supplémentaire de son travail.

Piwowar HA, Day RS, Fridsma DB (2007) Sharing Detailed Research Data Is Associated with Increased Citation Rate. PLoS ONE 2(3): e308. doi:10.1371/journal.pone.0000308

Voies de partage des données	Avantages	Limites
Données intégrées Articles soumis au peer-review	· Intégration maximale des données et de l'article : citable, recherchable · Paternité des données / crédits immédiat aux auteurs	· Données difficiles à trouver indépendamment de l'article (réservées à l'abonné selon le modèle éditorial) et dans une forme peu ou pas réutilisable
Données intégrées Supplementary files associés à un article	· Bonne intégration des données et de l'article · Format des données libéré des contraintes de rédaction de l'article · Paternité des données / crédits aux auteurs	· Taille souvent limitée à 10 GB · Peu de standardisation sur le signalement des fichiers «supplémentaires» · Identification des données indépendamment de l'article possible (via DOI) mais rare
Données déposées dans des entrepôts reconnus Liens réciproques entre l'article et les données (dépôt soit dans un entrepôt interne à la revue, soit dans un système externe)	· Entrepôts reconnus par une communauté disciplinaire · Données normalisées, standardisées, conservées de façon pérenne · Pas de restriction en volume · Liens réciproques sécurisés	· Existe dans quelques disciplines (biologie, sciences de la vie, sciences du sol, chimie) · Dépend du maintien de financement par les gouvernements (soumis aux aléas budgétaires)
Données publiées dans des Data Papers	· Paternité des données / crédits aux auteurs · Citation aisée · Réutilisation des données facilitée	· Interrogation sur la qualité : · du Peer-Review ? · des liens bidirectionnels entre données et Data Paper

Publier dans le Web des données

- Web des données :
  - Idée 1 : mettre sur le Web, des données dont la structure et la sémantique sont explicites.
  - Idée 2 : identifier chaque objet mis sur le Web avec une URI HTTP unique et persistante.
  - Idée 3 : lier les données entre elles indépendamment de leur emplacement sur le Web.
  - Idée 4 : une gigantesque base de données à l'échelle du Web, interrogeable via un langage de requête.

Publier dans le Web des données

Comment publier ?

- S’appuyer sur une stratégie commune d’attribution des URI pour garantir leur unicité
- Mettre à disposition des fichiers RDF
- Mettre ses données dans un entrepôt RDF ou « triple store »

Quel entrepôt RDF ?

- Pas nécessaire de tout centraliser dans un unique entrepôt
- Entrepôt disponible pour l’hébergement des

Comment trouver / réutiliser des données ?

Trouver / réutiliser des données ?

- Les sources bibliographiques classiques : bases de données, archives ouvertes, réseaux sociaux, moteurs de recherche n'intègrent actuellement quasiment pas ou très peu les datasets

- En pratique, les données sont donc le plus souvent encore repérées par l'article de recherche qui possède le lien (ou non) vers les données

Comment citer des données ?

La citation : un élément clé

- Assure une meilleure visibilité,
- Facilite la localisation, la découverte et l'accès aux données,
- Facilite la vérification/validation des résultats de recherche
- Participe à la pérennité des jeux de données (une bonne infrastructure de citation suppose un accès permanent aux données citées dans des entrepôts fiables)
- Permet la reconnaissance du travail de création des données (par les pairs, par les différentes métriques d'évaluation)

Culture de la citation des données : rôle du chercheur

Comment citer une donnée dans un article ?

_2

Rendre visible ~ partager les données Inra

Disponibilité

Des questions

Faut-il un entrepôt interne ?

•Avantages et enjeux, cahier des charges ?

Faut-il un annuaire des données Inra?

•Avantages et enjeux, cahier des charges?

Des orientations

Adopter une stratégie institutionnelle d'attribution des identifiants (DOI, URI)

Soutenir le partage et la publication des données

Recommandations - validation des bonnes pratiques

Des chantiers de mise en œuvre

Entrepôt, annuaire

Accompagnement : appui technique et juridique, formation, information

Préservation

Entrepôt et/ou annuaire interne Inra ?

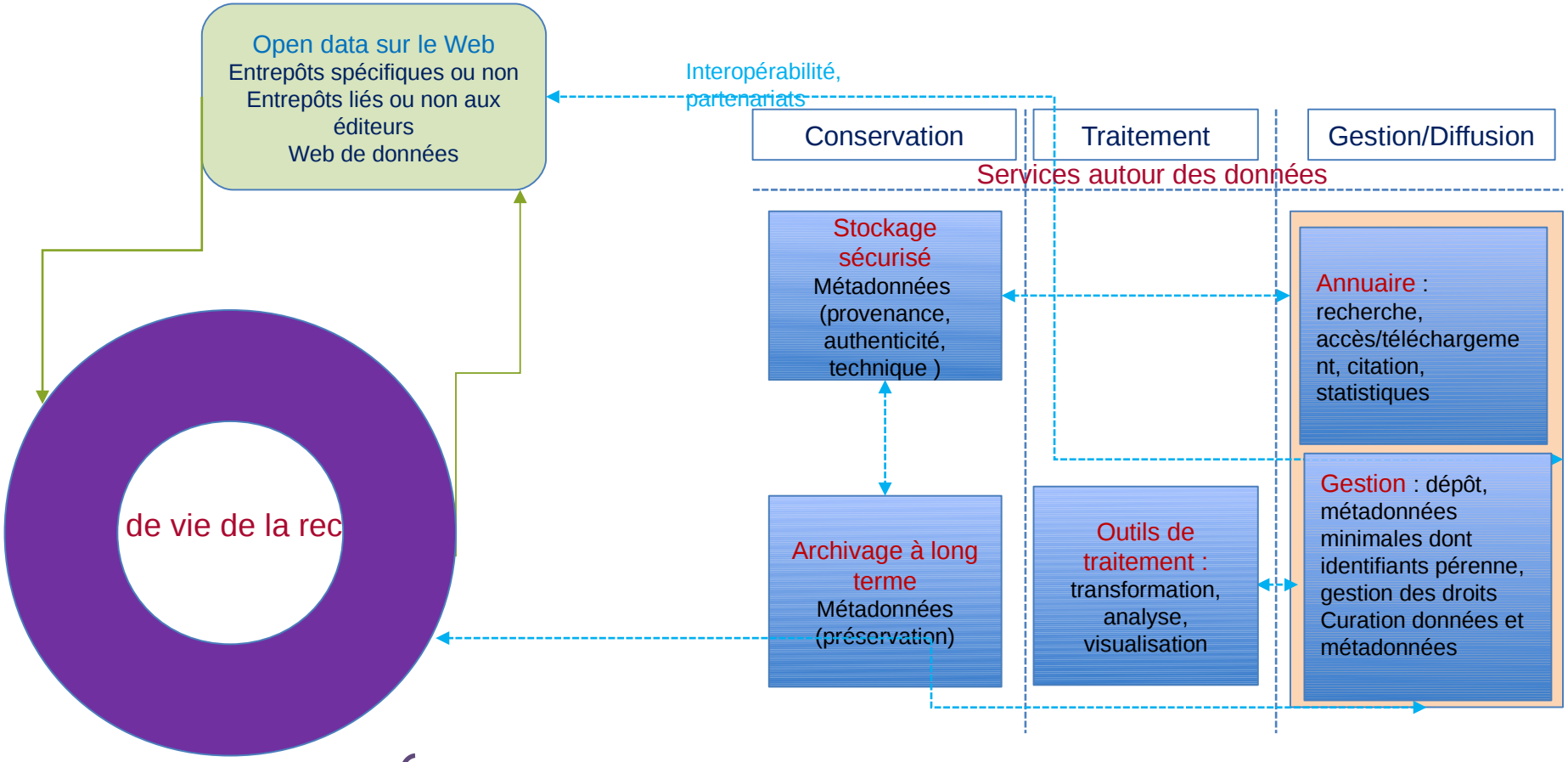
Avantages/enjeux

Avantages

- Valorisation des données produites
- Prise en compte de besoins spécifiques internes
- Soutien à la publication : les politiques éditoriales sont très hétérogènes et imparfaites
  - peu précises quand elles existent,
  - le dépôt est plus souvent préconisé qu'exigé

Disposer d'une plateforme interne Inra

- Quel cahier des charges respecter ?
- L'hypothèse Prodinra



Quelles données déposer? A quel moment du cycle de vie de la recherche?
Quels formats de fichier?
Quelles métadonnées fournir? Quels standards de métadonnées? Quels vocabulaires?
Quelle licence?

Connexions à décrire/formaliser dans le cadre du chantier partage des données

Connexions découlant de pratiques actuelles, à prendre en compte dans le chantier des données

Entrepôt interne – Cahier des charges

- **Mise en œuvre des fonctions et des services qui permettent/facilitent la citation des données.**
  - Exemples : identification par DOI, qualité, stabilité et accès pérenne des données et des métadonnées.
- **Interopérabilité avec certains entrepôts (entrepôts thématiques + les plus recommandés par les organismes de financement ou les éditeurs)**
  - Exemples : Dryad, Pangaea, Zenodo, etc.

Visibilité dans le Datation Citation Index de Thomson

- **Le DCI prend en compte les jeux (Dataset) les études de données (Data study) déposés dans un entrepôt reconnu (Repository).**
  - Une centaine d'entrepôts enregistrés dans le DCI en juillet 2013
- **La sélection (par des évaluateurs) au niveau de l'entrepôt inclue une série de critères**
  - Pérennité et stabilité, financement, examen par les pairs, lien avec la littérature de recherche, capacité et durabilité du stockage, contenu éditorial, citation des données et normes

Datation Citation Index

Points forts

- A ce jour, pas d'autres outils équivalents/concurrents pour la recherche de datasets en terme de couverture
- Intégration avec le Web of Knowledge : à partir d'une recherche sur un sujet on peut trouver des publications et des données, utiliser les facettes pour voir quelles sont les sources des données, les thématiques du WOS associées etc...
- Processus d'évaluation des jeux de données, en particulier sur la présence de métadonnées descriptives et le lien avec une publication

Des données dans

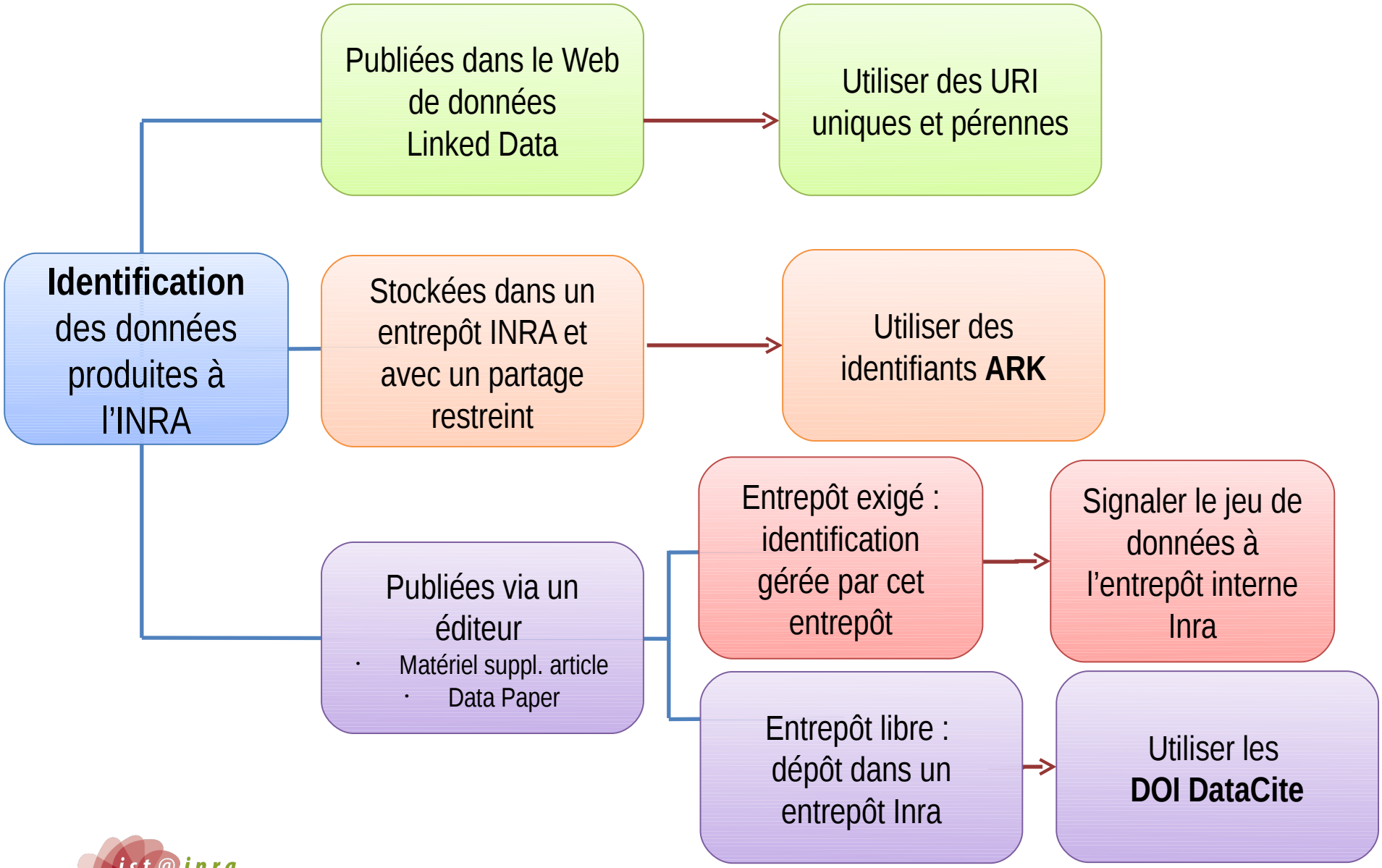
Les données sont déposées dans Prodinra

- Jeu de données en fichier joint de la notice d’un article : adaptations liées au stockage
  - Données non citables
- Jeu de données déposé en tant que produit au même titre qu’un article : plusieurs questions
  - identification, métadonnées spécifiques, stockage des données volumineuses, interopérabilité avec certains entrepôts externes;

Cela nécessiterait une évolution structurelle :

Quelles orientations au niveau institutionnel?

Une stratégie institutionnelle des identifiants?



Attribution des identifiants numériques

- **Caractéristiques attendus des identifiants numériques :**
  - Unicité,
  - Persistance.
- **Un préfixe commun pour chaque type d'identifiant :**
  - http://opendata.inra.fr pour les URI
  - Préfixe DOI délivré par DataCite
- **Des règles communes pour la génération de suffixes uniques**
  - Exemple pour les URI

Une indication claire des caractéristiques des

Plan de gestion des données

- document officiel, à établir au démarrage d'un projet de recherche,
- décrit la façon dont les données seront gérées pendant la phase de recherche, et une fois le projet terminé,
- décrit finement la nature des données, les méthodes d'obtention, les différents aspects de gestion des données, la création de métadonnées, la conservation immédiate et à plus long terme
- variable selon les disciplines et la nature des données produites.

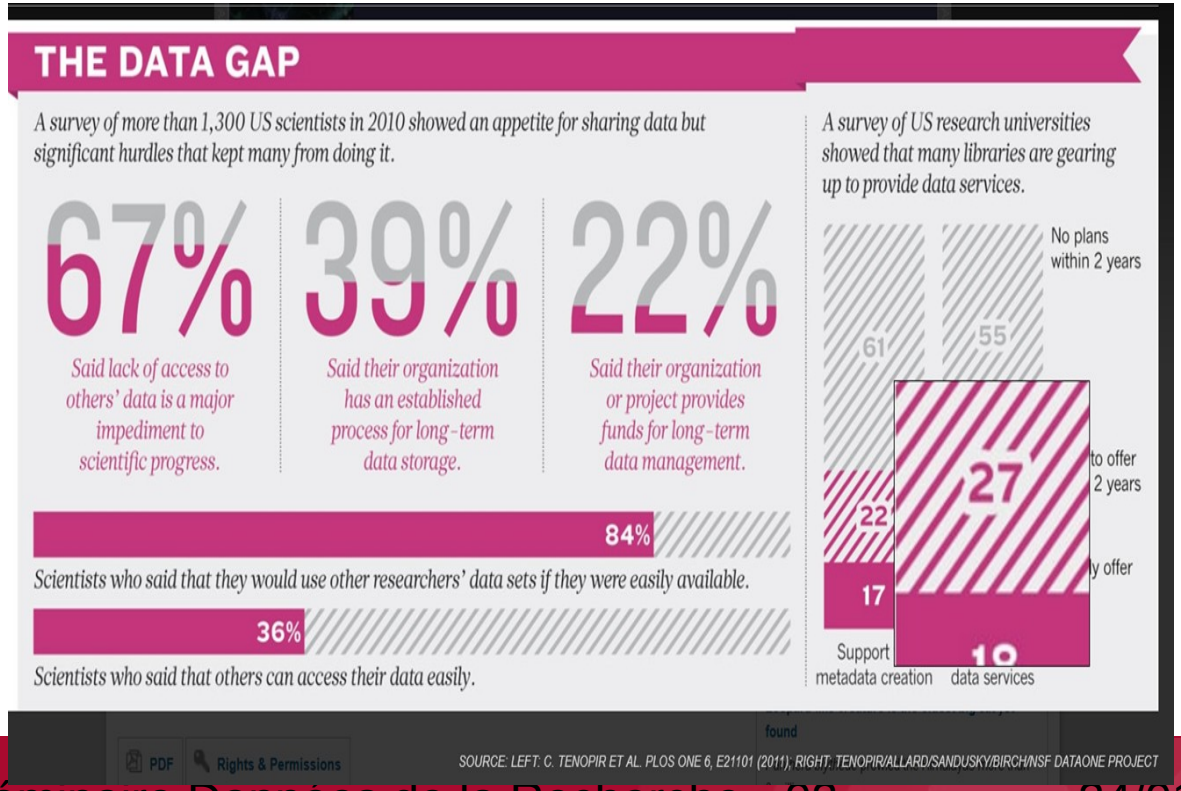
Plan de gestion – Projet de recherche

Description du **projet**	but de la recherche, organisations et personnel impliqués,
Description des **données collectées** :	nature et format des données, méthodes de collectes utilisées
Normes à appliquer pour les **formats de données et des métadonnée**s,	
Plan pour le **stockage à court terme** et la gestion de données	formats de fichiers, type de stockage, procédures de sauvegarde et de sécurité mises en place
Questions juridiques et éthiques	propriété intellectuelle, aspects de confidentialité
Accès et disponibilité des données	restrictions éventuelles
Dispositions pour **l'archivage à long terme et la préservation**	par exemple dépôt dans une archive
Définition des **différentes responsabilités**	personnes impliquées et rôles, procédure de contrôle pour le respect du plan de gestion ...

Soutenir le partage et la publication de données

Constat paradoxal …

- **Un large consensus en faveur du partage des données de la recherche mais des pratiques en fort décalage**



Des freins à lever …

- **Un large consensus en faveur du partage des données de la recherche mais des pratiques en fort décalage**

- **Crainte d'une mauvaise utilisation ou interprétation de leurs données**

- **Questions d'ordre juridique**
  - Données sensibles, liées aux personnes …
  - Violation de la propriété intellectuelle (perte de l'authorship)

Participer à une culture de citation

Souscrire aux principes de citation

- Importance: Data should be considered legitimate, citable products of research. Data citations should be accorded the same importance in the scholarly record as citations of other research objects, such as publications.

Faire des recommandations claires aux entrepôts et aux chercheurs de l'institut

Faire du lobbying auprès des agences de

Chantiers de mise en œuvre

* Mobilisation possible de compétences IST

Information, formation	Stratégies	Infrastructures	Actions vers l'extérieur
Plan de gestion des données *	Identifiants numériques *	Annuaire données Inra *	Lobbying
Propriété intellectuelle	Données à stocker	Entrepôt données Inra *	Implication dans des groupes de travail internationaux *
Publication de données * citables	Données à archiver	Guichet identifiants numériques *	
Citation des données *	Données à partager	Outils d'exploitation des données *	Partenariats (éditeurs scientifiques, entrepôts internationaux) *
Exploitation des données * (méthodes, outils)	Plan de gestion des données *		
Standards (métadonnées, vocabulaires) *			

Création Création Création

Rendre visible ~ partager les données Inra

	Producteur de données 	INRA SCIENCE & IMPACT
Disponibilité	Publier ses données, pour en obtenir reconnaissance et préserver un certain contrôle	Soutenir la publication des données Créer un entrepôt de données interne ?
Visibilité	Utiliser des identifiants uniques et pérennes comme le DOI Adopter de bonnes pratiques de citation	Promouvoir l'utilisation d'identifiants pérennes comme le DOI, soutenir et promouvoir de bonnes pratiques de citation
Interprétabilité	Produire des métadonnées de qualité facilitant l'interprétation des données, s'appuyant sur des standards existant	Aider à produire des métadonnées de qualité Promouvoir les liens entre publications et jeux de données
Réutilisabilité	Utiliser des formats standards, Assurer la préservation des données à long terme pour en faciliter l'utilisation secondaire	Mettre en place les conditions de curation, de préservation des jeux de données et des outils logiciels nécessaires à l'analyse et la réutilisation secondaire des données
Citabilité	Adopter un modèle de citation de données reconnu, Respecter les normes de métadonnées pour les jeux de données, Utiliser des identifiants pérennes tels que le DOI	S'engager au niveau de l'établissement sur des normes de citation de données uniformes (adaptées aux disciplines) Soutenir et promouvoir des identifiants pérennes
Curation	Élaborer/suivre des plans de gestion des données réalistes durables	Recommander l'usage des plans de gestion
Préservation	Collaborer avec les entrepôts de données	

Merci de votre écoute …

Groupe de travail IST :

Pascal Aventurier	Diane Le Hénaff
Caroline Dandurand	Virginie Lelièvre
Esther Dzalé Yeumo Kaboré (coord.)	Véronique Batifol
Dominique L'Hostis (coord.)	Sylvie Cocaud