CAHIER DES CHARGES BASE DE DONNEES

Pour la réalisation du

Cloud Environnemental au Bénéfice de

L'Agriculture en Auvergne (CEBA)

Challenge 1 – CAP 20-25

Version 20190409-JM

Table des matières

1. Présentation du projet

Ce document vise à définir le périmètre de la gestion des données du Cloud Environnemental au Bénéfice de l'Agriculture en Auvergne. Nous détaillerons donc les besoins des utilisateurs ainsi que les contraintes liées à ceux-ci.

Les définitions des termes marqués d'un (*) seront en annexe.

1.1 Périmètre du projet

Pour rappel, les partenaires du CEBA sont : CRAIG, GReD, BRGM, LaMP, GEOLAB, LMGE, ICCF, PIAF, LMV, UREP, GDEC, LPC, LIMOS, TSCF, Centre Michel de l'Hospital, Unité Expérimentale Herbipôle, Fédération des Recherches en Environnement, ATHOS Environnement, Weather Measures, Céréales Vallée, UNIVEGE.

La finalité est la création d'un service d'observation numérique de l'environnement et des agroécosystèmes (eau, sol, air, biodiversité) à l'échelle du territoire auvergnat pour répondre à différentes problématiques qu'elles soient scientifiques, culturelles ou socio-économiques, au bénéfice de l'agriculture.

Le "Cloud Environnemental" incarne l'ambition de créer un « grand » observatoire de l'environnement en Auvergne, en mettant à la disposition de la communauté scientifique un environnement numérique interconnecté valorisant les données environnementales existantes ou à acquérir. Apportant des fonctionnalités en matière de stockage, gestion, sécurisation des données environnementales mesurées sur les différents dispositifs instrumentés en région Auvergne, le « Cloud environnemental » offrira des facilités en termes d'organisation et d'extraction des informations, permettant ainsi de :

- ➔ Progresser dans notre compréhension des compartiments environnementaux face au changement global (évolutions, interactions)
- ➔ Comprendre l'impact sur les agro-écosystèmes et réciproquement
- ➔ Conceptualiser les interrelations entre les compartiments pour comprendre, modéliser et prédire le comportement des agroécosystèmes.

1.2 Périmètre de la gestion des données

Le CEBA a pour but de mettre à disposition des acteurs un système d'information (SI) qui répond aux besoins listés dans le « CR des interviews du CEBA » et dans « l'analyse des besoins du CEBA ». Plus le public visé est large (Collectivité territoriale, agriculteurs, centres de recherches, entreprises, associations…), plus la quantité de données à gérer sera grande et plus sa variété sera conséquente.

Concernant le stockage en lui-même, une partie sera consacrée à l'Open Data*, c'est-à-dire à des jeux de données publics ouverts à tous, et une autre sera destinée au stockage de jeux de données protégés ou privés.

Il sera aussi nécessaire de pérenniser la donnée, que ce soit la donnée brute présente sur les serveurs de fichiers, ou bien la donnée directement stockée en base.

Le concept associé à la collecte et au stockage de données de structures hétérogènes avec des flux hétérogènes (IoT, fichier, base de données) est celui de « data lake » que nous nous proposons de mettre en œuvre à l'aide de solutions open source.

1.3 Ressources du projet

De par la nature du projet qui est l'un des livrables du challenge 1 de CAP 20-25, les ressources informatiques qui permettront la mise en place du CEBA, c'est-à-dire les serveurs de stockage de fichiers, les serveurs de base de données et le serveur WEB, seront localisées au Mésocentre Clermont-Auvergne.

Concernant les technologies utilisées pour mettre en place les bases de données, nous utiliserons des logiciels libres.

1.4 Volumétrie des données

On peut estimer que nous serons sur une volumétrie d'environ un Téra octets (1000 Go) par an hors herbiers universitaires qui nécessiterait une volumétrie d'une centaine de Téra octets. (Cf : CEBA-Analyse-des-besoins paragraphe 2.3).

Certains de nos partenaires (CRAIG, BRGM) se proposent de nous fournir leurs services à travers des Webservices, ainsi, nous n'avons pas à stocker ces données sur la partie CEBA.

2. Description des besoins

2.1 Formats de données

Pour des résultats d'analyses ou des mesures réalisées par un opérateur, les données sont disponibles dans des fichiers, sur un ordinateur de travail. La majorité de ces fichiers sont des tableurs de différents types (Excel, CSV, etc.). Pour l'instant, lorsque ces fichiers nécessitent un partage, il s'effectue par clef USB ou email. Les propriétaires de ces fichiers sont très motivés pour prendre part à l'initiative CEBA pour pouvoir plus facilement échanger leurs données. De plus les outils qui pourront être proposés par le CEBA, telle que l'utilisation de DOI* (*Digital Object Identifier*) qui permet la traçabilité de jeux de données*, les intéresse fortement. Le DOI est un mécanisme unique d'identification de ressources numériques comme un film, un rapport, des articles scientifiques, des bases de données, des logiciels, etc. Il permet de protéger, d'un point de vue scientifique, les jeux de données.

Certains projets utilisent des réseaux de capteurs. Il s'agira ici de stocker, structurer, pérenniser, et rendre disponibles les données collectées par les capteurs de différentes natures (piézomètre, humidité du sol, dendromètre, station météorologique, etc.). Le CEBA pourra partager des données qui proviennent d'Acteurs n'utilisant pas les mêmes technologies de capteurs.

Enfin, une approche de certains Acteurs nécessite le partage de fichiers volumineux tels que :

- ➔ Images : En fonction de la taille initiale de l'image il pourra être intéressant de la présenter sous forme d'une miniature
- ➔ Vidéos : Comme pour les images, les vidéos pourront être présentées sous forme d'une image unique, avant d'etre parcourue. Il faudra définir les types de fichiers vidéo qui seront supportés par le CEBA.
- ➔ Audios : Les types de fichiers audios supportés devront être définis.

Il sera envisageable, au sein du CEBA, de visualiser des éléments cartographiques.

Les types de fichiers non cités précédemment seront traités au fur et à mesure de l'apparition du besoin. Ils pourront être traités d'une façon similaire aux fichiers précédents, ou ils nécessiteront une approche spécifique.

2.2 Sources de données

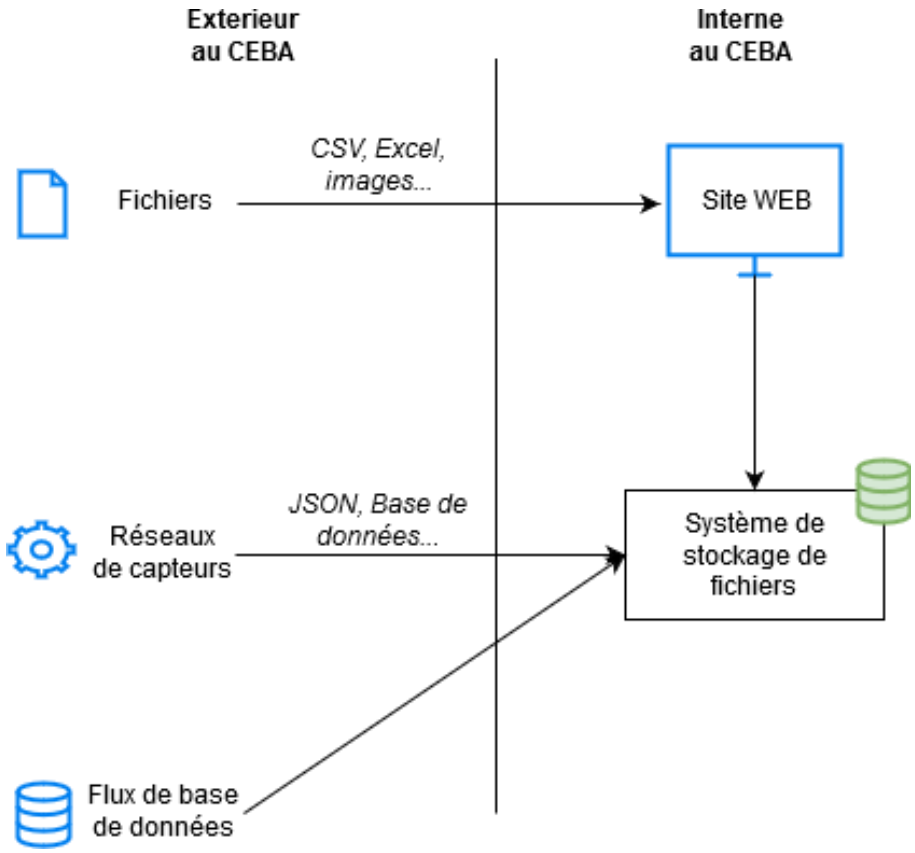

Figure 1 - Représentation des sources de données du CEBA

Comme l'illustre la *Figure 1* représentant les diverses sources de données et leurs acheminements, le CEBA sera capable de recevoir des données à partir de sources différentes. Les fichiers seront ensuite stockés sur un serveur dédié à cet usage et les données seront intégrées à la base de données (Icone verte sur le schéma).

2.3 Description des contraintes

La principale contrainte vient de la nécessité de stocker des données hétérogènes. En effet, certaines données nous parviendront directement structurées, d'autres semi-structurées. Il sera donc peut-être nécessaire de posséder à la fois une base SQL* et une base NoSQL*, pouvant communiquer l'une avec l'autre.

La seconde contrainte apparait lorsque l'on doit stocker des données privées. Il existe deux possibilités. Soit on crée un seul serveur de stockage qui hébergera les données publiques et privées, soit on crée deux serveurs de stockage, comme illustré sur le schéma ci-dessous.

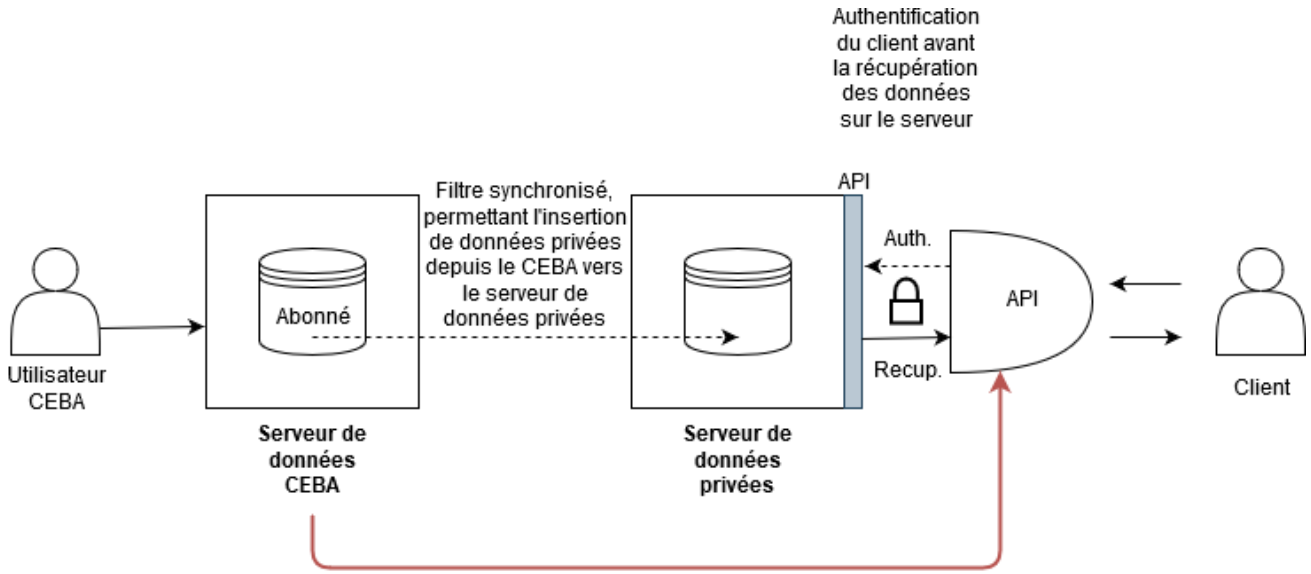

Figure 2 – Scénario de stockage de données privées dans le CEBA

La *Figure 2*, plus centré sur le stockage de données, illustre la possibilité de créer un second serveur de stockage de fichiers dédié, possédant sa propre base de données. Les clients extérieurs au CEBA pourront ainsi venir récupérer leurs données via une API développée par leur soin, depuis une interface externe. Il est important de noter que, dans le cas de la mise en place de ce genre d'infrastructure, cela mènera à de nouvelles contraintes.

2.4 Partager et publier la donnée

Dans notre infrastructure, il sera nécessaire de posséder un outil de partage et de publication des jeux de données. Cet outil pourra être transparent à la vue de l'utilisateur grâce au site WEB.

Il existe déjà une catégorie d'outil particulièrement intéressant appelé « Data Catalog* », ou catalogue de données. Il s'agit d'un emplacement centralisé où sont regroupées les informations sur les données contenues dans une base de données. Ces métadonnées associées seront de différentes natures telles que la structure, la qualité, la définition et l'utilisation.

Le principal objectif, et utilité, du catalogue de données est de permettre à tous les utilisateurs d'accéder aux jeux de données.

De plus, le propriétaire d'un jeu de données peut à tout moment enlever le partage tout public du jeu de données.

Le catalogue facilite également l'interopérabilité avec d'autres infrastructures de stockage de jeux de données, puisqu'un catalogue est capable de discuter avec d'autres infrastructures en s'appuyant sur des protocoles standards, ce qui facilite la visibilité des jeux de données.

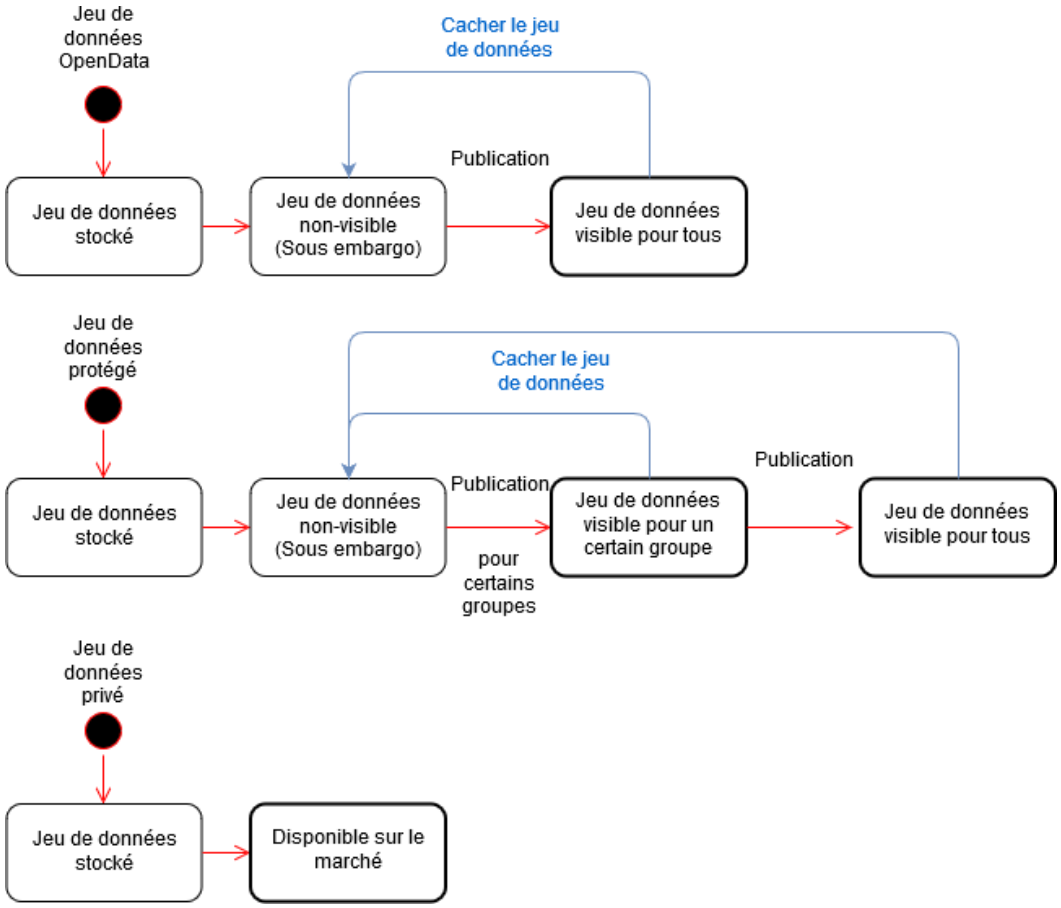

Figure 3 – Visibilité des jeux de données

La *Figure 3* décrit les différents états que pourront prendre les jeux de données au sein du CEBA. Lors de l'envoi d'un jeu de données, celui-ci se retrouvera directement stocké sur le serveur de fichiers. Si l'objectif est de le publier en OpenData ou bien seulement de le partager à quelques groupes de personnes, il sera possible de modifier sa visibilité en le publiant. Il est à noter qu'il sera possible de masquer la publication d'un jeu de données aux yeux des utilisateurs.

2.5 Aspect sécurité

Ici, la sécurité intervient sous deux aspects principaux : la donnée en elle-même afin de proposer une pérennisation efficace, et l'accès aux données.

Les serveurs de stockage posséderont des sauvegardes, c'est-à-dire des captures d'état à un instant bien précis. La fréquence de ces captures peut être amenée à être modifiée puisque le coût en espace et en maintenance augmente lorsque la fréquence augmente elle aussi.

Concernant l'accès à la donnée, des rôles seront attribués aux utilisateurs afin de différencier leurs permissions. Ces rôles seront gérés à la fois depuis le site WEB, mais aussi au sein de la base de données. Il y aura aussi un administrateur de base de données qui permettra de gérer la base de données et les rôles inclus dans celle-ci.

2.6 Réutilisation et gestion des versions

Il sera nécessaire de pouvoir ajouter une nouvelle version d'un jeu de données déjà existant tout en conservant les anciennes versions. Cela permettra de proposer un historique pour un jeu de données précis, mais aussi de pouvoir stocker le nombre de téléchargements par version de jeu de données.

Pour que cela soit efficace, il est probable que les jeux de données possédant une DOI soient impossibles à supprimer.

2.7 Propriété intellectuelle de la donnée

Différentes conventions nous lieront avec nos Acteurs et il est donc primordial de pouvoir leur garantir que nous ne trahirons pas le périmètre de distribution des données qu'ils nous confient.

Open Data signifie « Données ouvertes », ce sont des données dont l'accès, l'utilisation et la réutilisation sont publics et libres de droits.

L'Open Data se caractérise par 3 critères essentiels :

- Disponibilité et accès : les données doivent être accessibles, moyennant un coût de reproduction raisonnable, et pouvoir être téléchargées sur Internet. La forme des données doit être pratique et modifiable.
- Réutilisation et redistribution : les données doivent être fournies dans les conditions permettant leur réutilisation et leur redistribution, incluant le mélange avec d'autres ensembles de données.
- Participation universelle : tout le monde doit être en mesure d'utiliser, de réutiliser et de redistribuer les données. Il ne doit y avoir aucune discrimination à l'égard des utilisateurs (restrictions à certains secteurs par exemple) ou concernant les fins d'utilisation.

Il est possible de revendiquer la paternité des données publiées dans des archives ouvertes par le biais d'un DOI, qui lie de façon pérenne un document à son auteur.

Il existe des licences libres protégeant les propriétaires des données mais aussi les personnes hébergeant ces mêmes données.

➔ Creative Commons : Creative Commons apportent un équilibre à l'intérieur du cadre traditionnel "tous droits réservés" créé par les lois sur le droit d'auteur. Il existe différents niveaux de restrictions offrant le droit ou non à la personne copiant des données de les modifier ou non, par exemple. La citation du propriétaire dans l'utilisation des données est obligatoire.

- ➔ Open Database Licence : la licence Open Database permet à chacun d'exploiter publiquement, commercialement ou non, des bases de données; à condition néanmoins de maintenir la licence sur la base de données, et éventuellement, sur les modifications qui y sont apportées, et de mentionner expressément l'usage, s'il génère des créations à partir de celles-ci.

Pour les données produites par des acteurs privés, la politique de licence fera l'objet d'une étude spécifique.

3. Gestion des métadonnées

3.1 Nature de la métadonnée des jeux de données

Les métadonnées sont les données qui décrivent le jeu de données.

Suivant la source, les métadonnées des jeux de données seront différentes. Dans le cas de jeux de données provenant de capteurs, donc en flux continus, les métadonnées seront standardisées et potentiellement extraites de façon automatique à partir d'un fichier de configuration. Ces métadonnées répondront à minima aux questions classiques : Quoi, Quand et Où. Cependant, lors de l'ajout manuel d'un jeu de données sur la plateforme via le site WEB, il sera préférable de renseigner une liste de métadonnées nécessaire à sa publication. Cela permettra d'indexer correctement le jeu de données, et de le retrouver efficacement.

3.2 Proposition de métadonnées pour caractériser les jeux de données du CEBA

La directive INSPIRE apporte une liste de métadonnées très vaste à compléter lors de l'ajout d'un jeu de données au sein d'une plateforme de stockage. Afin d'optimiser le temps de chacun, l'inter-ZA a épuré cette fiche de métadonnées pour en ressortir l'essentiel qui conviendrait à tout le monde, tout en permettant à ce même jeu de données de rester moissonnable et disponible depuis l'extérieur du CEBA.

- **Title** : Titre du jeu de données
- **Abstract** : Description courte du jeu de données
- **Temporal_extent_name** : Précisions sur la date de la prise d'information
- **Start_date** : Date de début de la prise d'information
- **End_date** : Date de fin de la prise d'information
- **Spatial_extent_name** : Lieu de prise d'information
- **Topic_categories** : Catégories du jeu de données
- **Inspire_themes** : Thèmes INSPIRE
- **Gemet_keywords** : Mots clés « Gemet »
- **Other_keywords** : Autres mots clés représentant le jeu de données
- **Md_contact** : Informations et adresse mail de l'auteur
- **Lineage** : Remarque sur la qualité des données
- **Use_condition** : Périmètre d'utilisation du jeu de données

Il est nécessaire d'indiquer que cette liste n'est pas obligatoire pour insérer un jeu de données dans le CEBA. Les seules métadonnées obligatoires seront : Quoi, quand et Où.

Enfin, des thésaurus seront utilisés pour remplir et uniformiser ces métadonnées, tels que « Envthes » ou « Gemet ».

Des métadonnées spatiales devront donc être renseignées afin de pouvoir retrouver le jeu de données par une emprise géographique sur une carte.

La *Figure 4* illustre parfaitement cette saisie de métadonnées pour un jeu de données (il est possible de voir le résultat sur un catalogue à l'adresse indiquée en annexe).

title	abstract
Sédiments ruisseau ZATU	Collecte d'échantillons à différents niveaux (surface et premiers 5 cm) à différents points du site. Ces données ont été produite aussi dans le cadre du projet TREMPLIN. Les données contiennentla profondeur et le projet où ont été utilisés/produits les prélévements.

temporal_extent_name	start_date	end_date	spatial_extent_name
	01/12/2014	30/04/2017	Lachaux

topic_categories	inspire_themes
Informations géoscientifiques	ressources minérales--- http://inspire.ec.europa.eu/theme/mr, lieux de production et sites industriels--- http://inspire.ec.europa.eu/theme/pf

gemet_keywords	other_keywords
industrie minérale--- https://www.eionet.europa.eu/gemet/fr/concept/5268, sol contaminé--- https://www.eionet.europa.eu/gemet/fr/concept/1751, mobilisation de sédiments--- https://www.eionet.europa.eu/gemet/fr/concept/14850	

md_contact	lineage	use_condition
author=; principalInvestigator=; pointOfContact=da	Prélévement d'eau et de sol réalisé à différentes profondeurs.	This work is licensed under a Creative Commons Attribution 4.0 License (CC BY SA 4.0, https://creativecommons.org/licenses/by-sa/4.0/).

Figure 4 - Exemple de saisie de métadonnées

3.3 Scénarios d'utilisation

Dans la phase de prototypage, nous considèrerons trois applications illustrant des scénarios différents de génération, stockage et exploitation des données :

- L'Observatoire de l'Allier :

Les données envoyées par l'observatoire seront majoritairement des données publiques, qui seront accessibles en Open Data. Les jeux de données devront au minimum répondre aux questions : quand, quoi, où, bien qu'il serait préférable de respecter les normes INSPIRE en remplissant la fiche décrite dans la partie 3.2.

- ConnecSenS :

Les données sont envoyées sous forme de flux permanents provenant de réseaux de capteurs déployés sur des agro-écosystèmes en Auvergne. Les métadonnées associées à un flux de capteurs devront permettre de répondre aux questions : quand, quoi, où.

- Le Laboratoire d'Innovation Territorial (LIT) :

Les données sont collectées sur des parcelles privées instrumentées et leur spécificité vient de leur propriété intellectuelle. Il sera donc nécessaire de définir un accès spécifique à ce type de données. Les jeux de données devront aussi répondre aux questions : quand, quoi, où, bien qu'il serait préférable de satisfaire aux normes INSPIRE en remplissant la fiche décrite dans la partie 3.2.

Annexes

Lexique et définitions

CEBA : Cloud environnemental au bénéfice de l'agriculture

Open Data : Concept de partage, d'accès et d'usage des données, de manière libre et gratuite

DOI : Digital Object Identifier, sert à identifier une donnée de façon unique et pérenne

SQL : Langage de définition, manipulation et contrôle de données pour les bases de données relationnelles

NoSQL : Not Only SQL, désigne des bases de données qui ne sont pas fondées sur une architecture relationnelle

Data Catalog : Emplacement centralisé où sont regroupées toutes les métadonnées des jeux de données stockés par un organisme

Jeu de données : « Agrégation, sous une forme lisible, de données brutes ou dérivées présentant une certaine unité, rassemblées pour former un ensemble cohérent ». Cela peut donc consister en un enregistrement de données sous la forme d'un ou plusieurs fichiers électroniques, téléchargeables, citables (notamment par le biais d'un DOI) et intelligibles, ce jeu étant accompagné des métadonnées descriptives suffisantes

CRAIG : Centre Régional Auvergne-Rhône-Alpes de l'Information Géographique

UNIVEGE : Herbiers universitaires (UCA)

Adresse du résultat de la saisie de métadonnées pour un jeu de données de la Zone Atelier (Geonetwork):

http://meta.data-za.org/geonetwork/srv/fre/catalog.search#/metadata/ZATU_ca932ef8-2ed9-11e9-a271-71a2a9a47952