Guide

Les référentiels du Système d'information sur l'eau

Règles, contraintes et services

Titre	Les référentiels du Système d'information sur l'eau : règles, contraintes et services
Créateur	Laurent Coudercy
Contributeurs	René Lalement, Dimitri Meunier
Mots clefs	web ; interopérabilité ; identifiants ; référentiel
Résumé	Ce guide fixe les règles, contraintes et services qui s'appliquent pour la gestion des données de référence du Système d'information sur l'eau.
Éditeur	Onema
Version	1
Identifiant	urn:sandre:note-methodologique:GuideReferentiels::::ressource:1:::pdf
Date d'édition	Octobre 2016
Langue	FR
Droits d'usage	http://creativecommons.org/licenses/by-nc-sa/2.0/fr/

Historique du document

Action	Date	Par	Modifications	Version
Création	2015-07-09	Laurent Coudercy / DCIE	Création	0.1
Modification	2015-09-20	Laurent Coudercy / DCIE ; Dimitri Meunier, ST Sandre	Rajout des définitions ; homogénéisation du texte	0.2
Modification	2015-09-30	Laurent Coudercy / DCIE ; Dimitri Meunier, ST Sandre	Corrections diverses	0.3
Vérification	2015-10-20	GPS	Demande de quelques modifications	0.3
Modification	2016-07-22	Laurent Coudercy / DCIE ; Dimitri Meunier, ST Sandre	Intégration des éléments du cadre commun d'interopérabilité référentiel ; intégration des remarques du GPS du 20/10/2015	0.4
Validation	2016-08-23	Animateur GPS	Mise en forme au format ODF Modifications éditoriales, réorganisation du plan, ajout d'un préambule, de la loi République numérique et précisions sur l'objet du document	0.5
Validation	2016-10-14	GCiB	Validation	1.0
Approbation	2016-10-14	DG Onema	Approbation	1

Sommaire

1 Préambule

1.1 Le langage commun du système d'information sur l'eau

Le système d'information sur l'eau (SIE) est formé par un ensemble de dispositifs, processus et flux d'information, par lesquels les données publiques de l'eau sont acquises, collectées, conservées, organisées, traitées et publiées de façon systématique. Sa mise en œuvre résulte de la coopération de multiples partenaires, administrations, établissements publics, entreprises et associations, qui se sont engagés à respecter des règles communes définies par voie réglementaire ou contractuelle.

Du fait de la multiplicité de ces producteurs de données et des dispositifs de gestion des données (bases de données, etc.), il a été nécessaire dès la mise en place du système d'information, en 1992, de convenir d'un langage commun. C'est la raison d'être du Sandre, service d'administration nationale des données et des référentiels sur l'eau. Ce langage comporte notamment des données de référence, des dictionnaires de données et des scénarios d'échange. Il a été conçu pour permettre, de manière très opérationnelle, l'interopérabilité des dispositifs de gestion de données de l'eau.

Les observations exprimées lors de la conférence environnementale de septembre 2013 ont montré que malgré la cohérence interne au SIE permise par ce langage commun et la démarche d'ouverture des données, des progrès restaient à accomplir pour que ses données soient facilement utilisables par tous, au-delà des acteurs de l'eau souvent eux-mêmes producteurs de données.

Utiliser les données de l'eau, ce n'est pas seulement pouvoir les consulter, c'est aussi pouvoir les réutiliser avec d'autres finalités que celles qui ont justifié leur production, les croiser avec des données provenant d'autres domaines (la santé, l'agriculture, l'énergie, etc.) et leur appliquer des traitements pour produire de nouvelles connaissances : c'est en ce sens que la donnée, qui a un coût, peut aussi acquérir de la valeur. Il ne s'agit plus seulement de permettre une interopérabilité, au sein du SIE, des seuls dispositifs de gestion des données de l'eau entre eux, mais de permettre une *interopérabilité des données*, au-delà du SIE, pour susciter l'émergence d'un *data-écosystème*, avec de nouveaux utilisateurs capables d'usages innovants des données. Ce besoin d'une interopérabilité étendue exige à son tour un accès facilité au référentiel technique du SIE et un niveau de qualité élevé.

Au cœur de ce langage commun figurent les *données de référence*, appelées aussi *référentiels*, voire *master data*. Le SIE en a élaboré en 2011 la définition suivante :

> *« Certaines données sont utilisées dans de nombreux processus métier du SIE. Ces données appelées données de référence (ou données maître, ou encore données référentielles) existent* ***indépendamment des processus métiers*** *qui les consomment (toute masse d'eau a ainsi vocation à être référencée dans le SIE, qu'elle fasse ou non l'objet de mesure de son état, des pressions qu'elle subit ou d'actions spécifiques) et constituent les axes d'analyses des processus métier, c'est-à-dire qu'elles* ***permettent de croiser des données*** *provenant de processus métiers différents. Les stations d'épurations, les masses d'eaux bassin versants, les tronçons de cours d'eaux, les masses d'eaux superficielles, les aires d'alimentations des captages, les services publics de distributions d'eau*

sont autant d'exemples de données référentielles manipulées dans le cadre du SIE. »

Ces données de référence constituent une toute petite partie de l'ensemble des données de l'eau, mais elles jouent un rôle crucial à la fois pour la cohérence interne du SIE et l'utilisation de l'ensemble des données dans un cadre ouvert. Ce sont en effet ces données de référence, en tant que ***données pivot***, qui servent à croiser les données de l'eau entre elles et avec des données d'autres domaines.

Au sein du SIE, le Sandre a constitué au fil de ses travaux des règles de gestion de ces données de référence. L'ensemble constitue un corpus important mais dispersé (voir références au § 10) sur lequel reposent la construction, le partage et la diffusion des données de référence du SIE.

1.2 Le schéma national des données sur l'eau

La place des données de référence est affirmée dans le schéma national des données sur l'eau (SNDE), approuvé par un arrêté interministériel publié au JO du 24 août 2010 : ces données sont en effet une composante du *référentiel technique* du SIE qui doit être respecté par tous ses contributeurs, conformément au décret n° 2009-1543 du 11 décembre 2009, et en particulier du *référentiel des données*, qui comporte :

« 1. des spécifications des jeux de données et des services du système d'information sur l'eau ;

2. des règles relatives à l'établissement de ces spécifications et à leur emploi ;

3. des jeux de données de référence, portant notamment sur les thèmes de données figurant à l'annexe 14.1.1 ;

4. des règles d'administration de ces jeux de données de référence, relatives à leur création, leur mise à jour, leur mise à disposition et leur utilisation. »

L'Office national de l'eau et des milieux aquatiques est chargé de la définition et de la mise à disposition du référentiel des données. Son directeur général en approuve les éléments (spécifications, règles, jeux de données, services, jeux de données de référence, administration des jeux de données de référence).

Le SNDE ne définit pas davantage ces termes, mais donne une liste indicative de *jeux de données de référence* dans son annexe § 14.1.1 :

« 1. les éléments hydrographiques de surface, dont les bassins et sous-bassins hydrographiques et les hydro-écorégions ;

2. les aquifères ;

3. les masses d'eau de surface et d'eau souterraine, les sous-unités et les groupements de bassins établis pour l'application de la directive 2000/60/CE du Parlement européen et du Conseil du 23 octobre 2000 ;

4. les installations d'utilité publique pour l'assainissement et les points de captage utilisés pour la production d'eau potable ;

5. les installations de suivi environnemental pour la mesure des rejets et des prélèvements et pour l'observation de l'état quantitatif et qualitatif des eaux et des milieux aquatiques ;

6. les paramètres faisant l'objet de mesures ou d'observations et les méthodes correspondantes ;

7. les nomenclatures des pressions ;

8. les nomenclatures des réponses, en particulier des redevances, des aides et des zones protégées. »

1.3 Le cadre commun d'architecture des référentiels de données

Le **cadre commun d'architecture des référentiels de données**, publié par le Secrétariat général pour la modernisation de l'action publique, a pour ambition de fixer un premier corpus de règles de gouvernance des données de référence, afin qu'elles soient appliquées progressivement dans l'ensemble du système d'information de l'État. Ces règles intègrent des dimensions techniques, fonctionnelles, métiers et organisationnelles.

Ce cadre commun, dans sa V1.0 du 18 décembre 2013, définit ainsi les données de référence :

« Parmi les données collectées, traitées, manipulées, ou échangées au sein du système d'information des services publics, certaines ont des caractéristiques particulières, au nombre de cinq. Il est question alors de données de référence. Les cinq caractéristiques sont les suivantes :

*1. ces données sont utilisées **fréquemment par un grand nombre d'acteurs** internes ou externes (organisations, métiers, processus, applications…).*

*2. la **qualité de ces données est critique** pour un grand nombre de processus. Elle conditionne directement l'efficacité et l'efficience de ces processus, et donc plus globalement impacte le pilotage de l'action publique.*

***3. la sémantique** de ces données. C'est à dire la formalisation du sens et de la signification de ces données**, est partagée et relativement stable dans le temps**. L'unicité et la richesse sémantique de ces données sont recherchées pour simplifier les processus, optimiser leurs exécutions, et apporter plus de valeur aux bénéficiaires de ces processus. La portée de ces données, c'est-à-dire la couverture d'usage de ces données, est également un critère clé dans leurs utilisations, et des incompréhensions sur cette portée peuvent impacter également l'efficacité des processus. Il faut noter qu'une sémantique stable ne signifie pas qu'une donnée est stable. Certaines données de référence varient beaucoup et souvent dans le temps.*

***4. ces données ont une durée de vie qui va au-delà des processus opérationnels** qui l'utilisent. De fait, les données de contextualisation qui leur sont associées, c'est-à-dire leurs métadonnées, sont critiques.*

***5. la facilité d'accès et la disponibilité de ces données sont critiques** et conditionnent l'efficacité et l'efficience global des solutions mises en place pour utiliser / exploiter ces données : depuis n'importe où, tout le temps, et quel que soit le dispositif technique qui en a besoin.*

L'identification des données de référence est un sujet particulièrement sensible et conditionne l'efficacité des échanges et de l'exploitation de ces données

(identifiant unique et partagé). L'interopérabilité des dispositifs d'accès à ces données est une condition de succès. »

Par ailleurs, ce cadre commun définit les *référentiels de données* comme « *les outils informatiques nécessaires à la gestion de ces données dans le temps et leurs mises à disposition des autres applications, systèmes d'information ou utilisateurs. ».*

1.4 La Loi pour une République numérique

La loi n° 2016-1321 du 7 octobre 2016 institue un service public des données de référence, codifié par le nouvel article L. 321-4 du code des relations entre le public et l'administration. Cet article donne désormais une définition légale de ces données et prévoit la fixation par voie réglementaire de critères de qualité et d'une liste de ces données.

« Art. L. 321-4. – I. – La mise à disposition des données de référence en vue de faciliter leur réutilisation constitue une mission de service public relevant de l'État. Toutes les administrations mentionnées au premier alinéa de l'article L. 300-2 concourent à cette mission.

« II. – Sont des données de référence les informations publiques mentionnées à l'article L. 321-1 qui satisfont aux conditions suivantes :

« 1° Elles constituent une référence commune pour nommer ou identifier des produits, des services, des territoires ou des personnes ;

« 2° Elles sont réutilisées fréquemment par des personnes publiques ou privées autres que l'administration qui les détient ;

« 3° Leur réutilisation nécessite qu'elles soient mises à disposition avec un niveau élevé de qualité.

« III. – Un décret en Conseil d'État précise les modalités de participation et de coordination des différentes administrations. Il fixe les critères de qualité que doit respecter la mise à disposition des données de référence. Il dresse la liste des données de référence et désigne les administrations responsables de leur production et de leur mise à disposition. »

2 Objet et portée du document

L'objectif de ce guide est de permettre au système d'information sur l'eau de contribuer au service public des données de référence.

Il consolide et complète les règles de gestion des données de référence du SIE, sur la base des acquis du Sandre, et les met en relation avec le dispositif applicable au système d'information de l'État.

Il présente également les conditions qui doivent être satisfaites pour qu'une donnée « candidate » puisse être reconnue comme une donnée de référence du SIE.

Validé par le groupe de coordination inter-bassins et approuvé par le Directeur général de l'Onema, ce document est intégré au référentiel technique du SIE. Il sera révisé dans les mêmes conditions, en tenant compte de l'évolution des besoins et de la réglementation.

3 Les données de référence du SIE

3.1 Caractéristiques des données de référence

Les définitions des données de référence rappelées aux § 1.2, 1.3 et 1.4 concordent pour décrire leurs caractéristiques et les principes qui doivent guider leur organisation.

Dans la suite de ce guide, on considère que ces données doivent :

1. être utiles à un grand nombre d'utilisateurs ou de métiers ;
2. permettre de rattacher ou de positionner des données métier ;
3. être relativement stables dans le temps ;
4. disposer d'une sémantique connue, et limitée au strict nécessaire à la fonction de référence ;
5. disposer d'un identifiant stable, unique, non réutilisé, et géré, permettant d'identifier chaque objet sans ambiguïté ni doublon ;
6. être accessibles à tous librement, selon des modalités techniques adaptées aux usages ;
7. être de qualité connue et contrôlée ;
8. disposer de règles d'administration connues et contrôlées.
9. faire l'objet de contrôles qualité et d'un processus d'amélioration en continu.

Ces caractéristiques sont également des conditions qui doivent être satisfaites pour qu'une donnée « candidate » puisse être reconnue comme une donnée de référence.

3.2 Les catégories de données de référence du SIE

Les données de référence du SIE sont classées en trois grandes catégories, à savoir les données de référence, dites :

- **« alphanumériques »** : C'est une donnée sans composante géographique, regroupant des données peu évolutives, mais dont l'ensemble évolue régulièrement par ajout de nouvelles données. Parmi ces données de référence, on trouve les taxons et appellations de taxons, les intervenants et interlocuteurs, les paramètres, méthodes et unités de mesure, …
- **« géographiques »** : C'est une donnée, ayant une composante géographique (géométrie ponctuelle, linéaire, surfacique etc.), qui est utilisée pour porter les données métier directement (une station de mesure porte les mesures effectuées sur cette station), ou pour positionner d'autres données géographiques (la Bdcarthage sert de référentiel de positionnement à d'autres données, telles que les tronçons de cours d'eau classés, les périmètres de SAGE ou les bassins versants).
- **« nomenclaturales »** : Ce sont des vocabulaires contrôlés, sans composante géographique, très stables dans le temps, présentant un faible nombre d'items, qui sont utilisées comme valeurs autorisées de certains attributs. Les nomenclatures doivent si possibles être « fermées » c'est-à-dire comporter un item « non déterminé » ou « sans objet ».

Dans le **cadre commun d'architecture des référentiels de données** (V1.0), seuls les deux premiers types de référentiels sont considérés. La frontière entre ces catégories est d'ailleurs poreuse. Par exemple, la donnée de référence des obstacles à l'écoulement est une donnée de référence géographique ; mais elle peut être diffusée comme une donnée de référence alphanumérique, la localisation étant qu'un attribut parmi d'autres. De même, la différence entre une nomenclature et une donnée de référence alphanumérique est affaire de proportion dans la variabilité dans le temps.

3.3 Principales données de référence du SIE

Les principales données de référence du SIE (celles les plus utilisées, y compris à l'extérieur du SIE, et n'ayant pas leur équivalent en dehors du SIE) sont :

- BDCarthage® : référentiel géographique hydrographique, conforme avec la géométrie de l'hydrographie de la Bdcarto® ;
- BDLisa® : référentiel géographique hydrogéologique ;
- Paramètres : référentiel alphanumérique des paramètres de mesure ; fait référence au CAS *Chemical Abstracts System*® pour les substances chimiques, mais porte aussi sur d'autres paramètres (biologiques, physiques, etc.) ;
- Interlocuteurs/intervenants : référentiel alphanumérique des acteurs concernés par le SIE ; s'appuie sur la base de données Sirene®, pour les entreprises françaises, mais contient des références à des structures qui n'y sont pas enregistrées, et à des individus ;
- Stations de mesure (de la qualité des cours d'eau, plans d'eau et des eaux côtières, de la qualité des eaux souterraines, du niveau des nappes d'eau souterraine, des débits et hauteurs d'eau dans les cours d'eau) : référentiel géographique des stations de mesure du SIE ;
- Dispositifs de collecte : référentiel alphanumérique des organisations de collecte des données du SIE ;
- Appellations de taxon : référentiel alphanumérique des taxons et de leurs synonymes, utilisés par le SIE ; s'appuie sur le référentiel TAXREF® du muséum d'histoire naturelle, mais peut comprendre des taxons non encore codifiés par TAXREF® ;
- Zonages réglementaires : tous les zonages réglementaires du domaine de l'eau ;
- Masses d'eau (cours d'eau, plan d'eau, eau souterraine, côtières, de transition) et districts : référentiels géographiques des unités de suivi et de planification de la gestion de l'eau ;
- Ouvrages (obstacles à l'écoulement, stations d'épuration, …) : différents référentiels géographiques sur des ouvrages humains présentant des enjeux dans le domaine de l'eau, représentés sous forme de points.

Certaines de ces données sont utilisées y compris en dehors du SIE ; c'est en particulier le cas du référentiel des paramètres de mesure, utilisé aussi dans le domaine de la qualité de l'air, dans le domaine milieu marin, etc.

4 Règles applicables aux données de référence du SIE

4.1 Dictionnaires et scénario d'échange

Le principe d'interopérabilité sémantique impose que les données de référence s'appuient sur des dictionnaires et un scénario d'échange, qui définissent précisément leur sémantique et leur syntaxe, c'est-à-dire la façon de représenter ou de décrire des ressources ou objets du monde réel.

Ceci s'applique aux données de référence alphanumériques qui disposent de dictionnaires, de scénarios d'échange techniques et d'un scénario d'échange web. Ceci s'applique également aux données de référence géographiques, qui reposent sur des dictionnaires, des scénarios d'échange adaptés, et des flux de type WFS.

Règles du cadre commun d'architecture des référentiels de données
Règle RF1 : *La sémantique des données de référence est décrite, disponible, entretenue et partagée. Les objets métiers correspondant sont identifiés : leurs caractéristiques, leurs relations et leurs comportements (cycle de vie, états, événements) sont décrits et cette documentation est à jour par rapport à la réalité opérationnelle (processus et application).*
Règle RF2 : *Rechercher une sémantique riche et un haut niveau d'abstraction. Formaliser cette sémantique à l'aide d'une modélisation selon la notation UML 2.4.1 (ISO)*

4.2 Une sémantique minimale

Une donnée de référence ne doit comporter que le minimum d'attributs utiles pour sa fonction de référentiel, et si possible des attributs peu évolutifs. Elle n'a en particulier pas à être le réceptacle de données métiers spécifiques.

Les principaux attributs que doit comporter une donnée de référence sont :

- Ceux qui servent à identifier les objets décrits, pour assurer l'interopérabilité des données ; la qualité de cette information est critique pour l'ensemble des systèmes de données :
  - son identifiant (ou son code Sandre ainsi que son type),
  - sa géométrie, pour les référentiels géographiques,
  - les identifiants issus d'autres systèmes de données (agence, Sirene®, CAS®, TAXREF®,…) ainsi qu'une référence aux systèmes de données concernés,
  - son type, éventuellement.
- Ceux qui aident à identifier les objets décrits, pour les humains ; la qualité de cette information est importante pour l'administration des données :
  - une dénomination, voire plusieurs (version courte et longue par exemple),
  - des codes signifiants ou mnémoniques, éventuellement (ex : code de Pfafstetter, code poisson Onema),

- éventuellement une définition, indispensable pour les nomenclatures.

- Ceux qui permettent de connaître l'état des objets décrits ; la qualité de cette partie de la donnée est critique pour la cohérence du système de données :
  - le statut : gelé, validé (Cf. § 5.2),
  - les dates de création, gestion, gel,
  - les informations de généalogie.
- Ceux qui permettent le dé-doublonnement et la détection d'erreurs ; cette information complémentaire est utile pour la qualité de la donnée, au moment de sa création, mais n'est pas à mettre à jour obligatoirement :
  - pour une station, positionnée par ses coordonnées, on peut vouloir disposer aussi de sa commune de rattachement, ce qui permet un premier niveau de contrôle,
  - pour un taxon, on doit connaître l'auteur, la date de détermination …
  - pour un ouvrage de prélèvement, son gestionnaire, ses coordonnées géographiques.
  - ...
- Ceux qui expriment une relation métier avec d'autres données de référence du SIE ; la qualité de cette information, non référentielle, ne relève que de préoccupations métier.

5 Les identifiants

5.1 Généralités

Pour identifier de manière fiable – et reconnue par tous – un objet décrit par une donnée de référence, il faut doter cet objet (ou l'instance de l'objet, s'il peut évoluer dans le temps) d'un identifiant, unique, non utilisé par d'autres objets de la même donnée de référence. Ainsi cet identifiant, ou code, est localement unique et désigne un objet (ou son instance) unique dans une donnée de référence particulière.

Ainsi dans sa note de 2001 sur les règles de codification, le Sandre définit l'identifiant comme « *un identifiant est … un groupe de caractères …, employé pour désigner de façon unique un élément dans un ensemble, et ainsi le distinguer des autres éléments de cet ensemble* ».

Pour assurer que cet identifiant ne pose pas de problème de stabilité dans le temps, il doit autant que possible être non signifiant.

Ainsi le document de 2001 cité plus haut précise :

- « *un identifiant non signifiant est un identifiant qui ne comporte qu'une seule information : celle nécessaire à distinguer l'objet qu'il cible* ».
- « *un identifiant signifiant est un identifiant qui comporte tout ou partie de l'information contenue dans les éléments de données décrivant l'objet auquel il se rapporte* » et par construction peut être amené à évoluer quand les éléments qui le constituent évoluent, sans que l'objet évolue.

Ainsi l'identifiant d'une station de mesure superficielle continentale comporte en première partie le code du bassin hydrographique, qui est considéré comme non signifiant : quand les limites du bassin changent, ce point peut se trouver dans un autre bassin que celui indiqué dans son code, mais il n'y a pas lieu de changer pour autant l'identifiant de la station.

On a actuellement encore des codes Sandre comportant des parties signifiantes :

- code BSS, en cours de résolution
- code Hydro des cours d'eau

Par ailleurs, la BdCarthage® ne comporte pas d'identifiant pérenne pour tous les objets hydrographiques : seule une partie est codée, avec des codes signifiants.

5.2 Identifiant permanent, règle de gel

Un identifiant (ou code) est permanent, et ne doit pas être supprimé, ni surtout réutilisé pour identifier un autre objet (ou instance) d'une même donnée de référence. Si l'objet cesse d'être pertinent, son identifiant doit être *gelé*, mais pas supprimé.

Une note de 2013 rappelle les règles sur l'utilisation des codes gelés :

> *« La situation dans laquelle se trouve un objet à une date donnée représente son état. À sa création, l'objet est validé. … Un objet peut être mis à jour au cours du temps sur tout ou partie de ses attributs excepté sur la date de création de l'objet et sur son identifiant. Ces deux derniers restent invariants dans le temps. L'objet est gelé dans au moins l'une des conditions suivantes :*
>
> - ***La destruction** : C'est le fait que l'objet n'a plus d'existence dans le monde réel, ce qui amène à geler définitivement l'objet ou à modifier la date de fin de vie de l'objet. Il peut s'agir aussi de doublon comme 2 zones de pêche ayant les mêmes noms et localisées au même endroit, mais avec des codes différents. Prenons un autre exemple, celui de l'objet station de mesures de qualité des eaux superficielles et continentales, l'attribut date d'arrêt d'activité de la station de mesure permet de signaler si la station est toujours en activité ou pas. Dans ce cas, le statut de l'objet reste validé.*
> - ***La recodification** : C'est le fait de geler un objet parce qu'il comporte au moins une erreur de cohérence sur des attributs discriminants…. Il existe aussi des contres exemples, quand l'objet de change pas d'identifiant malgré la modification de ses attributs. C'est le cas … des contours des bassins qui changent sans que leurs codes soient modifiés.*
> - ***La division** : C'est le fait de geler un objet au profit de n nouveaux. Une masse d'eau peut avoir été découpée en deux masses d'eau distinctes. …*
> - ***La fusion** : C'est le fait de geler n objets pour créer un nouvel objet. C'est le cas des communes, il arrive que 2 communes fusionnent en une seule. »*

Ce mode de fonctionnement évite la réutilisation d'un code, et donc de faire pointer, selon la date d'utilisation, un même code vers deux objets différents.

Règles du cadre commun d'architecture des référentiels de données
Règle RF4 : *Séparer les données d'identités (ou d'identification métier), des identifiants des données de référence. Pour un objet métier : utiliser un identifiant de type URI : aisément partageable, non ambigus, non signifiant (donc ne contenant pas de données élémentaires à caractères personnels ou potentiellement confidentielles), non modifiables, non réaffectable, non supprimable et persistant.*

6 Les règles d'administration des données de référence

6.1 Plusieurs organisations

Dans la pratique, de multiples cas de figure d'administration des données de référence du SIE sont constatés, et ont été formalisés par le système d'information sur l'eau :

- **Administration centralisée** : l'ensemble des objets d'une même donnée de référence est administré au sein d'un seul système d'information. Un seul partenaire intervient dans l'administration de ce type de données de référence, mais il peut s'appuyer sur d'autres partenaires pour le contrôle de cette donnée de référence, ou s'appuyer sur une donnée de référence externe au SIE. Les paramètres, unités, méthodes, fractions et supports d'analyse, sont administrés de cette façon. En général, dans le cas du SIE, le responsable de l'administration centralisée est le Secrétariat technique du Sandre, mais pour la BDLisa®, il s'agit du BRGM, et pour la BDCarthage® de l'IGN. On est alors dans le pattern « *référentiel centralisé* » du cadre commun d'architecture des référentiels de données.
- **Administration répartie** : tous les objets d'une même donnée de référence ne sont pas administrés au sein d'un seul système d'information. Au contraire plusieurs partenaires interviennent dans l'administration de ce type de donnée de référence. De ce fait une règle de répartition des activités de gestion des objets d'une même donnée de référence doit être établie entre les différents acteurs participant à l'administration. Ce type d'administration répartie se décline en trois types d'organisation. On est alors dans le pattern « *référentiel de consolidation* » du cadre commun d'architecture des référentiels de données :
  - La gestion des données de références est **répartie de manière stricte** entre les différents acteurs, chacun gérant un sous-ensemble des données de référence. C'est par exemple le cas des stations de mesure qualité pour les eaux superficielles, gérées par une seule agence, selon leur territoire d'action.
  - Le cycle de vie d'une donnée de référence est **portée par un acteur principal,** qui a en charge la gestion de l'identification unique ; un ou plusieurs autres acteurs interviennent pour compléter la description de la donnée (exemple des stations d'épuration créées par les agences de l'eau et mises à jour par les services de police de l'eau), voire pour procéder à un changement d'état.
  - **Différents partenaires collaborent à la création**, mise à jour et archivage des données de référence, sans répartition permanente des objets entre eux. C'est par exemple le cas des sites industriels, pouvant être créés par les agences et les

services des installations classées, et mis à jour par ces mêmes acteurs indépendamment du créateur de la donnée.

- **Administration utilisant un référentiel tiers :** dans ce cas les informations d'une donnée de référence sont issues pour la plupart d'un autre référentiel distant, soit parce qu'il y a spécialisation du référentiel distant (on n'en retient qu'une partie), soit parce que ce référentiel distant doit être complété d'objets ne correspondant pas à son périmètre. C'est par exemple le cas du référentiel des appellations de taxon, par rapport à TAXREF®, ou du référentiel interlocuteurs s'appuyant en partie sur le fichier Sirene®, ou du référentiel des paramètres qui utilise le code CAS® pour les paramètres chimiques. On est alors dans le pattern « *référentiel esclave* » du cadre commun d'architecture des référentiels de données.

6.2 Un document d'administration par donnée de référence

Une donnée de référence est relativement stable au sein du système d'information. La donnée concernée aura une vie longue et peut être administrée par plusieurs acteurs répartis ou au fil du temps. Il est donc impératif de fixer des règles d'administration, afin d'homogénéiser les pratiques, par exemple de fixer les cas où un code doit être gelé.

Un document définissant les règles d'administration de chaque donnée de référence doit exister progressivement.

Pour les données de référence administrées directement par le secrétariat technique du Sandre, les documents d'assurance qualité du secrétariat technique du Sandre répondent au besoin.

Pour les autres données, et en particulier pour les données de référence à administration répartie, un document spécifique doit être tenu à jour, en accord avec les administrateurs de la donnée de référence concernés.

Ce document porte sur

- L'administration de la donnée de référence :
  - qui est responsable de son administration,
  - si cette administration est répartie, comment est gérée cette répartition,
  - qui diffuse ce référentiel.
- L'identification des objets décrits par la donnée de référence :
  - quel code, comment il est construit.
- La vie des objets de la donnée de référence :
  - les cas de gel autorisés,
  - les modalités de création, de modification des objets de la donnée de référence.
- Les spécifications de la donnée de référence :
  - quels dictionnaires et scénarios d'échange,
  - quels attributs font partie intégrante du référentiel.

Règles du cadre commun d'architecture des référentiels de données
Règle RM2 : *Les processus de mise à jour des données de référence sont décrits, publiés, entretenus et partagés. Ils identifient clairement tous les points d'acquisition des données et l'unique point de vérité. Les processus d'entretien des données sont alignés avec le cycle de vie métier des objets métiers.*

6.3 Contrôle qualité de la donnée de référence

« *La **qualité de ces données est critique** pour un grand nombre de processus* ».

Aussi une donnée de référence doit faire l'objet d'une démarche visant à assurer sa qualité et son amélioration continue, avec des exigences adaptées aux usages et aux contraintes de sa production.

- L'assurance qualité est la première étape pour assurer une qualité d'ensemble de la donnée de référence :
  - elle passe par la démarche qualité générale du Sandre,
  - c'est aussi le rôle assuré par Aquaref pour certaines données de référence.
- Elle s'appuie aussi sur la possibilité offerte aux usagers de signaler ce qui leur parait être une erreur :
  - la traçabilité des signalements et des suites qui leur sont données est un gage d'amélioration de la qualité.
- Elle passe aussi par des contrôles de qualité spécifiques, réalisés de manière régulière :
  - contrôles de conformité aux modèles, aux règles de codification,
  - contrôles de cohérence interne, entre données de référence, et à des règles métiers,
  - et si possible, contrôles de validité.

Dans le SIE, la règle est que la qualité d'une donnée est de la responsabilité de son producteur. Les résultats des contrôles qualité réalisés par le secrétariat technique du Sandre doivent donc être transmis aux producteurs, afin qu'ils améliorent cette qualité.

Ces éléments d'assurance qualité doivent être connus des utilisateurs et du public.

Règles du cadre commun d'architecture des référentiels de données
Règle RM3 : *Les indicateurs de qualité sont identifiés dans les dispositifs d'acquisition et de distribution des données de référence. Ils sont mesurés et publiés.*

7 La diffusion des données de référence

7.1 Diffusion en Open Data

Les données de référence du SIE doivent en règle générale être diffusées comme des données ouvertes (*Open Data*), c'est-à-dire être accessibles à tous, gratuitement, sans aucune restriction d'usage, sous licence ouverte, dans des standards ouverts aisément utilisables et exploitables par un système de traitement automatisé.

Des exceptions à la règle précédente peuvent être justifiées par la loi : limitation dans le positionnement précis des points de captage d'eau destinée à l'alimentation humaine, données personnelles pour les intervenants/interlocuteurs,…

Certaines contraintes conventionnelles ou réglementaires ont pu aboutir à des diffusions limitées : les conventions sur la BDCarthage® précédant celle de 2014 limitaient les droits d'usage de cette donnée de référence, même si chaque convention permettait d'avancer vers son ouverture complète, qui a été atteinte en 2014.

7.2 Point de vérité

La diffusion officielle publique doit être effectuée par le site du Sandre, qui est l'unique « point de vérité » pour les données de référence du SIE. Pour cela, le site du Sandre peut n'assurer que la redirection de données diffusées par ailleurs (ainsi les piézomètres sont issus du site ADES, mais sont présents sur le site du Sandre, par rediffusion automatique).

La diffusion du site du Sandre est ensuite relayée, par moissonnage de métadonnées, vers les serveurs nationaux de données (via data.eaufrance.fr, lui-même moissonné par data.gouv.fr) et pour les données géographiques vers www.geocatalogue.fr.

Évidemment, ces données étant en open data, tout acteur qui le souhaite peut les rediffuser.

Règles du cadre commun d'architecture des référentiels de données
Règle RM1 : *Unicité du point de vérité pour toutes données de référence.*

7.3 Diffusion dans différents formats

Afin de rendre ces données facilement réutilisables, les données de référence doivent être diffusées selon des modalités correspondant à de nombreux usages, allant de la simple consultation par une personne, à des outils assurant la synchronisation d'applications, en passant par des formats utilisables dans des applications locales ou sur le Web, voire des services de consultation répondant à des besoins spécifiques.

Il faut donc proposer un accès aux données de référence dans des formats répondant :

- à des besoins de simple consultation : présentation de l'objet de la donnée de référence ou de l'intégralité de la donnée de référence en HTML, voire en PDF, ou s'il s'agit d'une donnée géographique, sur une carte interactive en ligne ;
- à des besoins d'interopérabilité stricts entre outils : diffusion en téléchargement aux formats Sandre, disponibilité de flux normés Sandre en XML, diffusion des donnée de référence géographiques en WFS ;

- si possible, à des réutilisations plus locales, ou à des usages web : format simplifié tabulaire, format simplifié d'information géographique, JSON,…

Il faut veiller à progressivement diffuser les données de référence dans de nouveaux formats, adaptés aux pratiques nouvelles des utilisateurs : par exemple GML, RDF/XML, N3, …

Enfin, quand le besoin s'en fait sentir, il faut doter ces données de référence de services spécifiques : c'est par exemple le cas pour la BDCarthage®, pour laquelle des géotraitements permettent d'accéder à l'amont et l'aval d'un point des cours d'eau.

7.4 Accès à travers les URI

Le SIE s'est doté d'un répertoire d'identifiants uniques sous la forme d'URI (Uniform Resource Identifier).

Ce répertoire est administré et diffusé par le Sandre, sous l'autorité de l'Office national de l'eau et des milieux aquatiques. Il permet d'identifier de manière unique et universelle, i.e., à l'échelle mondiale du Web, toute ressource décrite par des données de référence du SIE. Il pourra être étendu progressivement à d'autres ressources du SIE (mesures, ...).

La structure d'un URI du SIE est formée selon le modèle générique suivant (version V1 du document définissant les URI) :

http://id.eaufrance.fr/{*type*}!{*version_col*}/{*identifiant*}!{*version_res*}#{*partie*}

- « Le nom id.eaufrance.fr est le nom de domaine de l'autorité attribuant les URI aux ressources d'intérêt du SIE ; il en garantit l'unicité ;
- {*type*} désigne une collection de ressources sur laquelle s'exerce l'autorité de nommage ; le type de collection correspond à l'une des valeurs possibles de la nomenclature Sandre n°373 ;
- {*version_col*} désigne une révision d'une collection de ressources ; si aucune version n'est indiquée, la version la plus récente est retenue ;
- {*identifiant*} désigne une ressource (de cette révision) de la collection ; la structure lexicale de l'identifiant est définie par le dictionnaire de données Sandre relatif à la collection ;
- {*version_res*} désigne une version (issue d'une révision) de la ressource ; si aucune version n'est mentionnée, la version la plus récente est retenue ;
- {*partie*} désigne une composante de la ressource (exemple : un concept d'un dictionnaire de données). »

Toute donnée de référence et objet d'une donnée de référence doit être accessible à travers ces URI.

Règles du cadre commun d'architecture des référentiels de données
Règle RF4 : *Séparer les données d'identités (ou d'identification métier), des identifiants des données de référence. Pour un objet métier : utiliser un identifiant de type URI[19] : aisément partageable, non ambigus, non signifiant (donc ne contenant pas de données élémentaires à caractères personnels*

ou potentiellement confidentielles), non modifiables, non-réaffectable, non supprimable et persistant.

7.5 Diffusion de l'intégralité de la donnée de référence

Une donnée de référence doit être accessible à tous, intégralement. On doit donc pouvoir accéder aussi bien aux objets valides qu'à ceux qui sont gelés. On doit pouvoir de même accéder aussi bien aux objets utilisés, qu'à ceux qui n'ont pas encore été utilisés (par exemple les stations sans mesure disponible), ou ceux dont on sait qu'ils ne sont plus utilisés.

7.6 Vues historisées

Malgré la grande stabilité des données de référence, elles subissent quand même des évolutions dans le temps. Or tous les systèmes d'information utilisant ces données de référence ne peuvent se mettre à jour en continu, et les banques de données hébergent des données historiques, constituées sur la base de versions anciennes des données de référence.

Il faut donc être capable d'accéder aux données de référence dans leur version ancienne, pour retrouver à quelle instance des objets de la donnée de référence se rattachent ces anciennes données. Le gel des instances des objets des données de référence permet de ne pas perdre cette information. Encore faut-il que ces anciennes versions soient disponibles soit sous forme de versions successives, soit avec gestion pour chaque élément de sa date de création et de gel, et la généalogie entre éléments.

Règles du cadre commun d'architecture des référentiels de données
Règle RA6 : *Toute action sur les données d'un référentiel doit être tracée et journalisée. Les données sont historisées. Les instances des objets métiers sont versionnées en fonction de leur cycle de vie. Les données ne sont pas détruites, mais marquées comme non active. La politique de conservation et d'archivage des données est définie.*

7.7 Métadonnées de la donnée de référence

Chaque donnée de référence doit faire l'objet d'une description sous forme de métadonnées ISO 19115.

Ces métadonnées doivent être disponibles sur le catalogue du Sandre. Elles doivent par ailleurs être disponibles sur data.eaufrance.fr, hormis pour les quelques données de référence non encore accessibles en Open Data.

8 Les relations avec les utilisateurs

8.1 Permettre les demandes de codification et de modification

Une donnée de référence est au service de nombreux outils et systèmes d'information métier, y compris des systèmes et outils non directement connus du secrétariat technique du Sandre.

Ces utilisateurs multiples ont le droit de demander la codification de nouveaux objets dans les données de référence, c'est-à-dire l'intégration d'une description de ces nouveaux objets ainsi que l'attribution d'un identifiant unique. Ils peuvent aussi signaler une erreur éventuelle, qui sera corrigée si besoin. Il faut donc fournir à tout utilisateur qui le souhaite (et pas seulement aux acteurs principaux du SIE), un outil en ligne leur permettant de demander une codification ou de proposer une modification.

Dans le cas des données de référence administrées par le secrétariat technique du Sandre, c'est l'outil Ogres/MDM qui joue ce rôle.

Pour la BDCarthage® et la BDLisa®, des outils similaires ont été mis en place par l'IGN et le BRGM respectivement.

Règles du cadre commun d'architecture des référentiels de données
Règle RM4 : *Chaque référentiel doit intégrer un dispositif d'alerte ou de signalement permettant à un utilisateur du référentiel de faire remonter au responsable du référentiel toutes anomalies sur les données détectées en aval (incomplétude, incohérence, doublon, amalgame, problèmes d'intégrité, etc.). Le processus de traitement des signalements doit être également décrit, à jour et publié. Il est recommandé également de rechercher son automatisation et donc son outillage.*

8.2 Assurer un délai raisonnable à toute demande de modification

Le besoin des utilisateurs de disposer d'un code pour identifier un nouvel objet d'une donnée de référence doit pouvoir être satisfait dans un délai raisonnable par rapport aux enjeux pour l'utilisateur, et aux possibilités de gestion de la donnée de référence, faute de quoi l'utilisateur créera ses propres codes d'objets de référence, ce qui compromet la cohérence des données. Pour cela le délai de réponse à une demande de codification ou de modification doit être connu et respecté.

Ainsi, pour les données de référence alphanumériques du Sandre, le secrétariat technique du Sandre s'engage à coder ou mettre à jour en moins de 20 jours ouvrés (en fait les délais observés sont actuellement inférieurs à la semaine dans la grande majorité des cas). Par contre, la Bdcarthage® est mise à jour une fois par an, la Bdlisa® tous les trois ans. Enfin certains objets de données de référence sont dépendants d'actes administratifs (zonages réglementaires, masses d'eau, …) et sont donc mis à jour dès qu'une modification se produit, avec selon les cas des versions successives (masses d'eau), ou une évolution de la donnée de référence (zonages réglementaires).

9 Glossaire

La terminologie est variable selon les sources et souvent fluctuantes. Les définitions ci-dessous sont celles employées dans le présent document ; cependant, dans les citations, les sens des mots correspondent à ceux des documents d'origine.

- **Administration des données de référence :** activité relative à la constitution, la mise à jour et la diffusion de données de référence, conformément à des règles (*dites règles d'administration de référentiel*).

- **Attribut** : donnée attachée à une représentation numérique d'une ressource, représentant une propriété élémentaire de celle-ci ; par exemple, une mesure est une ressource (un objet immatériel du monde réel), la date d'une mesure est une propriété qui est représentée par un attribut de la donnée représentant la mesure.
- **Catalogue** : ensemble de données (généralement de métadonnées) organisé de manière à pouvoir fournir des services de recherche, de consultation et de mise à jour.
- **Dictionnaire de données :** ensemble cohérent de descriptions, rassemblées dans un *catalogue*, des types de ressources représentées dans le SIE, indépendamment de la mise en œuvre de ces représentations par les applications informatiques qui les utilisent.
- **Donnée :** toute ressource numérique pouvant représenter ou décrire une ou plusieurs ressources, conformément à des spécifications relatives à la syntaxe (codage, format) et à la sémantique (signification).
- **Donnée de référence** : Cf. § 3.1 ; aussi appelée « référentiel », « master data », « enterprise master data ».
- **Donnée métier** : Donnée produite ou collectée pour satisfaire les besoins d'un métier particulier, généralement variable dans le temps et utilisant des données de référence ; par exemple, les données collectées au fil du temps (débit, hauteur de chute d'eau…) sur un barrage connu sont des données métier utilisant l'identité du barrage (invariable) comme une donnée de référence.
- **Ensemble de données** (anglais dataset) : plusieurs données regroupées dans un ensemble numérique structuré et homogène (base de données, fichier,…) ; on parle aussi de collection ou de jeu de données ; un ensemble de données est une donnée.
- **Format d'échange :** spécification relative au support numérique utilisé pour organiser, stocker et échanger la donnée ; par exemple, le format XML d'un flux de données d'un scénario, le transfert de fichier normalisé, un service web.
- **Identifiant** : donnée permettant de désigner une ressource de manière unique et non ambiguë, et donc de la distinguer des autres ressources.
- **Information géographique :** donnée disposant d'attributs de géolocalisation (ou donnée descriptive d'un objet localisé géographiquement).
- **Métadonnée** : donnée représentant une description d'une ressource, et dans le cas courant, d'un ensemble de données ou d'un document, par une série de propriétés comme le titre, la date de création, le créateur ou les droits d'utilisation de la ressource.
- **Modèle de données :** spécification formalisant, selon un métamodèle, des descriptions figurant dans les dictionnaires de données, nécessaires pour la mise en œuvre d'une application informatique ; par exemple, un diagramme de classes UML.
- **Objet du monde réel** : synonyme de « ressource ».
- **Référence** : donnée attachée à une représentation numérique d'une ressource, représentant une relation entre celle-ci et une autre ressource ; par exemple, « …au site de la Creuse à la Celle-Saint-Avant » est une référence représentant une relation de localisation sur un site de surveillance ; une référence à une ressource peut être codée au moyen d'un attribut ayant comme valeur l'identifiant de cette ressource.

- **Référentiel géographique** : Ensemble de données permettant de localiser des objets du monde physique ; par exemple, la BD Carthage®, le référentiel des sites de surveillance, le référentiel administratif.
- **Ressource** : entité matérielle du monde physique (objet, personne,...) ou immatérielle (organisation, concept, algorithme, ensemble de données,...) identifiable sans ambiguïté et pouvant bénéficier d'une représentation numérique ; aussi dénommée « objet du monde réel ».
- **Scénario d'échange** : spécification décrivant les modalités d'échange des données dans un contexte particulier ; par exemple, le scénario EDILABO, dans le contexte des relations entre commanditaires et laboratoires d'analyses.
- **Spécification** : document décrivant une catégorie de données et ses modalités d'échange ; par exemple, un dictionnaire de données, un modèle de données, un scénario ou un format d'échange.

10 Références

Note règles de codification, Sandre, 2001

Schéma national des données sur l'eau (SNDE), Onema - Ministère chargé de l'écologie - Partenaires du SIE, Août 2010

Sur le référentiel de données ; note pour discussion, DCIE 09/2010

Proposition d'organisation de la définition des modalités spécifiques d'administration des référentiels, Sandre, 2012

Note sur le gel, Sandre, 2013

Les identifiants uniques pour les données de référence du système d'information sur l'eau, SIE, 2016

Cadre Commun d'Architecture des Référentiel de données (V1.0), Direction Interministérielle des Systèmes d'Information et de Communication, 2013