Gestion des Données Ouvertes en Agriculture et Nutrition

Ce cours en ligne est le fruit d'une collaboration entre les partenaires de GODAN Action, y compris Wageningen Environmental Research (WUR), AgroKnow, AidData, l'Organisation des Nations Unies pour l'Alimentation et l'Agriculture (FAO), le Forum Mondial sur la Recherche Agricole (GFAR), l'Institut des Etudes du Développement (IDS), le Land Portal, l'Open Data Institute (ODI) et le Centre Technique de Coopération Agricole et Rurale (CTA).

GODAN Action est un projet de trois ans du Département pour le Développement International du Royaume-Uni pour permettre aux utilisateurs, producteurs et intermédiaires de données de s'engager efficacement avec les données ouvertes et maximiser leur potentiel d'impact dans les secteurs de l'agriculture et de l'alimentation. Nous travaillons en particulier à renforcer les capacités, à promouvoir des normes communes et les meilleures pratiques et à améliorer la manière dont nous mesurons l'impact. [www.godan.info]

MODULE 3 : OUVERTURE DES DONNÉES

LEÇON 3.3: Création et gestion des bases de données ouvertes

Objectifs et résultats d'apprentissage

Cette leçon a pour objectif de ;

- Donner un aperçu de la façon de créer un référentiel de données
- Ouvertes et expliquer les exigences relatives à sa gestion.
- Décrire l'importance et les différentes composantes de la curation des données

À la fin de cette leçon, vous devrez être en mesure ;

- D'identifier les étapes à prendre en compte lors de la création d'une Base de données ouvertes
- De sélectionner l'outil logiciel approprié pour la mise en place d'un référentiel en fonction de ses besoins
- De comprendre les exigences d'une gestion efficace d'une base de données ouvert
- De comprendre les différents éléments de la curation des données

Sommaire

Liste des illustrations

Liste des tableaux

1. Introduction

Dans la leçon 3.2, 'Gestion des ensembles de données dynamiques', nous avons examiné les diverses options de publication des données dynamiques et nous avons décrit les processus automatisés pour faciliter cette publication. Dans cette leçon, nous explorons les différentes options disponibles pour héberger un référentiel de données ouvertes. Ces solutions vont de l'hébergement commercial aux solutions libres et open source.

2. Création d'un référentiel de données ouvertes

En général, un référentiel de données est la pièce maîtresse d'un effort ouvert en matière de données et sert de lieu central pour trouver des données, de lieu de normalisation des pratiques et d'exemple de l'utilisation de ces données. D'un point de vue pratique, un dépôt sert de lieu central et consultable où les gens peuvent trouver des données. Certains logiciels de référentiel convertissent automatiquement les données d'un format à un autre, de sorte que même si vous ne pouvez fournir des données que dans un seul format (par exemple, CSV), ils génèrent XML, JSON, Excel, etc.

Certains logiciels permettent de visualiser des ensembles de données dans le navigateur, ce qui permet aux gens de les cartographier, de les trier, de les rechercher et de les combiner, sans qu'il soit nécessaire de connaître la programmation.

Certains logiciels de dépôt permettent la syndication, ce qui permet à d'autres organisations d'incorporer automatiquement vos propres données (par exemple, un organisme de transport d'État pourrait recueillir les données sur le transport de toutes les localités et les publier à nouveau).

En général, le logiciel de référentiel prend en charge soit le téléchargement de fichiers à stocker dans le référentiel, soit le pointage du référentiel vers une adresse de site Web existante où le fichier réside. La première fonctionne bien pour les petites organisations (localités, petites agences), tandis que la seconde fonctionne bien pour les grandes organisations ou les gouvernements, pour lesquels la centralisation des actifs peut être peu pratique.

Grosso modo, il y a trois types d'hébergement disponibles : l'hébergement commercial, l'hébergement autonome et l'hébergement gratuit. Le tableau 1 ci-dessous présente une liste de dépôts de données disponibles.

Tableau 1 Aperçu des référentiels de données

Nom	Type	Notes
ArcGIS Open Data	hébergé	
CKAN	open source	
DataHub	gratuit, hébergé	CKAN-powered
DKAN	open source	
GitHub	gratuit, hébergé	
JKAN	open source	
Junar	hébergé	
NuData	hébergé	
OpenData.city	gratuit & hébergé	CKAN-powered
OpenDataSoft	hébergé	
Open Data Catalog	open source	
Socrata	hébergé	

2.1. Établir les besoins et les exigences des utilisateurs

Les études qui ont examiné les causes de l'échec d'un projet de TI ont révélé que la " définition des exigences " est l'étape du projet la plus fréquemment citée qui a causé l'échec. Il est donc essentiel de comprendre ce que le dépôt devrait faire et de s'assurer qu'il répond aux besoins des intervenants pour assurer le succès du projet.

Identification des besoins:

- Comprendre les utilisateurs, ainsi que leur travail et son contexte.
- Le système en cours d'élaboration devrait aider les utilisateurs à atteindre leurs objectifs.
- L'identification des besoins est cruciale pour la prochaine étape.

Établir les exigences:

- En se basant sur les besoins identifiés, produire un ensemble d'exigences.
- Utiliser une approche du développement centrée sur l'utilisateur :
- Comprendre ce que le produit doit faire et s'assurer qu'il répond aux besoins des parties prenantes est absolument essentiel à son succès.

Quelles sont les exigences ?

Une exigence est un énoncé qui précise ce qu'un produit prévu doit faire ou comment il doit fonctionner. Traditionellement, on note deux types d'exigences :

- Les exigences *fonctionnelles* qui précisent ce que le système doit faire
- Les exigences *non fonctionnelles* qui précisent les contraintes qui pèsent sur le système ou son développement.

Le tableau 2 donne une liste plus élargie de catégories d'exigences.

Tableau 2 Aperçu des exigences

Catégories d'exigences	Description
Exigences fonctionnelles	Ce que le produit doit faire.
Exigences en matière de données	Le type, la volatilité, la taille/le montant, la persistance, l'exactitude et la valeur des quantités des données requises.
Exigences environnementales	Ou " contexte d'utilisation " - circonstances dans lesquelles le produit interactif doit fonctionner
Exigences des utilisateurs	Caractéristiques du groupe d'utilisateurs visé.
Exigences en matière d'utilisabilité	Les objectifs d'utilisabilité et les mesures associées.

En fonction des besoins et des services du dépôt, les institutions voudront ensuite évaluer les plateformes logicielles disponibles. Trois types d'options sont disponibles:

- **Logiciels libres :** Le logiciel est gratuit à télécharger, mais sa mise en œuvre et sa maintenance nécessitent généralement un certain niveau d'expertise. Un organe directeur central gère le code source, mais il est ouvert aux changements et aux améliorations de la communauté de développement (par exemple, CKAN, DKAN, JKAN).

- **Logiciels commerciaux :** Vous payez généralement le logiciel et, facultativement, les frais d'abonnement ou de consultation supplémentaires. Vous êtes propriétaire de l'utilisation du logiciel et, avec un abonnement, vous obtenez des mises à jour logicielles. Avec une interface de programmation, ou API, vous pouvez personnaliser le logiciel, mais le fournisseur du logiciel possède, crée et maintient le code source.

- **Modèle de service logiciel :** Un fournisseur de logiciels possède et distribue une plate-forme logicielle ou héberge et gère vos données pour vous. Dans ce modèle, le fournisseur de logiciels fournit des services supplémentaires contre rémunération et contrôle et met à jour le code source du logiciel (par exemple EPrints Services, Open Repository et bepress).

Les responsables de la mise en œuvre voudront choisir le logiciel qui correspond le mieux à leurs besoins et aux ressources disponibles (budget et personnel). Par exemple, les institutions qui n'ont pas une grande expertise technique peuvent vouloir examiner certains des services commerciaux disponibles. En termes de plates-formes logicielles libres, chacune d'entre elles possède ses propres forces.

2.2. Sélection du logiciel de référentiel

2.2.1. Logiciels libres

Il existe d'excellents programmes de référentiels de données libres qui sont des options solides pour les organisations techniquement averties, pour les organisations qui s'engagent à utiliser un logiciel libre, ou pour les organisations ayant le budget nécessaire pour engager un consultant pour déployer le logiciel.

CKAN: CKAN[1] est nominalement un acronyme de " Comprehensive Knowledge Archive "(Archives de la connaissance complète). Mais on ne l'appelle jamais que CKAN. CKAN, une création d'Open Knowledge basée au Royaume-Uni[2], est le logiciel de référentiel de données libre le plus couramment utilisé. Il est écrit en Python, et est le porte-étendard du logiciel de référentiel. Malheureusement, il est également connu pour être difficile à installer, bien que les images de Docker[3] l'aient considérablement simplifié.

Les utilisateurs de CKAN incluent notamment Data.gov[4], et la National Oceanic and Atmospheric Administration[5]. Les consultants de CKAN comprennent Open

[1] https://ckan.org
[2] https://okfn.org
[3] https://hub.docker.com/u/ckan/
[4] https://www.data.gov
[5] https://data.noaa.gov/dataset

Knowledge, Ontodia [6], et Accela [7], en plus de nombreux consultants indépendants. Les hôtes rémunérés de CKAN comprennent Open Knowledge et Ontodia. Veuillez consulter un site de démonstration de CKAN[8].

DKAN: DKAN[9] est un clone de CKAN, bien qu'il ne partage aucun code avec CKAN - il a été réécrit en PHP, comme un module Drupal. Pour une organisation qui utilise le système de gestion de contenu Drupal et qui souhaite également un référentiel de données, DKAN est une option particulièrement intéressante. Parmi les utilisateurs du DKAN, on peut mentionner entre autres le Département de l'agriculture des États-Unis[10],[11]. Veuillez consulter un site de démonstration DKAN[12].

JKAN: JKAN[13] est théoriquement basé sur CKAN, bien qu'il ne partage aucun code avec lui. JKAN a été créé par Tim Wisniewski[14], directeur des données de Philadelphie, comme un catalogue de données alimenté par Jekyll[15]. Notez que JKAN est un *catalogue* de données, pas un *référentiel*, c'est-à-dire qu'il stocke des liens vers des données et des métadonnées sur ces données, mais pas les données elles-mêmes. Les données peuvent être hébergées sur un serveur FTP, sur des sites Web d'agences, dans Amazon S3, dans Dropbox, ou n'importe quel endroit où l'on peut stocker un fichier pour accès public. La mise en place d'un site ne prend que quelques minutes. Veuillez consulter un site de démonstration de JKAN[16].

2.2.2. Hébergement commercial

Pour certaines organisations, l'hébergement commercial sera une option viable. Payer quelqu'un pour héberger vos propres données nécessite peu ou pas de connaissances techniques de la part de votre organisation, et l'hôte vous accompagnera tout au long du processus. Votre organisation n'aura pas à fournir d'infrastructure technique (par exemple, des serveurs) ni à savoir comment programmer. Il est toutefois important que vous examiniez attentivement les accords de niveau de service.

---

[6] https://opengov.com/open-data
[7] https://www.accela.com
[8] https://demo.ckan.org/pt_BR/
[9] http://www.nucivic.com/dkan/
[10] https://data.nal.usda.gov
[11] https://github.com/NuCivic/dkan-sites
[12] http://demo.getdkan.com
[13] https://jkan.io
[14] https://usopendata.org/2016/03/28/jkan/
[15] https://jekyllrb.com
[16] https://demo.jkan.io

ArcGIS Open Data: ArcGIS Open Data [17] est un nouvel arrivant dans le domaine, ayant été lancé fin 2014. ArcGIS Open Data est inclus dans un contrat ArcGIS Online - en raison de l'universalité de ce service parmi les municipalités et les états, il est effectivement gratuit pour ces clients existants. Cela en fait une option très attrayante pour les gouvernements ayant un faible niveau d'adhésion à un programme de données ouvertes, car elle élimine le coût d'un catalogue de données. ArcGIS Open Data n'est disponible qu'en tant que logiciel hébergé - il n'est pas possible d'en exécuter une copie sur vos propres serveurs.

Junar: Junar[18] fournit des plates-formes et des progiciels pour les entreprises, les gouvernements, les ONG et les universités, en mettant l'accent sur la collecte et l'analyse des données. Junar est bilingue, supportant les publics anglais et espagnol. Leurs prix s'adressent aux petites et moyennes entreprises, à partir de 10 000 $US environ. Le site de démonstration de Junar est disponible sur demande.

NuCivic Data: NuCivic Data[19] est basé sur DKAN, qui a été créé et est maintenu par nücivic. Il s'agit d'un fournisseur de milieu de gamme, en termes de prix - leurs tarifs sont beaucoup plus bas que ceux de socrata, mais plus chers, par exemple, que ceux de Junar.

CivicDashboards: La société de conseil en données ouvertes Ontodia fournit CKAN hébergé sous la bannière CivicDashboards [20]. Ils offrent un niveau gratuit, pour stocker un petit nombre d'ensembles de données. Leur prix est comparable à celui de Junar.

OpenDataSoft: OpenDataSoft[21] est une société française qui s'est récemment implantée sur le marché américain. Ils offrent un niveau gratuit (jusqu'à 5 ensembles de données, chacun d'un maximum de 20.000 enregistrements).

Socrata Open Data: Socrata [22] est le principal fournisseur de l'espace de stockage de données ouvertes, avec sa plate-forme Socrata Open Data. Socrata n'offre que des options hébergées - il n'y a aucun moyen d'exécuter le logiciel Socrata sur vos propres serveurs. C'est à la fois l'option la plus riche en fonctionnalités et la plus coûteuse, avec des plans se chiffrant à des centaines de milliers de dollars par an.

[17] https://hub.arcgis.com/pages/open-data
[18] http://www.junar.com
[19] https://getdkan.org
[20] http://www.civicdashboards.com
[21] https://www.opendatasoft.com
[22] https://socrata.com

2.2.3. Hébergement gratuit

Il y a quelques options disponibles pour l'hébergement gratuit de dépôts de données ouvertes. (Notez que les options libres sont également gratuites, mais nécessitent une installation, un serveur et du temps de maintenance). En général, il s'agit du niveau de service le plus bas fourni par les hôtes commerciaux.

DataHub: L'Open Knowledge Foundation fournit DataHub [23], un hôte de données CKAN gratuit. Il s'agit d'un grand dépôt collectif - les utilisateurs n'ont pas leur propre site, bien qu'il soit possible de n'inscrire que leurs propres données et de partager une adresse URL qui ne contient que ces ensembles de données.

GitHub: GitHub[24] n'est pas vraiment conçu comme un référentiel de données, mais il peut servir comme tel. Il n'a aucun des avantages d'un bon logiciel de référentiel (conversion de formats, récupération de données à partir d'URL distantes, etc.), mais il offre un aperçu de certains types de données, un suivi public des changements et constitue un endroit raisonnable pour stocker des ensembles de données.

Il offre un avantage significatif, qui est que GitHub - contrairement à tout autre logiciel de référentiel - fournit un mécanisme permettant aux personnes de proposer des modifications à vos ensembles de données, que vous pouvez accepter ou refuser, si elles détectent des erreurs ou des domaines à améliorer.

JKAN on GitHub: JKAN [25] est conçu pour être déployé sur GitHub, où le catalogue de données résultant peut être hébergé gratuitement. De cette façon, GitHub peut servir d'hébergeur gratuit sans sacrifier les subtilités d'un catalogue de données.

3. Gestion d'un référentiel de données ouvert

3.1. Interopérabilité et échange de données

Selon les principes 'FAIR', les données doivent être " trouvables, accessibles, interopérables et ré utilisables ". Les principes de FAIR Data servent de lignes directrices internationales pour une gestion des données de haute qualité. Une couverture plus approfondie de l'échange de données et des meilleures pratiques est abordée dans le module 4 : Échange de données ouvertes.

[23] http://datahub.io
[24] https://github.com
[25] https://how-to.usopendata.org/en/latest/The-Basics-of-Open-Data/DataRepositories/#jkan

Etre trouvables :

- les (méta)données se voient attribuer un identificateur globalement unique et éternellement persistant
- Les données sont décrites à l'aide de métadonnées riches
- Les (méta)données sont enregistrées ou indexées dans une ressource consultable
- Les métadonnées spécifient l'identificateur de données.

Etre accessibles :

- Les (méta)données sont récupérables de par leur identificateur à l'aide d'un protocole de communication standardisé
- Le protocole est ouvert, libre et universellement applicable
- Le protocole prévoit une procédure d'authentification et d'autorisation, le cas échéant
- Les métadonnées sont accessibles, même lorsque les données ne sont plus disponibles.

Etre interopérables:

- Les (méta)données utilisent un langage formel, accessible, partagé et largement applicable pour la représentation des connaissances
- Les (méta)données utilisent des vocabulaires qui suivent les principes de FAIR
- Les (méta)données comprennent des références qualifiées à d'autres (méta)données.

Etre ré-utilisables :

- Les méta(données) ont une pluralité d'attributs précis et pertinents
- Les (méta)données sont publiées avec une licence d'utilisation de données claire et accessible
- Les (méta)données sont associées à leur provenance
- Les (méta)données satisfont aux normes communautaires.

3.1.1. Publication des métadonnées

La nécessité de disposer de métadonnées claires, cohérentes et utilisables n'est pas nouvelle, mais représente toujours un défi pour de nombreux ensembles de données. L'exactitude des métadonnées est vitale non seulement pour leur facilité de repérage, mais aussi parce que le catalogage - de mauvaises métadonnées peuvent miner le référentiel lui-même. Koesten et al.[26] ont exploré les besoins de 20 professionnels des données en matière de recherche de données et de création de sens, y compris les aspects requis pour décider si un ensemble de données est pertinent ou non.

[26] Koesten, L, Kacprzak, E, Tennison, J and Simperl, E (2017) Trials and Tribulations of Working with Structured Data - a Study on Information Seeking Behaviour *CHI '17 Proceedings of the 2017 ACM SIGCHI Conference on Human Factors in Computing Systems* ACM, New York, USA. http://dx.doi.org/10.1145/3025453.3025838

Ils font la distinction entre trois dimensions : la pertinence (s'agit-il des données dont j'ai besoin ?), la facilité d'utilisation (puis-je les utiliser dans la pratique ?) et la qualité (dans quelle mesure les données sont-elles bonnes et faciles à utiliser ?). Les données doivent être accompagnées de descriptions de ces aspects, soit sous forme de métadonnées structurées, mais aussi sous forme de commentaires, d'études de cas, de rapports d'expérience, d'exemples d'utilisation, etc. Voir le tableau 3 ci-dessous:

Tableau 3 Les utilisateurs de métadonnées jugent utile de donner un sens aux données

Evaluation	Renseignements nécessaires sur
Pertinence	Contexte, couverture, but original, granularité, résumé, échéancier
Utilisabilité	Étiquetage, documentation, licence, accès, lisibilité par machine, langue utilisée, format, schéma, capacité de partage, etc.
Qualité	Méthodes de collecte, provenance, uniformité des données, etc. formatage/étiquetage, exhaustivité, ce qui a été exclu

3.2. Considérations relatives à la gouvernance

Une transparence ainsi qu'une gouvernance responsable sont essentielles dans une proposition de valeur des données ouvertes pour l'agriculture et la nutrition. Il est important de considérer quelles données ouvrir et comment. Dans quelle mesure les détenteurs de données doivent-ils rendre des comptes à la fois au côté de la demande et aux décideurs ? Comment les producteurs de données et les acteurs assurent-ils la qualité des données ? Qui sont les responsables chargés de rendre les données ouvertes ?[26]

Différents pays ou organisations auront différents modèles pour gouverner et administrer leurs activités, c'est-à-dire différents modèles de gouvernance. Par exemple, vous constaterez que certains pays sont plus décentralisés dans leur prise de décision, tandis que d'autres ont une administration publique plus centralisée. Ces modèles de gouvernance ont clairement un impact sur la façon dont les données ouvertes sont gouvernées - fournissant une large mosaïque de différentes formes de gouvernance des données ouvertes à travers le monde et rendant difficile l'identification des décideurs et des responsables du contrôle des données dans un pays donné.

Par exemple, si l'on veut accélérer l'ouverture des données agricoles, cela peut relever de l'autorité des gouvernements infranationaux (États, provinces, territoires ou même villes), alors que dans d'autres pays, l'agriculture est régie

[26] Brandusescu, A., Lämmerhirt, D. AND Verhulst, S. (2017) Mapping open data governance models: Who makes decisions about government data and how? https://blog.okfn.org/2017/02/16/mapping-open-data-governance-models-who-decidesand-how/

par le gouvernement central ou mise en œuvre par des partenariats public-privé. Les données du gouvernement peuvent être privatisées, alors que dans d'autres cas, elles peuvent relever de la responsabilité des autorités municipales ou régionales. Les responsabilités sont donc souvent réparties entre les niveaux administratifs et les organismes, ce qui influe sur la manière dont les données (ouvertes) sont produites et publiées.

Lorsque vous envisagez le modèle de gouvernance pour votre initiative de données ouvertes, dressez la carte du processus et de l'écosystème de gouvernance des données ouvertes en identifiant les principaux intervenants suivants, leurs rôles et responsabilités dans l'administration des données ouvertes, et en cherchant comment ils sont connectés :

- *Les décideurs* - qui dirigent/affirment le pouvoir de décision sur les données publiques dans les réunions, les procédures, la conduite, les débats, le vote et d'autres questions ?
- *Les détenteurs de données* - quelles organisations/organes gouvernementaux gèrent et administrent les données agricoles ?
- *Les producteurs de données* - quelles organisations/organes gouvernementaux produisent quel type de données agricoles et nutritionnelles ?
- *Les acteurs de l'assurance de la qualité des données* - qui sont les acteurs qui veillent à ce que les données produites respectent certaines normes de qualité ?
- *Les responsables/gardiens du contrôle des données* - qui contrôle la publication ouverte des données ?

3.3. Conservation et préservation

La conservation et la préservation des données combinent des politiques, des stratégies et des actions pour assurer le rendu le plus précis possible des données au fil du temps, quels que soient les défis envers la corruption des fichiers, de la défaillance des médias et des changements technologiques. La préservation numérique s'applique au contenu qui naît numérique ou qui est converti sous forme numérique.

Une fois la sélection du référentiel effectuée, la publication des données dans le référentiel peut s'avérer difficile en raison de plusieurs problèmes. La première question à traiter est la sélection des données à publier. Toutes les données ont-elles de la valeur, ou devrait-on seulement en fournir une sélection ? Étant donné que les ressources sont limitées, une forme de sélection des données doit être faite, mais quels critères doivent être utilisés pour identifier les données qui peuvent être publiées ? Toutes ces considérations sont articulées dans des politiques et des flux de travail qui, en réalité, sont en constante évolution.

Celles-ci seront développées en détail dans la leçon 3.5. L'objectif n'est donc pas d'établir des pratiques et des responsabilités précises, mais plutôt de documenter les politiques nouvelles et existantes dans le but de suivre les changements à mesure qu'ils surviennent.

Les personnes qui ont le rôle principal de gérer ou de " s'occuper " des données se présentent sous des formes très diverses [27] . Leurs titres de fonctions comprennent :

- conservateurs de données
- archivistes
- bibliothécaires
- bibliothécaires de données
- annotateurs.

Par exemple, les tâches d'un conservateur de données dans le contexte des biosciences peuvent inclure la gestion continue des données, la description intensive des données, l'assurance de la qualité des données, le travail d'infrastructure d'information en collaboration et le travail de normalisation des métadonnées. L'éventail complet des tâches et des responsabilités liées à la conservation des données pourrait ressembler à ce qui suit :

- Élaborer et mettre en œuvre des politiques et des services
- Analyser le contenu numérique pour déterminer quels services peuvent en être tirés
- Conseils aux créateurs de données et aux utilisateurs/ré utilisateurs
- Assurer la soumission des données à un dépôt d'archives
- négocier des accords
- Contrôle de la qualité des données
- Veiller à ce que les données soient structurées de la meilleure façon possible pour permettre l'accès, le rendu, le stockage et la maintenance
- Permettre l'utilisation et la ré-utilisation des données
- Trouver et récupérer des données grâce à la découverte et à l'extraction de données
- Planifier et mettre en œuvre la préservation (par exemple, assurer une routine de stockage et de sauvegarde appropriée, surveiller l'obsolescence)
- Veiller à ce que des politiques et des services soient en place pour s'assurer que les données soient viables, susceptibles d'être rendues, compréhensibles et authentiques
- promouvoir l'interopérabilité.

Le **modèle de cycle de vie de la conservation de la DCC [28]** (voir l'illustration 1) fournit une vue de haut niveau des étapes nécessaires à la réussite de la conservation et de la préservation des données depuis leur conception jusqu'à

[27] DCC – Digital Curation 101: Curate and Preserve. http://www.dcc.ac.uk/sites/default/files/documents/DC%20101%20Curation%20and%20Preservation.pdf

[28] http://www.dcc.ac.uk/sites/default/files/documents/publications/DCCLifecycle.pdf

leur réception. Le modèle peut être utilisé pour planifier les activités au sein d'une organisation ou d'un consortium afin de s'assurer que toutes les étapes nécessaires sont entreprises, chacune dans le bon ordre. Le modèle permet de mettre en correspondance des fonctionnalités granulaires, de définir les rôles et les responsabilités, et de construire un cadre de normes et de technologies à mettre en œuvre. Il peut aider à identifier les étapes supplémentaires qui peuvent être nécessaires, ou les actions qui ne sont pas requises par certaines situations ou disciplines, et à s'assurer que les processus et les politiques sont adéquatement documentés.

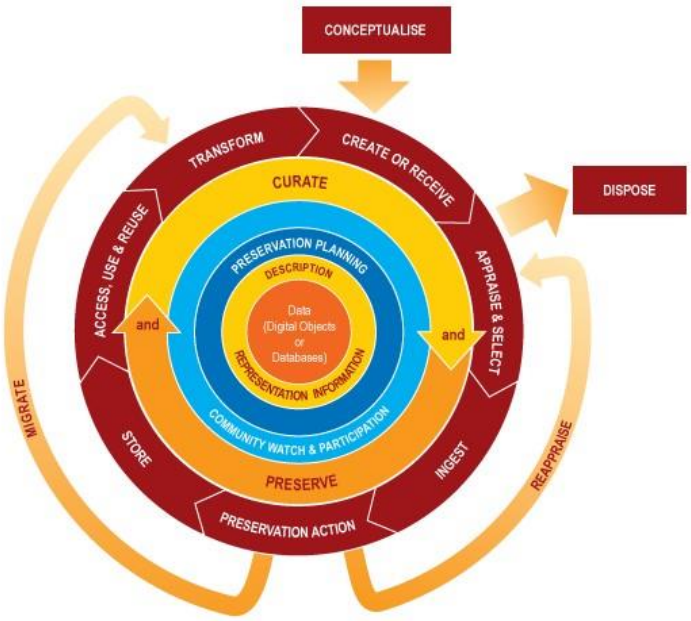

Illustration 1 Le modèle de cycle de vie de la conservation de la DCC

Tableau 4 Mesures prises dans le cadre du modèle de cycle de vie de la durée de conservation de la DCC :

Actions tout au long du cycle de vie	
Description et Renseignements sur la représentation	Attribuer des métadonnées administratives, descriptives, techniques, structurelles et de préservation, en utilisant des normes appropriées, pour assurer une description et un contrôle adéquats à long terme. Recueillir et attribuer l'information de représentation nécessaire à la compréhension et au rendu du matériel numérique et des métadonnées connexes.

Planification de la préservation	Planifier la préservation tout au long du cycle de vie de la conservation du matériel numérique. Cela comprend des plans de gestion et d'administration de toutes les mesures du cycle de vie de la conservation.
Veille communautaire et participation	Surveiller les activités communautaires appropriées et participer à l'élaboration de normes, d'outils et de logiciels communs.
Conservation et préservation	Connaître et entreprendre les mesures de gestion et d'administration prévues pour promouvoir la conservation et la préservation tout au long du cycle de vie de la conservation.
Actions séquentielles	
Conceptualisation	Concevoir et planifier la création de données, y compris la méthode de capture et les options de stockage.
Création ou réception	Créez des données comprenant des métadonnées administratives, descriptives, structurelles et techniques. Des métadonnées de préservation peuvent également être ajoutées au moment de la création. Recevoir les données, conformément aux politiques de collecte documentées, des créateurs de données, d'autres archives, dépôts ou centres de données et, au besoin, attribuer les métadonnées appropriées.
Évaluation et sélection	Évaluer les données et choisir pour la conservation et la préservation à long terme. Adhérer à des directives, des politiques ou des exigences juridiques documentées.

Ingestion	Transférer les données à une archive, un dépôt, un centre de données ou un autre dépositaire. Respecter les directives, les politiques ou les exigences juridiques documentées.
Action de préservation	Prendre des mesures pour assurer la préservation et la conservation à long terme du caractère officiel des données. Les mesures de préservation devraient garantir que les données restent authentiques, fiables et utilisables tout en préservant leur intégrité. Les mesures comprennent le nettoyage et la validation des données, l'attribution de métadonnées de préservation, l'attribution d'informations de représentation et la garantie de structures de données ou de formats de fichiers acceptables.
Stockage	Stocker les données de manière sécurisée en respectant les normes en vigueur.
Accès, utilisation et ré utilisation	Veiller à ce que les données soient accessibles aux utilisateurs et aux ré utilisateurs désignés, au jour le jour. Il peut s'agir d'informations publiées et accessibles au public. Des contrôles d'accès et des procédures d'authentification robustes peuvent s'appliquer.
Transformation	Créer de nouvelles données à partir de leur version originale, par exemple par migration dans un format différent, en créant un sous-ensemble, par sélection ou interrogation, pour créer de nouveaux résultats dérivés, peut-être pour publication.

Actions ponctuelles	
Elimination	Éliminer les données qui n'ont pas été sélectionnées pour une conservation et une préservation à long terme conformément aux politiques, aux directives ou aux exigences légales documentées. En règle générale, les données peuvent être transférées à une autre archive, un autre dépôt, un autre centre de données ou un autre dépositaire. Dans certains cas, les données sont détruites. La nature des données peut, pour des raisons juridiques, nécessiter une destruction sécurisée.
Réévaluation	Renvoyer les données qui ne répondent pas aux procédures de validation en vue d'une nouvelle évaluation et d'une nouvelle sélection.
Migration	Migrer les données dans un format différent. Ceci peut être fait en fonction de l'environnement de stockage ou pour assurer l'immunité des données contre l'obsolescence du matériel ou du logiciel.

Résumé

Un référentiel de données est la pièce maîtresse d'un effort ouvert en matière de données. Il sert de lieu central pour trouver les données, de lieu de standardisation des pratiques et d'exemple d'utilisation de ces données. D'un point de vue pratique, un dépôt sert d'endroit central et consultable où les gens peuvent trouver des données.

Avant de choisir une solution pour le référentiel, considérez d'abord les besoins de votre utilisation et établissez les exigences :

- Comprendre autant que possible les utilisateurs, ainsi que leur travail et le contexte de leur travail
- le système en cours d'élaboration devrait aider les utilisateurs à atteindre leurs objectifs
- S'appuyer sur les besoins identifiés et produire un ensemble d'exigences
- Utiliser une approche du développement centrée sur l'utilisateur
- Comprendre ce que le dépôt devrait faire et s'assurer qu'il répond aux besoins des intervenants.

Lorsque vous envisagez le modèle de gouvernance des données ouvertes pour votre initiative de données ouvertes, dressez la carte du processus et de l'écosystème de gouvernance des données ouvertes en identifiant les principaux intervenants suivants, leurs rôles et responsabilités dans l'administration des données ouvertes, et en cherchant comment ils sont connectés :

- décideurs
- détenteurs de données
- producteurs de données
- Acteurs de l'assurance de la qualité des données
- Les gardiens des données et les responsables de la gestion des données.

La conservation et la préservation des données combinent des politiques, des stratégies et des actions pour assurer le rendu le plus précis possible des données au fil du temps, quels que soient les défis de la corruption des fichiers, de la défaillance des médias et des changements technologiques.

Le modèle de cycle de vie de la conservation de la DCC fournit une vue d'ensemble graphique de haut niveau des étapes nécessaires à la réussite de la conservation et de la préservation des données depuis leur conception initiale ou leur réception. Le modèle peut être utilisé pour planifier les activités au sein d'une organisation ou d'un consortium afin de s'assurer que toutes les étapes nécessaires sont entreprises, chacune dans la séquence correcte.
.