# INCISO :

# Elaboration automatique d'un index de citations des revues espagnoles en sciences sociales.

José M. Barrueco (1)
Jose.Barrueco@uv.es

Julia Osca-Lluch (2)
m.julia.osca@uv.es

Thomas Krichel (3)
krichel@openlib.org

Pedro Blesa (4)
pblesa@dsic.upv.es

Elena Velasco (2)
elenavelascoarroyo@yahoo.es

Leonardo Salom (2)
leosamu@eui.upv.es

*(1) Biblioteca de Ciencias Sociales, Universidad de Valencia, 46022 Valencia*
*(2) Instituto de Historia de la Ciencia y Documentación López Piñero (Universidad de Valencia-CSIC) 46010 Valencia*
*(3)Palmer School, 720 Northern Boulevard, Brookville 11548-1300, USA*
*(4) Dept. de Sistemas Informáticos y Computación, Universidad Politécnica de Valencia, 46022 Valencia*

*Cet article est une traduction de Clotilde Roussel (INIST-CNRS) et Magali Rasolomanana (INIST-CNRS). Il a été révisé par Catherine GUNET (INIST-CNRS) et mis en ligne par l'équipe ARTIST.*

*Il est paru sous la référence originale :*

**INCISO:** ***Automatic Elaboration of a Citation Index in Social Science Spanish Journals***

**Mots-clés** : bibliométrie, recherche scientifique, littérature scientifique, périodique électronique, sciences sociales, évaluation, projet, index citation, Espagne, INCISO (Indice de Ciencas Sociales), logiciel, description système, architecture système

**Keywords :** bibliometrics, scientific research, scientific literature, electronic periodical, social sciences, evaluation, project, citation index, spain, software, system description, system architecture.

**Résumé :** Les index de citations sont des outils clés dans le système de communication scientifique pour deux raisons. Tout d'abord, ce sont une excellente source d'informations pour interroger la littérature scientifique puisqu'ils permettent de naviguer au moyen de liens entre les documents représentés par des références bibliographiques. Ensuite, ils permettent d'évaluer la production scientifique. Le comptage des citations est un processus courant pour évaluer la qualité d'un article scientifique. En Espagne, une telle évaluation n'est possible qu'en utilisant des outils élaborés par l'ISI mais dont la couverture des revues publiées hors des pays anglo-saxons est limitée. L'évaluation de la production scientifique espagnole se limite donc aux travaux publiés dans des revues internationales. Il n'existe aucun outil pour évaluer la recherche (notamment en Sciences humaines et sociales) publiée dans les revues locales. Dans le cadre du projet de recherche INCISO, nous étudierons la possibilité de créer automatiquement un index de citations. L'objectif du projet est de développer un logiciel permettant de générer automatiquement des index de citations et de créer un échantillon d'index de citations pour les Sciences sociales.

Cette recherche est financée grâce à la subvention HUM2004-05532 du ministère espagnol des Sciences et de l'Education.

# Introduction

Les revues scientifiques sont le principal moyen pour la communauté scientifique de communiquer les résultats de ses recherches. Le facteur d'impact des revues scientifiques est devenu un outil clé pour évaluer non seulement la diffusion et la visibilité de ces revues, mais aussi l'importance et la qualité de la recherche scientifique. Pour

calculer le facteur d'impact des revues dans une discipline donnée, il est nécessaire de constituer des bases de données bibliographiques dans lesquelles tous les travaux publiés dans les revues les plus importantes du domaine seront répertoriés. De plus, il faudra que ces bases de données contiennent des informations sur les références bibliographiques de chaque article afin de pouvoir établir des liens entre l'article citant et l'article cité. Enfin, le système doit pouvoir compter le nombre de fois où un article est cité. De telles bases de données s'appellent des index de citations. Actuellement, l'Institute for Scientific Information (ISI) publie trois index couvrant toutes les disciplines (Science Citation Index, Social Science Citation Index et Arts and Humanities Citation Index). Les données de ces index sont utilisées pour évaluer la recherche dans les universités du monde entier.

Les coûts élevés et la grande complexité technique qu'entraine la création d'index de citations ont freiné jusqu'ici, le développement de nouvelles bases de données qui pourraient être utilisées en complément des produits de l'ISI. Dans le cas des pays non-anglophones, un tel complément serait utile car l'ISI ne traite que des revues internationales en grande majorité de langue anglaise.

En 1983, Garfield a signalé que le facteur d'impact servait avant tout à la littérature anglo-saxonne et que donc toute évaluation fondée sur ce facteur d'impact n'était valable qu'au sein de cette communauté anglo-saxonne. Différents auteurs ont analysé les données du SCI et les ont comparées à la production scientifique des pays non anglo-saxons ; ils ont constaté que la discrimination envers ces pays était évidente. On peut observer ce problème dans les sciences dures et les technologies, mais il est encore plus marqué dans les sciences humaines et sociales, car dans ces domaines, les chercheurs publient souvent dans des revues nationales ou régionales car ces dernières sont davantage en rapport avec la portée locale de leur recherche. La recherche publiée dans les revues locales n'est donc pas couverte par l'ISI et ne peut être évaluée.

Le présent projet n'est pas le premier à vouloir développer des index de citations en Espagne ; il y a eu plusieurs tentatives depuis les années 1990 et notamment « l'Index de citations et les index bibliométriques des revues espagnoles en médecine interne et ses

spécialités » (Terrada et coll., 1991), « l'Index de citations de la documentation en espagnol » (Moya et coll., 1998), « l'Index de citations des revues espagnoles en sciences humaines » (Sanz et coll., 1998), « l'Index de citations des revues espagnoles en psychologie » (Tortosa et coll., 2002), « l'Index de citations des sciences économiques et des affaires » (Hernández et coll., 2003) et plus récemment « l'Index des sciences sociales » (Jiménez-Contreras et coll., 2004). Tous sont focalisés dans le cadre géographique spécifique de l'Espagne sur un secteur scientifique spécifique, avec une couverture temporelle délimitée.

La plupart des projets que nous avons cités ont malheureusement disparu par manque de financement mais tous partageaient les mêmes caractéristiques :

- ils étaient basés sur un enregistrement manuel des références et des citations,
- ils se concentraient sur une discipline concrète,
- ils utilisaient un échantillon réduit de revues (4-5 dans certains cas) et saisissaient aussi peu d'information que possible afin de réduire la charge de travail des opérateurs de saisie.

Notre conclusion est que la constitution d'index de citations généraux selon des moyens traditionnels demande des ressources beaucoup trop coûteuses pour généraliser de tels index au niveau national. Autrefois, seul l'ISI disposait des ressources nécessaires pour élaborer des index de revues papier. Toutefois, de nouvelles voies se sont ouvertes avec la généralisation de l'internet comme nouveau moyen de communication, avec la prolifération des revues électroniques au niveau national comme au niveau international et avec la possibilité de créer des index par des moyens automatiques. Si les articles étaient disponibles dans des formats numériques, un système informatique pourrait alors en extraire les références automatiquement. Avec un tel système les coûts seraient nettement réduits et de nouveaux index couvrant de nouveaux types de documents (par exemple la littérature grise) pourraient voir le jour.

Pour essayer de développer plus avant cette idée et en nous basant sur les travaux des auteurs décrits plus loin, nous avons décidé d'étudier la possibilité de développer un système informatique

capable de créer automatiquement des index de citations pour les publications espagnoles. Notre projet a obtenu une subvention de recherche de trois ans de la part du ministère espagnol de la science et de la technologie. Cette subvention a débuté en juillet 2005 et nous avons appelé le projet INCISO (Indice de Ciencias Sociales ou Index des Sciences sociales). INCISO avait pour but général de réduire les coûts du processus en remplaçant l'homme par un système informatique capable de constituer automatiquement un index de revues électroniques. Le projet avait deux objectifs principaux :

> 1) concevoir un système informatique pour élaborer un index de citations de manière automatisée. Le système pourra s'appliquer à des disciplines multiples mais il sera testé sur une sélection de revues espagnoles en sciences sociales.
>
> 2) élaborer et diffuser un index de citations pour les sciences sociales basé sur une sélection de revues espagnoles. Cet index sera mis à disposition de toute la communauté scientifique et sera librement accessible sur le site web du projet : http://inciso.openlib.org/.

La suite de cet article est organisée comme suit. La deuxième partie décrit certains autres projets de recherche de niveau international qui travaillent sur l'extraction automatique et les liens entre les références dans le but de constituer des index de citations. Dans la troisième partie, nous abordons la méthodologie et les étapes de notre projet. L'architecture d'INCISO est discutée dans la quatrième partie. La cinquième partie décrit l'état d'avancement du projet et conclut l'article.

# 1 Autres travaux sur le sujet

La généralisation des formats électroniques d'édition et de distribution des articles scientifiques a permis de développer de nouvelles fonctionnalités de recherche documentaire comme par exemple les recherches en texte intégral, les liens entre les références bibliographiques ou les index de citations autonomes. Notre projet

s'inscrit dans ce dernier aspect. Roth (2005) décrit plusieurs autres projets de recherche relatifs au développement des index de citations qui pourraient éventuellement concurrencer Science Citation Index (SCI). De la liste de Roth nous retenons deux groupes de projets : les projets commerciaux et les projets académiques.

Les projets commerciaux sont généralement menés par les sociétés d'édition afin de développer et d'améliorer leurs offres de services bibliographiques. D'un point de vue technique, ils utilisent les informations issues de documents et de références déjà disponibles dans les bases de données des éditeurs. De telles informations ont été générées au cours du processus éditorial et elles sont en général au format SGML ou XML. La grande qualité des données permet de développer des services à forte valeur ajoutée. Les défis techniques que ces projets doivent relever sont l'établissement de liens entre les références et le texte intégral sur différentes plates-formes et la gestion des droits d'accès aux documents. A titre d'exemple :

- Chemical Abstract propose une interrogation des références citées remontant à 1997. Chaque notice est liée à d'autres notices qui le citent correctement grâce à deux fonctionnalités : « Get related » (obtenir les références apparentées) et « Get citing references » (obtenir les références citantes), cette dernière fonctionnalité permettant aux utilisateurs de connaître le nombre de fois où un article a été cité. Voir http://www.cas.org/casdb.html,
- Scopus est généralement considéré comme un éventuel concurrent du SCI puisqu'il fournit des résultats de recherche qui incluent des résumés, des références citées et des liens vers les références citées (Roth 2005). Scopus est une initiative d'Elsevier, le géant de l'édition STM qui a récemment ajouté 13 millions de notices de brevets. Voir http://www.scopus.com.,
- CrossRef. est un service collaboratif d'établissement de liens entre les références qui fonctionne un peu comme un standard numérique. Il ne contient aucun contenu en texte intégral, mais plutôt des liens vers celui-ci grâce aux identifiants d'objets numériques (Digital Object Identifier-DOI) associés aux métadonnées des articles fournies par les éditeurs participants à CrossRef. Le résultat final est un

système de liens souple grâce auquel un chercheur peut cliquer sur une citation dans une revue et accéder à l'article cité. CrossRef a démarré en 2000 quand un groupement des principaux éditeurs scientifiques a fondé la Publishers International Linking Association, Inc (PILA) qui gère Crossref. Voir http://www.crossref.org.

Les projets académiques travaillent sur toute sorte de types de documents disponibles sur l'internet. Ils analysent les documents en texte intégral afin d'en extraire les références et les citations qui seront reliées aux documents d'origine s'ils sont disponibles sous forme électronique. Dans ce cas les données sont extraites automatiquement par des programmes informatiques. Ces projets ne sont pas tous d'un niveau de qualité similaire mais d'une manière générale leur qualité est inférieure à celle des projets commerciaux. Le principal enjeu de ces projets consiste à améliorer le processus technique d'extraction des métadonnées les plus pertinentes.

- Citebase permet d'interroger des documents dont les références ont été analysées et d'obtenir des résultats classés par facteur d'impact. Citebase est le fruit de l'Open Citation Project, projet bénéficiant du soutien du Joint Information Systems Committee du Royaume-Uni et de la National Science Foundation des Etats-Unis.
- CiteSeer est un système logiciel à double fonction, l'extraction de citations et leur implémentation dans le dispositif informatique, permettant ainsi de produire une base de données contenant plus de 200 000 documents indexés, avec plus de deux millions de références bibliographiques. Il a été développé dans les laboratoires de recherche de NEC par Steve Lawrence, Kurt Bollacker et C. Lee Giles. Voir http://citeseer.ist.psu.edu.
- CitEc est un index de citations dans le domaine des sciences économiques référençant les documents électroniques disponibles dans la bibliothèque numérique de RePEc. CitEc utilise une version modifiée du logiciel de CiteSeer pour établir des liens entre les références de documents disponibles en libre accès (principalement des documents de travail). Pour chaque notice dans RePEc, CitEc propose les fonctionnalités suivantes : « cited by » (cité par) si le

document a été cité par d'autres documents également disponibles dans RePEc et « get references » (obtenir les références) quand les références de l'article citant sont effectivement reliées aux documents cités. Voir http://netec.ier.hit-u.ac.jp/CitEc.

- Google Scholar est une base de données de littérature scientifique contenant des articles évalués par les pairs, des thèses, des livres, des prépublications, etc. provenant d'éditeurs scientifiques, de sociétés savantes et d'archives de publications électroniques. Google Scholar analyse et extrait automatiquement les citations et les présente sous forme de résultats séparés même si les documents auxquels ces citations font référence ne sont pas disponibles en ligne. Voir http://scholar.google.com.

## 2 Méthodologie et déroulement du projet

Les auteurs de cet article ont une grande expérience dans le développement d'index de citations autonomes puisqu'ils ont développé le service CitEc décrit ci-dessus. Avec ce nouveau projet, il va s'agir de transposer l'expérience acquise dans une discipline spécifique à des publications dans une langue autre que l'anglais issues de plusieurs disciplines, avec comme facteur commun le même pays de publication. Une grande partie du logiciel développé pour CitEc sera utilisé et testé dans ce nouvel environnement.

La méthodologie que nous allons suivre pour extraire et relier les données concernant les références, comprend sept étapes :

1. Il nous faut sélectionner les sources de données. Le système sera testé sur un échantillon de revues espagnoles en sciences sociales. Dans un premier temps, cet échantillon est réduit à dix revues représentant toutes les disciplines. La sélection a été faite selon les critères suivants : les revues doivent disposer d'une version électronique avec au moins quatre numéros publiés et un système d'évaluation par les pairs pour s'assurer de la qualité du contenu. Etant donné

que l'index va être créé automatiquement, il est essentiel que les revues disposent d'une version électronique. Le nombre de revues électroniques en Espagne est encore faible. Néanmoins il se développe rapidement comme on peut le voir dans le répertoire des revues électroniques espagnoles en sciences humaines et sociales (disponible à : http://citas.uv.es/DifusionRevistas/Revistaselectronicas/index.html). Y figurent les nouvelles revues créées exclusivement sous forme électronique et d'autres revues qui s'orientent vers le format électronique tout en maintenant une version papier. La sélection, basée sur la disponibilité ou non du format électronique, implique que d'importantes revues seront exclues de l'échantillonnage parce qu'elles n'existent que sous forme papier. Nous sommes conscients que les meilleures revues espagnoles sont exclues et que les résultats doivent être interprétés en conséquence et ne doivent pas être utilisés pour évaluer la recherche. La situation devrait s'améliorer à l'avenir au fur et à mesure que de plus en plus de revues passent à l'électronique.

2. Il nous faut obtenir les informations bibliographiques concernant les articles publiés dans les revues sélectionnées. A l'avenir, il serait souhaitable de collaborer avec les fournisseurs d'informations (les éditeurs) afin de définir des moyens automatisés pour alimenter le système. Cela signifie qu'il faut mettre en place des procédures permettant d'alerter INCISO quand de nouveaux articles sont publiés. Pour ce faire, nous utiliserons de nouvelles technologies dans le domaine des bibliothèques électroniques telles que le protocole OAI-PMH (Open Archives Initiative-Protocol for Metadata Harvesting), voir http://openarchives.org.
3. Les informations bibliographiques de chaque article ayant une adresse électronique renvoyant vers le document en texte intégral seront stockées dans une base de données mySQL. Ces documents seront considérés comme les documents citants. Un autre fichier de la base de données recueillera les métadonnées inhérentes aux documents publiés en Espagne dans le domaine des sciences sociales au cours des dix dernières années. Ces documents sont

potentiellement les documents cités. Ces métadonnées sont considérées comme faisant autorité (contrôlées) car elles proviennent de sources de qualité. Seules les citations faisant références à ces documents seront retenues et considérées comme de vraies citations. Toutes les autres citations seront écartées.

4. Pour chaque document citant, on télécharge le fichier contenant le document en texte intégral. A l'heure actuelle, INCISO ne gère que les fichiers en format pdf. Le fichier est converti au format ASCII afin que le texte puisse être facilement extrait et manipulé.
5. Une fois le fichier converti, on lance l'analyse syntaxique de l'ensemble du texte en vue d'identifier et de délimiter la section contenant les références bibliographiques. Si cette étape aboutit, il faut ensuite identifier chaque référence citée et la fractionner en fonction de ses différents éléments comme l'auteur, le titre, la revue, etc. C'est l'étape la plus importante, car l'efficacité du processus dépend essentiellement de la qualité et de la cohérence de ces résultats. Un problème majeur est que les références sont différentes selon les disciplines. La démarche adoptée par la plupart des projets décrits plus haut consiste à extraire de manière aussi précise que possible tous les éléments des références. Ces projets ont donc tenté d'analyser les notices de manière exhaustive. A notre avis, une telle analyse est compliquée et gourmande en ressources car la qualité des données source est très hétérogène. Notre approche est différente. Le système ne va identifier que les éléments de base de la référence et essaiera ensuite de localiser le document référencé dans la base de métadonnées contrôlées. S'il trouve le document, les bonnes métadonnées viendront compléter la référence.
6. Toutes les données extraites au cours des étapes précédentes sont stockées dans une base de données de références. Cette base de données servira à faire des études bibliométriques sur les résultats.
7. Le projet propose deux types de résultats. D'une part, l'index de citations qui sera utile pour évaluer la recherche en sciences sociales menée en Espagne et d'autre part, un

ensemble de documents techniques concernant le système qui sera d'un intérêt majeur pour la communauté des chercheurs dans le domaine des bibliothèques numériques. Tous les résultats seront publiés en libre accès sur le web.

INCISO va développer un système informatique pour réaliser de manière automatisée le processus décrit précédemment. La conception du système s'appuiera sur les principes fondamentaux suivants :

- la multidisciplinarité. Dans un premier temps, le système sera appliqué aux revues en sciences sociales mais reposera sur une architecture modulaire permettant d'adapter facilement de nouvelles fonctions au noyau de base afin de répondre aux besoins spécifiques des différentes disciplines.
- les logiciels libres. Le système sera complètement écrit en Perl. Les logiciels complémentaires nécessaires relèveront du logiciel libre comme par exemple ceux qui utilisent GNU ou des licences similaires. Le système fonctionnera sous DebianGNU/Linux sur une machine localisée à l'université polytechnique de Valence (Espagne) et utilisant mySQL comme système de gestion de base de données et Apache comme serveur web.
- l'autonomie et la continuité. Une des principales exigences à prendre en compte dans la conception du système est qu'il faudra que ce dernier puisse fonctionner avec le minimum de maintenance possible. Les systèmes actuels reposent sur le travail éditorial d'administrateurs ce qui nécessite des ressources financières pour les rémunérer. Si nous mettons en place un système automatisé au maximum et si nous parvenons à constituer une masse critique de documents, alors il se peut que certains éditeurs souhaitent contribuer au système en y versant leurs publications. Cela permettrait d'assurer un flux continu de documents et le système pourrait fonctionner par lui-même sous l'impulsion des éditeurs.
- l'ouverture. Les données produites seront accessibles à toute la communauté scientifique ainsi qu'à d'autres projets au niveau international. Le premier prolongement du projet pourrait concerner les revues publiées en Amérique latine.

Latindex ( http://www.latindex.org) est un annuaire de revues électroniques recensées par le CINDOC qui pourrait servir à sélectionner les revues de qualité à inclure dans INCISO.

# 3 Architecture du système

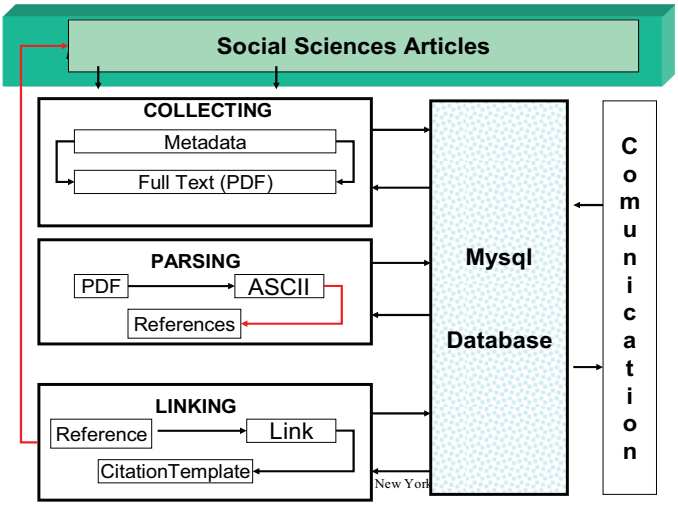


**Figure 1 : l'architecture d'INCISO**

Comme le montre la figure 1, l'architecture d'INCISO est fondée sur deux éléments principaux. Tout d'abord, nous travaillons sur un corpus d'articles publiés dans des revues espagnoles en sciences sociales. Nous avons constitué une banque de métadonnées contrôlées décrivant chacun des articles. Ces métadonnées sont stockées dans une base de données bibliographiques. Nous n'entrerons pas ici dans les détails de cette base de données car ce n'est pas le sujet du présent article. Ensuite, nous disposons d'une série de trois modules

correspondant aux trois étapes du processus d'établissement de liens entre les références et les documents (Barrueco, 2005) :

1. collecte des métadonnées et du texte intégral des documents,
2. analyse des documents afin de repérer là où se trouvent les références et extraire les éléments (auteurs, titre, etc.) de chacune de ces références,
3. création des liens entre les références et le document qu'elles représentent si celui-ci est disponible dans INCISO.

Il est important de noter que chaque module est tributaire de la production du module qui le précède. Ainsi, l'intégration de chaque document implique de valider les trois niveaux de traitement. Chaque document se voit attribué un statut correspondant à l'étape du processus où il se trouve. Le statut initial « nofulltex » (pas de texte intégral) et le dernier « linked » (lien établi) si tout se passe bien.

1. **Collecte**. La collecte implique trois étapes différentes : (1) collecte des métadonnées des documents, (2) téléchargement du texte intégral des documents et (3) conversion en un format prêt à être analysé par le système informatique. Les métadonnées des documents citant sont complétées avec l'URL du texte intégral des articles. Dans certains cas, les URLs fournies peuvent être erronées ou le serveur web peut être hors service lorsque le système tente d'accéder aux documents. Alors, un statut spécial est attribué aux articles et le processus s'arrête jusqu'à l'intervention manuelle du personnel éditorial qui vérifie et corrige le problème rencontré. Une fois que le fichier en texte intégral est sauvegardé sur le disque dur, nous commençons le processus de conversion. Dans un premier temps, nous vérifions si le fichier en texte intégral est compressé. Si c'est le cas, un algorithme de décompression est utilisé. Dans un second temps, nous vérifions le format du fichier. Actuellement, seuls les documents en pdf sont acceptés. Heureusement, le pdf est un format très répandu pour publier des articles scientifiques sur l'internet. La dernière étape consiste à convertir le document pdf en ASCII. Dans

ce but, nous utilisons pdftotext, le logiciel développé dans le cadre du visionneur Xpdf. Tous les fichiers pdf ne peuvent pas être convertis avec suffisamment de qualité pour permettre l'extraction. La qualité des fichiers pdf dépend principalement du logiciel utilisé pour créer les fichiers et également du bon codage des polices de caractères.

2. **L'analyse** syntaxique est l'étape la plus compliquée. Les auteurs utilisent divers formats pour leurs références et ces formats peuvent varier au sein d'un seul et même article. En outre, la manière dont les références sont indiquées dans les documents varie d'une discipline à l'autre. Compte tenu de l'importance de la phase d'analyse syntaxique, nous avons choisi de commencer avec un logiciel déjà testé plutôt que de développer un nouveau logiciel en recommençant à zéro. Notre choix s'est porté sur le logiciel développé pour le projet CitEc, qui a été décrit dans des articles comme celui de Lawrence (1999). Le logiciel de CitEc est capable d'identifier la partie du document contenant la liste des références. Ensuite, il peut séparer les différentes références de cette liste. Enfin, il procède à l'analyse de chaque référence pour en retrouver les différents éléments. Pour le moment, il n'identifie que l'année de publication, le titre et les auteurs. Cependant, ces éléments sont suffisants pour notre objectif. La qualité des références bibliographiques fournies dans les revues est variable. Par exemple, dans un même article, il est fréquent de trouver différentes formes du nom d'un même auteur, différentes formes du titre d'une même revue, etc. Nous utilisons les métadonnées contrôlées pour compléter les références et nous en améliorons la qualité avec des métadonnées provenant des éditeurs.
3. **Etablissement des liens**. Une fois que nous avons analysé les documents, l'étape suivante consiste à regarder si certaines des références extraites avec succès renvoient à des documents disponibles dans la base de données d'INCISO. Dans ce cas, il faut établir un lien entre les deux documents. Pour ce faire, nous comparons chaque référence

[NdT] – La phrase précédente ne mentionne que trois éléments.

analysée aux métadonnées contrôlées stockées dans la base de données bibliographiques d'INCISO. Actuellement, nous considérons qu'une référence représente un document d'INCISO quand :

a. l'analyse du titre de la référence et le titre dans notre collection de métadonnées sont assez proches ;
b. l'année de publication des deux items est identique ;
c. au moins un des auteurs des articles correspond aux auteurs de la notice de métadonnées.

Dans ce processus, nous prenons chaque référence, nous extrayons le titre analysé et nous le convertissons en une version normalisée appelée le titre clé. Ici tous les différents espaces et articles sont enlevés et toutes les majuscules sont converties en minuscules. Ensuite, nous sélectionnons dans notre base de données bibliographiques tous les documents qui contiennent dans leur titre tous les mots du titre clé de la référence. Tous les articles sélectionnés sont susceptibles de devenir le document cité. Dans un deuxième temps, nous calculons la distance de Levenshtein de chaque titre clé du document candidat avec le titre de clef de la référence. Si cette distance dépasse de 8 % la longueur du titre clef de la référence, le document est rejeté. Enfin, nous vérifions si l'année de publication des articles candidats et celle de la référence sont identiques. Si c'est le cas nous estimons que la référence correspond au document que nous avons. Les auteurs sont comparés uniquement quand le titre est court et qu'il ne permet pas de différencier les éléments. Les informations sur les citations sont stockées dans une table de la base de données mySQL. Cette base de données sera utilisée pour développer des indicateurs bibliométriques.

## Conclusions

Dans cet article nous avons décrit une méthodologie pour développer automatiquement un index de citations. En mettant en œuvre cette méthodologie, le projet d'INCISO va essayer de réduire les coûts élevés liés au développement d'index de

citations par des moyens traditionnels. En cas de succès, cela ouvrira la voie aux pays non-anglophones pour développer leurs propres index qui pourraient être utilisés comme complément de l'ISI dans l'évaluation de la recherche.

Actuellement, nous venons juste de commencer à développer le logiciel. On s'attend à avoir les premiers résultats en 2006. Alors une période d'évaluation commencera afin de déterminer si les résultats sont suffisamment bons pour permettre à la fois la recherche d'informations et l'extraction des indicateurs bibliométriques.

Il existe d'autres projets au niveau international opérant dans le même domaine. L'innovation d'INCISO réside dans l'utilisation d'une base de données de métadonnées contrôlées qui normalise les références extraites des documents.

## Bibliographie

[1] Barrueco, José Manuel, and Thomas Krichel (2005) "Building an autonomous citation index for grey Literature: RePEc, the economics working papers case" The Grey Journal, An International Journal on Grey Literature, vol. 1, no. 2, pp. 91–97

[2] Delgado López-Cozar, Emilio y otros (2005). INRECS: Índice de impacto de las revistas españolas de ciencias sociales. Biblio 3W, Revista Bibliográfica de Geografía y Ciencias Sociales, Vol. X, no. 574.

[3] Hernández Mogollón, Ricardo (2003). Citaedem.. Indice de citas de economía de la empresa. Memoria y resultados. Universidad de Extremadura.

[4] Lawrence, Steve, Kurt Bollacker, and C. Lee. Giles (1999) "Indexing and retrieval of scientific literature", proceedings of eighth International Conference on Information and Knowledge Management, CIKM99, pp. 139–146.

[5] López Piñero Jose María, Terrada María Luz. (1994). El consumo de información científica nacional y extranjera en las revistas médicas españolas: un nuevo repertorio destinado a su estudio. Medicina Clínica, vol.. 102, pp. 104-112.

[6] Osca-Lluch Julia and Haba Julia (2005). Dissemination of Spanish Social Sciences and Humanities Journals. Journal of Information Science, vol. 31, no. 3, pp.229-236.

[7] Osca-Lluch, Julia. (2005). Some considerations on the use the impact factor of scientific journals as a tool to evaluate research in psychology. Scientometrics, vol. 65, no.2, pp.189-197.

[8] Roth, Dana L. (2005) "The emergence of competitors to the Science Citation Index and the Web of Science", Current Science, 2005, vol. 89, no. 9. pp. 1531–1536.

[9] Tortosa, Francisco, Civera, Cristina, Osca-Lluch, Julia, Barrueco, José Manuel, Quiñónes, Elena, Peñareanda, María, Martinez, Francisco, López, Juan José (2005). "Creación de un índice de citas de revistas españolas de psicología". I Jornadas Españolas de Indicadores para la Evaluación de la Ciencia, Madrid. Disponible en: http://www.cindoc.csic.es/info/fesabid/25.htm