World Library and Information Congress: 69th IFLA General Conference and Council

1-9 August 2003, Berlin

Code Number: 032-F
Meeting: 126. Classification and Indexing
Simultaneous Interpretation: -

Mise en oeuvre de la CDU : des rayons de bibliothèque à un langage d'indexation structuré

Aida Slavic
University College London
United Kingdom
Grande-Bretagne

Résumé :

La CDU attire différents intervenants du secteur de l'information en raison de sa large application, de son vaste vocabulaire et de sa disponibilité en format électronique. Les systèmes modernes de recherche documentaire ont la nécessité mais également la capacité de réaliser des systèmes de récupération flexibles et interactifs. Le rôle de la classification dans de tels systèmes est d'apporter une structure fondamentale de la connaissance qui fournit l'organisation systématique des sujets et complète ainsi la recherche utilisant les termes du langage naturel. Il y a, cependant, des conditions spécifiques qui doivent être satisfaites pour une utilisation efficace de la classification, conditions qui sont mal connues hors du domaine des bibliothèque et mal appliquées dans les systèmes de bibliothèque. C'est particulièrement le cas pour les classifications synthétiques, telle la CDU, parce que ses éléments doivent être manoeuvrés par le système pour remplir différentes fonctions (affichage systématique flexible, feuilletage, recherche). Ce rapport récapitule les fonctionnalités les plus importantes de la CDU à prendre en considération pour sa mise en oeuvre. Des questions importantes sur la relation entre la CDU sous forme électronique - le Master Reference File - et un outil de classification (fichier d'autorité) qui peut être construit sur sa base, sont soulignées. Une meilleure compréhension de la fonctionnalité du système CDU peut améliorer ou faciliter sa mise en oeuvre et abaisser les coûts des systèmes de gestion, ce qui peut être utile aussi bien pour les utilisateurs potentiels que pour les systèmes déjà existants.

---

1. Contexte

Il y a actuellement dans le secteur de l'information plusieurs domaines d'activité qui rendent nécessaire de diffuser les connaissances relatives à la mise en oeuvre de la Classification décimale universelle, la CDU. Ces activités sont liées aux systèmes bibliographiques existants

ainsi qu'aux nouveaux utilisateurs du secteur non-bibliographique. Tout d'abord, il y a un grand nombre de bibliothèques, de services bibliographiques et de systèmes utilisant la CDU qui n'exploitent pas complètement la classification. On crée aussi toujours plus de portails d'information et de catalogues collectifs qui incluent différentes collections de ressource à un niveau national ou international. De plus en plus, les utilisateurs exigent eux aussi un procédé de recherche documentaire plus efficace et plus interactif que ceux que la majorité des OPACs tendent à offrir. La CDU est sous-jacente dans le système bibliographique de beaucoup de bibliothèques européennes et n'est pas correctement exploitée. En outre, la CDU peut fournir le support nécessaire dans un environnement multilingue et multialphabétique dans un espace global de l'information. Et, dans cet environnement, elle peut également être employée comme médiateur de correspondance entre les systèmes d'indexation mais ce potentiel est la plupart du temps gaspillé et reste inutilisé.

Malgré une documentation importante sur l'automatisation de la CDU, il reste encore beaucoup d'idées fausses chez les bibliothécaires comme chez les non-bibliothécaires au sujet de ce qui peut être réalisé avec des systèmes de classification tels que la CDU. Cet article essaiera donc de revisiter certaines des questions bien connues à la lumière des scénarios communs de mise en oeuvre basés sur la CDU en format électronique - le "Fichier de référence général de la CDU", UDC MRF (UDC Master Reference File). Le MRF est la forme électronique de la version standard de la CDU, possédée par le consortium : http://www.udcc.org/mrf.htm. Il est mis à jour annuellement et diffusé en janvier en ISO2709 ou comme fichier texte. Il peut être acheté sur la base d'un accord annuel de licence qui permet l'acquisition de la classification entière ou, depuis 2003, de certaines de ses parties seulement.

2. Politique de mise en oeuvre

La CDU est appliquée pour l'organisation et l'indexation de ressources électroniques, pages Web, documents imprimés et/ou objets. Indépendamment de l'application de la CDU et des normes de métadonnées qui vont être choisies pour porter les données de la classification, il y a quelques questions générales qui doivent être abordées. Le point de départ pour concevoir une politique de mise en oeuvre peut ainsi être établi autour des questions suivantes :

- Quelles sont les fonctions de la recherche matière qui doivent être assurées : recherche et feuilletage ou l'une de ces deux fonctions seulement ? Si:
  a) recherche et feuilletage : la transition aisée d'une fonction à l'autre sera-t-elle fournie?

  b) feuilletage seulement : va-t-il être possible de commencer le feuilletage depuis tout point dans la hiérarchie? Sera-t-il possible de suivre les liens "Voir aussi" dans les hiérarchies? La notation sera-t-elle être affichée avec l'intitulé (description de la classe)?
  c) recherche seulement : un index alphabétique approprié de recherche va-t-il être fourni? Sera-t-il possible de rechercher par indices aussi bien que par termes?
- La CDU sera-t-elle employée seule ou à côté d'un système d'indexation alphabétique (thésaurus ou vedettes-matière)?
  Si OUI : comment ces vocabulaires vont-ils être liés à la CDU : à travers les données d'autorité de la classification ou par un index de recherche seulement?
  Si NON : un index matière alphabétique doit être établi. Sera-t-il basé sur l'UDC MRF seulement? Comment est-il prévu de l'accroître, de le maintenir? Quelle sera sa forme : index alphabétique simple, index de chaînes, index relatif?
- Y a il un plan pour mettre en valeur la collection et pour l'inclure dans un portail d'information plus vaste (multilingue?) dans lequel la CDU devra être reliée par

mapping à un autre système d'indexation? Est-il prévu d'assurer une classification automatique dans le futur?

- Comment envisage-t-on qu'une structure, un contenu et une syntaxe de remplacement peuvent appuyer la classification? Les métadonnées sont-elles incluses ou autonomes? Quelle norme/format de métadonnées prendra en charge l'index CDU et quels sont leurs éléments/champs qui vont servir de support à l'utilisation de la classification? Quel genre de format/codage est disponible pour la CDU?
- Comment la politique de catalogage/indexation et la norme de métadonnées vont-elles relier des données matière différentes (personnes, événements, périodes, sujets noms communs) : seront-elles réparties dans différent champs/éléments, comment ces champs seront-ils ordonnés et reliés pour former des index de recherche et être employés par le logiciel de recherche ; pour lesquels de ces sujets la CDU sera-t-elle employée?
- Y aura-t-il un fichier d'autorité matière, et comment l'architecture des métadonnées va-t-elle se relier à la description des document et au fichier d'autorité? Le fichier d'autorité va-t-il être maintenu à l'extérieur du système, ou va-t-il être partagé par différents systèmes ou encore employé pour des fonctions telles que le mapping et des recherches à travers différentes collections?

Certaines de ces questions peuvent être plus pertinentes que d'autres, selon le but du système, mais il est certainement intéressant de constituer une liste de conditions basées sur la politique choisie. La plupart de ces points ne sont pas nécessairement difficiles à mettre en application.

Indépendamment du choix du système d'indexation, il existe une étape importante mais souvent négligée : l'accord sur une politique d'indexation. Ce point n'est pas spécifique à la classification, ni bien sûr à la CDU, et est en dehors du sujet précis de cet article. Cependant, une directive ou un document sur le sujet, indépendamment d'être de bon sens, est une question primordiale pour le succès d'un système et son efficacité dans la récupération des ressources. Les classifications laissent toujours la liberté de choix aux classificateurs, et c'est plus encore le cas avec une classification synthétique. Bien que l'existence d'un fichier d'autorité de classification puisse aider à assurer la cohérence et le contrôle de l'indexation, il y a toujours des règles de politique générale à donner. Des décisions et des principes doivent être dégagés en ce qui concerne l'exhaustivité et la spécificité dans l'indexation. En outre, des points tels que le traitement des personnes et des noms propres qui peuvent être ajoutés à un indice, et tels encore que les lieux et les événements comme sujets doivent être considérés. La CDU peut également contenir l'information qui est habituellement renfermée, en format MARC et dans d'autres formats de métadonnées, dans d'autres champs/éléments de l'autorité tels que *langue de la ressource*, *public*, *forme* et *format* ou *couverture*. Il est nécessaire de décider si la répétition de ces éléments dans l'indice CDU peut être utile ou pas.

Dans la politique d'indexation, on devra faire particulièrement attention aux points propres à la CDU. Souvent, dans les principes directeurs et recommandations des métadonnées, les indexeurs sont amenés à croire que la classification devrait être employée au niveau de spécificité le plus élevé possible [1]. Si cette politique peut bien fonctionner avec des classifications plus petites et énumératives telles que la classification décimale de Dewey (CDD/DDC), elle produit des résultats gênants et indésirables une fois appliquée à la CDU, qui est trois fois plus grande, fortement synthétique, et peut produire énormément de termes d'indexation. En outre on peut devoir prendre des décisions concernant l'ordre de citation dans les indices CDU synthétisés comme la possibilité de modifier cet ordre afin de produire un arrangement utile des ressources. Enfin, si un index matière alphabétique de la classification

est créé, les règles devraient aussi en être enregistrées. Les procédures pour le traitement des homonymes, des synonymes, des termes composés et celles pour relier les termes associés devraient être discutées comme faisant partie de la conception du système ou être au moins prévues comme devant être résolues plus tard au cours du processus.

3. Mise en oeuvre de la CDU : conditions fonctionnelles et de système

Il y a deux manières d'appliquer la CDU :
a) en utilisant uniquement des indices simples, ou des indices construits pré-combinés comme indices simples ; b) en utilisant un index (structuré) synthétique.
Selon la portée et l'objectif de l'utilisation de la classification, les deux approches soulèvent des problèmes qui doivent être résolus par des méthodes de mise en oeuvre spécifiques. Certains problèmes d'exécution et de maintenance qui ont été mentionnés sont liés à la manière dont les données sont rendues disponibles dans l'UDC MRF, d'autres sont liés à la façon dont les indices vont être employés dans le système de recherche documentaire. Ces deux aspects sont abordés ci-dessous. Tandis que le premier ensemble de problèmes peut être plus ou moins allégé en préparant une exportation différente et plus riche des données de classification, le second dépend de la création d'outils appropriés pour gérer et contrôler l'utilisation de ces données.

3.1 Mise en oeuvre avec notation simple et non-synthétique

L'approche la moins compliquée de l'emploi de la CDU recouvre à la fois l'utilisation d'indices simples seulement, et l'utilisation d'indices construits traités comme des indices simples. L'édition standard de la CDU, avec son ensemble actuel de 66 149 indices de classification, peut être employée en choisissant de ne se préoccuper que des indices simples de la classification. Ceux-ci peuvent être pris à partir des tables principales ou des tables auxiliaires communes du MRF et ils seront suffisamment détaillés pour satisfaire beaucoup d'utilisateurs. En d'autres termes, la CDU peut fonctionner comme taxonomie simple ou classification énumérative. Cet aspect est souvent exploité pour l'organisation des rayonnages dans les plus petites bibliothèques, particulièrement en Europe centrale, où la CDU est employée dans les bibliothèques publiques et scolaires. Les portails thématiques sur Internet qui déploient la CDU tendent eux aussi à l'employer de cette façon [2]. Appliquée comme une classification énumérative et non-synthétique, la CDU permet en effet d'atteindre l'objectif simple du feuilletage systématique. Elle a alors une fonctionnalité très semblable à celle de la CDD, la seule différence étant que la CDU a un vocabulaire plus important et plus spécifique, et ne contient pas autant de termes énumératifs, termes composés prêts à l'emploi, que la Dewey.

Un fichier CDU avec des indices simples uniquement n'exige pas beaucoup d'effort de mise en oeuvre. La notation de classification étant dans ce cas un simple texte se composant d'indices et d'une ponctuation sans signification (un point décimal) après chaque troisième chiffre. Les indices sont automatiquement classés correctement par n'importe quel système informatique. Plus souvent, cependant, on peut trouver des indices CDU créés d'une manière pré-combinée, mais traités comme notation simple. C'est souvent le cas avec les systèmes de bibliothèque, la plupart du temps en raison de la manière dont les formats MARC avaient enregistré les données de classification comme de simples chaînes de caractères, sans tenir compte de ce que leur contenu est un terme simple ou un terme d'index structuré pré-combinée. Le classement correct de ces indices est dès lors difficile et il en résulte en général un ordre systématique perturbé qui ne suit pas l'ordre des sujets *du plus large au plus*

précis/du général au spécifique, ordre qui est primordial pour permettre la fonction de feuilletage. En outre, de tels enregistrements ne permettent la recherche que du premier élément de la notation, les autres ne pouvant pas être employés pour la recherche.

L'utilisation de la CDU comme classification énumérative (avec des indices simples ou avec des indices construits traités comme des indices simples) peut fort bien servir son but principal si les intitulés des indices (libellés, descriptions) sont ajoutés au système de récupération de sorte que, derrière les indices, les termes soient disponibles pour la recherche et soient ajoutées à l'affichage systématique dans l'interface utilisateur. L'UDC MRF est une bonne source pour ces termes d'index, qui peuvent être récupérés non seulement dans le champ intitulé (libellé) mais également à partir des notes et des exemples des combinaisons [3].

3.1.1 Problèmes et recommandations

Source des données. En employant l'UDC MRF comme source des données de classification, il convient de noter que tous les indices fournis ne sont pas simples, ce que doivent garder à l'esprit ceux qui souhaitent mettre en oeuvre la CDU à ce niveau d'utilisation [4]. Dans les tables principales il y a un nombre restreint mais inconnu d'entrées se composant d'une combinaison d'un indice principal simple et d'un auxiliaire commun, comme *94(680) Histoire de l'Afrique du Sud*. Ces entrées ne sont pas marquées en tant que telles dans la base de données. En outre, il y a des indices qui sont bien la combinaison de deux indices principaux ou auxiliaires tels que la combinaison de regroupement dans *562/569 Paléozoologie systématique* ou des regroupements d'indices auxiliaires communs pour le temps dans *"321/324" Saisons*.

Extraire des nombres simples automatiquement à partir de l'UDC MRF peut, donc, ne pas être si aisé. Les combinaisons des indices principaux et des auxiliaires spéciaux sont indiquées grâce à un champ spécial, alors que, comme mentionné ci-dessus, la combinaison d'un indice principal et d'un auxiliaire commun telle que 94(410) n'est pas marquée pour le traitement automatique. C'est un inconvénient qui devrait être corrigé par la suite.

Les nouvelles applications, particulièrement celles qui permettent d'accéder aux ressources de l'information sur Internet, utilisent l'UDC MRF afin d'extraire des indices simples ou des sélections d'indices avec leurs intitulés. Le MRF exporté pour la distribution aux éditeurs et aux bibliothèques est appelé "user MRF" (UMRF) et il ne contient aucun champ administrateur ni même de données propres à la gestion de la base de données. Il n'y a par conséquent pas assez de données à extraire automatiquement ; par exemple, on y trouve seulement les indices principaux simples et pas les entrées qui sont une combinaison d'indices principaux et d'auxiliaires spéciaux, car cette information n'est pas disponible dans l'UMRF.

Il serait souhaitable de procurer aux développeurs les données complètes du MRF ainsi que le manuel MRF qui fournirait toutes informations sur la structure et le contenu des champs [5]. Le consortium UDC devrait fournir plus de choix en termes de formats de données. Les conversions dans les différents formats MARC, par exemple, faciliteraient l'importation des données vers les systèmes de bibliothèque basés sur le MARC. Ces aspects, comme quelques changements dans la base de données du MRF sont actuellement à l'étude [6][7]. Il y a quelques champs du MRF qui n'ont jamais été complètement employés, comme celui prévu pour les termes d'index qui était laissé vacant pour être complété par les différents éditeurs de la CDU. Les nouveaux utilisateurs de la CDU apprécieraient sans doute cette valeur ajoutée à

la classification et c'est un autre secteur d'amélioration que devront aborder les propriétaires de la CDU.

Fonctions de récupération. Normalement la notation expressive de la CDU permet que les hiérarchies soient liées à la notation dans son ensemble sans besoin d'aucune adaptation spéciale pour le classement et l'affichage. La troncature à droite mènera au niveau supérieur de la classe qui peut être exploité pour élargir la recherche. Par exemple, la recherche de *004.415 #* donnera des résultats qui incluent toutes les divisions qui suivent. Cela fonctionnera également pour des indices construits traités comme une simple chaîne de caractères. Cependant, comme précisé par Buxton et Riesthuis, la troncature à droite ne conduit pas toujours à la catégorie plus large. Par exemple la catégorie supérieure de *563.4 Spongiaires (Eponges)* n'est pas 563 mais *562 Invertébrés en général* (Buxton, 1990, Riesthuis, 1998). C'est souvent le cas avec un emploi de regroupement (c'est-à-dire quand une classe est définie comme une extension couvrant un certain nombre de classes à la suite comme *562/569*), mais cela peut aussi se produire ailleurs. C'est plus une exception qu'une règle et, bien qu'il soit graduellement corrigé par révision, ce point reste un trait de la classification qui ne peut être correctement contrôlé par la seule application de la CDU sans un certain contrôle des hiérarchies comme de la notation elle-même.

Si l'on choisit d'employer indépendamment des indices auxiliaires communs (par exemple lieux, temps, personnes, etc.) comme des indices simples, au même titre que les indices principaux, cela réclamera une attention particulière car ces indices vont contenir des symboles arbitraires et seront automatiquement classés avant les indices principaux. Leur ordre va ainsi être différent de celui que recommande le système CDU. Une solution est alors de saisir les données de la classification en utilisant les préfixes qui serviront à indiquer l'ordre de classement et n'apparaîtront pas à l'affichage.

Ceux qui mettent en oeuvre la classification avec des indices simples uniquement devraient considérer que le niveau de spécificité est très restreint dans cette utilisation de la CDU et que la nécessité d'employer des combinaison d'indices peut apparaître très vite dans les principales classes qui sont entièrement à facettes. C'est le cas, par exemple, à *821 Littérature* et *94 Histoire*. Avec la tendance des révisions actuelles conduisant la CDU vers une structure davantage à facettes, cette situation se produira plus fréquemment. Afin de faire une différence entre, par exemple, la *littérature anglaise 821.111* et la *littérature américaine,* on doit employer l'auxiliaire commun de lieu *(73) Etats-Unis d'Amérique*. De même, pour obtenir l'indice pour l'histoire des différents pays, on doit employer l'indice de l'histoire 94 et l'auxiliaire commun de lieu de façon à indiquer le pays, et au besoin l'auxiliaire commun de temps pour rendre la période, par exemple *94(410)"16" : Histoire des îles britanniques au 17e siècle*. C'est la raison pour laquelle les bibliothèques utilisent des indices pré-combinés bien qu'elles tendent à ne pas les traiter en tant que tels dans leur système.

Gestion de la classification. Si on veut les utiliser pour le feuilletage dans l'interface utilisateur (OPACs, portails ou passerelles par sujets), les indices CDU devraient être employé avec leurs intitulés (descriptions) ou devraient être complètement omis, auquel cas la hiérarchie devrait être affichée par l'indentation de la description de la classe ou par l'intermédiaire d'une autre aide graphique. La plupart des nouveaux utilisateurs de la CDU tendent à choisir une application simple pour éviter les conflits et les problèmes additionnels dans le classement et l'affichage. Cependant, même quand un système peut traiter l'interclassement des indices, comme c'est le cas avec la CDU simple, il peut être nécessaire de gérer la classification comme un fichier d'autorité séparé. Cela permet l'addition et la

gestion de toutes les données nécessaires pour l'usage en combinaison avec des indices de classification afin de permettre les fonctions de feuilletage et de récupération :

a) les intitulés des indices (description)

b) les termes de recherche qui ne sont pas présents dans l'intitulé mais permettent la recherche et le positionnement dans la hiérarchie

c) les références "voir aussi" qui ont de l'intérêt pour le feuilletage

d) l'établissement d'une hiérarchie relative, indépendante de la notation afin de contrôler les variantes occasionnelles dans la notation de la hiérarchie

e) le classement des auxiliaires communs utilisés comme indices simples, où le symbole sera employé pour l'affichage et pas pour l'informatique et le classement.

3.2 Mise en oeuvre avec notation synthétisée et pré-coordonnée

La CDU peut être distinguée des autres classifications bibliographiques en raison de son dispositif synthétique puissant. L'avantage de la classification synthétique est qu'elle permet de couvrir un nombre illimité de sujets et leurs combinaisons avec une quantité limitée de concepts simples. Les dispositifs synthétiques rendent la classification plus accueillante et extensible et plus puissante pour l'indexation. La synthèse, cependant, ajoute à la complexité du système de classification qui exige alors plus de connaissance des règles de syntaxe et plus d'aide en termes d'outils de gestion. Il est utile de se rappeler que les dispositifs synthétiques vont être exploités encore plus à l'avenir et seront encore facilités par la "facétisation" de la CDU [8] [9]. Beaucoup de classifications se fondent sur une sorte de synthèse dans leurs listes, principalement pour économiser de l'espace. La CDU est cependant équipée de mécanismes fiables pour assurer et contrôler une synthèse illimitée à plusieurs niveaux :

a) entre indices d'une ou plusieurs classes principales, en utilisant des symboles qui expriment les relations entre deux sujets

b) entre les indices principaux des classes et un ou plusieurs auxiliaires communs

c) entre une classe principale et un ou plusieurs indices auxiliaires spéciaux

d) entre un ou plusieurs indices auxiliaires communs

e) entre un ou plusieurs indices auxiliaires spéciaux

f) entre les indices principaux de la CDU et un autre vocabulaire externe

g) entre les indices principaux et toute autre extension alphabétique utilisés pour des spécifications ultérieures.

Des indices pré-combinés tendent à être construits pendant l'indexation et sont rarement énumérés dans la classification. Au cours du processus de révision, des concepts composés sont régulièrement supprimés du système et sont remplacés par une combinaison de concepts simples. Un indice CDU une fois appliqué pour l'indexation est donc mieux compris comme étant un terme d'indexation structuré et pré-coordonné qui a son vocabulaire et sa syntaxe comme tout autre langage d'indexation pré-coordonné. La signification de chaque élément demeure la même à l'extérieur comme dans les combinaisons et peut être recherchée comme dans une recherche post-coordonnée. Par exemple, on emploiera les auxiliaires communs *(73) Etats-Unis d'Amérique et "18" 19e siècle* dans un nombre illimité de combinaisons comme : *94"18"(73) Histoire -- 19e siècle -- Etats-Unis*, ou *821.111(73)"18" littérature américaine -- 19e siècle,* ou *dans 321.7"18"(73) Politique -- Démocratie -- 19e siècle -- Etats-Unis*. Par conséquent, la recherche de *(73)* récupèrera chaque item lié aux Etats-Unis, et la recherche de *"18"* tout ce qui est lié au 19e siècle, quel que soit le sujet.

Quand on l'utilise avec ses pleines possibilités synthétiques, il y a deux conditions principales pour maîtriser la CDU : a) le classement des indices complexes; b) la recherche de chaque

élément individuel qui intervient dans la construction des indices pré-combinés. Le classement des indices CDU simples et pré-combinés permet la présentation par sujets du général au spécifique. Le système de classification réalise ceci par la combinaison des règles de classement et des règles utilisées pour construire la séquence dans un indice pré-combiné. La gestion de la CDU réclame donc le contrôle de chaque indice, qu'il soit employé seul ou en construction dans un indice pré-combiné. Ce contrôle devrait se fonder sur des données de classification formatées de telle façon que chaque élément de l'indice construit puisse être identifié par le système indépendamment des indicateurs de symboles et de facette qui sont employés pour son affichage et quelle que soit sa place dans un indice CDU pré-coordonné.

Ce sont les raisons pour lesquelles l'utilisation d'une classification, et tout particulièrement de la CDU, dépend des outils rendus disponibles pour réaliser cette fonctionnalité avec le moins d'inconfort possible pour les catalogueurs, outils grâce auxquels la complexité d'une notation est uniquement manipulée par le système.

3.2.1 Problèmes et recommandations

Source des données. Dans le processus de classification, l'UDC MRF est la source des indices simples qui sont employés pour construire les termes d'indexation pré-combinés. Le fait que les données de classification sont disponibles en format électronique permet seulement d'éviter une partie de l'édition manuelle et du travail répétitif, mais l'aide véritable dans l'indexation des documents est d'avoir accès aux indices pré-combinés et d'assurer leur réutilisation facile. Cela est habituellement réalisé par la création d'une base de données de classification ou d'un fichier d'autorité qui se développe avec son application. Les indices construits qui existent dans le MRF, à la fois dans les indices mêmes de la classification et dans les exemples des combinaisons, peuvent être employés comme source prête à l'emploi de termes d'indexation pour enrichir le fichier d'autorité de classification. Ces indices n'ont pas leurs éléments structuraux codés et un travail manuel d'édition peut être nécessaire pour les rendre entièrement fonctionnels dans l'outil existant. Le véritable outil d'indexation devient alors le fichier contenant les indices CDU pré-coordonnés qui ont été établis au cours du processus d'indexation de la collection. L'utilisation et la gestion aisées de ce fichier est donc primordiale pour le bon usage de la classification. Le but principal est qu'une fois créée, une vedette CDU pré-coordonnée peut être liée à un nombre illimité de références dans une collection sans beaucoup de travail manuel d'édition.

Selon la politique d'indexation et de mise en oeuvre, le nombre total d'indices utilisé dans une collection peut être sensiblement plus petit ou sensiblement plus grand que le MRF lui-même. Si la politique est d'employer la classification avec d'autres langages d'indexation, sa fonction de regroupement et d'agglomération sera plus importante et donc le nombre d'indices CDU pré-combinés pourra ne pas excéder trois à quatre mille même pour une collection de quelques centaines de milliers de ressources. Cette approche est caractéristique de certaines grandes bibliothèques publiques ou de bibliothèques universitaires de taille moyenne dans les pays d'Europe de l'Est comme la Hongrie, la Croatie, la Slovénie, etc. Les bibliothèques qui emploient la CDU comme langage d'indexation et de récupération principal avec une politique pour exprimer précisément la spécificité des différentes références peuvent avoir des dizaines de milliers d'indices pré-coordonnés différents à contrôler. C'est le cas par exemple pour la bibliothèque universitaire centrale de Bucarest qui a des centaines de milliers d'indices, ou pour *ETH-Bibliothek*, la bibliothèque de l'école polytechnique fédérale de Zurich, avec environ 60 000 indices pré-combinés différents [10].

Une organisation des indices CDU pré-combinés correctement structurés et codés peut être une ressource utile dans l'échange d'information et peut être partagée, adaptée et développée par de nombreux participants. Ce genre d'outil de référence fournit une garantie documentaire pour les concepts qui sont employés et peut être une ressource de valeur pour développer de nouveaux vocabulaires ou, évidemment, pour réviser la CDU elle-même [11].

Fonctions d'aide à la récupération. Les conditions fonctionnelles pour le feuilletage et la récupération des indices construits ont été récapitulées par Buxton en 1990. Il a souligné les fonctions suivantes : la nécessité de pouvoir rechercher les indices avec tous les symboles qui peuvent être employés ; la capacité de classer les indices pré-combinés d'UDC, avec possibilité de recherche tronquée ; la capacité de rechercher séparément chaque auxiliaire commun, et celle de rechercher les indices intermédiaires quand les indices CDU pré-combinés contiennent un regroupement (extension) avec la possibilité de tronquer à l'intérieur d'un indice. Buxton a proposé de casser les chaînes des indices construits au moins par un espace mais a également suggéré le remplacement des symboles de la CDU par des lettres à l'exemple du système AUDACIOUS [12]. La recherche post-coordonnée des éléments CDU et leur combinaison avec des termes d'index est particulièrement importante pour les classificateurs. Chercher des expressions comme "lapin" ET "6 #", ou "7 #" ET "technique" avec ou sans troncature est en effet la meilleure manière pour se positionner à un endroit précis dans la classification.

La plupart de ces problèmes peuvent être résolus si la classification est mise en application pour traiter des indices CDU codés de façon pré-coordonnée. Cela permettrait d'établir facilement les règles pour le classement des indices pré-combinés. Ainsi, afin d'exprimer la hiérarchie des sujets, les indices CDU devraient être classés selon un ensemble spécifique de conventions. Par exemple, *73 Arts plastiques* ; *73+75 Arts plastiques et Peinture* ; et *73/75* (séquence couvrant dans la classification 73 Arts plastiques, 74 Dessin et 75 Peinture), réclament d'être classé dans l'ordre suivant : *73+75 > 73/75 > 73*. C'est parce que chaque symbole a sa place dans l'échelle du général au particulier, et parce que deux indices principaux liés par "+" et "/" donnent des classes de sujets plus larges qu'un indice simple de la classe. Cependant, deux classes principales liées par ":" (deux points), par exemple *73:75 Relations entre les arts plastiques et la peinture*, représentent toujours un sujet plus étroit qu'un indice simple et cette combinaison doit donc être classée après l'indice lui-même, par exemple *73 > 73:75*. Cet ordonnancement purement intellectuel doit être soutenu par le système qui assurera le traitement de ces index structurés sans se baser sur les symboles utilisés pour la représentation visuelle et l'affichage.

Une autre raison d'avoir l'accès et le contrôle pour chaque partie d'un indice CDU structuré est liée à la flexibilité dans les combinaisons des éléments. Indépendamment de la recommandation générale de citer les auxiliaires communs dans l'ordre *temps, ethnie, lieu, forme, langue*, l'ordre dans lequel les indices CDU peuvent être combinés (i.e. l'ordre de citation) est flexible. Selon l'intention dans la présentation et la distribution, certaines collections peuvent vouloir fournir des approches différentes pour leurs utilisateurs. En histoire, par exemple, il est possible d'avoir l'ordre suivant : indice principal, temps, lieu, soit *94"18"(410) Histoire -- 19e siècle -- Iles britanniques*, qui va présenter l'histoire par périodes puis par pays, alors qu'un autre affichage peut permettre de regrouper l'histoire par pays : *94(410)"18" Iles britanniques -- 19e siècle*. Quand on peut établir l'accès et le contrôle pour chacun des éléments constitutifs, il est possible de fournir différents affichages pour répondre aux préférences des utilisateurs.

Cependant, si les indices pré-combinés sont conservés et contrôlés comme de simples chaînes de caractères, il peut être encore possible de rechercher des indices simples dans une chaîne pré-coordonnée et d'employer la CDU pour la recherche post-coordonnée. Cela peut être réalisé en utilisant un programme spécialement écrit qui permettra la déconstruction des indices en leurs éléments constitutifs selon des algorithmes extraits à partir de la syntaxe de la CDU. Riesthuis a suggéré un ensemble d'algorithmes qui peuvent être employés pour écrire un programme qui décomposerait les indices CDU [13][14].

Gestion de la classification. Comme on l'a souvent souligné, la mise en oeuvre de la CDU comme données d'autorité structurées et correctement codées est primordiale pour permettre toutes les fonctions que la classification peut remplir dans la recherche documentaire. De nombreuses bibliothèques ont développé des outils de classification basés sur leur propre expérience et leurs besoins en contrôle d'autorité et jusqu'ici ces outils sont la plupart du temps des solutions "propriétaires" et des exemples de bonne pratique. Dans ces systèmes, la classification est généralement liée et "mappée" à d'autres systèmes d'indexation ou à son propre index alphabétique matière. Quelques bibliothèques développent des thésaurus ou des systèmes de vedettes-matière basés sur des données CDU existantes. Les nouveaux utilisateurs en dehors du domaine des bibliothèque comptent gérer des systèmes de classification sous forme de données d'autorité et souhaitent trouver des fichiers d'autorité CDU disponibles pour le partage. Ceux qui mettent en application la CDU avec indices synthétisés et pré-combinés peuvent choisir une des approches suivantes :

- prévoir de structurer et de coder les éléments séparés de la CDU à l'intérieur de la description bibliographique/des métadonnées
- maintenir la classification en tant que données séparées avec des liens à un système de recherche documentaire
- avoir à la fois un index structuré dans la description bibliographique/métadonnées et un riche fichier d'autorité de classification maintenu séparément.

La première approche aiderait pour le classement et la recherche des indices CDU, mais ne parviendra pas à fournir un lien entre les indices, leur description et les termes de l'index de recherche, et elle ne permettra pas non plus de mettre en application les références "voir aussi" ou d'assister les catalogueurs puisque les indices devront toujours être resaisis et réentrés dans le système. Les deuxièmes et troisième approches qui maintiennent des données de classification dans des fichiers séparés sont beaucoup plus efficaces. La troisième a l'avantage d'être la plus sûre puisque les indices CDU peuvent être correctement traités et échangés même lorsque le fichier d'autorité est détaché.

Les données d'autorité de classification dont il s'agit ici comportent non seulement un fichier de contrôle simple pour assurer les points d'accès et l'uniformité des vedettes CDU, mais aussi un outil entièrement fonctionnel qui sert à contenir, contrôler, maintenir et partager ces données. Son but serait de rendre illimitées l'utilisation et la ré-application des données CDU et donc d'économiser temps et effort dans la classification des ressources.

4. Importance des formats de classification

Plus le niveau de formalisation des données est élevé, plus une classification devient puissante. En même temps, elle devient moins appropriée à la manipulation humaine qui exige plus d'intermédiation et des mécanismes plus sophistiqués pour la mettre en oeuvre et l'exploiter. C'est typiquement le cas avec les indices CDU synthétiques quand ils sont utilisés pour une indexation pré-coordonnée. Actuellement il n'y a aucun format CDU admis ou

proposé capable de faire totalement face à toutes les demandes de manipulation des indices pré-combinés. Il y a, cependant, des formats et des analyses de structure des données disponibles qui peuvent aider en rassemblant l'information nécessaire pour créer une base de données CDU opérationnelle vraiment apte à remplir des fonctions telles que le contrôle d'autorité de la classification, la gestion, l'échange et le partage des données, la recherche documentaire et les fonctionnalités appropriées d'un outil pour classifier.

La première source à mentionner est la structure de données telle qu'elle est utilisée dans l'UDC MRF, qui est disponible dans le manuel MRF, et expliquée dans plusieurs articles et sur les pages web du consortium UDC [15][16][17][18]. Mais la structure de données du MRF ne comporte pas toute l'information qui doit être adaptée pour atteindre convenablement tous les objectifs exigés. Il manque en particulier la structure requise pour automatiser entièrement le maniement des différents éléments d'informations dans un indice CDU pré-combiné.

Une autre source utile pour aider à se faire une idée des données qu'il peut être nécessaire d'inclure dans un format de classification est le format MARC21 pour les données de classification [19], développé en 1991 et mis à jour en 1995 par la bibliothèque du Congrès, afin de gérer les systèmes de classification du Congrès (LCC) et Dewey (DDC). Mais là encore, comme ces deux classifications sont énumératives et non synthétiques, ce format ne répondra pas aux besoins de la CDU exposés ici. Notamment, les champs pour enregistrer un indice de classification (c'est-à-dire une vedette de classification) permettent seulement l'entrée d'une simple chaîne de caractères. Sous d'autres aspects cependant, ce format peut être une source utile pour déterminer les données nécessaire : intitulé, notes d'application, instructions, exemples de combinaisons, hiérarchie relative (classe supérieure), termes d'index et structure d'autres informations telles que description, références "voir aussi", champ remplace/remplacé par, etc.

Plus récemment, à l'initiative du Comité permanent UNIMARC, un format de classification UNIMARC a été développé. Ce travail a commencé après une étude préliminaire - conditions requises pour un format des données de classification, de E. W. Woods, 1994 - dont les recommandations incluaient l'applicabilité pour différents classifications, pour des demandes multilingues, pour des fonctions de contrôle d'autorité, etc. Par la suite, un format abrégé (Concise UNIMARC Format for Classification Data [20]) a été rendu disponible pour le débat public en 2000, et est toujours à l'état de projet et sous une forme non définitive.

On s'attendait à ce que ce nouveau format accorde plus d'attention aux classifications synthétiques telles que la CDU, ce qui signifie fondamentalement la prise en compte du traitement et de la manipulation des éléments structuraux de la CDU qui sont le plus grand obstacle à son exploitation appropriée dans les OPACs de bibliothèque. Mais, dans son état actuel, il n'offre pas plus que ce qui est déjà rendu disponible par le format du MARC 21, et ne prévoit ainsi toujours pas un traitement approprié des vedettes de classification pré-coordonnées. Si ces détails étaient fournis, il répondrait aux besoins d'accès multidirectionnel, d'une recherche aisée et d'un classement exact. En particulier, le format n'aborde pas particulièrement les problèmes liés aux indices CDU pré-combinées, notamment la recherche des éléments signicatifs séparément ou la gestion des changements globaux des éléments composants. C'est le principal inconvénient de ce nouveau format. Le reste des besoins liés aux notices CDU est en revanche couvert, tels la description, les termes d'index, notes et notes d'application, exemples, "voir aussi", tout comme le sont les données de gestion.

Construire une structure de données plus complexe pour les classifications assurant une meilleure fonctionnalité d'application sera certainement un investissement rentable. Ce qui s'est produit jusqu'ici est que les développeurs dans les bibliothèques qui emploient des systèmes propriétaires, ainsi que les utilisateurs de la CDU en dehors des bibliothèques, ont été en meilleure position que les bibliothèques utilisant des systèmes standard. Tandis que dans le premier cas les systèmes sont développés selon les structures localement définies de conception et de données, les bibliothèques avec des systèmes standard basés sur le MARC ont été confrontées à l'hésitation, voire au refus des fournisseurs pour mettre en application des changements qui apporteraient des déviations des structures de données officielles du MARC. Beaucoup de bibliothèques ont essayé de négocier avec les vendeurs pour permettre de scinder les indices CDU dans les données bibliographiques. Ce point a été discuté en début d'année par le groupe de discussion sur le forum CDU [21]. Quelques collègues ont réussi avec INNOPAC mais non sans avoir eu à se tourner vers le bureau de gestion d'USMARC d'abord et d'avoir obtenu l'autorisation d'employer pour la CDU le sous-champ avec code "'x" pour séparer chaque élément du symbole CDU. D'autres collègues ont été moins chanceux avec, par exemple, le système Vubis, pour des demandes semblables.

Ce que nous pouvons au moins conclure est qu'aucune des situations mentionnées n'est bonne. Le meilleur résultat serait une structure de données standard, flexible et assez complète qui pourrait atteindre tous les objectifs. Ceci présenterait une bonne occasion pour le consortium UDC de rendre l'UDC MRF disponible dans un tel format permettant ainsi aux utilisateurs d'obtenir à la fois la structure et des données prêtes pour l'exécution. La communauté plus large des métadonnées a déjà admis que les outils d'organisation de la connaissance sont mieux gérés comme données indépendantes, externes à l'enregistrement lui-même. L'intérêt des langages pré-coordonnés est habituellement perdu si les relations syntaxiques ne sont pas codées. La CDU est recommandée comme standard d'organisation de la connaissance dans de nombreuses normes de métadonnées telles que le Dublin Core, ou les Learning Object Metadata, Encoded Archival Description etc. La plupart de ces normes permettent le codage de la source d'où provient le terme, celui du terme lui-même, ainsi que l'utilisation d'URIs (Uniform Resource Identifiers - identifiants uniformes de ressource) au cas où les données de classification sont exposées et le réseau accessible à différentes applications. Beaucoup d'efforts sont en effet maintenant concentrés pour assurer le permanence de ces identifiants et permettre aux sujets et autres données d'autorité d'être partagées et mieux exploitées. Il s'agit là d'une tendance générale et l'expertise dans le contrôle d'autorité des bases de données bibliographiques aura certainement une chance d'être déployée dans un environnement plus large. Considérer cette tendance peut aider pour prendre les décisions de mise en oeuvre qui peuvent initialement exiger plus d'effort mais devraient rapidement rapporter des dividendes.

5. Remarques conclusives

Il est courant de décrire et d'analyser la Classification décimale universelle comme un langage d'indexation autonome avec ses propres limites et avantages indépendants des contraintes de sa mise en oeuvre et de son mauvais usage. Mais l'idée que les outils bibliothéconomiques d'organisation de la connaissance sont des outils tout prêts et immédiatement disponibles qui vont résoudre tous les problèmes de la recherche documentaire, est graduellement remplacée par des approches plus pragmatiques. Les questions de la politique d'indexation, du coût de la formation et de la mise en oeuvre sont de plus en plus prises en compte et on admet aujourd'hui largement que l'efficacité et les coûts de maintenance des systèmes de

classification, par exemple, dépendent non seulement de la disponibilité des données de classification dans un format électronique, mais également des outils et du système de recherche construits autour.

On admet généralement aussi qu'une fois produites, les données matières basées sur n'importe lequel des systèmes d'indexation sont trop chères pour pouvoir être gaspillées. Les systèmes d'information modernes ont la capacité et la puissance technologiques d'employer et de combiner différents outils et techniques pour se compléter afin de réaliser des résultats satisfaisants. Il est aujourd'hui courant d'inclure les coûts pour faire correspondre et relier diverses applications, divers formats et structures de données dans tout système. Les bons systèmes tendent à changer et à s'adapter, mais aussi mêlent et assortissent différentes approches et différentes fonctions pour mieux atteindre leurs objectifs. Un exemple extrême de cette approche est, par exemple, GERHARD (German Harvest Automated Retrieval and Directory) qui a employé les données d'UDC MRF, un fichier d'autorité CDU de bibliothèque universitaire, un récupérateur de pages Web et un programme de traitement du langage naturel pour construire un outil automatique de classification.

Le choix de la CDU devrait être fondé sur sa capacité d'échelle, son ouverture à l'expansion, sa flexibilité et son aptitude à venir être complétée par d'autres systèmes. Dans un système d'information, une classification atteint mieux son objectif lorsqu'elle est mise en application avec un index alphabétique ou un langage d'indexation alphabétique. La classification est une robuste structure de base de la connaissance qui fournit un cadre sémantique coordonné, et des relatons hiérarchiques, subordonnées et collatérales parmi ses concepts. Elle peut donc servir d'outil non dépendant du langage pour le contrôle du vocabulaire ainsi que d'outil complémentaire de recherche documentaire pour permettre le feuilletage et la navigation par sujets selon des modes interactifs avancés.

Les classifications sont des langages d'indexation fortement formalisées et, comme tels, sont capables d'atteindre différents objectifs. C'est particulièrement vrai pour la CDU qui n'a pas été seulement créée pour l'organisation des rayonnages de bibliothèque. Elle possède un vaste vocabulaire, une structure de base et une grammaire suffisamment flexible pour s'ajuster à différentes applications. Les classifications sont cependant des outils professionnels par excellence qui exigent expertise et travail d'intellectuel. Mais cette condition peut être réduite considérablement si des solutions technologiques appropriées sont rendues disponibles. La CDU, avec sa notation pré-combinée et sa numérotation structurée, est un candidat idéal au stockage, à la gestion et à l'emploi dans une configuration informatique. L'un des préalables au développement d'outils informatiques pour utiliser et exploiter correctement la CDU est un format approprié de classification, qui doit être créé en ayant à l'esprit les classifications synthétiques et à facettes. Les développements dans ce domaine sont d'une importance capitale pour les usages actuels et futurs de la classification.

6. Notes et références

[1] Subject data in the metadata record: recommendations and rationale: a report from the ALCTS/CCS/SAC/Subcommittee on Metadata and Subject Analysis. July 1999. http://www.govst.edu/users/gddcasey/sac/MetadataReport.html

[2] Exemples de portails thématiques qui appliquent la CDU simple :
Social Science Information Gateway (http://www.sosig.ac.uk),
Directory of Network Resources - NISS (http://www.niss.ac.uk/subject2/new95udc.html)

[3] Il y a plus de mille indices de classe (2285), dans le MRF 2002, qui contiennent au moins un exemple ou plus de combinaison avec leur description, et 6912 indices ont une ou plusieurs références "voir aussi"

[4] La dernière édition du MRF (2002) contient 48 962 indices simples principaux, 17 187 combinaisons d'indices principaux et d'auxiliaires spéciaux, et 11 138 indices auxiliaires communs qui sont des combinaisons de symboles et de nombres

[5] UDC Consortium *Master Reference File Manual*, version Mai 2003, est écrit pour assurer la maintenance de la bdd MRF CDS/ISIS et la diffusion du MRF. Le Manual (26 pages, PDF) peut être téléchargé depuis (http://www.udcc.org/mrf.htm)

[6] Riesthuis, G. J. A. "Some thoughts about the format of the Master Reference File database. *Extensions & Corrections to the UDC*, 23 (2001), 15-22

[7] Riesthuis, G. J. A. "A Revised Format for the Master Reference File Database". *Extensions & Corrections to the UDC*, 24 (2002) 11-15

[8] La politique actuelle de révision de la CDU vise à restructurer les listes sur la base de l'analyse de facette, ce qui limitera la répétition inutile et le nombre de composés dans la classification

[9] McIlwaine, I. C. "UDC in the twenty-first century", The future of classification, edited by Rita Marcella and Arthur Maltby, Aldershot: Gower, 2000, 93-104

[10] La bibliothèque universitaire centrale de Bucarest, Roumanie, emploie la CDU pour une indexation très précise.
ETH-Bibliothek à Zurich emploie environ 60.000 indices construits auxquels ont été reliés plus de 400 000 termes correspondants en anglais, français et allemand.
Voir : Hug, H.; Noethiger, R.. "ETHICS: an online public access catalogue at ETH-Bibliothek, Zurich". *Program*, 22, 2(1988), 133-142

[11] German Harvest Automated Retrieval and Directory - GERHARD http://www.gerhard.de, a employé les données du fichier d'autorité de Zurich (ETH-Bibliothek), 60 000 indices CDU pré-combinés au total, ce qui avec l'index sujet a donné 500.000 lignes de texte. Voir : Möller, G. etc. "Automatic classification of the World Wide Web using Universal Decimal Classification", 23rd International Online Information meeting, London, 7-9 December 1999: proceedings. Editor Brian McKenna. Oxford: Learned Information Europe, 1999. 231-237

[12] Buxton, A. B. "Computer searching of UDC numbers", *Journal of Documentation*, 46 (3) 1990, 193-217.

[13] Riesthuis, G. J. A. "Decomposition of complex UDC notation", Knowledge organization for information retrieval: proceedings of the Sixth International Study Conference on Classification Research, London, 16-18 June 1997. The Hague: International Federation for Information and Documentation (FID), 1997, (FID 716), 139-143.

[14] Riesthuis, G. J. A. "Decomposition of UDC-numbers and the text of the UDC Master Reference File", Structures and relations in knowledge organization: proceedings of the Fifth International ISKO Conference, Lille, 25-29 August 1998. eds. W. Mustafa el Hadi, J. Maniez, S. Pollitt. Würzburg: ERGON Verlag, 1998, (Advances in knowledge organization 6), 221-228.

[15] Strachan, D. "UDC revision work in FID", The UDC: essays for a new decade, edited by Alan Gilchrist and David Strachan. London: ASLIB, 1990, 1-10.

[16] Strachan, D.; Oomes, F. M. H. "The UDC Master Reference File (MRF)", *Extensions & Corrections to the UDC*, 15 (1993). The Hague : UDC Consortium, 1993, 19-28.

[17] Strachan, D.; Oomes, F. M. H. "Universal Decimal Classification update", *Cataloging and Classification Quarterly*, 19(3-4) 1995, 119-131.

[18] Riesthuis, G. J. A. The UDC Master Reference File. 64th IFLA General Conference, August 16 - August 21, 1998. http://www.ifla.org/IV/ifla64/157-158e.htm

[19] MARC 21 Concise Format for Classification Data : http://www.loc.gov/marc/classification/eccdhome.html

[20] Concise UNIMARC Format for Classification Data : http://www.ifla.org/VI/3/p1996-1/concise.htm]

[21] See the thread 'computers handling UDC numbers' at the udc-forum archive at http://www.jiscmail.ac.uk/lists/udc-forum.html, February 2003.