Science Ouverte : sauvegarder, visualiser et partager vos données

Régis Witz
Maison Interuniversitaire des Sciences de l'Homme - Alsace, Université de Strasbourg
rwitz@unistra.fr

Julia Sesé
Direction du Numérique, Université de Strasbourg
julia.sese@unistra.fr

Ana Schwartz
Direction du Numérique, Université de Strasbourg
ana.schwartz@unistra.fr

Stéphanie Cheviron
Service des bibliothèques, Université de Strasbourg
scheviron@unistra.fr

Vincent Lucas
Direction du Numérique, Université de Strasbourg
lucas@unistra.fr

Résumé

Le travail scientifique représente bien plus que des publications. En effet, les échecs et expérimentations non publiés constituent une réelle richesse de connaissances inexploitées.

Ainsi, conserver et diffuser ces données de recherche permet d'étayer les succès, mémoriser les erreurs et révéler des idées encore inexplorées. Sauvegarder, référencer et exposer ces données est un gage de pérennité et de traçabilité, sur lequel se fonde la méthode scientifique basée sur la reproductibilité et la réutilisation des résultats.

Dans le contexte de la science ouverte, l'Université de Strasbourg s'investit dans le développement de la plateforme POUNT, écosystème libre, modulaire et interopérable d'accès aux savoirs :

- *Sauvegardez et versionnez vos données (documents, images, vidéos, modèles 3D) ;*
- *Structurez vos données en respectant les standards de votre discipline, et étendez ces standards avec vos métadonnées personnalisées ;*

- *Partagez vos données avec un identifiant unique, un modèle de citation, configurez les droits de lecture et de modification pour vos communautés ;*
- *Valorisez vos données et visualisez-les de manière fluide, enrichissez-les avec des informations contextuelles et des hyperliens.*

Cet article présente le socle technologique permettant d'implémenter ces fonctionnalités ainsi que l'infrastructure sous-jacente. Il détaille également la visualisation des données, basé sur l'exemple de modèles 3D. Enfin, il décrit l'ouverture à la communauté par le respect des standards et la distribution du code sous licence libre.

La conclusion résume ces fonctionnalités et présente les améliorations envisagées comme la possibilité pour chacun de contribuer au projet ou de déployer sa propre instance de POUNT, capable de s'interconnecter entre elles et avec diverses sources de données.

Mots-clefs

POUNT, science ouverte, pérennité, métadonnées, valorisation, transdisciplinarité, ouverture à la communauté

1 Introduction

Que deviennent les données de nos chercheurs ? Quand un doctorant finit sa thèse, ses données de recherche sont-elles perdues pour son laboratoire ? Un article de recherche est-il reproductible par la communauté ? Autant de questions qui montrent l'importance cruciale et l'enjeu de la gestion et de l'exposition des données pour l'avancée scientifique.

Le travail scientifique représente bien plus que les résultats contenus par les publications, les échecs et expérimentations non publiés sont une réelle richesse de connaissances inexploitées. Conserver et encourager la diffusion de ces données permet à la fois d'étayer ces succès, de mémoriser les erreurs à ne pas reproduire, ainsi que de révéler des idées encore inexplorées. Les sauvegarder et les diffuser est un gage de pérennité et de traçabilité : cela aussi bien en termes de reproductibilité que de réutilisation pour la communauté scientifique et pour le grand public.

Après la mise en place d'une archive ouverte institutionnelle univOAK [1][2], l'Université de Strasbourg continue son engagement dans la diffusion des résultats de recherche via POUNT, la Plateforme Ouverte Numérique Transdisciplinaire. POUNT propose un écosystème d'accès aux savoirs où chaque chercheur peut :

- *Sauvegarder :*

  Déposer des données (documents, images, vidéos, modèles 3D, etc), conserver un historique, étiqueter des versions, le tout avec des fichiers contenant les données brutes ou des formats de compression sans perte.

- *Structurer :*

  Décrire ses données en respectant les standards, référentiels DataCite [3], Dublin Core [4] et Darwin Core [5], nomenclatures, taxonomies comme la base de paléobiologie [6], etc. et étendre les standards avec des métadonnées personnalisées.

- *Partager :*

  Générer un identifiant unique, un modèle de citation, configurer les droits de lecture et de modification pour ses communautés et publier. L'interopérabilité et l'accessibilité sont assurées dans le respect des protocoles d'échanges standards, une API RESTful ouverte [7] et documentée (cf. Swagger/OpenAPI [8]).

- *Valoriser :*

  Exposer ses données avec des modules de visualisation dédiés, tel que 3DHOP [9], permettant un affichage fluide et multi-résolutions ou par téléchargement des données brutes. Enrichir ses modèles avec des informations contextuelles pouvant intégrer des hyperliens.

Cet article détaille des propositions sur la manière de structurer et de partager vos données en passant par les structures standards de données, les formats de stockage et de visualisation, ainsi que l'architecture déployée.

2 Structurer vos données

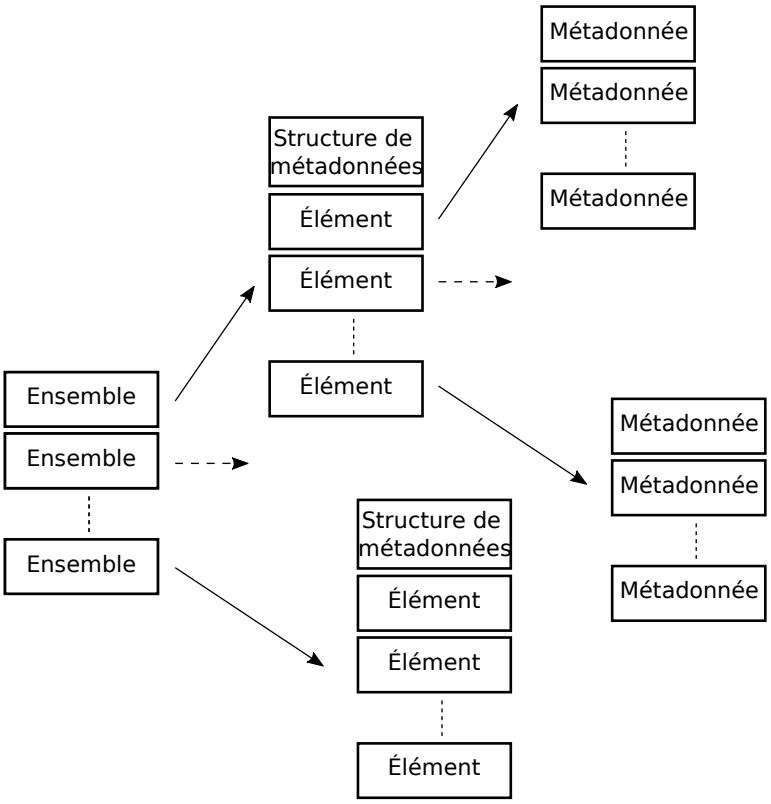

Figure 1 - Concepts :ensemble, structure de métadonnées, élément, métadonnées

POUNT est un outil pour aider les chercheurs à structurer et à diffuser leurs données. Ainsi, pour permettre d'organiser les données 4 concepts (cf. figure 1) ont été définis :

- Les e*nsembles :*

  Chaque ensemble correspond à une liste d'éléments de même nature. Par exemple une collection de minéralogie, des fouilles archéologiques, etc. Tous les éléments d'un ensemble ont la même structure de métadonnées.

- Les *structures de métadonnées* :

  C'est la définition unique de tous les éléments contenus dans un ensemble. Chaque champ est défini comme obligatoire ou facultatif, et est typé : texte, entier, date, fichier binaire, image 3D, vidéo, etc.

- Les *éléments* :

  Chaque élément représente un échantillon, par exemple une roche, un sarcophage, etc. Chacun contient les métadonnées définies par la structure de métadonnées de l'ensemble.

- Les *métadonnées* :

  Ce sont les informations stockées. Chaque métadonnée est la plus atomique possible. Son caractère obligatoire ou non, ainsi que son type, sont définis par la structure de métadonnées de l'ensemble.

Le but de cette organisation est de former des ensembles cohérents contenant des informations homogènes, pouvant être référencées et analysées de façon automatisée.

2.1 Métadonnées et standards

Le créateur d'un ensemble est chargé de spécifier la structure de métadonnées qui en constituera le modèle.

La structure des métadonnées est libre, mais doit comporter au minimum un titre.

Il existe des standards de structures de métadonnées associées à des thématiques et domaines de recherche spécifiques, tels que :

- DataCite [3]: format de métadonnées pour l'identification précise des ressources et leur citation ;
- Dublin Core element set [4]: format de métadonnées générique composé de quinze propriétés de base facultatives relatives au contenu (titre, sujet, description, source, langue, relation, couverture) à la propriété intellectuelle (créateur, contributeur, éditeur, gestion des droits) et à l'instanciation (date, type, format, identifiant de la ressource) ;
- Darwin Core [5]: dérivé du Dublin Core et spécialisé pour les données de biodiversité.

À la création d'un ensemble, POUNT suggère d'utiliser par défaut le modèle spécifié par DataCite car celui-ci est concis, compréhensible et adapté à toutes les disciplines. Les chercheurs peuvent adapter et enrichir cette structure en fonction de leurs besoins.

2.2 Autres métadonnées des ensembles

Un ensemble est également décrit par des métadonnées spécifiques, telles que :

- son titre ;
- sa description ;
- ses droits d'accès : visibilité et droits de modification pour certains groupes d'utilisateurs spécifiques ou pour le grand public.

Une fonctionnalité permettant aux utilisateurs de partager leurs modèles de métadonnées entre eux est actuellement en cours de développement. Cela permettra de favoriser le travail en groupe, d'enrichir et de propager les bonnes pratiques de structuration des données de recherche.

3 Sauvegarder et partager vos données

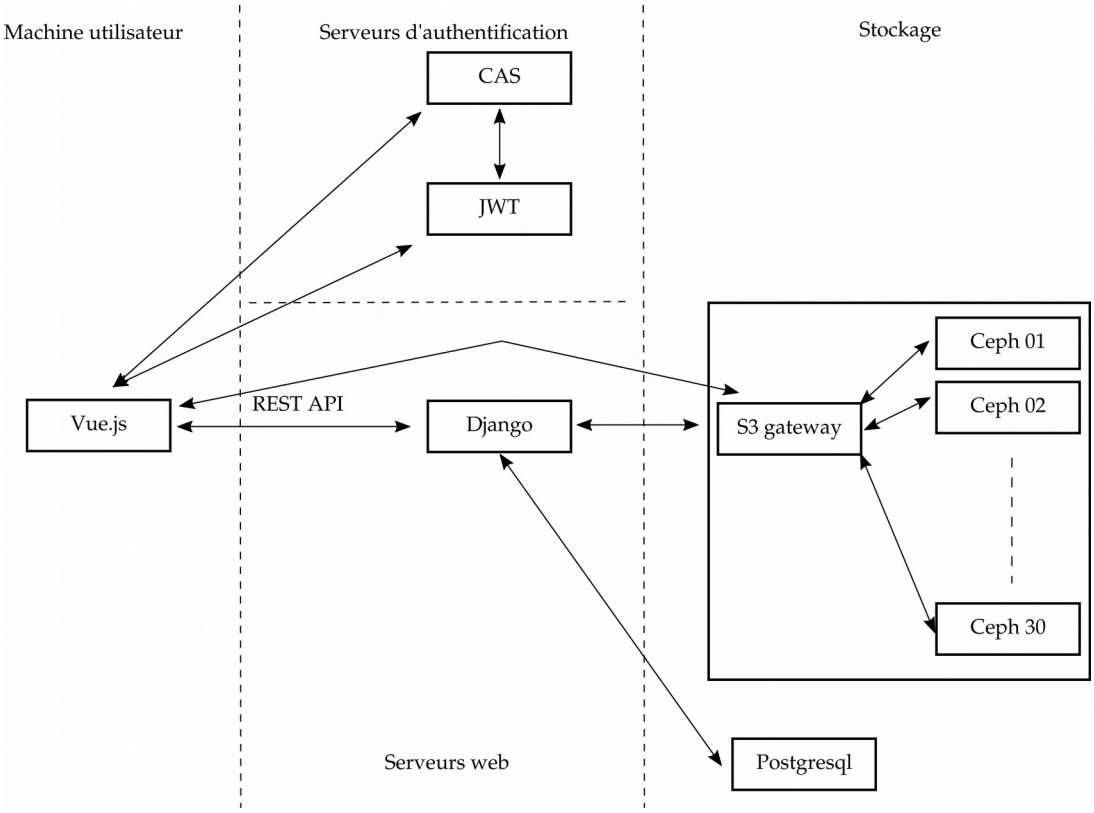

Figure 2 - Architecture

L'architecture de POUNT se compose de trois modules principaux : du stockage, une API REST et une interface web.

3.1 Stockage

Le stockage retenu, Ceph, présente l'avantage d'être extensible, distribué et redondé. Cela permet une évolution à la fois en termes de volumétrie, de performances et d'éviter la présence d'un unique point de défaillance (single point of failure).

L'accès au stockage se fait par l'API REST S3 [10][11], qui permet non seulement d'avoir une interface d'accès standardisé, mais également de pouvoir accéder directement aux données via les passerelles S3. L'utilisateur profite ainsi de performances optimales, améliorant ainsi le débit et la qualité de service.

3.2 API REST

Au cœur de POUNT réside une API REST, proposant toutes les fonctionnalités d'accès et de modification des ensembles, des éléments, des modèles de métadonnées, ainsi que de la gestion des utilisateurs et des groupes. Cette API est implémentée en python et servie par un serveur Django et un serveur web nginx. Cette API sert de routeur pour accéder aux données stockées dans le cluster Ceph, aux données utilisateurs stockées dans la base de données Postrgesql. Cette API sert également de contrôleur pour valider les données saisies et la validité des tokens d'authentification.

Le but est de garder ce serveur modulable afin de pouvoir le connecter avec d'autres types de stockage et d'agréger plusieurs espaces de stockage comme d'autres serveurs POUNT, ou d'autres solutions.

Cette API est spécifiée via un fichier de description swagger [8]. Un serveur web expose cette spécification permettant à la fois la diffusion de cette API et la création de requêtes simples. Cela permet à tout client ou script tierce d'interroger l'API REST de POUNT et ainsi d'afficher les données d'un ensemble sur un site tiers, ou encore de partager et d'exporter des ensembles.

3.2.1 Authentification

Dans la configuration actuelle, l'authentification est réalisée de façon centralisée avec un serveur CAS-JWT. L'évolution de l'authentification est de pouvoir utiliser :

- une fédération d'identités Shibboleth pour les membres de la communauté universitaire ;
- une authentification 0Auth, permettant aux utilisateurs non universitaires de s'authentifier à travers leurs comptes de messagerie ou de réseaux sociaux.

Le but est de garder le mécanisme d'authentification indépendant et configurable.

En regard du traitement de ses données utilisateurs, POUNT n'a aucun but commercial et est en accord avec le Règlement Général sur la Protection des Donnéesi (RGPD) [20] [21].

3.2.2 Autorisation

Chaque élément est publié avec des droits, qui peuvent être restreints en lecture/écriture pour certains utilisateurs ou groupes d'utilisateurs. Cela permet de gérer des situations comme celles de données sous embargo, ou celles du travail non encore publié d'une équipe de recherche.

3.3 Interface web

L'interface web est réalisée avec les technologies standards du Web telles que HTML5, CSS3 et javascript avec le framework Vue.js [12].

Cette interface fonctionne dans le navigateur de l'utilisateur afin de :

- requêter l'API REST de Django et l'API S3 afin de récupérer les listes d'ensembles et d'éléments disponibles ;
- afficher une visualisation des ensembles et éléments.

Ainsi, la majorité du travail est déporté sur chaque poste utilisateur permettant d'augmenter la capacité de passage à l'échelle de POUNT.

De plus, grâce à Vue.js, POUNT est implémenté sous forme de composants indépendants, légers et réutilisables. Cela rend l'interface web évolutive, capable d'afficher des métadonnées variées telles que du texte, des images, des vidéos, des pistes sonores, des modèles 3D, etc.

Une fonctionnalité permettant de générer automatiquement le code source nécessaire à inclure tout ou partie de la page d'un élément ou d'un ensemble dans un contexte externe est actuellement en cours de développement. Ainsi, les chercheurs pourront valoriser leurs données dans d'autres contextes que POUNT.

Chaque donnée est visualisée à l'aide d'une visionneuse appropriée. Le choix de cette visionneuse dépend du format du fichier, mais aussi de la nature de ce qu'elle représente. Par exemple, l'acquisition d'un ancien manuscrit, ou une orthophotographie peuvent toutes deux être des images, mais peuvent requérir chacune une visionneuse spécifique différente. POUNT peut mettre à profit les métadonnées de chaque élément, ou les métadonnées de configuration de l'ensemble concerné, pour choisir une ou plusieurs visionneuses appropriées, améliorant ainsi l'expérience utilisateur ainsi que la valorisation de chaque donnée.

POUNT peut aussi agréger sous forme visuelle plusieurs métadonnées d'un même élément, ou la même métadonnée d'éléments différents. Par exemple, la page d'un élément comporte un espace dédié à la manière de citer la donnée en question. Ce modèle de citation peut être configuré par le propriétaire de l'élément, en en combinant les métadonnées telles que le titre, l'auteur, la date de publication, le DOI, etc. D'autres visionneuses permettant de combiner des métadonnées sous forme graphique (tels qu'histogrammes, camemberts, timelines, etc) sont actuellement à l'étude.

3.3.1 Visionneuse 3DHOP

Dans l'interface web, les modèles 3D sont affichés via la visionneuse 3DHOP [9]. Implémentée en HTML et Javascript, elle s'intègre très facilement à n'importe quelle page web.

Elle propose plusieurs modes de caméra, ainsi que plusieurs outils intéressants, tels que des outils de mesure, des plans de coupe, ou l'affichage de points d'intérêt. 3DHOP sera aisément extensible au gré des besoins des utilisateurs de POUNT, afin de l'adapter au mieux à des scénarios de visualisation spécifique, et toujours mieux mettre en valeur chaque donnée.

4 Formats de fichiers

POUNT a deux objectifs principaux relatifs aux données :

- sauvegarder chaque donnée de manière pérenne ;
- afficher rapidement chaque donnée dans le cadre d'une expérience web interactive et réactive.

Pour les données volumineuses comme les modèles 3D, remplir ces deux objectifs requiert l'utilisation de deux formats de fichier différents.

4.1 Formats de stockage

Pour la pérennité des données, le format de stockage doit être ouvert, interopérable, stable et sans perte lors du codage et décodage. Pour l'interopérabilité, sa spécification doit être claire et exhaustive afin d'éviter les implémentations aux comportements hétérogènes.

Pour les modèles 3D, POUNT a fait le choix du format COLLADA *digital asset exchange* [13] comme format de stockage. Ce format est bien établi, supporté par de nombreux outils et est actuellement l'unique recommandation du Centre Informatique National de l'Enseignement Supérieur (CINES) [14] dans le domaine de la 3D.

4.2 Formats de visualisation

Un format de visualisation doit offrir le meilleur compromis entre trois considérations :

- La vitesse de téléchargement qu'il permet. Dans des conditions équivalentes, plus le fichier est de petite taille, plus il sera téléchargé rapidement. Les techniques de compression et de décompression ont une influence directe sur cette taille ;
- Une fois téléchargé, le fichier doit pouvoir être visualisé rapidement et de manière fluide, avec de bonnes performances ;
- Les possibilités de mise en valeur qu'il offre : sources d'éclairage, richesses des textures, transformations, filtres, et ainsi de suite.

Les formats de visualisation dépendent également de la visionneuse utilisée. Par exemple, 3DHOP [9] utilise les fichiers *Polygon File Format*(PLY) [18] ainsi que la compression Nexus [19] pour les fichiers multi-résolutions. Ces formats permettent un transfert et un affichage rapide des données.

5 Partage

POUNT est une plateforme ouverte qui veut faciliter les échanges, diffuser les savoirs, préserver et valoriser les données qu'elles aient une valeur patrimoniale ou pas. POUNT offre à l'ensemble des disciplines une plateforme adoptant les principes FAIR [22], c'est-à-dire rendre les données :

- Faciles à trouver (Findable) ;
- Accessibles (Accessible) ;

- Interopérables (Interoperable) ;
- Réutilisables (Reusable).

De fait, partager la connaissance favorise l'innovation, l'avancée de la recherche en mutualisant les efforts.

Il en va de même pour le code de POUNT [7][15], qui est librement accessible et modifiable sous licence GNU Affero GPL [16], afin de permettre à n'importe quelle institution de se l'approprier. En effet, les développements de POUNT sont orientés afin de ne pas avoir de dépendance vis-à-vis d'une architecture ou des services d'une université spécifique. De conception ouverte, POUNT utilise systématiquement des logiciels libres, fait preuve de transparence et de pédagogie sur sa conception, son implémentation, ses objectifs ainsi que les alternatives existantes [17].

Le but est de permettre l'installation de POUNT dans chaque établissement le désirant et de créer une communauté pour que ce projet puisse évoluer et grandir en y intégrant de nouveaux services.

6 Conclusion

Pour conclure, POUNT est un logiciel pour stocker, partager et diffuser les données. Cela demande un travail à la fois sur la structuration des données, le format de stockage et de visualisation. Dans cet article, nous avons détaillé l'architecture choisie pour mettre en place ce service, dont notamment la mise en place d'un API REST permettant à tous de lister, examiner et télécharger les ensembles de données.

Ce type d'application résulte une demande croissante de la part des chercheurs, laboratoires et établissements d'enseignements supérieurs. Cela confirme les politiques de science ouverte, aussi bien dans le contexte national [23] [24] qu'européen [25] et mondial [26]. Ce projet s'inscrit donc dans l'idée de partage universel des savoirs et vise à valoriser, fédérer et rendre accessible les données de recherche au plus grand nombre. Cette plate-forme est une solution libre, gratuite, ouverte, modulaire et interopérable. De plus, une évolution envisagée est de permettre à chaque université de créer sa propre instance de POUNT et même de l'interconnecter afin de devenir une plate-forme d'échange inter-universités.

Bibliographie

[1] Schwartz Ana, Rege Adeline, Joncour Sylvain et Gerber Antoine. JRES 2017. Ouvrir les résultats de la recherche universitaire : retour d'expérience du site Alsace

[2] Plateforme univOAK: https://univoak.eu

[3] DataCite: détail du format de métadonnées : hthttps://schema.datacite.org/

[4] Dublin Core : vocabulaire du web sémantique utilisé pour exprimer les données dans un modèle RDF. : https://www.dublincore.org/

[5] Darwin Core : extension du Dublin Core pour la bio-informatique : http://rs.tdwg.org/dwc/

[6] The Paleobiology Database : https://paleobiodb.org

[7] POUNT : code source de l'API : https://git.unistra.fr/community/pount-api/

[8] Swagger / Open API Initiative (OAI) : projet open source provenant de l'ouverture à la Linux Foundation de la spécification Swagger. Framework pour définir et créer des APIs RESTful : https://www.openapis.org/

[9] Marco Potenziani, Marco Callieri, Matteo Dellepiane, Massimiliano Corsini, Federico Ponchio er Roberto Scopigno. 3DHOP: 3D Heritage Online Presenter. Computers & Graphics, volume 52, 2015, pages 129-141

[10] RADOS Gateway - Ceph documentation : https://docs.ceph.com/docs/bobtail/radosgw

[11] Amazon S3 REST API : https://docs.aws.amazon.com/AmazonS3/latest/API/Welcome.html

[12] Site du framework Vue.js : https://vuejs.org/

[13] COLLADA : Format de fichier d'échanges utilisé pour la 3D : https://www.khronos.org/collada/

[14] Centre informatique national de l'enseignement supérieur : recommandations de formats : https://www.cines.fr/archivage/des-expertises/les-formats-de-fichier/

[15] POUNT : code source de l'interface web : https://git.unistra.fr/community/pount-front/

[16] Licence libre GNU GPL : https://www.gnu.org/licenses/licenses.fr.html

[17] HumaNum : liste des services proposés par cette Très Grande Infrastructure de Recherche en sciences humaines : https://www.huma-num.fr/services-et-outils

[18] PLY : Polygon File Format, format de description 3D : http://paulbourke.net/dataformats/ply/

[19] Nexus : suite d'outils pour la 3D multi-résolution : http://vcg.isti.cnr.it/nexus/

[20] Règlement général sur la Protection des Données (RGPD) : hhttps://eur-lex.europa.eu/legal-content/FR/TXT/HTML/?uri=CELEX:32016R0679/

[21] Commentaires de la CNIL sur la RGPD : https://www.cnil.fr/fr/rgpd-par-ou-commencer

[22] Principes « FAIR » : https://www.go-fair.org/fair-principles/

[23] Loi n ◦ 2016-1321 pour une République numérique : https://www.legifrance.gouv.fr/affichTexte.do?cidTexte=JORFTEXT000033202746

[24] Plan pour la science ouverte : http://www.enseignementsup-recherche.gouv.fr/cid132529/le-plan-national-pour-la-science-ouverte-les-resultats-de-la-recherche-scientifique-ouverts-a-tous-sans-entrave-sans-delai-sans-paiement.html

[25] Open science and its role in universities : a roadmap for cultural change : https://www.leru.org/publications/open-science-and-its-role-in-universities-a-roadmap-for-cultural-change

[26] Certification CoreTrustSeal : https://www.coretrustseal.org/why-certification/requirements/