

RÉPUBLIQUE FRANÇAISE
Ministère de la culture et de la
communication
Direction générale des patrimoines
Service interministériel des Archives de
France

**Note d'information DGP/SIAF/2014/007 du 7 octobre 2014
relative à la mise en œuvre d'un environnement de création, de gestion et de
diffusion de vocabulaires (listes d'autorité, thésaurus, terminologies)**

Le directeur chargé des Archives de France

à

Mesdames et Messieurs les directeurs des services des Archives nationales
Mesdames et Messieurs les directeurs des services départementaux d'archives
sous couvert de Mesdames et Messieurs les préfets de région
et de Mesdames et Messieurs les préfets de département
Mesdames et Messieurs les responsables des services d'archives communales
sous couvert de Mesdames et Messieurs les maires
Mesdames et Messieurs les responsables des services d'archives régionales
sous couvert de Mesdames et Messieurs les présidents de région

Depuis 2008, les Archives de France participent activement au programme HADOC (Harmonisation de la production des données culturelles), piloté par le département des systèmes d'information patrimoniaux de la direction générale des patrimoines. Ce programme, transverse à l'ensemble des directions du ministère de la Culture et de la Communication (MCC), vise à accroître la qualité des données et à améliorer les processus de production, en agissant selon trois axes :

- l'harmonisation des données, qui doit faciliter la constitution, à plus ou moins long terme, de référentiels partagés (listes d'autorité, thésaurus, terminologies, etc.), véritable ossature du futur système d'information documentaire du ministère ;
- l'harmonisation des processus qui doit conduire à la mise en œuvre d'une nouvelle chaîne de production permettant de pallier l'obsolescence fonctionnelle et technique de l'actuelle chaîne de production ;
- la généralisation des normes et des standards de production.

Les développements les plus récents autour du web sémantique ont conduit l'équipe projet à intégrer cette problématique dans la manière d'envisager la production et la mise à disposition des données. C'est ainsi que la plate-forme conçue en 2010 par le Service interministériel des Archives de France (SIAF) pour la diffusion, dans le format SKOS, du *Thésaurus pour la description et l'indexation des archives locales anciennes, modernes et contemporaines*, a été étendue à d'autres vocabulaires et s'est vu adjoindre de nouvelles fonctionnalités, en 2012, dans le cadre du programme HADOC, pour devenir le site d'exposition des vocabulaires scientifiques et techniques du ministère en accès ouvert (voir note d'information DGP/SIAF/2012/05 en date du 12 novembre 2012). La plate-forme de diffusion des vocabulaires du ministère est accessible à l'adresse : <http://data.culture.fr/thesaurus/>.

Parallèlement, une réflexion a été engagée sur la mise en œuvre d'un nouvel environnement de gestion des vocabulaires scientifiques et techniques, d'où le développement de l'application GINCO (Gestion Informatisée de Nomenclatures Collaboratives et Ouvertes), objet de la présente note d'information.

Qu'est-ce que GINCO ?

GINCO est un outil collaboratif et open source de gestion de thésaurus et de vocabulaires contrôlés, qui permet d'exporter ceux-ci au format SKOS (Système Simple d'Organisation des Connaissances), modèle de données pour échanger et relier des systèmes d'organisation de connaissances sur le Web, comme les thésaurus, les nomenclatures, les taxonomies¹, les classifications ou les systèmes de vedettes.

L'application GINCO est conforme à deux standards :

- la norme ISO 25964. Norme de gestion de thésaurus monolingues et multilingues, la norme ISO 25964 révisé les deux normes antérieures ISO 2788:1986 (thésaurus monolingues) et ISO 5964:1985 (thésaurus multilingues). Son premier volet, consacré à la gestion de thésaurus², a été publié en 2011 et son second volet, relatif aux questions d'interopérabilité entre vocabulaires et abordant en particulier la question des alignements,³ a été publié en 2013.
- le standard SKOS, mentionné ci-dessus, pour la présentation des vocabulaires sur le Web de données.

Après une étude exhaustive des solutions de thésaurus existantes et pouvant convenir à l'implémentation des exigences de la norme ISO 25964, c'est vers la création d'un outil spécifique que le choix s'est porté. L'application a été développée en code ouvert. Elle est sous licence CeCILL⁴ et disponible sur une forge logicielle (<https://github.com/culturecommunication/ginco>). Il est donc possible de l'utiliser et éventuellement de contribuer à ses futurs développements.

GINCO est une application professionnelle multi-vocabulaires et multi-utilisateurs intégrant des services web qui permettent à des applications clientes (de saisie, de contrôle ou de diffusion) d'interroger les vocabulaires. La procédure d'installation est

1 Hiérarchie de termes organisée généralement avec une relation de généralisation/spécialisation (par exemple : une division du monde en continents, de chaque continent en pays).

2 ISO 25964-4: 2011 -- *Information et documentation -- Thésaurus et interopérabilité avec d'autres vocabulaires -- Partie 1: Thésaurus pour la recherche documentaire*

3 ISO 25964-2: 2013 -- *Information et documentation -- Thésaurus et interopérabilité avec d'autres vocabulaires -- Partie 2: Interopérabilité avec d'autres vocabulaires*

4 Licence de logiciel libre de droit français élaborée par le Commissariat à l'énergie atomique (CEA), le Centre national de la recherche scientifique (CNRS) et l'Institut national de recherche en informatique et en automatique (Inria).

donc technique. Toutefois, afin de faciliter la prise en main sur un poste de travail à des fins de test, l'équipe projet a fait réaliser une machine virtuelle (VM) préconfigurée, disponible sur la forge, qui reproduit l'ensemble des fonctionnalités de l'application (https://github.com/culturecommunication/ginco/blob/master/doc/VM_INSTALL.md).

L'application GINCO a été développée en deux versions : la première version dite V1 a été mise en production au MCC en juin 2013. La seconde version dite V2, qui intègre en particulier les fonctions d'alignement avec d'autres vocabulaires, a été déployée en mars 2014. Initialement en intranet, GINCO est désormais accessible en extranet à des contributeurs déclarés (<https://ginco.culture.fr/>), tout comme les espaces de travail collaboratifs Sémaphore. Des experts externes issus d'établissements publics, de collectivités territoriales, de laboratoires de recherche, etc. peuvent ainsi participer à l'élaboration ou à l'enrichissement des vocabulaires du MCC, en soumettant notamment leurs propositions d'évolution directement dans la plateforme. Déjà disponible au format SKOS, le *Thésaurus pour la description et l'indexation des archives locales* a été intégralement repris dans ce nouvel environnement. Désormais, toutes les mises à jour du *Thésaurus* se feront donc avec GINCO, qui devient donc le « back-office pour la production » de la plate-forme de diffusion. La liste d'autorité relative aux supports et procédés photographiques élaborée pour faciliter la lecture et la recherche sur la base BORA Archives Photo (<http://www.archivesdefrance.culture.gouv.fr/static/7455>) est également en cours de conversion en SKOS et devrait être prochainement rendue disponible sur le front-office de GINCO qu'est la plate-forme de diffusion des vocabulaires du MCC (<http://data.culture.fr/thesaurus/>).

Aperçu des fonctionnalités de GINCO

L'application GINCO offre les fonctionnalités de base d'un outil de gestion de thésaurus : création de termes et de concepts, relations d'équivalence entre termes, relations de hiérarchie (mono- ou polyhiérarchie) et relations d'association entre concepts, gestion du multilinguisme, typage des relations et des notes, interrogation au moyen d'un moteur de recherche plein texte enrichi (basé sur SOLR⁵), éditions alphabétiques et hiérarchiques. Mais GINCO offre également de nombreuses fonctionnalités avancées :

- la distinction entre les concepts et leurs valeurs lexicales (termes), conformément à la norme ISO 25964. Chaque concept est lié à un ou plusieurs termes, dont au moins un est le terme préférentiel. Cette séparation entre les concepts et les termes qui les représentent permet la gestion d'un certain nombre d'informations au niveau de la valeur lexicale, comme la source ou le statut de validation ;
- la possibilité d'ajouter des attributs personnalisés sur les concepts et sur les termes ;
- la gestion des statuts des concepts et des termes (candidat, validé, rejeté, déprécié, etc...) ;
- la gestion d'un historique de modification sur les vocabulaires que l'on peut exporter en CSV ;
- la gestion des concepts complexes (regroupement de plusieurs autres concepts) ;
- la gestion des tableaux de concepts permettant d'implémenter le mécanisme de facettes dans un thésaurus ;

5 Plate-forme logicielle de recherche s'appuyant sur le moteur de recherche Lucene, créé par la Fondation Apache et distribuée et conçue sous licence libre.

- la gestion du marquage des versions sur les vocabulaires (mais sans possibilité de revenir à une version précédente) ;
- la possibilité de faire des alignements, c'est-à-dire d'établir des relations entre les concepts d'un vocabulaire et ceux d'un autre ;
- l'import de données dans des formats et modèles de représentation standards (CSV, XML, SKOS) ;
- l'export de données en SKOS ou dans un format XML spécifique permettant de capturer toute la richesse des données ;
- la possibilité, pour un utilisateur, d'adresser à l'administrateur d'un vocabulaire des demandes d'évolution de ce vocabulaire (proposition de nouveaux termes, etc.). Les services d'archives qui souhaitent soumettre des propositions d'évolution du *Thésaurus pour l'indexation des archives locales* peuvent le faire en envoyant une demande via la plate-forme de diffusion depuis l'onglet Contact sur la page d'accueil de chaque vocabulaire (<http://data.culture.fr/thesaurus/>).

Dans le cadre du programme HADOC, l'application GINCO se veut plus qu'un simple outil de création et de gestion de vocabulaires contrôlés. L'objectif principal du projet reste l'utilisation des nouvelles possibilités apportées par la norme ISO 25964 et les technologies du web sémantique pour passer d'une approche où les vocabulaires « métier » sont enfermés dans des outils de production ou de diffusion, à la création d'un Référentiel unifié permettant d'offrir aux usagers un accès unique et cohérent à l'ensemble des ressources terminologiques produites par le MCC pour en faciliter la réutilisation.

L'outil GINCO présente également un grand intérêt pour les services d'archives, pour lesquels il importe de s'ouvrir aux évolutions en cours et d'envisager leur présence sur le web sémantique et plus spécifiquement sur le web de données. L'application GINCO, avec ses fonctionnalités d'alignement vers d'autres vocabulaires et ressources, pourrait permettre aux ressources archivistiques de gagner en visibilité en étant mieux référencées sur la toile. Malgré sa technicité, GINCO n'est pas réservé aux experts. Il a vocation à être téléchargé par les services d'archives non équipés et ayant un besoin en matière d'alignements sur des référentiels locaux. Pour les services d'archives qui sont déjà équipés avec des logiciels de gestion des thésaurus, GINCO peut être utilisé en complément des solutions existantes. Il est tout à fait possible d'importer dans GINCO les exports aux formats CSV, XML ou SKOS réalisés à partir des outils du marché pour les enrichir avec les propriétés complémentaires proposées par la norme ISO 25964 (gestion des groupes de concepts, alignements avec d'autres vocabulaires ou ressources, etc.). GINCO pourra enfin être utilisé dans le cadre du futur portail national interministériel des archives (PNIA) dont il constituera ainsi l'une des briques.

Pour faire suite au développement de GINCO, un nouveau travail de refonte de l'outil de diffusion des vocabulaires du ministère de la Culture et de la Communication a été engagé, la plate-forme étant appelée à s'enrichir progressivement de l'ensemble des ressources terminologiques issues de l'application GINCO. Outre ces évolutions, il a été décidé de déposer l'application de diffusion en code ouvert, comme un complément logique à l'outil de gestion, sur la même forge (<https://github.com/culturecommunication/ginco-diff>).

Le directeur, chargé des Archives de France

Hervé LEMOINE