

Guide pratique d'ouverture des données du Parc Naturel Régional du Vercors



par Michel Roche

Document réalisé pour le Parc Naturel Régional du Vercors, avec la participation de la Région Rhône-Alpes, et publié sous licence Creative Commons CC-BY qui vous autorise à copier, modifier et distribuer le présent document en citant l'auteur : Parc Naturel Régional du Vercors (<http://parc-du-vercors.fr>).

Ce texte est sous licence Creative Commons Attribution 4.0 International. Pour accéder à une copie de cette licence, merci de vous rendre à l'adresse suivante <http://creativecommons.org/licenses/by/4.0/> ou envoyez un courrier à Creative Commons, 444 Castro Street, Suite 900, Mountain View, California, 94041, USA.

Table des matières

1. Ouvrir les données, qu'est-ce que c'est ?	3
1.1 Approche.....	3
Pourquoi ouvrir des données ?.....	3
Trois exemples d'utilisation innovante des données :.....	3
Limites légales de l'ouverture.....	4
1.2 Où publier ?.....	6
La plateforme data.gouv.fr ?.....	6
D'autres lieux intéressants :.....	6
2. Que publier ?	6
2.1 Que sont les jeux de données ?.....	6
2.2 Privilégier les formats ouverts.....	7
3. Licences de publication	7
3.1 Les licences.....	7
3.2 Comment choisir la licence ?.....	8
Critère : originalité des données.....	9
Critère : possibilités de réutilisation offertes.....	9
Faut-il vraiment restreindre certains usages ?.....	9
3.3 Cas des données produites avec des tiers.....	11
Identifier la paternité des données.....	11
Accord de publication sous la licence X, Y ou Z.....	11
Texte possible de l'accord.....	11
3.4 Conseils sur deux licences.....	12
Le Domaine Public.....	12
La Licence Ouverte.....	12
4. Comment publier les données ?	13
4.1 Jeu de données.....	13
4.2 Ressources (les données proprement dites).....	15
5. Exemples de publications	16

1. Ouvrir les données, qu'est-ce que c'est ?

1.1 Approche

Pourquoi ouvrir des données ?

Le Parc du Vercors s'est engagé en Janvier 2012 dans une démarche d'ouverture des données qu'il est amené à créer et compiler dans ses multiples activités. L'accès aux données publiques détenues par les administrations au sens large (État, Collectivités territoriales...) est un droit pour tout citoyen, limité par des considérations de vie privée ou du caractère sensible de certaines informations.

Ouvrir les données détenues, produites ou coproduites par le Parc du Vercors est donc une manière valorisante de satisfaire aux exigences de la loi et de faire preuve de transparence envers les citoyens. C'est également une manière élégante de rendre au public ce qu'il a contribué à produire.

La loi CADA impose de donner l'information à qui la demande, l'Ouverture des données présente les informations disponibles à tous, et y accède qui veut.

Cette distinction dans la mise à disposition des données introduit un biais fondamental dans la manière d'accéder à celles-ci. La loi impose à minima de donner l'accès, mais la démarche peut être fastidieuse, lente, les données fournies sur un support mal adapté, et au final difficilement réutilisables. S'engager dans une démarche d'ouverture des données, revient à rendre ces mêmes données publiques et accessibles au plus grand nombre, en diminuant le plus possible les contraintes d'accès de façon à ce que les citoyens y aient accès sans plus avoir à effectuer de démarche administrative.

Le potentiel de création induit par l'ouverture et la mise à disposition de données est immense. Grâce à l'ouverture des données, chercheurs, écoles, administrations, citoyens peuvent disposer aisément d'un grand nombre de jeux de données pour mener à bien recherches ou études sur des domaines que l'on n'imagine peut-être même pas, ou en croisant des jeux de données d'une manière nouvelle.

L'égalité d'accès pour tout le monde aux données est assurée par la publication des données, et par leur placement sous une licence ouverte. Des données fermées peuvent être mises à disposition du public, mais la mise à disposition ne pourra jamais se faire avec le même degré d'égalité d'accès, il y aura toujours une contrepartie à prévoir pour le fournisseur des données : paiement, adhésion à un organisme, publicité, etc.

Trois exemples d'utilisation innovante des données :

1. Cartographie d'une épidémie de choléra

Londres, 1854 : une épidémie de choléra éclate à Londres dans un quartier pauvre. Un médecin, le Dr. Snow n'est pas convaincu par l'explication de la transmission par voie aérienne, les miasmes, pourtant pratique puisqu'elle désignait les pauvres comme victimes de leur mode de vie moralement dépravé. Il eut l'idée de dresser une carte du quartier et de ses puits, puis de noter sur les blocs d'habitation une barre pour chaque décès. Sa cartographie était assez édifiante : la distribution des morts se faisait autour d'un puits de Broad Street à quelques exceptions près, dont il trouva une explication rationnelle par la suite. Il parvint à convaincre les autorités locales de démonter le levier de la pompe du puits et l'épidémie se calma en quelques jours. C'est l'analyse de données produites par une enquête minutieuse et leur mise en forme cartographique qui a permis d'identifier la source du mal.

2. Une multinationale du big data

Une histoire plus récente, celle d'une compagnie publicitaire née au XX^e siècle et qui fit fortune au tout début du XXI^e grâce... aux données. La force d'une régie publicitaire est de savoir quelle publicité présenter à quelle personne et au bon moment si possible. L'astuce de ce publicitaire ne tient pas à l'ouverture des données de ses concurrents ou des états dans lesquels cette société a prospéré, mais à leur collecte privée, tous azimuts, de façon à se constituer de nombreux jeux de données... qu'il ne reste plus qu'à croiser avec des algorithmes bien écrits. C'est en croisant un grand nombre de données extrêmement

variées saisies par les utilisateurs de ses services (moteur de recherche, mail, documents, sondages, formulaires, réseau social, etc.) qu'elle parvient à présenter à ces mêmes utilisateurs des publicités ciblées vendues à prix d'or (près de 10 milliards d'euros de bénéfices en 2013).

Cet exemple ne montre évidemment pas un usage des données ouvertes vers les citoyens ou les utilisateurs des services, mais illustre en revanche l'ampleur des résultats que peuvent apporter l'utilisation des croisements de données.

3. La découverte vient de l'inattendu

En 2012, un jeune américain, Jake Andraka gagna un prix scientifique destiné aux étudiants en proposant une nouvelle méthode de détection du cancer du pancréas. Ce test, encore loin d'être parfait, constitue cependant une piste intéressante d'amélioration des tests existants. Ce qui est nous importe ici, c'est qu'il a pu mener ses recherches en utilisant des articles scientifiques mis à disposition librement grâce à l'action d'un activiste des libertés sur Internet : Aaron Swartz, un peu auparavant. Ce dernier a en effet utilisé l'accès de son université pour télécharger un grand nombre d'articles et le mettre à disposition librement sur d'autres serveurs.

Aux États-Unis la plupart des publications scientifiques sont accessibles via Jstor, un organisme de diffusion payant. De nombreuses écoles, universités, centres de recherche disposent d'abonnements permettant à leurs étudiants ou employés d'accéder facilement à cette base de données, mais pour les autres citoyens cet accès est très fortement limité, leur interdisant de fait d'utiliser efficacement cette base de données.

La publication libre (bien que non souhaitée par Jstor, qui a intenté un procès contre cette publication) a ici permis à un jeune étudiant d'inventer une nouvelle piste de test de détection, piste que personne n'avait exploré de cette manière jusqu'alors. L'ouverture des données permet de s'exposer à des réutilisations surprenantes et innovantes que l'on n'aurait pas obtenues en limitant l'accès par l'argent, l'appartenance à une institution, ou tout autre moyen.

Limites légales de l'ouverture

La loi, tout comme la délibération du Parc identifie assez rapidement quelques limites à la transmission de certaines données (Cf. article 6 de Loi n° 78-753 du 17 juillet 1978, modifié en 2013) dans certaines situations, notamment :

- a) Au secret des délibérations du Gouvernement et des autorités responsables relevant du pouvoir exécutif ;
- b) Au secret de la défense nationale ;
- c) A la conduite de la politique extérieure de la France ;
- d) A la sûreté de l'Etat, à la sécurité publique ou à la sécurité des personnes ;
- e) A la monnaie et au crédit public ;
- f) Au déroulement des procédures engagées devant les juridictions ou d'opérations préliminaires à de telles procédures, sauf autorisation donnée par l'autorité compétente ;
- g) A la recherche, par les services compétents, des infractions fiscales et douanières ;
- h) Ou, sous réserve de [l'article L. 124-4](#) du code de l'environnement, aux autres secrets protégés par la loi ;

II.-Ne sont communicables qu'à l'intéressé les documents administratifs :

-dont la communication porterait atteinte à la protection de la vie privée, au secret médical et au secret en matière commerciale et industrielle ;

-portant une appréciation ou un jugement de valeur sur une personne physique, nommément désignée ou facilement identifiable ;

-faisant apparaître le comportement d'une personne, dès lors que la divulgation de ce comportement pourrait lui porter préjudice.

L'article cité du code de l'environnement apporte une précision importante pour une partie des données produites par le Parc du Vercors :

2° A la protection de l'environnement auquel elle se rapporte ;

La loi précise en outre que lorsque le jeu de données à transmettre comprend des données non communicables à des tiers il est possible d'occulter ou de disjoindre ces dernières de façon à rendre le document transmissible.

Les limites précisées dans la loi concernent le droit d'accès aux données publiques dont peut faire état chaque citoyen, et l'obligation de réponse de l'administration concernée. Si la loi précise que le droit d'accès aux données ne peut plus s'exercer une fois celles-ci publiées, l'administration ayant publié volontairement ses données ne peut faire abstraction des contraintes de publication citées.

Application au cas du Parc

Parmi les contraintes de publication citées, nous identifions les suivantes comme pertinentes compte-tenu du fonctionnement et des données produites par le Parc.

- Les données portant atteinte à la vie privée, portant appréciation sur une personne physique ou divulguant des comportements
- Les données pouvant porter atteinte à la protection de l'environnement

Analyse des contraintes et axes de prévention et résolution

Données portant atteinte à la vie privée

Ces données sont assez faciles à identifier, mais peuvent se nichier dans des jeux de données complexes et il convient d'être attentif lors de chaque publication. Notamment à propos des données géographiques comme par exemple l'occupation des sols, des zonages divers, ou des données statistiques sur la population dont les jeux de données pourraient contenir des noms ou données pouvant formellement identifier des personnes. Il faudra alors avant publication anonymiser les jeux de données en supprimant les colonnes permettant l'identification des personnes.

Protection de l'environnement

La loi impose également que la publication des données ne puisse porter atteinte à l'environnement auquel elles se rapportent. Le texte reste très vague dans la définition du fait de porter atteinte ce qui laisse donc une grande marge d'appréciation... et d'erreur possible.

Parmi les atteintes possibles provenant de la publication de données, il a été fait état de la publication de données géographiques trop précises concernant des espèces protégées : ce type de publication pouvant entraîner une surfréquentation touristique. Cette crainte peut malgré tout se mitiger en publiant des localisations au pas plus large, ou en proposant déjà une synthèse selon différents critères (milieux, zonages, etc.) que l'on maintiendra la plus exhaustive possible de façon à permettre la réutilisation du jeu de données. Il faudra conserver à l'esprit qu'il faut malgré tout publier des données les plus brutes possibles, publier des données déjà traitées fait perdre une grande partie de l'intérêt de l'ouverture des données.

1.2 Où publier ?

Une fois les contraintes précédentes identifiées et traitées le cas échéant, où publier les données à ouvrir ?

La plateforme data.gouv.fr ?

Aujourd'hui, après 3 ans d'existence de data.gouv.fr la plate forme semble être arrivée à maturité et bien être le lieu privilégié de publication pour les administrations et collectivités françaises. Toute ouverture de donnée devrait donc donner lieu à la création d'une fiche sur ce site. Le système de data.gouv.fr permet d'envoyer le jeu de données directement sur leur site, ou bien de n'y indiquer que le lien permettant de la télécharger. Cette souplesse est très appréciable car elle rend possible la publication de données même à des personnes ou institutions ne disposant pas d'un serveur en ligne capable de les héberger. Il est à noter que l'inscription et l'hébergement des données sur data.gouv.fr sont gratuits.

Un autre critère intéressant pour choisir le lieu de dépôt est la fréquence de mise à jour des fichiers de données. Les fichiers fréquemment mis à jour seront avantageusement publiés sur le site du Parc sur lequel vous avez à la fois maîtrise et habitude. Des fichiers mis à jour peu souvent peuvent être envoyés sur le site de data.gouv.fr.

D'autres lieux intéressants :

Le site web du Parc bien sûr. Dans le cas où les données sont publiées sur data.gouv.fr il suffit de publier un lien vers la bonne page du site. Dans le cas contraire où les données sont publiées sur le site du parc, c'est sur le site de data.gouv.fr que la source de données sera publiée non pas directement, mais sous forme de lien vers la donnée à charger sur le site du Parc.

En fonction du jeu de données, d'autres sites dédiés au thème de la donnée peuvent s'avérer intéressants pour que l'information sur la publication touche les personnes potentiellement intéressées.

Une campagne de communication sur le thème des données ouvertes avec des liens fonctionnels pourra constituer un bon complément à la publicité apportée autour de la publication. Ne négligez pas vos partenaires réguliers, surtout s'ils ont contribué à la production des données d'une quelconque manière, qu'elle soit financière ou par ses compétences.

2. Que publier ?

2.1 Que sont les jeux de données ?

Tous les jeux de données que vous avez produites ou compilées peuvent être publiées. Un jeu de données peut être un fichier texte ou csv (données texte séparées par des virgules ou tabulations prêtes à être importées dans un tableur pour traitement), un fichier de couches shp pour un système d'information géographique, un ensemble de photographies...

Il sera possible de distinguer deux choses dans un jeu de données : la base de données elle-même, sa présentation, sa manière de collecter les informations, et les données elles-mêmes : les températures pour un fichier météo, les images pour une collection d'images, etc. Cette distinction pourra amener à choisir des options différentes pour l'application des licences ensuite.

2.2 Privilégier les formats ouverts

Les données informatiques sont présentées, rangées selon des conventions pour que les autres utilisateurs ou programmes parviennent à les lire. Dans la logique d'ouvrir les données, les formats de présentation des données privilégieront donc l'utilisation de formats ouverts. Citons les principaux :

Type de donnée	Formats conseillés	Formats à proscrire
Texte + images	.txt - .odt (Open Document) - .pdf	.doc - .docx
Tableaux de données	.txt - .csv - .ods	.xls - .xlsx
Son	.ogg - .flac - .mp3*	.wmv
Image	.png - .png	.gif
Image de grande taille	.ecw*	
Vidéo (conteneurs)	.ogg - .ogv - .mkv - .avi* - .mp4*	.mov - .wma
Vidéo (codecs)	Xvid, h263, h264	QuickTime
Fichiers de couche pour SIG	.shp	

Les extensions de fichiers suivies d'une astérisque ne sont pas des formats libres, mais sont répandus et peuvent être lus avec des Logiciels Libres, ce qui les rend acceptables.

Les formats de compression admissibles sont : gzip, bzip, tar, et zip.

Note à propos des fichiers pdf : si vous utilisez un générateur de pdf non libre comme celui de l'éditeur Adobe, vérifiez bien à sa création que vous avez réglé les options de manière à obtenir une compatibilité maximale du pdf généré. Régler les options ainsi permet de générer un pdf au format ouvert. Dans tous les cas, vérifiez toujours votre document avec d'autres logiciels que ceux de l'éditeur de manière à vous assurer que celui-ci est bien lisible par tous les lecteurs de pdf.

3. Licences de publication

3.1 Les licences

Le choix de la licence détermine ce qu'il est possible de faire avec les données publiées. Seules certaines licences permettent de qualifier les données de données ouvertes. Les licences trop restrictives limitant certaines réutilisation ne sont pas considérées comme compatibles avec des données ouvertes par l'Etalab. Le site de publication data.gouv.fr propose le choix entre 7 licences au moment de la publication des jeux de données.

- Licence Ouverte
- Open Data Commons Public Domain Dedication and Licence
- Open Data Commons Open Database License
- Open Data Commons Attribution License
- Creative Commons CCZero
- Creative Commons Attribution
- Creative Commons Share-Alike

Cette liste permet de visualiser 3 familles de licences, selon leur rédacteur : Etalab (Licence Ouverte), Open Data Commons (<http://opendatacommons.org>) et Creative Commons (<http://creativecommons.fr/>). Bien que les licences rédigées par opendatacommons.org soient dédiées aux données, les licences de creativecommons (rédigées au départ pour les œuvres artistiques et/ou relevant de la propriété intellectuelle) peuvent tout à fait s'appliquer à des jeux de données, notamment depuis leur version 4 qui a consisté justement en une adaptation du texte aux jeux de données (databases). Afin de mieux cerner les différences entre ces licences, présentons leurs caractéristiques principales sur un tableau.

Licence	Domaine public	Paternité	Redistribution d'une version modifiée	Redistribution à l'identique	Licence de redistribution imposée	Pas de contrainte de redistribution, même commerciale et/ou fermée.
Licence Ouverte		X	X			X
Open Data Commons Public Domain Dedication and Licence	X		X			
Open Data Commons Open Database License ODC		X	X		X	
Open Data Commons Attribution License ODC-BY		X	X			X
Creative Commons CCZero CC0	X		X			X
Creative Commons Attribution CC-BY		X				X
Creative Commons Share-Alike CC-BY-SA		X			X	

À quelques détails près selon le rédacteur, nous pouvons dégager trois familles de licences :

- les licences de domaine public : elles placent les données dans le domaine public, et deviennent ouvertes à toute réutilisation, la mention du créateur des données n'est pas requise.
- Les licences d'attribution de paternité : la Licence Ouverte en fait partie, tout comme les licences de type Attribution. L'attribution de paternité est la seule contrainte imposée par la licence lors de la réutilisation des données.
- Les licences imposant une contrainte lors de la redistribution : que cette contrainte soit la redistribution sous la même licence, ou l'interdiction de l'utilisation commerciale. Bien que ces licences assurent l'ouverture des données, elles peuvent représenter une contrainte à la réutilisation des données, nous détaillons les contraintes ci-après et décourageons l'emploi de telles licences.

3.2 Comment choisir la licence ?

La licence doit être choisie en fonction des données à ouvrir et publier, mais également avec un souci de ne pas compliquer la gestion des données par la suite. Nous rechercherons donc à utiliser le moins de licences différentes, et à utiliser la licence qui n'exerce que des contraintes minimales à la réutilisation.

À ce niveau, il importe de bien faire la distinction entre le fichier de données (la base de données) et les données elles-mêmes. Ces deux parties peuvent être placées sous des licences différentes. Cette distinction a été introduite en droit européen par la directive 1996/9/CE¹. Cette distinction rajoute un niveau de complexité qui n'est pas forcément souhaitable ni pratique à utiliser et peut conduire à des problèmes d'interprétation. En pratique, si la licence attribuée est suffisamment permissive on pourra se contenter de placer l'ensemble sous une licence unique, c'est ce qui se passe par exemple sur la plate forme data.gouv.fr.

Voici deux exemples pour mieux saisir la distinction entre la base de données et les données :

- Un jeu de photographies : lorsque vous publiez un jeu de photographies, l'ensemble de ces photographies constitue la base de données et peut posséder une licence et un producteur représentant votre travail de compilation des données, chaque image pourrait posséder son propre producteur (le photographe) et des conditions d'utilisation différentes. En pratique on préférera publier ensemble des images pouvant être publiées sous la même licence, et dont les photographes auront abandonné les droits au Parc.

- Des données météo : le fichier de données représente le travail de compilation des données issues des capteurs. Actuellement le Parc, en collaboration avec le LTHE, participe à la création des données elles-mêmes. L'originalité du fichier de données réside donc à la fois dans le travail de compilation et dans celui

¹ Voir en annexe une copie de celle-ci : Directive_96-9-CE_du_11-03-1996_fr.pdf
ou sa version en ligne : <http://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:31996L0009&from=FR>

de création des données. Il en irait autrement si par exemple ces données étaient compilées à partir de données ouvertes publiées par une autre partie, dans ce cas l'originalité du travail se situerait uniquement dans la compilation des données, c'est à dire dans la base de données elle-même.

La plate-forme data.gouv.fr ne fait pas cette distinction et englobe le jeu de données et les données elles-mêmes sous le terme « information ». Cette position clarifie l'utilisation des jeux de données en évitant d'introduire plus de complexité que nécessaire, mais impose une réflexion honnête au moment de choisir la licence. En effet si les données ne sont pas originales mais que seule l'est leur compilation, l'attribution de paternité appliquée à l'ensemble deviendrait au mieux exagérée, au pire une usurpation sur l'identité du producteur.

Critère : originalité des données

L'originalité des données est un très bon indicateur pour décider quelle licence appliquer à celles-ci. Par originalité, on entend le caractère unique des données, leur mode de production : a-t-on utilisé un système ad-hoc mis en place pour leur production ou simplement utilisé des données publiques ou produites par un autre organisme public.

Dans le cas où les données sont des données originales, produites par le Parc du Vercors, il sera possible d'utiliser une licence imposant l'attribution de paternité. Dans le cas contraire ce sont les licences de type domaine public qui sont adaptées.

Données originales	Données récupérées, compilées...
Licence Ouverte, CC-BY	CC0, Domaine Public

Ainsi on peut considérer que des images seront toujours des données originales pour peu que le photographe ait cédé ses droits au Parc (que ce soit par contrat, ou parce que la photographie a été réalisée dans le cadre d'un contrat de travail avec le Parc). Les données produites par le Parc, éventuellement en collaboration avec d'autres organismes, peuvent être considérées comme originales. Des données provenant de l'INSEE ou d'autres organismes producteurs de données qui auraient été compilées avec comme seul critère la couverture géographique du Parc ne devraient pas être considérées comme originales ni donner lieu à une attribution de paternité au Parc du Vercors.

Critère : possibilités de réutilisation offertes

Que souhaite-t-on autoriser comme usages et réutilisation des données ?

Puisque nous nous plaçons dans une perspective d'ouverture des données, toutes les réutilisations doivent être permises. Toutes les licences limitant la réutilisation des données introduisent des biais juridiques qui complexifient l'analyse des possibilités de réutilisation, et fragilisent finalement la réutilisation des données d'un point de vue juridique.

Nous déconseillons donc l'utilisation de licences limitant les possibilités de réutilisation par des clauses telles que « redistribution à l'identique » (Share-Alike ; SA) ou « utilisation non commerciale » (Non-Commercial ; NC). Ce point est discuté de façon plus détaillée dans le chapitre suivant.

Faut-il vraiment restreindre certains usages ?

Il est ressorti de discussions avec des techniciens ou chargés de mission du Parc du Vercors que l'ouverture de certains jeux de données pourraient générer des réutilisations pouvant poser problème. Quelques exemples de craintes liées à l'ouverture des données :

- Réutilisation de photos dans des publications à caractère commercial
- Réutilisation des données cartographiques des sentiers balisés par le Parc : concurrence possible avec les Carto-Guides du Parc.

Les questions qui se posent expriment toutes la crainte d'une utilisation à caractère commercial des données libérées qui aurait pu se monnayer si le Parc n'avait pas abandonné ses droits. Dans le cas du Carto-Guide, cette crainte se double d'une possible mise en concurrence sur un produit phare des éditions du Parc du Vercors.

L'objet du présent document n'étant pas de dissenter sur les modes de financement des activités du Parc du Vercors, nous laisserons de côté la question de la pertinence de ces craintes et de leur incidence possible sur le budget du Parc. Concentrons-nous sur les moyens de résoudre, ou pas, la question grâce à l'utilisation de clauses restrictives dans la licence.

1. Premier exemple : les photos

Si le Parc du Vercors décide de publier un jeu de photos, c'est avec la volonté de diffuser ces images le plus largement possible, qu'elles finissent par constituer un jeu d'images « officielles » sur différents aspects du territoire. Ce sont donc des images destinées à être largement réutilisées, sur de nombreux supports, et pouvant servir d'illustration commune et facile à utiliser.

On a donc toutes les chances de les retrouver dans des publications internes, de partenaires du Parc, de structures locales (Offices de Tourisme, Communes, Syndicats, associations...), journaux, magazines, sites web. Nous nous trouverions ici dans des cas d'utilisation tout à fait conformes à l'esprit de la publication : faciliter la communication et la promotion du territoire.

Mais l'on pourra également retrouver les images dans des publications telles qu'un livre contenant des photographies. Or la plupart du temps, la conception d'un tel livre exige de l'éditeur qu'il achète des droits sur les illustrations utilisées, générant des revenus au détenteur des droits des images en question. La publication de photographies sous licence Ouverte ou Domaine Public ferme tout espoir de générer un tel revenu à partir des images ainsi publiées.

Afin d'ouvrir la base de photographies en se ménageant la possibilité de générer du revenu sur celles-ci en cas d'utilisation commerciale on pourrait être tenté de limiter les réutilisations commerciales en ajoutant une clause « Non Commerciale » NC.

Les problèmes principaux posés par une licence NC sont les suivants¹ :

- Les jeux de données risquent de devenir incompatibles avec un ensemble toujours croissant de données libres, même si l'on a autorisé la distribution de jeux modifiés.
- La clause NC peut empêcher légalement des usages et bénéfiques que vous auriez pu vouloir autoriser malgré tout. Ceci pourrait vous demander de redistribuer sous une nouvelle licence les données à un organisme souhaitant réutiliser les données dans un cadre précis. Exemple : les GTV publient un guide faisant usage d'une photographie d'un jeu libéré sous licence CC-BY-NC ; pour des raisons budgétaires l'association décide de vendre ce guide afin de couvrir les coûts de publication ; cela devient une utilisation commerciale de l'image, interdite par la licence. Bien qu'il soit peu probable que le Parc intente un procès contre les GTV pour cette utilisation, l'utilisation de la clause NC a fragilisé juridiquement une réutilisation de l'image, pourtant conforme aux objectifs initiaux de sa publication.
- Ce type de licence finit par apporter du crédit aux problèmes de copyright actuels en voulant distinguer les usages. Cela peut impliquer d'avoir à résoudre des questions juridiques pour savoir si oui ou non telle réutilisation est possible.

Avec une clause « Partage à l'identique » (Share-Alike ; SA) à la place, on empêche pratiquement aussi efficacement une réutilisation commerciale, ce type de réutilisation étant intrinsèquement incompatible avec une licence ouverte. Mais le problème de la fragilisation de certaines réutilisations reste tout aussi présent. Pour conserver l'exemple de réutilisation pour un guide des GTV, la clause SA leur impose de conserver la même licence de publication, ce qui peut constituer une gêne non souhaitée pour eux.

1 Argumentaire inspiré de <http://freedomdefined.org/Licenses/NC>

2. Deuxième exemple : le Carto-Guide

Pour illustrer la crainte exprimée sur la publication des données cartographiques, prenons l'exemple d'un éditeur de cartes souhaitant utiliser ces données cartographiques pour fabriquer un guide concurrent des Carto-Guides du Parc.

Nous ne jugerons pas ici du bien-fondé de cette peur de la concurrence et la tiendrons pour un risque réel dans le raisonnement qui suit.

L'utilisation de la clause NC ou SA comme précédemment interdirait à l'éditeur d'utiliser librement les données cartographiques pour éditer lui-même un guide. De ce point de vue là, ces clauses remplissent l'objectif. Mais elles rendront plus délicates de la même manière la réutilisation des données par des organismes de promotion (Offices de Tourisme, associations telles que GTV) en les plaçant sur une ligne juridique étroite encadrant leur utilisation des données, ce qui est contraire à l'objectif initial de la publication ouverte : que par l'ouverture de ces données, les publications de promotions puissent donner des informations exactes et à jour.

En résumé, l'utilisation de ces clauses restrictives manque l'objectif d'ouverture sur un point essentiel : la libre réutilisation des données. Les données sont ouvertes... mais presque. Il faut se poser des questions juridiques sur chaque réutilisation. Finalement, l'utilisation des clauses SA ou NC, bien loin de répondre au problème, complexifient la question juridique à un point où seul un procès sur un cas donné et précis pourrait apporter une réponse certaine à la légalité de réutilisation sur ce cas donné.

Nous déconseillons donc d'utiliser cette voie juridique pour répondre au risque concurrentiel. La réponse doit être apportée de façon politique, en choisissant d'assumer ce risque (publier ouvert dans le domaine public ou sous licence ouverte avec attribution – LO ; CC-BY) ou en conservant l'avantage concurrentiel que vous procurent ces données (*ie* en ne publiant pas ces données, tout simplement).

3.3 Cas des données produites avec des tiers

Il arrive régulièrement que le Parc du Vercors se retrouve à produire des données en collaboration avec des organismes tiers, ceux-ci apportant leur expertise technique par exemple pour aider le Parc à mener à bien son projet.

Dans l'objectif de pouvoir ouvrir ces données par la suite, le Parc devra ajouter une clause autorisant leur ouverture dans les conventions ou contrats de ces projets. La clause devra préciser les deux points importants :

Identifier la paternité des données

Les données produites dans le cadre de ce contrat seront réputées avoir été produites par : « Le Parc du Vercors », ou « Le Parc du Vercors en collaboration avec le Laboratoire ZZZ » Cette dénomination devra figurer dans le texte de la licence, et figurer dans le descriptif des données fournies sur data.gouv.fr, l'organisme collaborateur reconnaissant par ailleurs que le producteur officiel des données sur ce site sera le Parc du Vercors seul pour des raisons de simplicité technique.

Accord de publication sous la licence X, Y ou Z

L'organisme collaborateur donnera par la convention son accord pour publier les données produites sous la licence X, Y ou Z, voire un ensemble de celles-ci si la licence finale n'est pas encore choisie et qu'il est possible de laisser ce point flou.

Texte possible de l'accord

Tierce partie : le Labo Z ; paternité des données : « Parc du Vercors et Labo Z » ; licence : Licence Ouverte

Les données produites dans le cadre de cette collaboration entre le Laboratoire Z et le Parc du Vercors seront publiées sous la licence : Licence Ouverte, dont le texte se trouve ici : https://wiki.data.gouv.fr/images/9/9d/Licence_Ouverte.pdf. Les données seront réputées produites par le « Parc Naturel Régional du Vercors et le Labo Z » dans les termes de la licence. Le Parc Naturel Régional du Vercors se chargera de la compilation et de la publication des données, notamment via la plate forme data.gouv.fr en veillant à respecter le présent engagement sur la licence et l'auteur des données. Le Labo Z comprend que les données apparaîtront comme produites par le seul Parc du Vercors sur l'interface de la plate forme pour des raisons techniques, mais que le nom complet des producteurs apparaîtra en entier dans les descriptions du ou des jeux de données.

Variantes :

Texte de la licence : il peut être proposé de l'imprimer en annexe de la convention.

Producteur : Il peut être possible de créer un producteur ad-hoc pour assurer une égale publicité aux deux organismes si cela s'avère nécessaire.

3.4 Conseils sur deux licences

Le Domaine Public

Placer les données dans le domaine public revient à abandonner tous vos droits sur celles-ci et d'en faire par là un bien public, commun.

La loi française ne permet pas de placer des œuvres dans le domaine public juste par intention de l'auteur, c'est le temps (70 ans après sa mort) ou l'extrême simplicité des données qui les placent automatiquement dans le domaine public.

Il est cependant possible pour un producteur de données de renoncer à tous ses droits sur la base de données ainsi publiée, ce que l'on peut réaliser en appliquant la licence CC0¹.

La Licence Ouverte

La Licence Ouverte d'Etalab présente de nombreuses qualités :

- Elle assure la paternité des jeux de données tout en laissant la plus grande liberté de réutilisation.
- Elle est concise : trois pages. Elle est facile à lire et aisément compréhensible, même sans grande culture des textes juridiques.
- Elle est très ouverte et autorise tous les types de réutilisations, et satisfait donc pleinement au projet initial.
- Elle a été pensée pour les données et pas seulement les œuvres de création et s'applique donc naturellement aux jeux de données.

C'est cette licence que je recommande vivement d'appliquer pour l'ensemble des données à ouvrir.

Licences compatibles : Creative Commons Attribution (BY/paternité) et Open Data Commons Attribution (ODC-BY)

La licence CC précise en outre qu' « aucune restriction supplémentaire (légale ou technique), interdisant à d'autres de faire ce que la présente licence permet, ne peut être introduite. » ce qui restreint la possibilité d'application de restriction digitale des droits (DRM) aux données réutilisées.

1 Voir texte de la licence sur le pdf joint : CCO-fr.pdf

4. Comment publier les données ?

Pour réussir la publication de jeux de données, il est nécessaire d'entourer le jeu de données de toutes les métadonnées descriptives utiles à leur compréhension, améliorant leur visibilité et permettant de se faire une idée des traitements possibles. Il faudra également veiller à ce qu'elles soient facilement accessibles.

Le site data.gouv.org organise les données de la façon suivante : on y publie un jeu de données qui lui-même va contenir une ou plusieurs données. Le jeu de données contient les métadonnées utiles à l'indexation, la recherche et permettant d'évaluer l'intérêt des données pour leur réutilisation. Chaque fiche de donnée contient les informations d'accès à la donnée, de format de fichier, et du contenu de celui-ci.

<http://data.gouv.fr>

S'y connecter comme utilisateur membre de l'organisation « Parc Naturel Régional du Vercors ». Dans le cadre de cette étude, l'organisation et un utilisateur ont été créés, les identifiants sont détenus par le chargé de mission, l'organisation « Parc Naturel Régional du Vercors » a été certifiée comme organisme public reconnu.

4.1 Jeu de données

1. Titre

Le titre doit décrire les données en quelques mots. Il ne doit pas contenir d'indications temporelles ni être trop long.

2. Description

La description du jeu de données doit intégrer tous les mots-clés pertinents pour la recherche du jeu de données. Elle pourra contenir également des liens vers d'autres ressources ou informations complémentaires, typiquement une page du site du Parc décrivant les données ou le programme dans lequel elles s'inscrivent.

3. Licence

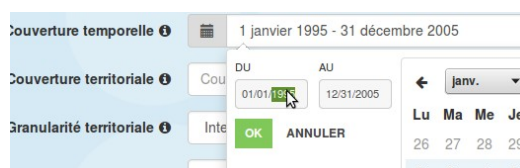
Choisir une licence ouverte, de préférence la Licence Ouverte.

4. Fréquence

La fréquence indique le niveau de détail temporel à l'intérieur de la période couverte. Si les données sont par exemple des données par mois, indiquer « mois ». Cette indication n'implique pas que vous devez mettre à jour les données tous les mois !

5. Période de couverture

Indiquer de date à date la période concernée. S'il s'agit d'une année entière, tant pis indiquer du 1^{er} janvier au 31 décembre de la même année.



The image shows a user interface for selecting a temporal coverage period. The main field is labeled 'Couverture temporelle' and contains the text '1 janvier 1995 - 31 décembre 2005'. Below this, there are two input fields: 'DU' with the value '01/01/2005' and 'AU' with the value '12/31/2005'. To the right of these fields is a dropdown menu currently set to 'janv.'. Below the date fields are two buttons: a green 'OK' button and a grey 'ANNULER' button. At the bottom right, there is a small calendar grid showing the days of the week (Lu, Ma, Me, Je) and the numbers 26, 27, 28, 29.

6. Couverture territoriale

La couverture territoriale indique le périmètre géographique couvert par les données, cela peut être la France, la région Rhône-Alpes, le département de la Drôme ou de l'Isère.

Malheureusement le territoire du Parc Naturel du Vercors ne fait pas partie de la liste, l'amélioration a été demandée. En attendant il est suggéré de saisir les deux départements, ou seulement l'un des deux si le jeu de données n'en concerne qu'un seul.

7. Granularité territoriale

On demande d'indiquer ici, comme la fréquence vis à vis de la période de couverture, le niveau de détail contenu par les données : s'agit-il de communes, de pays, canton, ou même va-t-on jusqu'au POI (Point d'Intérêt) ?

8. Tags – mots clefs

Le système des mots-clefs permet de raccrocher le jeu de données aux autres abordant les mêmes thèmes. Les mots-clefs sont déjà nombreux, essayez au maximum de réutiliser les existants !

Prenez un peu de temps pour aller voir les mots-clefs utilisés sur des jeux de données similaires à ceux que vous vous apprêtez à publier.

9. Privé

Lorsque vous cochez cette case, le jeu de données n'est pas visible sur le site, il n'est pas publié. Utilisez ce mode lors de la création de vos jeux de données et de la mise en ligne des données correspondantes. Vous reviendrez décocher la case pour publier le jeu de données une fois qu'il sera prêt.

10. Valider

Une fois ce travail terminé, validez la fiche du jeu de données. Vous devriez voir une fiche ressemblant à celle-là :

The screenshot shows a data.gov.fr page for 'Occupation du sol 1995-2005'. The page header includes the data.gov.fr logo, the text 'Plateforme ouverte des données publiques françaises', and navigation links like 'Comment ça marche?', 'Données', 'Tableau de bord', 'Dataconnexions', 'Etablab', and 'CADA'. A search bar with 'Recherche' and a 'Thématiques' dropdown menu are visible. The main content area features the title 'Occupation du sol 1995-2005', a note that the data is from a certified public service, and a description: 'Fichiers shape d'occupation des sols sur le Parc Naturel Régional du Vercors réalisés en 1995 et 2005. La troisième source de données (à paraître) montrera l'évolution sur la décennie concernée.' Below this is a 'Ressources' section with a '+ Ajouter une ressource' button. On the right, there is a 'Producteur' section with a green logo of a tulip and a bird, and a red 'PRODUCEUR CERTIFIÉ' seal. The text below the logo reads: 'Le Parc Naturel Régional du Vercors, à cheval sur l'Isère et la Drôme, regroupe 85 communes et 5 villes portes (Grenoble, Crest, Romans, Saint-Marcellin et Vinay) et abrite...'

4.2 Ressources (les données proprement dites)

Vous pouvez maintenant ajouter une ou plusieurs données au jeu de données en cliquant sur le + « Ajouter une ressource »

La recommandation d'Etalab est de fournir des fichiers les plus denses possibles c'est à dire contenant le maximum de données. En effet, il est plus pratique de traiter un tableau contenant 50 colonnes qu'à n'en utiliser que 20 plutôt que de devoir traiter 10 fichiers différents pour obtenir les mêmes informations.

Bien entendu, selon la présentation des données de départ cet objectif est plus ou moins aisé à atteindre, il faudra donc mettre en balance la facilité de traitement obtenue par un retravail des données, et le travail nécessaire pour y parvenir ainsi que le risque d'introduction d'erreurs induit par la manipulation.

1. Titre

Le titre de la ressource précisera de préférence les critères, les unités, la granularité qui décrit le mieux la ressource, plutôt que de recopier le titre du jeu de données.

2. Description

Celle-ci est optionnelle. Vous pouvez l'utiliser pour apporter des précisions sur le contenu des données, sur les méthodologies d'acquisition, etc.

3. Type

Fichier uploadé :

Choisissez dans votre ordinateur le fichier à envoyer sur le site.

Fichier distant

Indiquez l'URL du fichier dans le champ.

API

Si la publication n'est pas un jeu de données proprement dites, mais un ensemble de méthodes pour accéder à des données que vous publiez. Les ressources jointes seront donc plutôt de la documentation expliquant comment accéder aux données. Exemple d'API : « Prochains départs des trains Francilien » publiée par la SNCF et qui permet d'interroger directement leur base de données des trains au départ dans ce réseau. Cette forme est plutôt adaptée aux données en temps réel, ainsi qu'à certaines réutilisations.

4. Format du fichier

Une fois le fichier envoyé ou l'url de celui-ci saisie, le système tentera de deviner les données relatives au fichier.

Attention, si votre fichier est compressé par zip, le format indiqué sera « zip » et son type mime « application/zip » ou « application/zip-compressed » sans tenir compte du contenu réel du fichier (image ecw, jpg, document pdf, odt, ods etc.), c'est pourquoi il est important de le préciser dans la description.

5. Taille du fichier

Si le système ne l'a pas devinée, indiquez la taille du fichier en octets (attention cela peut faire de grands nombres)

6. Somme de contrôle

Si vous avez envoyé le fichier, le système calculera de lui-même un condensé « sha1 » du fichier : contrôlez son intégrité en appliquant le même condensé sur le fichier original présent sur votre ordinateur (sur un système Unix ou Linux la commande devrait ressembler à 'sha1sum monfichier.zip'). Le calcul de cette somme de contrôle permet de s'assurer que le fichier n'a pas été détérioré pendant son transfert.

7. Valider

Voilà, vous y êtes, votre ressource est en ligne !

L'Etalab propose un guide de publication actualisé sur la page suivante :

https://wiki.data.gouv.fr/wiki/Guide_de_publication

5. Exemples de publications

Des exemples de publication sont disponibles en ligne sur la page de l'organisation Parc du Vercors de data.gouv.fr :

<https://www.data.gouv.fr/fr/organizations/parc-naturel-regional-du-vercors/>