Étude de faisabilité d'un service générique d'accueil et de diffusion des données simples : recueil des besoins et des contraintes des usagers

Collège Données De La Recherche

Étude de faisabilité d'un service générique d'accueil et de diffusion des données simples

Livrable de synthèse 1 : Recueil des besoins et des contraintes des usagers

Comité pour la Science Ouverte / MESRI

2 novembre 2020

MINISTÈRE DE L'ENSEIGNEMENT SUPÉRIEUR, DE LA RECHERCHE ET DE L'INNOVATION
Liberté
Égalité
Fraternité

DATACTIVIST

OUVRIR LA SCIENCE !

Accès rapide

Contexte de l’étude

Méthodologie

Préambule : Périmètre de l’étude

Grands principes, facteurs d’adhésion

↳Parcours producteur : Dépôt

↳Parcours réutilisateur : Recherche

↳Parcours gestionnaire : Modération et administration

Un besoin pour tous les acteurs : L’accompagnement

Informations complémentaires

Contexte de l'étude

Émergence institutionnelle de la Science Ouverte

Évolutions récentes du cadre réglementaire et institutionnel :

- La Loi pour une République Numérique instaure la possibilité de l'ouverture des publications et le principe de l'ouverture des données de la recherche financées à 50% au moins sur fonds publics (2016)
- Le programme européen Horizon 2020 rend obligatoire l'ouverture des données de recherche quand cela est possible (2017)
- Le Plan National pour la Science Ouverte rend obligatoire l'ouverture des données de recherche issues de programmes financés par appels à projets quand cela est possible (2018)
- L'Agence Nationale pour la Recherche rend obligatoire le Plan de Gestion des Données pour les projets financés (2019)

Les initiatives Science Ouverte

- Comité pour la Science Ouverte :
  - Collège Données de la recherche
  - Groupe de travail ***Service générique d'accueil et de diffusion des données simples***

La présente étude s'inscrit dans l'action "Développer un service générique d'accueil et de diffusion des données simples" de l'axe "Structurer et ouvrir les données de la recherche" du Plan National pour la Science Ouverte.

Étude de faisabilité d'un entrepôt générique

1 EXPRESSION DES BESOINS ET DES CONTRAINTES

Objectifs :

- préciser le périmètre du dispositif par l'analyse des retours d'expérience et la rencontre des différents acteurs français et européens ;
- identifier les besoins fonctionnels.

Contenu du livrable:

- périmètre du dispositif ;
- synthèse des besoins fonctionnels ;
- cahier des charges pour les fonctionnalités primaires et additionnelles du service.

2 BANC DE TESTS LOGICIELS

Objectifs :

- analyser les offres logicielles et assurer les benchmarks pour argumenter les choix ;
- analyser et proposer les ergonomies et les services à valeur ajoutée indispensables à l'appropriation en lien avec les besoins fonctionnels identifiés.

Contenu du livrable:

- synthèse des benchmarks sur les outils ;
- préconisations pour le déploiement au regard des infrastructures numériques possibles aux échelles régionales et nationales.

3 PROPOSITION DE SCÉNARII DE MISE EN ŒUVRE

Objectifs :

- étudier les différents degrés de mutualisation et leur impact sur l'appropriation, la faisabilité technique au regard des contraintes d'interopérabilité, d'architecture ;
- mesurer les coûts en s'appuyant sur un nombre restreint de scénarii :
- évaluer le besoin et les coûts en matière de déploiement.

Contenu du livrable:

- description des différents scénarii de mutualisation ;
- description des coûts des différents scénarii retenus.

Étude de faisabilité d'un entrepôt générique

4 RÔLES ET RESPONSABILITÉS

Objectifs :

- définir les rôles et responsabilités du dispositif, présenter des éléments sur le niveau et les procédures de curation ;
- proposer une démarche réfléchie et progressive pour mettre en place un entrepôt de qualité qui remporte l'adhésion.

Contenu du livrable:

- tableau des rôles et des responsabilités, schéma des différentes couches, cartographie, schéma de workflow sur le cycle de vie et disposition des acteurs dans les différentes étapes ;
- description de la démarche permettant l'adoption et la montée en charge progressive d'un service donnant accès à des données de qualité et des fonctionnalités à valeur ajoutée.

5 COMMUNICATION

Objectifs :

- prise de connaissance et partage des travaux vers la communauté de gestion des données, les établissements.

Contenu du livrable:

- mise en forme de l'étude pour leur diffusion et publication sur ouvrirlascience.fr , diffusion des rapports du GT CoSO, présentation aux journées nationales de la Science Ouverte (JNSO), diffusion à l'échelle internationale (traduction EN).

Bibliographie et entretiens

Cette étude se nourrit des travaux menés depuis une dizaine d'années sur les services d'accueil et de diffusion des données de la recherche, et les services connexes de l'écosystème des données de la recherche :

Études et enquêtes sur les entrepôts de données, menées dans des contextes différents :

- **National** :
  - Enquête données de la recherche en SHS Rennes 2 (2017) ;
  - Étude COPIST n°2 (2018) ;
  - Enquête chercheurs Inria (2019).

- **International** :
  - Étude comparative des services nationaux menée par Hugo Catherine pour le GT.

Retours d'expérience d'opérateurs de services connexes ou similaires :

- **Archives ouvertes** : entretien avec l'équipe du CCSD qui opère le service HAL ;

- **Entrepôts spécialisés** : entretien avec l'équipe de la TGIR Huma-Num qui opère notamment les services Nakala et Isidore ;

- **Entrepôts généralistes** : entretiens avec les équipes de DANS, Zenodo, JISC Open Research Hub, pour une présentation des services et un retour d'expérience sur la construction de ces services.

Méthodologie

Méthodologie

Une méthodologie de recueil des besoins et contraintes des usagers

- **participative**, afin de bénéficier de l'expertise de la communauté de l'ESR,
- issue des méthodes **agiles** de développement logiciel.

Principe

Les participants décrivent des *user stories :*

- "en tant que <X>, je souhaite <Y> afin de <Z>"
- en adoptant le point de vue d'un **producteur** de données,
- ou d'un **réutilisateur** de données.

Objectifs

Les user stories amènent les participants à

- **distinguer** leurs besoins et
- à les **motiver**,
- pour ensuite en établir le niveau de **priorité**.

Les user stories s'inscrivent dans des parcours utilisateur, qui dessinent les fonctionnalités attendues ou souhaitées d'un entrepôt pour différents profils d'usager.

La synthèse des user stories est complétée par des verbatims d'entretiens.

Calendrier et adaptation

Calendrier initial

4 ateliers collaboratifs étaient prévus :

- Un atelier à Paris avec les **membres du Groupe de Travail**, impliqués dans différentes initiatives et réseaux locaux, nationaux et internationaux ;
- Trois ateliers à Marseille, Nantes, Strasbourg, organisés en collaboration avec les points de contact locaux, avec des **acteurs de terrain** exerçant dans différents laboratoires et disciplines.

Adaptation au contexte sanitaire

- Le 1er atelier à Paris s'est tenu le 6 mars 2020.
- Les 3 **ateliers en région** ont dû être **annulés**,

Les 3 ateliers annulés ont été **remplacés** par **11 entretiens individuels** semi-dirigés d'environ **2 heures** avec un panel de personnels de l'ESR choisis pour leur diversité de profils (fonctions, structures, disciplines scientifiques, état d'avancement d' éventuels projets d'entrepôts de données).

26 participants

- 15 à l'atelier de Paris ;
- 11 en entretien semi-directif de mai à juin.

***256 user stories* recueillies**

- 82 *user stories* collectées en atelier,
- 174 *user stories* extraites des entretiens

Caractérisation du panel des participants

Disciplines scientifiques

- **12** participants issus des Sciences Humaines et Sociales : archéologie, histoire, information documentaire, linguistique…
- **14** participants issus de Science - Technique - Médical : biologie, chimie, informatique, médecine, neurosciences, physique, sciences de l'univers…

Positions

- Corps : conservateur des bibliothèques, CR, DR, IE, IR, MCF, MCU-PH, PR
- Fonctions : chargé de mission data et science ouverte, chercheur, data librarian, directeur d'unité, enseignant-chercheur, référent intégrité scientifique, responsable de l'information scientifique, responsable des données de la recherche, responsable du pôle numérique, responsable innovation, responsable science ouverte, responsable technique, VP délégué science ouverte et données de la recherche

Structures et tutelles

- CEA, CINES, CNRS, INRAe, Inria, Inserm, IRD, MNHN
- Aix-Marseille Université, EHESS, Sorbonne Université, Université de Bretagne Occidentale, Université de Lille, Université de Lorraine, Université de Montpellier, Université de Nantes, Université de Rennes 1, Université de Strasbourg, Université Paris-Diderot
- services centraux (SCD), structures fédératives (institut convergences, OSU), unités mixtes (UMR, UMS)

État d'avancement des entrepôts de données

- entrepôts existants : EHESS, INRAe, IRD, CIRAD, Science Po…
- études finalisées : GRICAD, MNHN, U. Lorraine, U. Strasbourg
- études en cours : CNRS, INRIA
- études préliminaires, réflexion ou pas de démarche engagée : Aix-Marseille U., Sorbonne U., U. Bretagne Occidentale, U. Lille, U. Montpellier, U. Nantes, U. Rennes 1, U. Paris-Diderot

Vue générale sur les *user stories*

Sur les 256 user stories recueillies…

Provenance :

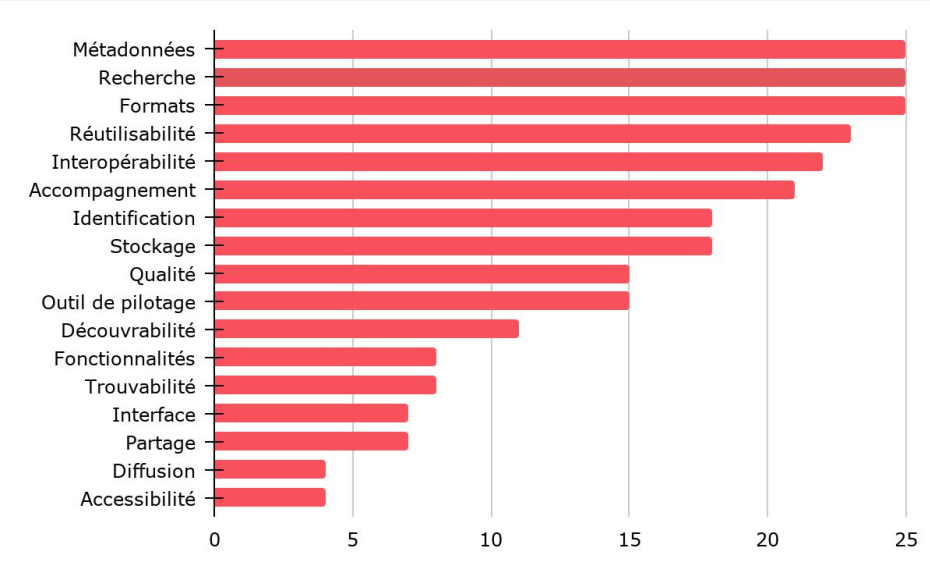
- 82 collectées en atelier ;
- 174 extraites des entretiens.

Point de vue adopté :

- **160** stories du point de vue du **producteur** et/ou du **gestionnaire** de données ;
- **96** stories du point de vue du **réutilisateur** des jeux de données déposés dans un entrepôt.

Les user stories ont fourni le matériau de base à la présente synthèse.

Les thèmes les plus discutés :



(nombre de user stories concernées)

Préambule

Périmètre de l'étude

Questions sur le positionnement et le besoin d'un service générique national d'entrepôt des données de la recherche

Au début de l'atelier et des entretiens, les participants ont formulé un certain nombre de questions et d'objections sur l'ouverture des données, le choix des données à ouvrir et les modalités de cette ouverture.

Les questions et objections se rapportent à deux grands thèmes :

- Quelles données devons-nous, pouvons-nous, voulons-nous ouvrir, et quand ?
- Pourquoi déposer les données sur un service générique et national ?

Nous avons pris le temps d'écouter ces questions et objections et de clarifier le périmètre de l'étude.

Les questions et les objections reflètent la diversité :

- des profils des participants,
- des pratiques disciplinaires,
- des initiatives en cours dans les établissements et organismes,
- des initiatives dans les disciplines (TGIR),
- des contraintes réglementaires (données personnelles, propriété intellectuelle, secret industriel…),
- des opportunités et des risques perçus dans les communautés scientifiques (valorisation économique, course à la publication…)

Apporter des réponses précises aux questions et lever les objections sera une condition nécessaire au succès d'un éventuel service d'entrepôt des données de la recherche.

Quelles données ouvrir ?

Les participants se demandent quelles données ils doivent, peuvent, veulent ouvrir et déposer sur un entrepôt, quand et comment :

- Des données avec une date de péremption, dont la pertinence décroîtra avec le temps, voire qui deviendront inexploitables ou caduques ?
- Des données qui peuvent être reproduites, issues de simulations ?
- Des données correspondant à des résultats négatifs, issues d'expérimentations qui n'ont pas permis de répondre à la question initiale ?
- Des données personnelles (médicales, entretiens, enquêtes...) anonymisées ou agrégées ? Avec quelles modalités, quelles bonnes pratiques, quelles garanties ?

*" Il y a une crainte de la communauté sur la perte de la publication si ils mettent en ligne leurs données. Ils n'ont pas encore conscience que l'ouverture des données de terrain va être le vecteur de notoriété et de citation et que les données peuvent être **partagées en accès restreint avant l'article** fondateur. "*

*" Nous avons de plus en plus de **demandes** de faire les soutenances à **huis clos** et de ne pas publier les thèses. On fait signer des engagements de confidentialité, qui peuvent aller d'un à quatre ans. Très souvent, les chercheurs demandent que les **données** restent **confidentielles**. "*

Pourquoi un service générique et national ?

Les participants s'interrogent sur la **pertinence** de déposer les données de la recherche sur un entrepôt :

- **générique**, plutôt qu'adapté au plus près des besoins des communautés disciplinaires,
- **national**, plutôt qu'européen (a minima) ou international, dans un contexte de circulation des connaissances et de collaboration mondialisées.

*" Toutes les bases de données [disciplinaires] qu'on utilise sont **internationales**. Je ne vois pas l'intérêt spécifiquement sur mon sujet d'avoir une infrastructure nationale. "*

*"Avoir une infrastructure **européenne** est nécessaire. "*

Les avantages d'un entrepôt national

Pour les participants, un entrepôt national est pertinent en tant qu'infrastructure :

- **neutre** vis-à-vis des établissements, qui lui permet d' être un lieu de dépôt des données adapté aux projets partenariaux et évitant les conflits de multi-appartenances institutionnelles,
- **mutualisée** à l'échelle de la communauté de l'ESR français, qui permet d'éviter la multiplication des initiatives, la dispersion des efforts et le gaspillage des moyens.

*" Les multi appartenances [institutionnelles] sont infernales à gérer. Un entrepôt national **neutre** peut permettre de régler certains problèmes."*

"Il faut éviter de gaspiller les énergies et vouloir faire son truc à soi avec son logo. C'est une dépense d'énergie, d'impact, de tout parce que la concurrence n'est plus CNRS versus université mais recherche française versus Elsevier, Google etc.
S'il y a dispersion des forces, on aura les outils qu'ils nous donneront et pas les outils qu'on aura choisis."

Les besoins exprimés couvrent tout le cycle de vie des données

Les besoins pour un entrepôt évoquent des **besoins connexes** :

- stockage de la donnée dès sa création (acquisition, génération),
- non-perte de données,
- traitement et analyse des données collectées,
- préservation et stockage à long terme,
- curation de la donnée,
- archivage

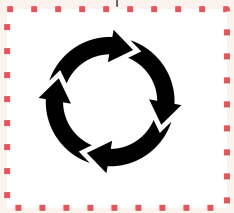
Ces besoins sont réels mais dépassent le périmètre de notre étude.

Grands principes
facteurs d’adhésion

Un entrepôt de confiance

Un entrepôt de qualité

Un entrepôt responsable



Grands principes, facteurs d'adhésion

DATACTIVIST

Un entrepôt de confiance

Les participants attendent d'un entrepôt de données qu'il soit fiable et sûr, et qu'il garantisse :

- la **pérennité** des données et services,
- l'**intégrité** des données stockées,
- l'**innocuité** des données téléchargées.

La confiance des usagers peut être renforcée par la **certification** de l'entrepôt, par exemple CoreTrustSeal.

*" Les entrepôts institutionnels et nationaux doivent viser la **non-perte des données**.*
La première ambition est le stockage, avec des backups professionnels.
La deuxième ambition est l'indexation, il faut être capable de retrouver nos données."

" Ce qui manque actuellement, c'est d'assurer la pérennité des services qui sont rendus.
***Si je pars du laboratoire, ma plateforme de diffusion de données va disparaître** car il n'y aura personne pour la maintenir. "*

Un entrepôt de qualité

Les utilisateurs souhaitent que l'entrepôt soit garant d'une qualité élevée sur :

- les métadonnées ;
- les données à un niveau superficiel (le déposant étant responsable in fine de son contenu).

Les moyens évoqués pour garantir cette qualité sont :

- un important travail de **modération** pour vérifier la qualité des dépôts ;
- un **accompagnement** des déposants pour améliorer la qualité de leurs dépôts et diffuser les bonnes pratiques ;
- l'attribution de **labels** de qualité délivrés par une institution ;
- des fonctionnalités de **relecture ouverte** des dépôts.

Un entrepôt responsable

Les participants ont exprimé leur souci que l'entrepôt soit opéré avec un haut niveau d'exigence et de transparence sur la responsabilité sociale et environnementale :

- informer précisément sur le lieu d'hébergement des données,
- utiliser un hébergement vert (green data center) afin de limiter l'empreinte environnementale,,
- accompagner et inciter les déposants à une certaine sobriété numérique,
- permettre de spécifier une date de péremption sur un dépôt.

*" Pour chaque institution, stocker ses données de manière pérenne a un avantage mais la redondance et la **multiplicité** des diverses **initiatives** n'est pas **écologique**, et il faut réfléchir au fait qu'on n'est pas obligé de tout garder."*

Parcours producteur
Dépôt

Un dépôt réellement simple

Les participants ont exprimé des attentes fortes sur la simplicité du dépôt d'un jeu de données.

> *" Si on veut **inciter** les gens à déposer leurs données, ça doit être **simple**. Les chercheurs n'ont **pas assez de temps** pour faire leur cœur de métier, on veut bien faire ce qui a un **vrai sens scientifique** mais on veut que ce soit simple "*

Besoins

Niveau 1 (Essentiel) :

- interface simple, claire et ergonomique pour permettre un dépôt rapide ;

Niveau 2 (Important) :

- saisie assistée (auto-complétion) pour les auteurs et leurs affiliations, les projets, grâce à l'intégration des référentiels métier et de répertoires d'identifiants : ORCID, IdRef, OpenAIRE, AuréHAL etc ;

S’identifier

Sélectionner les données

Renseigner les métadonnées généralistes (dont DOI) et spécifiques

Spécifier les droits d’accès et la licence

Renseigner les financeurs et les projets

Déclarer les ressources liées

Éditer un jeu de données déposé

Accéder à mon espace personnel

Parcours producteur : dépôt

DATACTIVIST

1. S'identifier sur l'entrepôt

L'usager se connecte à l'entrepôt avec un identifiant dédié ou en ré-utilisant un identifiant existant :

- identifiant sur un service tiers, par Single Sign On : HAL, ORCID, OpenID Connect…
- identifiant institutionnel (de l'établissement), via une fédération d'identité, comme la Fédération Éducation-Recherche Renater.

La possibilité de réutiliser un compte existant permet à l'usager de ne pas être bloqué dès le début du dépôt.

Besoins

Niveau 1 (Essentiel) :

- Single Sign On : HAL, ORCID, OpenID Connect ;
- Fédération Éducation-Recherche Renater

Niveau 2 (Important) :

- Single Sign On : autres services (ex: GitHub) ;
- Fédérations d'identité autres

2. Sélectionner les données

L'usager sélectionne les fichiers qui constituent le jeu de données qu'il veut déposer sur l'entrepôt.

- L'arborescence des fichiers et leurs métadonnées sont conservées ;
- L'entrepôt accepte des dépôts de taille conséquente (taille maximale par fichier, nombre maximal de fichiers et taille totale d'un jeu de données) ;
- Les types de données acceptés sont clairement indiqués ;
- Le dépôt peut se résumer à un lien vers des fichiers sur un autre entrepôt (disciplinaire ou institutionnel) ;
- Les jeux de données volumineux peuvent être envoyés par synchronisation distante (ex: Data Capture Model de Dataverse)

Besoins

Niveau 1 (Essentiel) :

- conservation de la structure des fichiers
- limites élevées sur la taille maximale d'un fichier, le nombre maximal de fichiers, la taille totale maximale d'un jeu de données
- documentation claire sur les types de fichiers acceptés et les limites de taille

Niveau 2 (Important) :

- envoi des fichiers par synchronisation distante

3. Renseigner les métadonnées génériques

L'usager renseigne les métadonnées indispensables au dépôt d'un jeu de données :

- Titre,
- Description,
- Type de données,
- Date de publication,
- Auteurs avec leurs affiliations et leurs rôles,
- Identifiant permanent du jeu de données : Digital Object Identifier (DOI), qui peut être attribué par l'entrepôt lors du dépôt.

Besoins

Niveau 1 (Essentiel) :

- métadonnées essentielles,
- attribution de DOI

Niveau 2 (Important) :

- auto-complétion des auteurs et affiliations par intégration de référentiels,
- inventaire de rôles spécifiques pour les auteurs

4. Renseigner les métadonnées spécifiques

L’usager peut renseigner des métadonnées spécifiques et plus complexes pour une description plus fine du jeu :

- Version du jeu de données ;
- Langue et encodage des données ;
- Mots-clés issus de référentiels métiers ;
- Champs optionnels spécifiques disciplinaires.

Une documentation de ces métadonnées spécifiques permet de favoriser leur diffusion et leur homogénéité.

Besoins

Niveau 1 (Essentiel) :

- schémas de métadonnées flexibles,

Niveau 2 (Important) :

- saisie assistée (auto-complétion ou sélection) par intégration de référentiels et nomenclatures ;
- documentation sur les métadonnées spécifiques

5. Spécifier les droits d'accès et la licence

L'usager spécifie la licence et les droits et conditions d'accès de son jeu de données :

- ouvert ;
- sous embargo, avec date d'expiration ;
- restreint à un groupe d'usagers ou sous conditions ;
- fermé (pour faciliter l'ouverture à terme).

La possibilité de déposer les données en accès fermé ou restreint, par exemple pour pouvoir les corriger avant leur ouverture, est une demande forte des usagers.

Les droits et conditions d'accès, et la licence, doivent être fixés en cohérence avec les conditions des financeurs et la politique de l'institution.

Les valeurs saisies doivent être cohérentes avec le contenu du Plan de Gestion des Données, pour les projets qui en ont un.

Besoins

Niveau 1 (Essentiel) :

- sélecteur de licence,
- mécanisme de contrôle d'accès,

6. Renseigner les financeurs et projets

L'usager déclare les projets dans le cadre desquels les données ont été produites, les financeurs et supports de financement (ANR, ERC, Horizon 2020...) :

- nom du financeur et du programme,
- nom du projet, acronyme et numéro du financement.

Ces champs permettent de :

- lister les jeux de données produits dans le cadre d'un projet pour le suivi administratif (production de rapport d'activité de projet, d'équipe, de laboratoire),
- créer des collections de jeux de données par projet afin d'augmenter la visibilité des travaux,
- générer des tableaux de bord pour le pilotage

Besoins

Niveau 1 (Essentiel) :

- collection de jeux de données,

Niveau 2 (Important) :

- saisie assistée (auto-complétion ou sélection) par intégration de référentiels et nomenclatures (ex: OpenAire pour les projets européens, AuréHAL pour les projets ANR) ;
- tableaux de bord

7. Déclarer les ressources liées

L'usager déclare les ressources liées au jeu de données déposé :

- autres **jeux de données**, sur le même entrepôt ou un autre,
- **publication** sur les archives ouvertes (dont HAL),
- **code source** sur les forges publiques (GitLab/GitHub),
- **plan de gestion de données** du projet,
- **documentation** technique (ex: protocole de codage),
- **cahier** de laboratoire virtuel,
- ...

en renseignant leurs identifiants pérennes (DOI, handle) et/ou des URL pérennes.

Besoins

Niveau 1 (Essentiel) :

- champs dédiés permettant de stocker ces identifiants et liens ;

Niveau 2 (Important) :

- vérification par API, à la saisie, de l'existence des ressource renseignées, pour un nombre réduit de services essentiels (HAL, forge...)

Niveau 3 (Utile) :

- vérification par API pour un bouquet large de services (dont autres entrepôts)

A. Éditer un jeu de données déposé

Le déposant souhaite pouvoir revenir autant de fois que nécessaire sur son dépôt :

- modifier les métadonnées ;
- modifier les droits d'accès au jeu de données ;
- remplacer des fichiers ;
- supprimer des fichiers ;
- dé-publier un jeu de données.

La possibilité de dé-publier un jeu de données intéresse les producteurs de données, notamment pour corriger leurs travaux, mais peut constituer un obstacle pour les réutilisateurs des données.

Besoins

Niveau 1 (Essentiel) :

- gestion des versions pour les jeux de données ;
- génération de nouveaux DOI ;
- liens entre les versions pour proposer à la consultation la plus récente (par défaut)

Niveau 3 (Utile) :

- affichage des différences entre versions successives ;

B. Accéder à mon espace personnel

L'usager dispose d'un espace personnel lui permettant de :

- voir l'ensemble de ses dépôts et des jeux de données dont il est auteur ;
- modifier son "profil chercheur" public, similaire au CV HAL ;
- consulter les métriques de diffusion (citations, réutilisations) de ses jeux de données ;
- visualiser les indicateurs d'activité qui lui sont demandés dans les rapports d'activité et les demandes de financement.

Les indicateurs et statistiques doivent être **exportables** pour faciliter leur utilisation.

Les participants souhaitent que cet espace soit **personnalisable** et que les indicateurs soient manipulables sous la forme de graphiques, tables...

Besoins

Niveau 1 (Essentiel) :

- espace personnel ;
- profil public ;
- métriques de diffusion
- indicateurs d'activité

Niveau 2 (Important) :

- export des indicateurs et statistiques

Niveau 3 (Utile) :

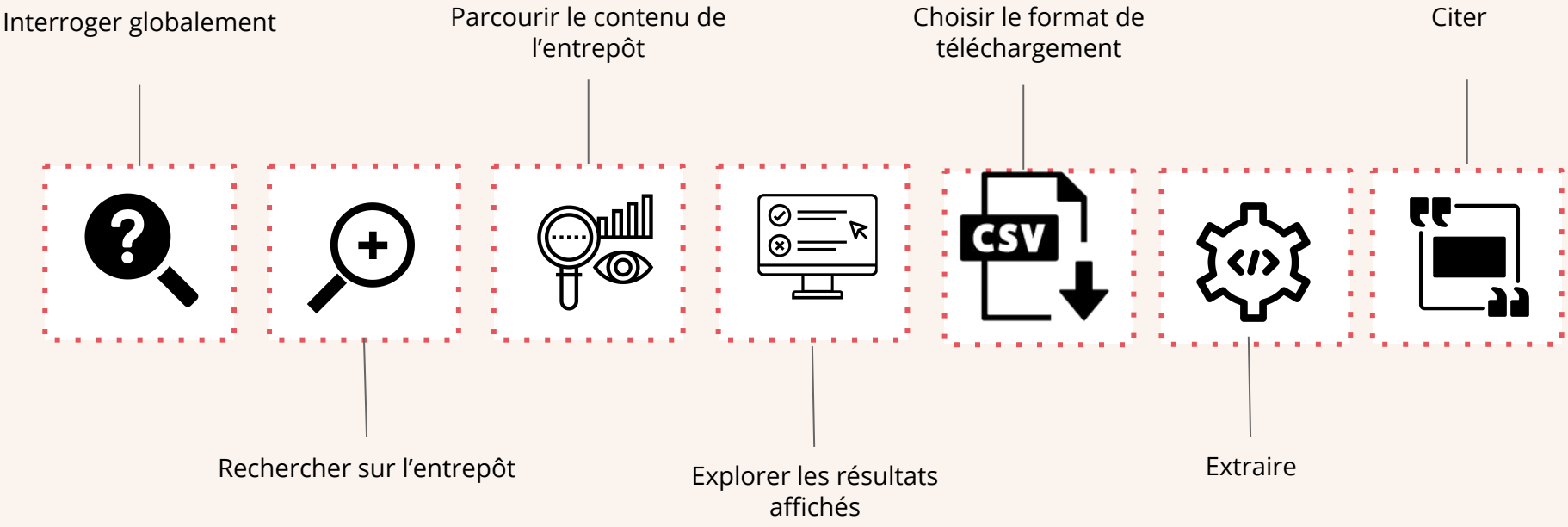
- personnalisation des visualisations d'indicateurs

Fonctionnalités avancées

Les participants ont évoqué des fonctionnalités avancées qui dépassent le cadre d'un entrepôt simple mais seraient séduisantes et donc susceptibles à terme d'augmenter l'adhésion des usagers :

- dépôt depuis un autre service (site éditeur, entrepôt disciplinaire) ;
- dépôt depuis les logiciels métiers pour préserver toutes les métadonnées spécifiques: provenance, conditions de production (heures de relevé, calibrage des instruments...), identifiants d' échantillons ou de gènes...
- récupération des métadonnées embarquées dans les fichiers pour le dépôt
- définition de différents types de versions de données (ex : brutes, traitées, corrigées) ;
- définition et exécution de scripts de pré-traitement (ex: OCéRisation) ;
- conversion automatique vers des formats plus pérennes ou universels,
- nettoyage des données pour les rendre *FAIR (facilement trouvables, accessibles, interopérables, réutilisables)*
- fonctionnalités de croisement et d'alignement automatique de certaines données avec des référentiels ou des champs d'autres jeux de données ;
- interface de dépôt joint pour déposer simultanément la publication, le code source et les données associées ;
- outils d'intégration (embed) et d'éditorialisation des données (CMS) ...

Parcours réutilisateur
Recherche



Parcours réutilisateur : recherche de données

1. Interroger globalement

Les ré-utilisateurs souhaitent utiliser un outil de **recherche** qui serve de point d'accès unique à l'entrepôt national et aux entrepôts spécifiques (disciplinaires ou institutionnels) existants ou à venir.

*" C'est toujours **frustrant** si on n'arrive pas à trouver des choses dont on sait qu'elles sont dans la base ; si on ne sait pas, on passe à côté de données. [...] Le moteur de recherche [du site de l'université] ne me renvoie pas ce que je cherche, je suis obligée de passer par Google pour trouver la bonne page. "*

Besoins

Niveau 1 (Essentiel) :

- l'entrepôt est moissonnable (pour l'utilisation d'autres moteurs de recherche comme ceux de Google)

Niveau 2 (Important) :

- l'entrepôt moissonne d'autres entrepôts ou
- un outil national qui catalogue les données produites (pouvant être hébergées ailleurs) ou
- un méta-moteur de recherche s'appuyant sur les moteurs de recherche d'entrepôts généralistes et spécifiques (ex: NARCIS de DANS)

2. Rechercher sur l'entrepôt

Les ré-utilisateurs ont accès, sur l'entrepôt, à des outils de **recherche** et d'**interrogation** performants et précis:

- sur les métadonnées ;
- sur les données (recherche plein texte),
- par filtres (dont facettes) ;
- avec une assistance à la saisie (auto-complétion)

Besoins

Niveau 1 (Essentiel) :

- moteur de recherche,

Niveau 2 (Important) :

- auto-complétion par intégration de référentiels ;

Niveau 3 (Utile) :

- indexation plein texte

3. Parcourir le contenu de l'entrepôt

Les ré-utilisateurs parcourent le contenu de l'entrepôt au travers :

- de datavisualisations ou
- de collections thématiques éditorialisées.

Des collections ou des vues dédiées permettent de mettre en avant les jeux de données avec un niveau de FAIRness élevé.

Besoins

Niveau 1 (Essentiel) :

- data-visualisation simples sur les métadonnées ;
- curation : collections thématiques, institutionnelles, de niveau de qualité vérifié (FAIRness) ;

Niveau 2 (Important) :

- data-visualisations avancées (ex : cartes géographiques, nuages de mots) ;

Niveau 3 (Utile) :

- data-visualisations exportables et intégrables

4. Explorer les résultats affichés

L'usager **personnalise** le module de recherche et l'affichage des résultats :

- les résultats s'affichent en miniature ou en liste, ordonnés selon un critère, par pages…
- l'usager peut sélectionner plusieurs jeux de données parmi les résultats afin de créer des paniers en vue de leur extraction ;
- il prévisualise le contenu d'un jeu de données (vue partielle),
- il visualise intégralement et il explore un jeu de données sans devoir le télécharger.

Besoins

Niveau 1 (Essentiel) :

- paramétrisation de l'affichage des résultats ;;

Niveau 2 (Important) :

- prévisualisation de formats de données simples ;

Niveau 3 (Utile) :

- prévisualisation de formats de données avancées ;
- création de paniers de jeux de données ;
- visualisation du contenu intégral de formats de données simples.

5. Choisir le format de téléchargement

Face à l'hétérogénéité des données, le réutilisateur souhaite avoir le choix des formats de téléchargement pour faciliter la réutilisation.

Les moyens d'y parvenir sont :

- **Définir dans la politique d'accueil** des formats privilégiés / recommandés pour les déposants ;
- Fournir des outils de **conversion** de formats

Les versions produites par conversion doivent être explicitement marquées pour avertir du risque de perte d'information sur la donnée.

Besoins

Niveau 1 (Essentiel) :

- définition et documentation de la politique d'accueil des formats privilégiés ;

Niveau 2 (Important) :

- outils de conversion de formats courants ;

Niveau 3 (Utile) :

- outils de conversion entre un plus grand nombre de formats.

6. Extraire des données

Le réutilisateur extrait les données :

- d'une collection, sous-collection sans limitation de volume,
- d'un panier d'export créé en amont,
- de manière manuelle avec une interface simplifiée,
- de manière automatique grâce à des points d'accès : SPARQL, API, IIIF, JSON, GraphQL… .

Les extractions sont ré-utilisées dans le cadre de :

- la reproduction d'article scientifique,
- le croisement ou l'augmentation des données,
- l'opération d'autres traitements : deep learning, IA…
- en tant que matériel pour les sciences participatives,
- en tant que matériel éducationnel (Mooc, webinaire…).

Besoins

Niveau 1 (Essentiel) :

- API CRUD ;

Niveau 2 (Important) :

- endpoint SPARQL, IIIF, JSON, GraphQL
- autres APIs ;

*" C'est très important d'avoir des **API** pour récupérer des données sous toutes ses formes. [Sur cette base] on a développé un projet de **science participative**"*

7. Citer un jeu de données

Le réutilisateur cite facilement le jeu de données dans le style de son choix et communément d'usage dans sa communauté.

Il choisit entre :

- une citation minimale reprenant les **métadonnées** de description obligatoires (Dublin Core, DataCite...).
- une citation plus complète reprenant d'autres métadonnées spécifiques à la description d'un jeu de données (nature, type, version...).

Pour y parvenir, il utilise un **outil de formatage** proposant différentes normes de style. Par exemple : DOI Citation Formatter.

La citation interagit facilement avec d'autres outils bibliographiques (interopérabilité avec Zotero, Mendeley...)

Besoins

Niveau 1 (Essentiel) :

- Modèle de citation dans les styles les plus courants (APA, Chicago...) ;

Niveau 2 (Important) :

- Interopérabilité avec les logiciels bibliographiques les plus courants et ouverts (ex: Zotero)

Fonctionnalités avancées

Les participants ont évoqué des fonctionnalités avancées qui dépassent le cadre d'un entrepôt simple mais seraient séduisantes et donc susceptibles à terme d'augmenter l'adhésion des usagers :

- sauvegarde des requêtes pour les rejouer ultérieurement ;
- sauvegarde de paniers de jeux de données ;
- création de favoris ou marque-pages ;
- recommandation de jeux de données basées sur son profil auteur et son historique de recherche ;
- alertes lors de la mise à jour ou lors de nouvelles réutilisations de jeux de données suivis ;
- veille, avec alertes lors de la publication de nouveaux jeux de données sur une thématique suivie ;
- suggestions basées sur les recherches d'utilisateurs de profil similaire ;
- espace de commentaire, discussion, interaction autour des jeux de données avec les auteurs et les autres membres de la communauté ;
- fonctionnalités collaboratives : fork, pull request…
- partage sur les réseaux sociaux : Facebook, Twitter, LinkedIn, ResearchGate, Academia…
-

Fonctionnalités communes entre les parcours

Certaines fonctionnalités citées dans le parcours producteur sont pertinentes pour le réutilisateur de données :

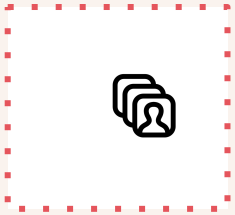
- consulter les indicateurs de diffusion et de réutilisation d'un jeu de données ;
- être alerté des nouvelles réutilisations d'un jeu de données ;

Certaines fonctionnalités citées dans le parcours réutilisateur sont pertinentes pour le producteur de données :

- pré-visualiser ou visualiser intégralement un jeu de données pour vérifier le dépôt ;
- ré-exporter un jeu de données dans un format simple.

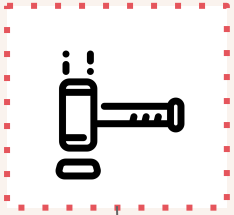
*"J'utilise Google Scholar et le Web of Science pour voir les citations sur mes papiers : qui les a cités ou réutilisés ? Cela permet de maintenir une bibliographie mais aussi de **valoriser** nos productions pour **justifier** des demandes de financement [pour] continuer à travailler dans un domaine dans lequel la communauté a montré un intérêt"*

Parcours gestionnaire
Modération et administration



Administrer les utilisateurs et leurs droits

Mesurer la qualité



Modérer

Évaluer selon des critères de qualité

Bénéficier d'un accompagnement sur les aspects juridiques

Suivre l'activité

Parcours gestionnaire : modération et administration

DATACTIVIST

1. Gérer les utilisateurs et leurs droits

Les gestionnaires attendent de l'outil une gestion administrative des droits d'éditions et la possibilité de créer des groupes d'utilisateurs.

- La gestion des **rôles**
  - administrateur,
  - développeur,
  - relecteur,
  - ...
- et des **droits associés**
  - ajouter,
  - éditer,
  - valider,
  - supprimer,
  - dé-publier.

Besoins

Niveau 1 (Essentiel) :

- Interface simple ;
- Gestion fine des rôles et droits ;

2. Mesurer la qualité de l'ensemble des jeux de données

Le gestionnaire veut disposer d'outils pour mesurer et évaluer la qualité des jeux de données.

Il a besoin d'un tableau de bord reprenant :

- des indicateurs sur les jeux déposés (FAIRness...) ;
- un circuit de modération avec différents rôles et leurs activités (relecture, modération, validation du contenu...) ;
- des alertes sur les nouveaux dépôts en vue d'une modération.

De manière générale, il ressort que les usagers privilégient la **qualité** des jeux de données plutôt que leur quantité.

Besoins

Niveau 1 (Essentiel) :

- Circuit de modération ;
- Alertes sur les nouveaux dépôts ;

Niveau 2 (Important) :

- Indicateurs sur les jeux de données

3. Modérer des jeux de données

Le gestionnaire souhaite la mise en place d'une **modération gérable** sur :

- les **métadonnées** : vérification des affiliations, de l'indexation…
- le **contenu** : la mise en conformité du format, l'innocuité du système, le degré de FAIRisation….

Pistes soulevées pour la modération à différents degrés :

- Responsabilité du déposant ouvert au dialogue ;
- Attribution du **rôle** modérateur dans l'entrepôt et une interface dédiée à cette activité : des cellules dédiées soit au niveau de l'institution soit au niveau de la communauté thématique ;
- Relecture ouverte (open peer-reviewing) avec des outils sociaux interactifs (par ex. Slack)

Besoins

Niveau 1 (Essentiel) :

- Interface de modération ;
- Contrôle d'innocuité ;

Niveau 2 (Important) :

- Indicateurs de FAIRisation
- Vérification des métadonnées par intégration de référentiels ou API

4. Évaluer selon des critères de qualité

Le gestionnaire a besoin d'indices pour évaluer la qualité générale d'un jeu.

Des indices obtenus via différentes opérations de certification telles que :

- une procédure de labellisation institutionnelle
- une évaluation par la communauté extérieure (open reviewing)
- une conformité technique :
  - FACILE pour valider les formats
  - REACH Registration, Evaluation and Authorisation of Chemical...
  - ...

Des indices variables selon la nature du jeu et son contexte de production.

Besoins

Niveau 1 (Essentiel) :

- Définition de labels ;

Niveau 2 (Important) :

- Mécanisme d'évaluation ouverte ;
- Vérification de la conformité

5. Bénéficier d'un accompagnement sur les aspects juridiques

Le gestionnaire souhaite être accompagné pour être en mesure d'aider a minima les différents acteurs :

- **le producteur de donnée** sur l'attribution d'une licence, les enjeux de la propriété intellectuelle, l'accord de tous les autres producteurs pour le dépôt…
- **d'autres gestionnaires** sur les principes d'anonymisation, le respect du cadre légal de son institution, le respect des données personnelles…
- **le réutilisateur** sur la compréhension des différentes licences et l'incitation à la réutilisation…

Le gestionnaire souhaite acquérir ces nouvelles compétences ou bénéficier d'un service d'appui juridique.

La bonne gestion juridique des jeux traités est un gage de qualité supplémentaire lors de son dépôt.

Besoins

Niveau 1 (Essentiel) :

- Accompagnement juridique ;
- Accompagnement technique ;
- Mise en réseau des gestionnaires

6. Suivre l'activité

Le gestionnaire peut suivre son activité et rendre des comptes à son institution sur l'activité globale de ses membres.

Il accède à un **tableau de bord** avec ces fonctionnalités attendues :

- **traçage** de la réutilisation des jeux de données ;
- **indicateurs statistiques** ;
- **vue globale et paramétrable** de l'état d'une collection / d'une institution ;
- des indicateurs exportables et affichés sous formes de **tableaux** et de **graphiques** ;
- des **indicateurs bibliométriques** (nombre de téléchargements, réutilisations...) et **altmetrics ;**
- des observations **exportables**.

Le choix des indicateurs peut se baser sur la **recommandation COUNTER** for Research Data.

Besoins

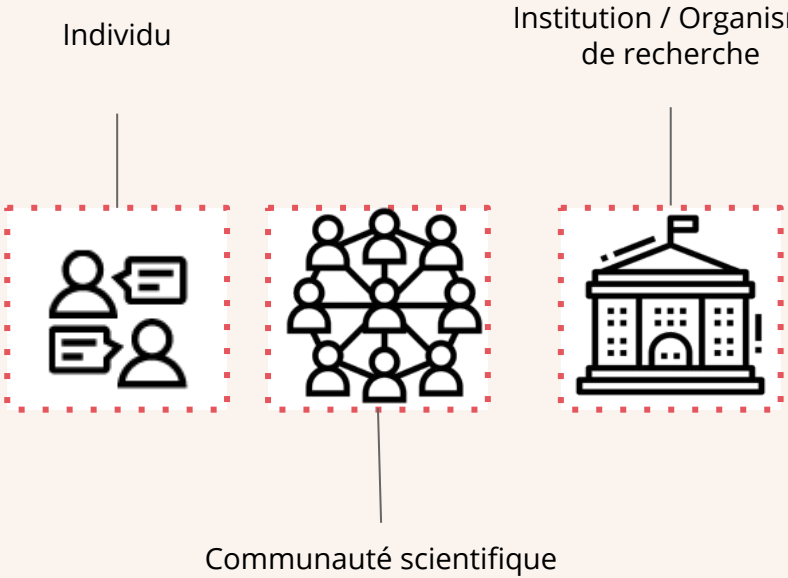
Niveau 1 (Essentiel) :

- Tableaux de bord institutionnels ;
- Vue consolidée conforme à la recommandation COUNTER for Research Data ;

Niveau 2 (Important) :

- Export des indicateurs et statistiques ;

Un besoin pour tous les acteurs
L'accompagnement



Un besoin pour tous les acteurs : l'accompagnement

Importance de l'accompagnement

Les participants souhaitent bénéficier d'aides pour s'approprier de manière optimale le nouvel outil.

Un accompagnement :

- **Adapté à la diversité des métiers** impliqués : *data librarian*, enseignant chercheur, ingénieur de recherche, directeur de recherche, doctorant, informaticien…
- **Abordant** les fonctionnalités de l'outil et ses enjeux : du dépôt à la fairisation des données, la modération…
- **Multi-forme** : sur un support adéquat selon l'audience concernée : recommandations techniques, tutoriels, aides contextuelles, FAQ, guides pratiques, cellules d'appui, d'ateliers de sensibilisation, réseaux d'experts, collaborations…

Il s'agit d'un facteur clé de succès dans ce projet.

Accompagner l'individu sur l'utilisation

Le déposant a des fortes attentes dans un système d'accompagnement à **échelle humaine** lui facilitant la gestion et la diffusion des données dans le cadre de son activité de recherche pour le délester d'une charge de travail administrative.

L'usager souhaite :

- des **outils facilitateurs** : des suggestions automatiques, la récupération des données…
- de la **documentation** claire et accessible (tutoriel, guide pratique…)
- des **interlocuteurs directs** ou des **relais** au sein de leur organisme de recherche sur les questions relatives :
  - au dépôt
  - à la réutilisation
  - à la gestion et la diffusion des données de la recherche.

L'usager attend que les interlocuteurs possèdent un **bagage professionnel** minimum dans sa **discipline** pour faciliter le dialogue.

*" La trame doit être mise en place facilement pour un accompagnement initial puis **ponctuel au fil du projet**. Il y a des métiers à créer dans ce cadre. Certaines compétences (juridiques et IST) existent déjà mais il reste à les **mettre en connexion.**"*

Accompagner la communauté scientifique sur l'adhésion

La communauté scientifique a des attentes sur la **conduite au changement** pour adhérer à un nouvel outil et s'inscrire dans le mouvement de la science ouverte.

Elle souhaite un accompagnement adapté à leurs pratiques et usages déjà en vigueur au sein de leur discipline thématique.

Elle souhaite :

- une cellule d'accompagnement : un **aiguillage** au niveau des institutions;
- des **réseaux d'experts** pour échanger les pratiques;
- formaliser et standardiser certaines pratiques (notamment sur les schéma de métadonnées) pour faciliter le partage des jeux de données;
- des **recommandations** adaptées à leur spécificité;
- des actions de **sensibilisation** sur la gestion, la diffusion et la réutilisation des données dans leur discipline.

*" Certaines communautés sont perdues face à un type de données. L'idée serait d'imaginer un **aiguillage** et de leur proposer des solutions. L'institution locale ou nationale est là pour leur dire qu'en structurant leurs données de telle manière, ils pourront déposer dans tel entrepôt."*

Accompagner l'institution / l'organisme de recherche sur le déploiement

Les institutions ont besoin de soutien et d'accompagnement sur les spécificités suivantes :

- un **positionnement** de l'outil national pour l'intégrer dans sa feuille de route.
- une **aide technique** et **humaine** sur
  - le déploiement de l'outil et son interopérabilité avec d'autres systèmes présents,
  - le développement de fonctionnalités spécifiques : module de téléversement, modération, tableaux de bord...
  - les aspects juridiques liés aux données.
- une démarche de **sensibilisation** sur les enjeux de la science ouverte auprès de l'institution et de ses composantes (aide sur la conduite du changement)
- un **réseau** d'experts : administrateur de la donnée, Groupe de Travail au sein du CoSO..; sur lequel s'appuyer.

> *" S'il y a un entrepôt de données, il faut [des personnes] qui accompagnent et dialoguent [...], qui fassent **tourner le moteur**."*

Conclusion

- Les usagers ont des attentes élevées sur trois grands principes : **confiance**, **qualité** et **responsabilité** de l'entrepôt.
  Les réponses données à ces attentes à travers la mise en œuvre de l'entrepôt sont des facteurs très forts d'adhésion.

- Les besoins techniques sont nombreux et variés, selon les usagers, les communautés disciplinaires, les établissements.
  Les fonctionnalités essentielles correspondent globalement au périmètre des solutions logicielles d'entrepôt de données et d'archives ouvertes que les usagers utilisent déjà.
  Les fonctionnalités importantes et utiles correspondent globalement à des fonctionnalités que les usagers trouvent dans d'autres outils, et qu'ils aimeraient trouver dans une nouvelle offre d'entrepôt de données.

- Au-delà des besoins techniques et fonctionnels, les usagers voient **l'accompagnement** (des usagers individuels, des communautés scientifiques et des établissements) comme une **condition indispensable à l'adoption et au succès de l'entrepôt**.