Universidad Politécnica de Valencia
Licenciatura en Documentación

CitEc : un agente para la extracción automática de citas en Economía

José Manuel Barrueco

Proyecto Final de Carrera

Dirigido por Vicente Julián Inglada (DSIC)

13 de febrero de 2002

Agradezco la colaboración de Silvia Giner
Silvia.Giner@uv.es en la corrección
de este trabajo

Agradezco también la ayuda de Thomas Krichel
krichel@openlib.org y el soporte técnico
de Satoshi Yasuda yasuda@ier.hit-u.ac.jp

Índice General

Capítulo 1

Introducción

La ciencia moderna es una actividad social en la que no existe el trabajo individualista de una sola persona. Los avances y descubrimientos son fruto del trabajo de grupos de investigadores que ponen en común sus trabajos por encima de limitaciones geográficas o temporales.

Esa puesta en común se lleva a cabo a través del denominado sistema de publicación de la ciencia, que (15) define como:

> el conjunto de elementos y pautas que sostienen, regulan y perpetúan el proceso por el que los investigadores hacen accesible de modo "oficial" al resto de la comunidad científica sus pretensiones de contribuir al acervo científico. El fruto que resulta de la operación de este sistema es la literatura científica.

Un trabajo científico no puede considerarse como tal hasta que no haya sido publicado y puesto a disposición del resto de la comunidad científica. Esos elementos y pautas para llegar a la publicación se han ido perfeccionando desde el siglo XVII cuando aparecieron las primeras revistas científicas. Así por ejemplo en la redacción de los trabajos existe un estilo literario propio de la ciencia, que aunque no está explícitamente recogido en ningún tratado es fielmente seguido por los autores, ya que es condición indispensable para que éstos aparezcan en las revistas. La propia estructura del documento científico también ha variado hasta alcanzar la forma particular que tiene en la actualidad. Aunque existen diferencias dependiendo de las disciplinas, un documento científico se puede caracterizar por los siguientes elementos (15):

- **Título**. Es una frase que encabeza y presenta el trabajo mediante una descripción breve de su contenido. Se espera que sirva para que el lector pueda determinar si le interesa o no continuar leyendo el documento.

- **Autores**. Mención explícita de los nombres de los responsables intelectuales del documento con indicación, en muchos casos, de sus lugares de trabajo.
- **Resumen**. Texto breve que suele describir el problema tratado y enumerar los resultados más destacados que se reivindican en el texto principal.
- **Palabras clave**. Una serie de palabras o conceptos cuyo objetivo es destacar los puntos por los que el trabajo se conecta con la investigación de su disciplina.
- **Texto**. Es el núcleo esencial del documento científico, el portador del contenido.
- **Referencias bibliográficas**. Contiene la lista de todos los trabajos que han sido citados en el texto, enumerados bajo un formato abreviado establecido, que posibilita la identificación y localización del correspondiente trabajo.

Este último punto es un nuevo indicador del carácter social de la ciencia y una de las características que diferencian la literatura científica de otras representaciones literarias. Igual que no existe el investigador individual, tampoco puede existir un documento aislado. Si un documento científico no es más que la representación en un soporte de los hallazgos y descubrimientos que los autores han realizado, es lógico que éstos hagan referencia explícita a cuáles han sido sus puntos de referencia, sus colaboradores, etc. De esta forma, un documento ocupa un nodo en la inmensa red de la literatura científica, conectado a través de las referencias bibliográficas con un número mayor o menor de nodos relacionados temáticamente con él.

Gráficamente se podría representar esta red como se ve en la figura 1.1. Cada nodo representa un documento. Las flechas representan las relaciones marcadas por las referencias bibliográficas. De esta forma, las flechas que parten de los nodos son referencias que ha realizado el autor de los mismos a otros documentos publicados con anterioridad. Las flechas que llegan a los nodos serían citas que tales nodos han recibido. Así por ejemplo el documento **E** referencia dos trabajos: **F** y **G**, mientras que a su vez es citado por otros dos **C** y **D**. Lógicamente las relaciones son unidireccionales, un documento no puede citar a la vez que es referenciado por otro.

Para el objeto de este trabajo es importante marcar desde el principio la diferencia entre citas y referencias puesto que habitualmente se confunden y tienden a usarse indistintamente. (16) señala que desde los trabajos de Krauze y Hillinger en 1971 la distinción entre ambos términos ha quedado claramente establecida: citas son aquellas menciones que una publicación recibe de otras posteriores mientras que referencias son aquellas menciones que una obra hace a otras anteriores.

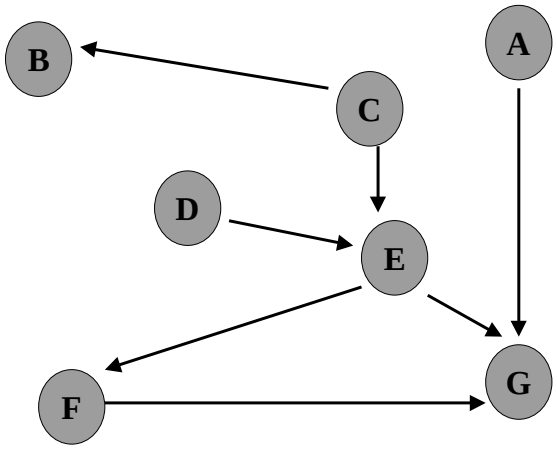


Figura 1.1: Enlaces de documentos a través de citas

De lo dicho se puede intuir la importancia de las listas de referencias en los documentos científicos. Debido a este papel preponderante que juegan en el sistema de comunicación de la ciencia, han sido objeto de innumerables estudios desde disciplinas como la bibliometría, las ciencias de la documentación o la sociología de la ciencia.

- Desde la Bibliometría se ha utilizado el estudio de las citas entre otras cosas para estudiar el consumo de la literatura científica en un país, disciplina, etc. así como para cuantificar la repercusión o impacto que ha tenido un determinado trabajo en el resto de la comunidad científica. Todos los trabajos bibliométricos han sido posibles gracias a las investigaciones de Garfield en los años 60. Este autor es el creador de los índices de citas (*Science Citation Index*), una de las aportaciones más importantes hechas a la Documentación durante el pasado siglo. Una descripción de sus trabajos puede verse en (10).
- Desde la Sociología de la Ciencia se ha estudiado la motivación de los científicos a la hora de realizar citas. La pregunta que se intenta responder es ¿por qué se cita un trabajo?. Para ello generalmente se recurre al estudio del contexto en el que se realiza la cita. (15) en su excelente trabajo recopila una serie de motivaciones que van desde las de clásicos como Ziman (las referencias articulan las contribuciones individuales en un corpus de conocimiento consensuado y común) o Merton (las citas son una parte fundamental del sistema de recompensas de la ciencia, ser citado es un signo de reconocimiento), hasta trabajos más modernos como los de Cozzens donde se insiste en la multidimensionalidad de la motivación para citar incluso al mismo trabajo. Sintetizando se podrían establecer tres motivos por los cuales se cita: atribuir ideas a sus autores originales recompensando así su trabajo, autojustificación del propio trabajo al indicar las obras en las que está basado y abreviar la exposición de teorías ya conocidas mediante la remisión a los documentos originales.
- Desde las Ciencias de la Documentación se han utilizado las referencias como un instrumento adicional para la recuperación de información. Dado que son representaciones abreviadas de documentos relacionados temáticamente, son un medio ideal para navegar por la literatura científica partiendo de un documento relevante. Otra aproximación ha sido la realización de mapas conceptuales de una determinada disciplina como por ejemplo ha hecho (18).

Autores como (7) han imaginado una base de datos universal de citas que permitiera unir cualquier trabajo científico escrito en la historia con los documentos a los que hace referencia. Este autor describe un sistema en el cual cualquier documento estaría disponible y podría ser localizado a través de Internet. La base de datos incluiría información sobre citas y sería exhaustiva y

actualizada. Sin embargo su propuesta requiere que los autores o instituciones proporcionaran información sobre las citas en un formato específico. Esto es muy costoso de hacer en el futuro y muy difícil de lograr sobre el material del pasado, por lo que no es sorprendente que su propuesta nunca se haya hecho realidad.

La elaboración de índices de citas por procedimientos tradicionales, es decir, con un equipo humano recopilando información desde los documentos impresos e introduciéndola en una base de datos implica elevados costes. De hecho, los únicos existentes hasta el momento son los del Institute for Scientific Information de Estados Unidos. La principal carencia que presentan los mismos se refiere a la cobertura de obras escritas en idiomas diferentes del inglés. Fuera del ámbito anglosajón este tipo de herramientas es prácticamente inexistente. Por poner el caso de nuestro país no existe ningún repertorio similar, de forma que es imposible saber cuáles son los documentos, autores o revistas más citadas y por consiguiente más representativas de una disciplina. Y lo que es peor tampoco existe por parte de las instituciones que teóricamente deberían encargarse de ello, como el CINDOC, proyectos de llevarlo a cabo a pesar de que la comunidad científica española lo reclama.

Al menos una de las condiciones que plantea Cameron en su trabajo sí que se está cumpliendo. Cada vez son más los trabajos que están disponibles en formato electrónico a través de Internet. El número de investigadores que están colocando sus publicaciones en sus páginas web o archivos de documentos está creciendo cada día. Es el caso por ejemplo de arXiv[1] en Física o CogPrints en ciencias del conocimiento. Para estudiar la interoperabilidad de dichos archivos (9) se ha creado recientemente la Open Archives Initiative (OAi)[2].

Con la existencia de los documentos en formato electrónico ha surgido la posibilidad de tratar los textos automáticamente para extraer las referencias sin intervención de personas. De esta forma, si un programa es capaz de determinar dónde se encuentra la sección de referencias en un documento, individualizar cada una de ellas y diferenciar cuál es el título, los autores, la fuente donde ha sido publicado y cuál es su dirección electrónica cabría la posibilidad de crear un enlace entre ambos documentos que nos permitiera ir de uno a otro a través del hiperespacio electrónico.

Esta posibilidad de enlazar una referencia con el texto completo del documento al que representa es lo que los anglosajones llaman **reference linking** o enlace de referencias.

El proceso de enlace de referencias plantea los siguientes problemas o etapas:

1. Conversión del documento original a un formato que pueda ser analizado.

[1] `http://www.arXiv.org`
[2] `http://www.openarchives.org`

Por ejemplo de PDF a ASCII o de HTML a XML.

2. Analizar el documento para encontrar las listas de referencias. Seguidamente identificar cada referencia y los elementos que la componen. Además es necesario identificar la parte del documento donde se ha realizado la cita. Es decir, aislar la frase o frases con las cuales el autor se ha referido a esta obra en cuestión.

3. Comprobar si la obra citada se encuentra disponible en formato electrónico y en caso afirmativo realizar el enlace.

Esta idea está siendo objeto de múltiples estudios en la actualidad. Existen iniciativas como ResearchIndex que es capaz de construir verdaderos índices de citas de forma autónoma (autonomous citation indexing) y que ha sido probado con éxito en el área de informática o proyectos como el del Cornell Digital Library Research Group que está trabajando con la extracción de citas desde entornos poco homogéneos, etc.

Aquí se ven los dos tipos de enlaces de referencias que se pueden dar en estos momentos. El universo de documentos electrónicos se podría clasificar en dos tipos:

1. Documentos accesibles libremente en la red a través de webs institucionales o personales. Están caracterizados por su gran heterogeneidad en cuanto a formatos de presentación física e incluso de presentación formal.

2. Documentos accesibles desde archivos como los citados anteriormente e incluso desde los servidores de las editoriales de las revistas científicas. Estos se caracterizan por todo lo contrario, se trata de documentos muy homogéneos en todos los sentidos.

Lógicamente el trabajo de extracción de referencias y de establecimiento de enlaces plantea problemas muy diferentes en ambos casos.

El presente trabajo se inserta en esta línea de investigación tendente a utilizar las referencias como un instrumento para la recuperación de la información. El tipo de enlace de referencias que se plantea aquí está a medio camino entre los vistos antes, porque si bien se trabajará con lo que aparentemente se puede ver como un sólo archivo, los documentos tienen una procedencia muy variada con lo cual se mezclan los problemas de ambos modelos.

El conjunto de datos elegidos para realizar la investigación es RePEc (Research Papers in Economics)[3]. En el siguiente capítulo se hace un análisis bastante detallado de RePEc.

[3] `http://www.repec.org`

En resumen, el objetivo de este trabajo es **experimentar la posibilidad de llevar a cabo un enlace de citas entre los documentos disponibles en RePEc**. Es importante señalar que solamente se intentará hacer un enlace interno, es decir, a documentos que estén disponibles en RePEc. Los enlaces externos a documentos ajenos a RePEc, incluso si éstos están en formato electrónico, serán objeto de futuros trabajos.

Para llevar a cabo este objetivo se ha estudiado el software generado por proyectos similares y se han reutilizado algunos de sus algoritmos para desarrollar un sistema nuevo llamado **CitEc**.

Para concluir el trabajo se estructura en las siguientes partes: Primero se hace una descripción detallada del conjunto de datos con el que se va a trabajar. En segundo lugar se plantea un análisis sumario de los trabajos que están desarrollando actualmente en este mismo área otros grupos de investigación. El tercer capítulo se dedica a estudiar las posibilidades de adaptar el software generado por estos proyectos para nuestros objetivos. Una vez concluido que ninguno de los programas se adapta totalmente a los objetivos de este trabajo, se pasa a presentar una nueva propuesta: CitEc. Se describe su estructura general, las base de datos subyacente para la gestión de la información y los módulos utilizados para llevar a cabo las tres funciones básicas del enlace de referencias: recolección de documentos, parsing de las referencias y enlace con los documentos citados. Finalmente se presentan una serie de conclusiones y de trabajo pendiente que será abordado en futuras investigaciones.

Capítulo 2

Descripción de RePEc

RePEc (**Research Papers in Economics**) (`http://www.repec.org`) es el conjunto de documentos electrónicos seleccionado para desarrollar nuestro trabajo de enlace de referencias, por ello se debe en primer lugar definir qué es RePEc, cuál es su historia, qué tipo de documentos contiene, cuántos son, quienes los utilizan o cómo se puede acceder a ellos.

2.1 ¿Qué es RePEc?

RePEc es un conjunto de herramientas conceptuales, protocolos, normas y software cuyo objeto es la distribución electrónica y descripción bibliográfica de los documentos científicos producidos en el seno de una disciplina académica, en concreto la Economía.

2.2 ¿Cómo nació RePEc?

Los comienzos de RePEc se remontan a 1993. Sus orígenes se encuentran en el proyecto WoPEc (Working Papers in Economics) financiado por eLib (Electronic Libraries Programm) en el Reino Unido. Este proyecto tenía como objeto inicial la creación de una base de datos de prepublicaciones o documentos de trabajo en formato electrónico accesibles gratuitamente en Internet. Todo ello con la intención de facilitar la comunicación entre científicos por medios gratuitos, alternativos a los canales comerciales tradicionales.

El aumento espectacular en el número de prepublicaciones electrónicas que se produjo en los años 1995-96 puso de manifiesto la dificultad de elaborar tal base de datos desde una perspectiva centralizada: una única institución recopilando

información, en la línea de las bibliotecas tradicionales. Así se comenzó a buscar la colaboración de aquellos departamentos que producían working papers, solicitándoles que remitieran a WoPEc las referencias de los nuevos documentos que fueran publicando. Una persona se encargaba entonces de recibirlos, corregirlos en insertarlos en el servidor. Varios departamentos respondieron a esta llamada al constatar las ventajas publicitarias que les aportaba el figurar en una base de datos que por entonces ya era ampliamente conocida en su disciplina.

No obstante el problema no estaba resuelto: ¿Qué pasaría cuando la financiación de eLib acabara? Si no se disponía de fondos para pagar el mantenimiento que este sistema implicaba, el proyecto no podría continuar. Aquí se planteaba una doble alternativa, mientras la magnitud que estaba alcanzando el proyecto hacía imposible su continuidad sin financiación, el objetivo incuestionable era permanecer al margen del mercado comercial, proporcionando un producto gratuito. Por tanto había que buscar soluciones sostenibles de futuro. Se concluyó que dichas soluciones pasaban por la descentralización, es decir, distribuir los costes del sistema entre el mayor número posible de miembros. Como la propia filosofía de Internet, era necesario que el sistema continuara funcionando en el futuro, aunque alguno de sus elementos desapareciera. Para ello se pensó en una descentralización total, en la cual el ideal sería que cada departamento catalogara sus propios documentos para que después esas descripciones bibliográficas se pudieran intercambiar con el resto de departamentos a través de la red. Se generaría así una masa de datos bibliográficos de dominio público que cada departamento o institución podría utilizar en la forma que quisiera. Además no habría ningún *big brother* que controlara el sistema ya que cada departamento sería dueño exclusivamente de sus propios datos pudiendo modificarlos o borrarlos en cualquier momento.

Con la intención de discutir esta idea, en Mayo de 1997 WoPEc convocó una reunión de colaboradores en Guildford (Reino Unido) a la que asistieron representantes de Suecia y los Paises Bajos. Esta reunión es clave, ya que en ella nació el concepto de RePEc y se sentaron las bases del sistema. Allí se perfilaron las dos normas que forman el centro de RePEc: el **Protocolo de Guildford** (11) el cual establece las normas de colaboración entre las instituciones participantes y **ReDIF** (Research Documents Information Format) (12) que establece un formato para la descripción bibliográfica de documentos.

Desde entonces RePEc ha crecido de forma espectacular hasta alcanzar actualmente casi 200 departamentos colaboradores. Entre ellos están los principales productores de working papers a nivel internacional, el NBER (National Bureau of Economic Research) de Estados Unidos, así como importantes instituciones a nivel mundial en el ámbito de la Economía como son los Bancos Federales de Estados Unidos (FedinPrint) o el CEPR (Center for Economic Policy Research).

2.3 ¿Cuál es el objetivo de RePEc?

El objetivo básico de RePEc es crear un sistema de intercambio de información bibliográfica y documentos electrónicos entre las instituciones académicas, con el fin de facilitar el acceso por medios electrónicos a los últimos resultados de investigación en las diferentes áreas de la Economía.

Si bien el objeto principal de RePEc son los documentos de trabajo, pues es en esta tipología documental donde primero se recogen los resultados de investigación, con el tiempo ha evolucionado para englobar también artículos aparecidos en revistas e incluso software de distribución gratuita.

Mientras que la distribución de información bibliográfica debe ser siempre gratuita, en el caso de los textos completos de los documentos, los departamentos o instituciones editoras pueden optar entre distribuirlos gratuitamente o no. En el último caso se encuentran, por ejemplo, los artículos publicados en revistas de la editorial Springer. Aquí la información bibliográfica, hasta donde lo permiten las limitaciones de copyright, es libre, pero el acceso al texto completo está restringido a subscriptores del servicio LINK.

2.4 Introducción del modelo RePEc

El funcionamiento básico de RePEc se estructura en torno a dos normas, una que establece los principios y reglas de colaboración entre los participantes en el sistema, llamada Protocolo de Guildford y otra, denominada ReDIF, que define el formato para la descripción bibliográfica de los documentos electrónicos. Aunque ambas normas fueron creadas desde la perspectiva de la Economía, son independientes de la disciplina y pueden ser fácilmente adaptadas a otros campos. De hecho existe un proyecto basado en RePEc llamado ReLIS (Research in Library and Information Science) accesible en `http://dois.mimas.ac.uk` Igualmente, aunque el objeto inicial de RePEc es la distribución de documentos, el esquema puede servir también para la distribución de otro tipo de información, como datos sobre investigadores, instituciones, etc.

ReDIF es un formato de datos simple, basado en la estructura "campo: valor", e inspirado en los *templates IAFA* (Internet Anonymous FTP Archives) . En este sentido, RePEc no sigue un estándar existente o emergente en cuanto a las descripciones bibliográficas, sino que se decidió crear un nuevo formato por razones que se verán más adelante. La ventaja de esta decisión es que el formato ReDIF es mucho más flexible que cualquiera de los existente y puede ser fácilmente modificado a medida que vayan surgiendo nuevas necesidades. Por otro lado es lo suficientemente simple como para que pueda ser utilizado por personal no especializado, por ejemplo, personal de administración de los

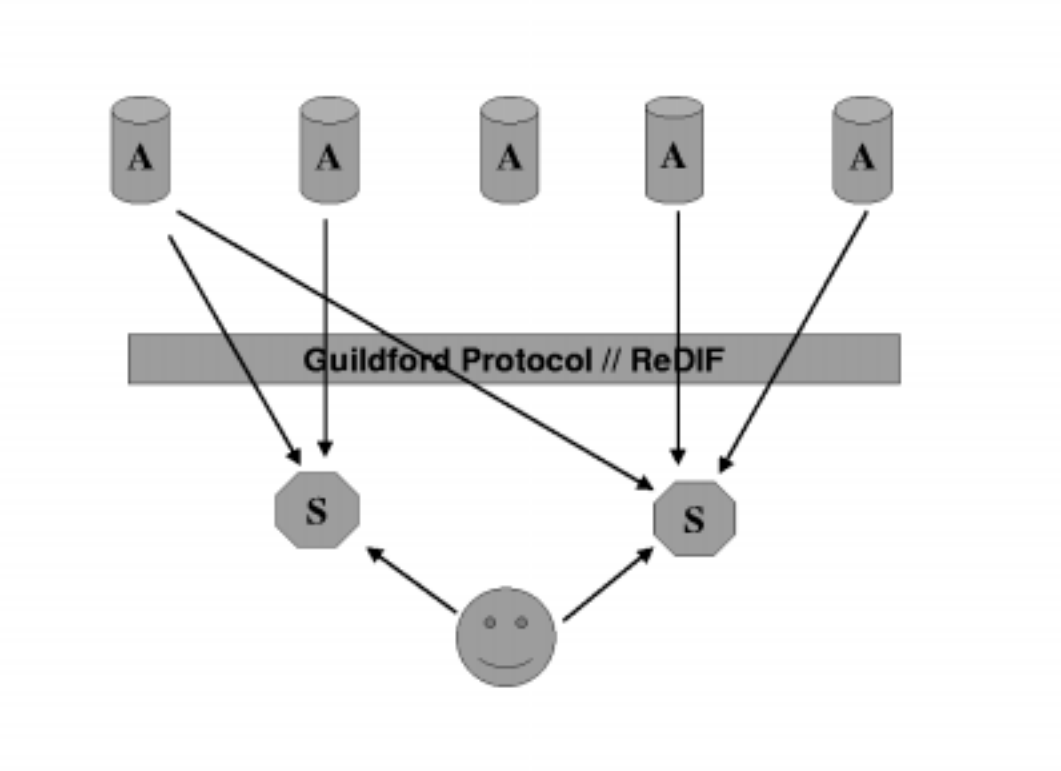


Figura 2.1: Modelo RePEc

departamentos que en muchas ocasiones son los encargados de actualizar las listas de documentos publicados.

El Protocolo de Guildford, cuyo nombre hace referencia a la ciudad donde fue creado, proporciona un conjunto de reglas para la publicación e intercambio de documentos y metadados en la red. Podría ser implementado por cualquier grupo o individual que deseara intercambiar documentos en Internet. Establece dos niveles para la participación de los departamentos, el de archivo, de carácter pasivo, simplemente proporciona información, encargados de crear y almacenar la información bibliográfica y en algunos casos los documentos propiamente dichos. El de servicio, activo, ya que toma la información de los anteriores y construye una aplicación de utilidad para los usuarios finales. Finalmente, los propios usuarios quienes acceden al sistema a través de los servicios con ayuda de un navegador web o un cliente de correo, etc. dependiendo de cada caso.

Gráficamente este modelo está representado en la figura 2.1.

2.4.1 ¿Qué son los archivos?

Archivos son aquellos departamentos o centros de investigación que proporcionan información bibliográfica sobre los documentos que publican. Toda esta información es de dominio público y está accesible gratuitamente en la red. Puede ser copiada y/o distribuida para propósitos de investigación por cualquier persona, relacionada o no con RePEc. El texto completo de los documentos puede ser también distribuido junto con las referencias si el departamento así lo desea.

Desde un punto de vista técnico, un archivo es simplemente una estructura de directorios y subdirectorios en un servidor FTP o HTTP, donde se almacenan ficheros en formato ASCII conteniendo los datos bibliográficos y en ocasiones ficheros PDF, PS, etc. conteniendo el texto completo de los mismos. Esta estructura debe seguir unas reglas fijas establecidas en el Protocolo de Guildford, con objeto de permitir a los robots el acceso a los datos.

2.4.2 ¿Quien pone orden?

Aunque el objetivo de RePEc es mantener una descentralización máxima, es obvia también la necesidad de una cierta intervención que garantice la continuidad del sistema. Así, de la coordinación se encarga un archivo central, denominado **core site**. Este archivo está accesible en la dirección `ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all`

Sus funciones son:

- Mantener actualizada la documentación del sistema. Los dos principales documentos son las definiciones del Protocolo de Guildford y la especificación ReDIF. Además se encarga de mantener las paginas del servidor web donde se ofrecen guias con información para nuevos miembros, documentación del software distribuido por el sistema, etc.
- Regulación de la toma de decisiones entre los miembros del grupo. La comunicación entre los distintos participantes se realiza a través de una lista de discusión: RePEc-Admin@jiscmail.ac.uk
- Incluir y en su caso retirar miembros (archivos) del sistema.
- Asignar códigos de identificación a cada nuevo archivo. A cada departamento que desea colaborar con RePEc se le asigna un código único compuesto de tres letras. El asunto de los códigos es clave para el desarrollo de RePEc y sobre ellos se volverá más adelante.
- Distribución de software para el funcionamiento de los archivos y servicios. Básicamente este software incluye programas para controlar la sintaxis de

las descripciones bibliográficas, para realizar la copia o mirror de información desde los archivos a los servicios y para crear presentaciones a través del web utilizando los datos almacenados en los archivos. Todo el software es mantenido y actualizado por el archivo central, de forma que cualquier departamento participante, aunque carezca de personal informático, pueda crear un archivo o servicio. El lenguaje de programación utilizado es **perl**. Los programas han sido diseñados inicialmente para funcionar en máquinas UNIX, aunque existe una preocupación creciente por crear versiones en otros sistemas operativos como por ejemplo Windows.

2.4.3 ¿Qué son los servicios?

La información ofrecida por los archivos o departamentos en formato ReDIF es de poca o ninguna utilidad para los usuarios finales tal cual está en los servidores. Así, es necesario que algunos departamentos, que harían las funciones de intermediarios, tomen la información y le incorporen un determinado valor añadido para hacerla accesible a los usuarios. Esta es la finalidad de los servicios. Ese valor añadido puede adoptar distintas formas: la más simple sería convertir las descripciones bibliográficas del formato ReDIF original a un formato legible por una persona, como por ejemplo páginas HTML. También puede consistir en crear un índice de los datos que permita realizar búsquedas, etc. O realizar una selección, comentario y presentación de determinados documentos existentes en el sistema. En general cualquier archivo puede crear su propio servicio utilizando datos propios y/o procedentes del resto de archivos. Así nos podemos encontrar con departamentos que ofrecen archivos, el caso más frecuente; departamentos que ofrecen servicios, y departamentos que ofrecen ambas cosas a la vez.

En resumen, nos encontraríamos ante una colaboración descentralizada, con multitud de servidores de información, que comparten sus datos con todos los demás servidores, de forma que cualquier miembro del grupo puede hacer uso de ellos para crear un servicio final al usuario.

Como ejemplos de servicios tenemos:

- WoPEc y BibEc accesibles en las direcciones `http://netec.mcc.ac.uk/WoPEc.html` y `http://netec.mcc.ac.uk/BibEc.html` son los servicios más antiguos. Son dos bases de datos que ofrecen información sobre documentos de trabajo en formato electrónico e impreso respectivamente. Ambos pueden utilizarse de dos formas: o bien ojeando las paginas html estáticas que contienen las distintas series de los archivos participantes y una clasificación JEL (Journal of Economic Classification) de los mismos, o bien utilizando alguno de los motores de búsqueda ofrecidos. Actualmente se puede elegir entre dos bases de datos: una utilizando SWISH++

Codigo	Nombre	Editor
nep-cfn	Corporate Finance	Evgenia Stupina
nep-cmp	Computational Economics	Nienke Oomes
nep-dcm	Discrete Choice Models	Josep M Camacho Cabiscol
nep-dge	Dynamic General Equilibrium	Christian Zimmermann
nep-ecm	Econometrics	Sune Karlsson
nep-ets	Econometrics-time series	Yong Yin
nep-eff	Efficiency and Productivity	Vania Sena
nep-env	Environmental Economics	Francisco S. Ramos
nep-eec	European Economics	Marco Catenaro
nep-evo	Evolutionary Economics	Murat Yildizoglu
nep-exp	Experimental Economics	Reinhard Sippel

Figura 2.2: Ejemplo de listas NEP y sus editores

y otra utilizando un sistema SQL (MySQL). En cuanto a número de documentos, WoPEc ofrece información sobre más de 50.000 working papers y artículos en texto completo y BibEc proporciona datos bibliográficos sobre más de 85.000 documentos impresos, con información detallada sobre las formas de obtener copias de los mismos.

- NEP son las siglas de New Electronic Papers. Es un servicio de alerta a través de listas de distribución, donde se anuncian semanalmente las novedades aparecidas en RePEc. Se compone de más de 30 listas temáticas de novedades, especializadas en áreas concretas de la Economía, como las que aparecen en la figura 2.2.

  Cada una de ellas está moderada por un investigador de prestigio en el tema tratado. El funcionamiento de NEP es el siguiente: semanalmente se extraen las novedades del sistema y se envían a los moderadores de cada lista. Por el momento las novedades sólo incluyen working papers cuyo texto completo esté accesible electrónicamente. Los editores se encargan de seleccionar los documentos que se ajustan al ámbito temático de su lista y distribuirlos entre los subscriptores. Para más información sobre NEP pueden consultarse las páginas del proyecto en la dirección `http://nep.repec.org`.

- HoPEc. Este es un ejemplo de mezcla de servicio y archivo al mismo tiempo. HoPEc permite a los autores de documentos registrarse como tales y asociar su nombre a las obras que han escrito y que se encuentran en RePEc. Además se ofrece información contrastada sobre cada uno de ellos como por ejemplo un nombre normalizado, un lugar de trabajo, direcciones electrónicas, etc. Es responsabilidad de cada autor mantener sus datos al día a través de un formulario web que se puede consultar en la página web: `http://hopec.repec.org` Actualmente ya son más de dos mil los autores que se han registrado de esta forma.

  Los problemas que viene a solucionar HoPEc son varios. En primer lugar

la estandarización de los nombres. En cualquier catálogo o base de datos se puede comprobar cómo una misma persona se encuentra con diferentes formas de su nombre lo cual acarrea pérdidas de información. En segundo lugar permite descargar los *templates* de documentos de información relativa a los autores. Es frecuente ofrecer por cada documento la dirección de trabajo del autor, su email, etc. Ahora bien, esta es una información estática, se crea cuando se redacta el registro y es muy difícil que se actualice en el futuro. Por ello cuando el autor cambiaba de institución esos datos se vuelven obsoletos y dejan de cumplir su función. HoPEc intenta responder a ese problema. En la idea de la base de datos relacional que intenta representar RePEc, HoPEc ocupa la parte del autor que se enlaza con el resto de datos a través del handle de cada autor.

Los datos generados por HoPEc se distribuyen a través de un archivo especial que solamente contiene *templates* de personas para que puedan ser utilizados por el resto de servicios.

2.5 ¿Cómo se regulan las relaciones entre archivos y servicios?

Como ya se dijo anteriormente, a través de dos normas:

2.5.1 El Protocolo de Guildford

Establece una serie de normas y recomendaciones que deben seguir todos los archivos que forman parte del sistema. El no seguimiento de las mismas haría imposible la comunicación entre los elementos del sistema.

Si se define un archivo como una estructura definida de directorios y subdirectorios accesibles en un servidor FTP o HTTP, lógicamente la parte más importante del protocolo es aquella donde se describe tal estructura. A grandes rasgos el protocolo establece que: todo archivo debe ser identificado por un código único (archivo_id) asignado por el **core site** y compuesto por tres letras. Este código constituye el elemento raíz de la estructura de directorios. Los archivos a su vez identificarán las distintas colecciones o series de documentos que publiquen con un código de seis letras (series_id). Cada serie dispondrá de un subdirectorio donde se almacenarán las descripciones bibliográficas. Dentro del archivo todos los ficheros con extensión .rdf contendrán datos en formato ReDIF.

Veamos un ejemplo. El National Bureau of Economic Research (NBER) proporciona un archivo RePEc, con código **nbr**, en la dirección `http://nberws`.

`nber.org/RePEc/nbr`. En este archivo se ofrece información sobre la colección "NBER working papers", código **nberwo** en el subdirectorio `http://nberws.nber.org/RePEc/nbr/nberwo` Este subdirectorio contiene una serie de ficheros ASCII con datos bibliográficos sobre los documentos.

2.5.2 ReDIF

La primera pregunta que se nos presenta al hablar de ReDIF es ¿por qué un nuevo formato de descripción de documentos? ¿por qué no utilizar alguno de los ya existentes como por ejemplo el MARC o el Dublin Core?

En primer lugar por la sencillez. Normalmente las descripciones bibliográficas van a ser creadas por personal no especializado, personal de administración, por ello debe ser un formato lo más fácil de utilizar posible. De esta forma se descartarían formatos como el MARC, difícil de entender incluso por bibliotecarios.

En segundo lugar, debe ser un formato interdisciplinar, que no sea "propiedad"de un determinado grupo de usuarios. Es el caso del BibTeX entre la comunidad de físicos y matemáticos.

En tercer lugar, debe ser un formato fácilmente actualizable para adaptarlo a nuevas necesidades. Debe ser algo dinámico, donde no haya que esperar la decisión de varios comités para incluir una modificación.

Finalmente existen también razones históricas. Como ya se ha dicho RePEc surgió a partir de WoPEc. WoPEc desde el año 1995 venía utilizando el formato IAFA (Internet Anonymous FTP Archive) de la IETF (Internet Engineering Task Force). Este formato pensado inicialmente para la descripción de recursos accesibles a través de servidores FTP contó con un amplio apoyo de la comunidad Internet durante los años 1995 a 1997, cuando comenzó a cobrar fuerza el Dublin Core. Al no encontrar entre los formatos de descripción de recursos existentes alguno que se adaptara a las necesidades mencionadas se decidió construir uno propio que seguirá las lineas básicas del formato IAFA. De hecho ReDIF es una evolución o interpretación de este y en gran parte compatible con él. Como él se basa en una sintaxis sumamente sencilla de: "campo: contenido del campo".

En un mismo fichero pueden almacenarse varios registros del mismo tipo. Cada registro debe ocupar un párrafo de texto, es decir, no están permitidas lineas en blanco dentro de los registros. Debe comenzar con un campo obligatorio: **Template-Type:** . El contenido define el tipo de documento que se va a describir y por consiguiente los campos aplicables a dicho tipo de documento. El resto de los campos no importa el orden en que aparezcan. Ahora bien, ReDIF utiliza el concepto de clusters, por lo que los datos referidos a un mismo

objeto deben aparecer juntos en la descripción. Un cluster se puede explicar de la siguiente forma: existen una serie de datos que son comunes en varios casos, por ejemplo, cada persona mencionada en un registro puede tener un nombre, un número de teléfono, una dirección postal, etc. De igual forma, cada organización mencionada tendrá un nombre, un número de teléfono, una dirección postal, etc. A ese conjunto de elementos que son comunes a la persona y a la organización es a lo que se denomina cluster. Así un cluster sería un conjunto de campos enlazados lógicamente en un registro y que se refieren a un objeto determinado. Todos esos campos deben aparecer seguidos dentro del registro. ReDIF diferencia tres clusters: "person", "organization" y "file". Seguidamente se presenta un ejemplo del cluster persona:

```
Author-Name: Joe Smith
Author-Email: JoeSmith@some.uni.edu
Author-Postal: PO Box 123, Smith Street,
 Somewhere In The Universe, 987654
Author-Phone: +99 456-321123
Author-Homepage: http://www.some.where.edu/~JoeSmith
```

El registro se completa además con campos como por ejemplo, title, abstract, creation-date, classification, keywords, handle. Mención especial merece este último ya que es un campo obligatorio cuyo contenido es un código que identifica unívocamente a cada objeto descrito en RePEc. Esta es la característica más importante del formato ReDIF, ya que una simple cadena de caracteres nos proporciona información sobre el archivo, la serie y el número de un objeto determinado. Por otro lado a través del *handle* se puede crear una estructura relacional que enlace los distintos elementos que forman la descripción de un documento. El handle se compone de: la cadena "RePEc" seguida de ":", el código del archivo al que pertenece el objeto, ":", el código de la serie donde se ha publicado, ":",un número de orden asignado por el archivo. Así tendríamos:

RePEc:bon:bonnsf:a452
RePEc:die:calsdi:9338

Finalmente, un registro ReDIF completo quedaría como sigue:

```
Template-Type: ReDIF-Paper 1.0
Title: A Discussion of the Reliability of Results Obtained with
 Long-Run Identifying Restrictions
Author-Name: St-Amant, P.
Author-Email: St-Amant@uv.es
Keywords: CENTRAL BANKS ; MONETARY POLICY ; MACROECONOMICS
Length: 14 pages
Classification-JEL: C1 ; E27
Creation-Date: 1998
```

```
File-URL: http://www.nber.org/papers/w6490.pdf
File-Format: application/pdf
File-Restriction: Access to the full text is restricted. Look
 up http://www.nber.org/wwpfaq.html for details, or write to
 feenberg@nber.org. If you have no access to the full text you will be
 shown an abstract page instead. Anyone browsing the
 NBER working paper database from a site with a TLD in a non-OECD, non-OPEC
 country will be offered full text downloads for any paper for which the
 full text is available online, but only if their DNS does reverse name
 lookup.
Handle: RePEc:nbr:nberwo:6490
Price: \$5.00 per paper (plus \$10.00 postage for orders outside U.S.)
```

Este registro ilustra cómo la colección no está necesariamente limitada a documentos gratuitos. Para aquellos documentos de pago ReDIF permite especificar en texto libre las condiciones de acceso a los mismos. Para esta finalidad se dispone del campo **File-Restriction**.

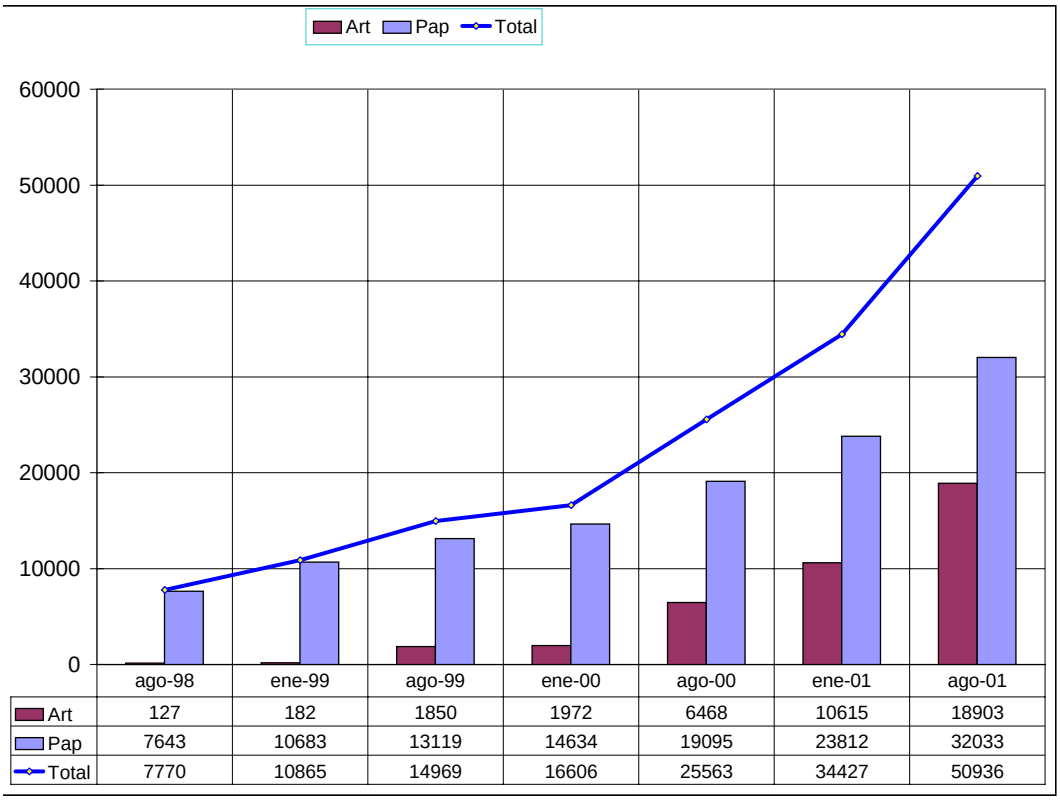
2.6 ¿Cuántos documentos hay en RePEc y cuánta gente los utiliza?

Para acabar esta introdución en la que se ha intentado hacer un cuadro del funcionamiento de RePEc, se incluye la evolución que ha tenido esta en los últimos años, tanto en lo que se refiere a número de documentos como a número de usuarios.

En las figuras 2.3 y 2.4 aparece la evolución en el número de documentos a texto completo disponibles en RePEc desde agosto de 1998 hasta la fecha. No se ofrece aquí datos sobre aquellos documentos de los que solamente existe información bibliográfica.

El número tanto de artículos como de working papers ha ido creciendo de forma lineal durante todo este periodo, pasando de 7643 en agosto de 1998 a 32000 en el mismo mes de 2001.

Ahora bien el crecimiento de ambos tipos ha sido muy desigual, el número de artículos ha crecido mucho más rápidamente en los últimos meses que el número de papers. En la figura 2.4 se puede apreciar como el porcentaje de papers respeto del conjunto de documentos ha caído de forma sustancial a partir de enero de 2000 en beneficio del de artículos. Esto se debe a una política de colaboración de RePEc con algunas editoriales como por ejemplo la American Economic Association que ha hecho que estas distribuyan la información sobre sus publicaciones a través de RePEc. Desgraciadamente muchos de estos documentos solo estarán disponibles a través de subscripción, no de forma gratuita.



	ago-98	ene-99	ago-99	ene-00	ago-00	ene-01	ago-01
Art	127	182	1850	1972	6468	10615	18903
Pap	7643	10683	13119	14634	19095	23812	32033
Total	7770	10865	14969	16606	25563	34427	50936

Figura 2.3: Número de documentos en RePEc

Esto lo se comprobará más adelante ya que si no se puede acceder a ellos no se podrá llevar a cabo el enlace de sus referencias.

En la última figura se ve como ha crecido el uso de uno de los servicios de RePEc: WoPEc. Dada la descentralización y diversidad de los servicios es muy difícil cuantificar el total de uso de RePEc si no se hace de forma fragmentada como en nuestro caso. En los datos que se ofrecen se reflejan las conexiones registradas por los logs de los servidores web de las tres mirrors de que consta WoPEc: Reino Unido, Estados Unidos y Japón.

El uso de WoPEc se ha multiplicado por diez en estos cuatro años, pasando de 60.000 consultas al mes en enero de 1997 a casi 700.000 en enero de 2001. El crecimiento ha sido prácticamente constante durante todo el periodo, excepto una pequeña caída durante la segunda parte de 1998, con una tendencia alcista que se ve mucho más marcada a partir de febrero de 1999.

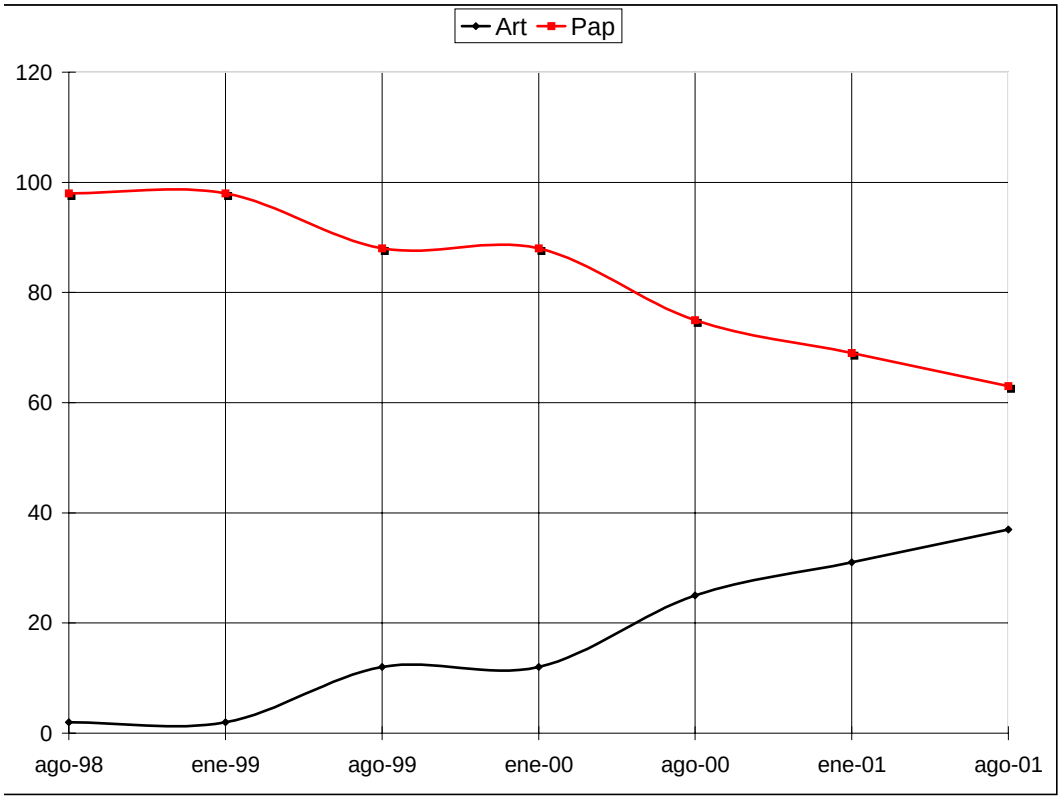


Figura 2.4: Porcentaje de artículos frente a papers

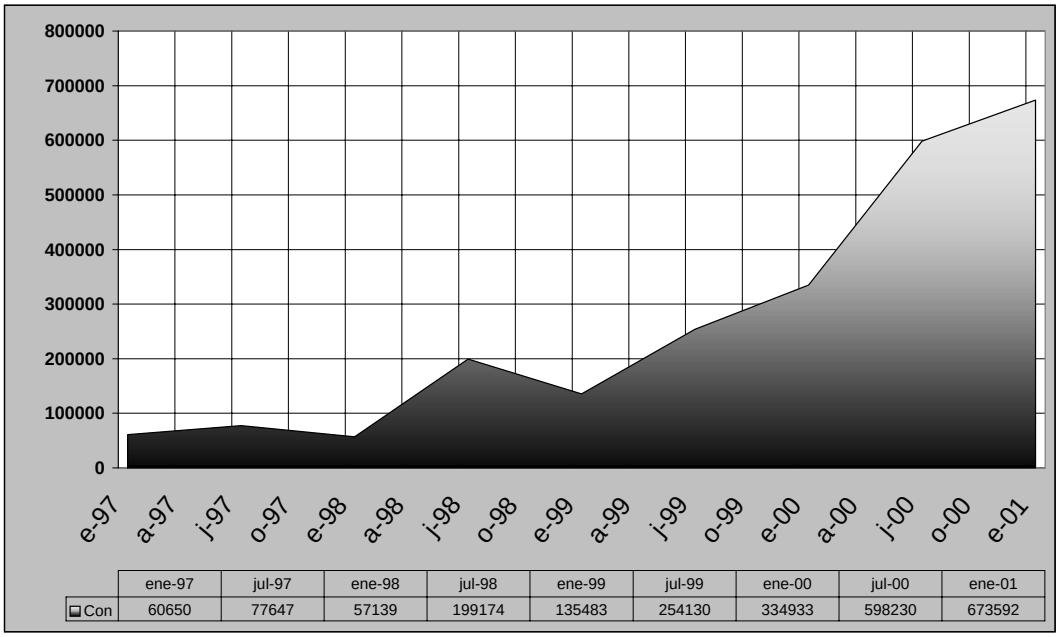


Figura 2.5: Conexiones a WoPEc

Capítulo 3

Trabajos relacionados

Como ya vimos en la introducción, en los últimos años ha crecido el interés por estudiar el proceso de *reference linking* o, lo que es lo mismo, la creación de enlaces entre las referencias de los documentos disponibles en formato electrónico. Las iniciativas que están liderando estos estudios provienen de dos áreas: la industria editorial y las comunidades de archivos electrónicos.

En el primer caso las editoriales comerciales se han dado cuenta del enorme potencial y valor añadido que pueden incluir en sus productos si incluyen enlaces entre los artículos que están ofreciendo en versión electrónica. El mayor exponente de este área de trabajo es el proyecto CrossRef.

Por otro lado están las comunidades académicas que han desarrollado archivos de informes técnicos, documentos de trabajo e incluso artículos publicados en revistas (RePEc en Economía, CogPrints en Ciencias del Conocimiento o ArXiv y CERN en Física). También ellas han visto las ventajas de enlazar, en la medida de lo posible, los documentos de que disponen.

La investigación en este área es un tema de gran actualidad como lo demuestra el interés que están depositando en ella instituciones del prestigio de la universidad de Cornell, la Universidad de Southamtom en el Reino Unido que está experimentando en la creación de enlaces en el archivo ArXiv, la empresa NEC donde se ha desarrollado **ResearchIndex** un auténtico índice de citas para informática pero que podría ser aplicable a cualquier otra disciplina o el CERN (Centre European pour la Recherche Nuclear) que está enlazando sus propios informes de investigación.

En la presente sección se plantea un breve análisis de estos proyectos con el objeto de situar en contexto el trabajo que se describe en este documento.

3.1 ResearchIndex

ResearchIndex ha sido desarrollado en el departamento de investigación de la empresa NEC por Steve Lawrence, Kurt Bollacker y C. Lee Giles. Es el punto de referencia obligado para cualquier trabajo sobre identificación de citas en documentos electrónicos.

Aunque ha sido pensado para trabajar con documentos en el área de informática podría ser adaptado a cualquier disciplina, como se intentará demostrar a lo largo de este trabajo. Se puede ver **ResearchIndex** en funcionamiento en la página `http://csindex.org`. Se trata de una base de datos con más de 200.000 documentos indizados y más de dos millones de referencias.

ResearchIndex no es exactamente un software para enlazar referencias sino que tiene un objetivo mucho más amplio , es un verdadero índice de citas construido de forma automática. Es lo que los autores denominan un "autonomous citation index". Se menciona aquí porque el paso previo tanto para la elaboración del índice como para el enlace de referencias es necesariamente la extracción de las referencias de los documentos, así como la identificación de cada uno de los elementos de que constan. En este apartado de análisis ResearchIndex hace un extraordinario trabajo.

Entre las muchas funciones que este software aporta se podrían destacar las siguientes (14):

- Localización de documentos científicos en la web. Para ello actúa como un metabuscador utilizando múltiples buscadores como Altavista a los cuales envía palabras clave junto con términos específicos para limitar el ámbito de la búsqueda a documentos de carácter científico (por ejem. papers, conference, proceedings, etc.)

- Indización del texto completo de los documentos encontrados en formato PDF,PostScript o HTML.

- Extracción de información bibliográfica de los documentos. Incluye algoritmos y técnicas de aprendizaje para detectar automáticamente datos como el título y los autores de los documentos que indiza.

- Identificación de las referencias dentro de los artículos. Identifica la sección que contiene la bibliografía y es capaz de aislar cada una de las referencias. Además es capaz de determinar si dos referencias con formatos diferentes se refieren al mismo documento.

- Identificación de los distintos datos que contiene una referencia. Entre otros: año de publicación, título y autores.

- Identificación del contexto donde se ha producido la cita. Utiliza la clave que encabeza cada referencia para buscar en el cuerpo del documento su equivalente y de esta forma aislar la frase utilizadas por el autor en la cita.
- Identificación de documentos relacionados utilizando la información sobre citas disponible. Ante cada documento el sistema sugiere al usuario una serie de documentos alternativos en función del número de referencias que tienen en común, de los documentos que lo han citado, etc.
- Es capaz de analizar las redes de literatura científica con objeto de identificar cuales son las autoridades en una materia, entendiendo por tales aquellos documentos que reciben mayor número de citas. Igualmente es capaz de determinar las revisiones, que serían aquellos trabajos que contienen un elevado número de referencias.
- Puede identificar las autocitas, comparando el autor del documento con los autores de las referencias.
- Permite a los usuarios corregir errores en la base de datos, convirtiéndose así en un sistema interactivo.
- Tiene la posibilidad de mantener un perfil de los usuarios, de tal forma que puede recomendar, bien por correo electrónico o a través del propio web, nuevos documentos que se ajusten a dichos perfiles. Estos perfiles pueden ser actualizados directamente por el usuario o bien por el propio sistema mediante técnicas de inteligencia artificial.

En el apartado técnico se trata de un conjunto de programas escritos en **perl** y **C++** que funciona en máquinas UNIX. El código fuente se distribuye gratuitamente para fines no comerciales. Es necesario compilarlo una vez que se hayan instalado en el sistema una larga lista de módulos **perl** y programas requeridos. Todos ellos son también de distribución gratuita.

3.2 OpCit: Open Citation Project

OpCit es un proyecto financiado conjuntamente por el Join Information Systems Comittee del Reino Unido y la National Science Foundation de Estados Unidos. El centro inglés participante es el Intelligence, Agents, Multimedia Research Group de la Universidad de Southampton, mientras el participante americano es el Cornell Digital Libraries Research Group de la Universidad de Cornell. Esto hace que ambos proyectos colaboren estrechamente en el desarrollo de software y servicios. Mientras el primero está centrado en colecciones de documentos con unas características bastante regulares, en concreto

trabajan con un sólo archivo de documentos: arXiv, Cornell se está especializando en colecciones más pequeñas pero también mucho más irregulares, con varios estilos de referencias, con documentos formateados sin reglas establecidas y principalmente en formato HTML. (5). En sus propias palabras intentan aportar un trabajo de enlace de referencias a la parte científica y académica del web.

3.2.1 Cornell Digital Libraries Research Group

El trabajo desarrollado por Cornell está descrito en informes como (5; 6; 2). Según éstos intentan crear una arquitectura para el enlace de referencias que se compondría de dos niveles: el *nivel de enlace de referencias* sobre la web, que se encargaría de proporcionar suficientes datos para una variedad de servicios de valor añadido, o como ellos los definen *aplicaciones de enlaces de referencias*. Un ejemplo de tales aplicaciones sería la creación de interfaces de usuario para navegar por la red de referencias.

El nivel de enlace de referencias debe, dada una determinada referencia:

1. Buscar su correspondiente clave en el texto del documento.
2. Hacer un análisis de la referencia para determinar qué obra es, si el documento que representa se encuentra en formato electrónico, si es posible establecer un enlace con ella, etc.
3. Proporcionar acceso a todos estos datos para que puedan ser usados por las aplicaciones.

Por su parte las aplicaciones de enlaces de referencias deben:

1. Convertir la clave de la referencia en el texto del documento en un enlace que apunte al texto completo. Por ejemplo en HTML o PDF convirtiéndola en un ancla hipertextual

Por el momento el trabajo del grupo se ha centrado en el primer apartado, lo que sería las fases de análisis, acceso y distribución de los datos.

La aplicación práctica de esta base teórica la han llevado a cabo con la revista **D-Lib**[1]. Esta es una revista electrónica que nació en 1995. Está especializada en el área de bibliotecas digitales y tiene una periodicidad mensual.

Su objetivo era alcanzar un 80% de precisión en el proceso de extracción de información. Este es el mínimo aceptable para que se puedan crear servicios

[1] `http://www.dlib.org`

de valor añadido con los datos extraídos. A la hora de evaluar los resultados han creado dos indicadores en función de los dos tipos de errores posibles en el análisis de los documentos. El primero sería un error en la extracción de la información bibliográfica del documento que está siendo analizado. El segundo serían los errores en el análisis de las cadenas de referencias que contiene el documento. De esta forma tenemos como primer indicador el **item accuracy** que es el número de elementos analizados correctamente, dividido por el número total de elementos en el item. Entre los elementos que se intentan identificar están: el título del documento, cada uno de los autores, el año de publicación y los contextos de las referencias. El segundo indicador es el **reference accuracy**. Es el porcentaje de los elementos de una referencia que son analizados correctamente. Estos elementos incluyen: título, cada autor, año, contextos y URL si existe.

Los resultados que han obtenido para estos indicadores se acercan bastante a los objetivos previstos, con un 75% en el caso del **item accuracy** y un 70% de **reference accuracy** sobre un conjunto representativo de referencias.

3.2.2 Intelligence, Agents, Multimedia Research Group

La parte inglesa de OpCit está trabajando en establecer enlaces entre los documentos disponibles en el archivo arXiv. ArXiv es el archivo de prepublicaciones que existe en Internet más antiguo. Fuel diseñado por Paul Ginsparg en Los Alamos National Laboratory a comienzos de los años 90. En la actualidad almacena casi la mitad de la literatura que se genera actualmente en Física de Altas Energías. Almacena más de 150.000 documentos a texto completo que pueden ser descargados gratuitamente desde cualquiera de los más de 15 mirrors que existen por todo el mundo. Una idea de la importancia del archivo la da el hecho de que más de 35.000 personas lo consultan diariamente.

El trabajo de OpCit esta consistiendo en enlazar internamente las referencias de los documentos depositados en arXiv. En este caso, el disponer de los documentos originales, da la posibilidad de manipularlos para insertar determinada información que luego servirá para crear enlaces. El trabajo no se ha limitado solamente a las citas sino que se ha planteado una reforma global del sistema, desde que el autor remite el documento hasta la forma de acceder al servicio por parte de los usuarios. Así la identificación de las citas sería un elemento más del proceso.

En arXiv los autores remiten sus documentos al archivo. Una vez recibidos los ficheros son convertidos a varios formatos (PDF, PostScript, DVI, etc.). En la identificación de las citas comenzaron trabajando sobre los documentos en PDF, pero dado que el formato de remisión es casi unánimemente el TeX, OpCit se ha beneficiado de ello para hacer un análisis de este formato en lugar

del PDF. Así evitan los errores de conversión PDF a ASCII que llegaban al 20%.

En estos momentos ya existe una versión interconectada de todo el archivo y se ha diseñado un interfaz que permite navegar por los documentos usando las referencias. Puede verse en `http://arabica.ecs.soton.ac.uk/cgi-bin/search_tj`.

Una vez interconectado todo el archivo, OpCit cuenta con un servicio duplicado de arXiv con nuevas facilidades. No obstante está siendo muy poco utilizado por los usuarios quienes prefieren acudir al sitio tradicional. El reto que se plantea ahora es devolver los datos de OpCit para que puedan ser integrados a su vez en arXiv. Para ello se ha pensado utilizar las facilidades que aporta la OAi (Open Archives Initiative)[2].

Esta es una iniciativa para facilitar el acceso a los archivos de documentos electrónicos existentes en las distintas disciplinas. El objeto de OAi sería crear una inmensa biblioteca digital de material científico y técnico. Cualquier institución puede participar. Para ello simplemente debe poner en marcha un servidor siguiendo un protocolo determinado a través del cual se dará acceso a sus documentos. Esto es lo que ha hecho el grupo con el archivo arXiv. Ahora además pretenden ampliar la información disponible, exportando también los datos sobre citas. Para ello es necesario utilizar un formato de metadatos que no existía en arXiv, donde se ha dado muy poca importancia hasta el momento a la información bibliográfica. El formato que se va a adoptar es el AMF (Academic Metadata Format) descrito en (13).

Una vez que el proyecto dispone de una importante base de datos de citas el siguiente paso que han emprendido es la realización de estudios bibliométricos sobre la disciplina. Están investigando la aplicación de indicadores como el factor de impacto a archivos de documentos y están estudiando las prácticas de los usuarios al manejar el sistema.

Finalmente el otro punto que están investigado es la aplicación de las citas para ordenar los resultados de las búsquedas realizadas sobre el servicio. Es decir, que ante una pregunta los documentos devueltos en primer lugar fueran los que mayor número de veces hayan sido citados.

En cuanto al software que han hecho público se compone de una serie de módulos **perl** que permiten extraer listas de referencias a partir de documentos originales (exclusivamente en formato TeX o LaTeX). Básicamente incluyen:

- **Extract::Reference** inserta una marca al comienzo de cada referencia en un fichero TeX o LaTeX, después lo convierte a DVI y finalmente a ASCII para producir una lista de referencias.

[2] `http://www.openarchives.org`

- **Parser::Citation** extrae los metadatos (autores,revista, volumen, número, etc.) de una referencia en concreto. Tiene la limitación a la hora de aplicarlo a otras disciplinas de que solo analiza referencias en el formato utilizado en Física, donde no aparece el título del documento y el título de la revista aparece abreviado.
- **Parser::arXivCite** es un caso especial del anterior. Procesa los identificadores de documentos utilizados en arXiv que se encuentren en una referencia
- **Parser::arXivRef** procesa múltiples citas dentro de una sola referencia. Esta es una práctica habitual en documentos de Física.

3.3 CERN

En la misma línea de trabajo el CERN Document Server ha anunciado el 1 de Noviembre de 2001 que han llevado a cabo el enlace de todos los documentos que distribuyen, principalmente prepublicaciones en el campo de la Física. En total ello supone más de dos millones de enlaces desde sus documentos a revistas electrónicas u otras prepublicaciones. Una descripción del proyecto apareció en (8).

La versión completamente operativa del sistema puede verse en `http://weblib.cern.ch/`. Por el momento ofrecen información tanto de las referencias de un documento como de las citas que éste haya recibido. Todo ello en un sistema perfectamente integrado con lo que los usuarios venían utilizando hasta ahora.

Al igual que en el caso anterior esta iniciativa se beneficia de la particular forma de construir las referencias que existe en Física. Es un formato muy estandarizado en el que es habitual referenciar solamente los autores y la publicación donde ha aparecido el trabajo en formato abreviado.

En la actualidad el CERN almacena más de 170.000 documentos a texto completo. Todos ellos están en formato PDF. Los autores pueden enviar los trabajos en cualquier formato incluyendo MS Word, LaTeX, etc. Todos los documentos son convertidos a PDF. Esto proporciona una gran uniformidad en el resultado, pues todos los ficheros habrán sido creados con el mismo procedimiento. La extracción de los datos se ha realizado sobre estos ficheros en PDF. Se ha convertido cada documento a formato ASCII y después se ha analizado para extraer las referencias. La tasa de efectividad de este proceso se sitúa en el 91%, incluyendo en los errores tanto documentos que no están en inglés como aquellos que no contienen una sección de referencias.

De los documentos analizados se han extraído casi tres millones de referencias, el 80% de las cuales se han podido analizar correctamente. Por análisis correcto

se entiende que el sistema ha sido capaz de identificar la revista donde ha sido publicado el documento. En 1.937.162 referencias se ha podido establecer un enlace entre la referencia y el documento electrónico al que representa. En este caso se han hecho enlaces a aquellas revistas disponibles en la biblioteca del CERN o a otras prepublicaciones almacenadas en el propio servidor.

Finalmente resaltar que se ha realizado una integración total de los datos extraídos en el registro bibliográfico del documento analizado elaborado por la biblioteca. Para ello se ha utilizado el formato MARC: por cada referencia encontrada se ha añadido un campo 909 al correspondiente registro MARC. La dirección electrónica, si existe, se ha colocado en el subcampo $x. Esta es una característica que la hace diferente del resto de iniciativas vistas.

3.4 CrossRef

Las editoriales comerciales se dieron cuenta desde muy pronto de las ventajas que les aportaría el tener enlazados todos los documentos que publican. Como respuesta a esta necesidad surgió en 1999 el proyecto CrossRef[3] puesto en marcha por la Publishers International Linking Association (PILA). Esta es una empresa que engloba a más de ochenta editores y proveedores de servicios de resúmenes de todo el mundo. No obstante, el primer prototipo del sistema surgió de una iniciativa de Wiley y Academic Press en colaboración con la International DOI Foundation.

El funcionamiento de CrossRef es bastante sencillo. No tiene nada que ver con el resto de iniciativas que hemos visto ya que aquí las referencias ya están disponibles como parte de la edición del texto. Además se dispone de abundante metadatos por lo que no es necesario realizar un análisis del documento para encontrarla. Los editores participantes aportan información bibliográfica sobre los documentos que publican a una base de datos común. Por cada artículo se requiere información básica como título de la revista donde se ha publicado, volumen, número, etc. Pueden aportar además otra información adicional según su elección. Además a cada artículo se le asigna un DOI[4] (Digital Object Identifier) y un URL asociado para recuperar el texto completo de los documentos. El intercambio de información entre editoriales y CrossRef se realiza utilizando el formato XML. Es obligación de cada editorial mantener actualizada la correspondencia entre DOIs y URLs de cada documento.

La base de datos contiene en estos momentos casi cuatro millones de documentos. Cuando una editorial va a publicar un nuevo artículo, debe por un lado remitir sus datos bibliográficos a CrossRef y por otro tomar la lista de refe-

[3] `http://www.crossref.org`
[4] `http://www.doi.org`

rencias e interrogar la base de datos de CrossRef para ver si están disponibles en formato electrónico. Para ello las referencias son enviadas a un *reference resolver* que se encargará de devolver el DOI del documento citado si existe. Este código se incluirá en el artículo publicado. En el futuro cuando alguien quiera acceder al documento citado será el sistema DOI quien se encargue de convertir ese código en la dirección de Internet donde se almacene el texto completo del documento. Adicionalmente este proceso se podría realizar también de forma dinámica, en el momento en que un usuario seleccionara una referencia de la bibliografía del artículo una vez publicado.

El resultado de una interrogación a CrossRef solamente será un código que se resolverá en un url que apunta al texto del documento. Será competencia de cada editorial vigilar si el usuario que está intentando acceder al documento dispone de una subscripción o en su caso establecer un sistema de pagar por ver para cada usuario individual. En definitiva, CrossRef no tiene que ver nada con control de accesos a los documentos. Igualmente no tiene nada que ver con los documentos en si mismos ya que lo único que almacena son sus metadados.

La utilidad de CrossRef en nuestro trabajo es bastante limitada. En primer lugar no existe un software de análisis de referencias ya que los editores trabajan con los documentos originales de los cuales ya tienen disponible en un formato utilizable (normalmente SGML) la información bibliográfica. Igualmente sucede con las referencias que generalmente estarán señalizadas en el documento utilizando algún tipo de marcas. Por otro lado todos los documentos que se distribuyen son de pago, por lo que el usuario que intente acceder a los mismos se encontrará con un sinfín de barreras de carácter económico.

3.5 Posibilidad de utilizar estos trabajos

Antes de optar por diseñar un nuevo sistema para el enlace de referencias en RePEc, se ha estudiado la posibilidad de utilizar o adaptar algunos de los programas creados por las iniciativas mencionadas en la sección anterior. Lamentablemente todos tienen el inconveniente de que el software desarrollado está pensado para la disciplina a la que se aplican. De los cinco proyectos vistos solamente ResearchIndex tiene un carácter interdisciplinar. Puestos en contacto con los autores, inmediatamente nos han enviado la última versión disponible del programa.

En la misma línea se contactó con el CERN Document Server explicando nuestro proyecto y pidiendo su colaboración. Nos contestaron remitiéndonos por correo electrónico algunos programas, por ejemplo el encargado de identificar los elementos que componen cada referencia. Si bien existe la posibilidad de que nos enviaran el resto, la utilidad de los mismos es muy limitada ya que

están excesivamente centrados en el caso particular de su archivo. Por otro lado la forma de construir las referencias en Física es completamente diferente a la utilizada en Economía. Básicamente los físicos acostumbran a indicar solamente el título de la revista abreviado y los datos fuente del fascículo en el que ha aparecido. Un ejemplo sería:

```
[3] Review of Particle Properties, Eur. Phys. J., C15 (2000)
```

Las referencias en Economía son mucho más complejas como se verá más adelante. Por todo ello desde el primer momento se ha descartado este software.

El mismo problema se plantea con OpCit y su software desarrollado para el enlace de arXiv.

En conclusión **ResearchIndex** es el único software del que disponemos para evaluar su adecuación a nuestros objetivos.

3.5.1 ¿Se podría aplicar ResearchIndex a nuestro proyecto?

En este punto se plantean dos alternativas. Una sería estudiar a fondo el programa y adaptarlo para que respondiera a las necesidades de este trabajo. La ampliación básicamente consistiría en mejorar el soporte para metadatos que ya está en fase inicial en el caso de documentos locales, es decir, que no han sido recuperados de la red. En segundo lugar habría que tener en cuenta el carácter descentralizado de nuestra colección, que nos obliga a generar una base de datos utilizable por los servicios.

El problema que se plantea en la adaptación es que puede ser muy complicada dada la falta de documentación absoluta del programa. El punto positivo es que se podrían hacer sugerencias de mejoras a los autores, si estos estuvieran interesados en seguir el trabajo con **ResearchIndex**.

Una segunda alternativa sería tomar los algoritmos utilizados para localizar las referencias en los documentos y aplicarlos directamente sobre nuestras publicaciones. En este trabajo sólo se necesita una parte muy pequeña de **ResearchIndex** dado que este es un verdadero índice de citas con gran cantidad de funciones que quedan fuera de nuestro objetivos, al menos por el momento. Lo más fácil sería tomar esa parte y olvidarse del resto.

A favor de esta segunda opción juega además el hecho de que a lo largo de la instalación y estudio de **ResearchIndex** han aparecido infinidad de problemas:

1. Falta de actualizaciones. La última versión es de Septiembre de 1999. Según las páginas personales de los autores su investigación parece cen-

trarse ahora más sobre el proyecto **inquirus** relacionado con la localización de información científica en Internet.

2. Su instalación es muy compleja ya que necesita multitud de módulos **perl** para funcionar.

3. Falta absoluta de documentación. La única ayuda que se ofrece es un fichero donde se señalan brevemente los pasos a seguir para la instalación y más brevemente aún para su funcionamiento. Estos dos puntos convierten al software en poco popular.

4. Dedica muchos recursos a distribuir procesos entre diferentes máquinas. Según la documentación esto agiliza el proceso de descarga y conversión de los documentos PDF o PostScript a ASCII, pero es totalmente inútil si solamente se dispone de una máquina.

5. Todo el trabajo de coordinación lo realizan una serie de demonios que gestionan la comunicación entre las máquinas. En general es un sistema muy inestable. Cualquier error en la ejecución de un programa, que por lo demás son frecuentes, obliga a reiniciar todos los demonios. En este sentido reiniciar todos al mismo tiempo usando el script **rc** hace que la máquina se caiga.

A la vista de esta serie de problemas e inconvenientes se optó por crear un sistema propio, utilizando hasta donde fuera posible los algoritmos de ResearchIndex. Es así como nace **CitEc**, un agente para la identificación de enlaces de referencias en Economía.

Capítulo 4

CitEc : Un agente para Economía

4.1 Características generales

El entorno en el que trabaja CitEc está formado por el conjunto de datos disponibles en RePEc. En la arquitectura de RePEc descrita en el capítulo dos, CitEc se integrará con una doble vertiente de archivo y de servicio. Como servicio aportará como valor añadido información para llevar a cabo el enlace de referencias. Como archivo distribuirá esta información a través de *templates* especialmente diseñados para citas.

CitEc está compuesto por diferentes módulos que permiten el desarrollo de sus funciones. Por principios CitEc esta diseñado para trabajar de forma autónoma sobre el universo de información de RePEc. La forma en que lo hace se puede ver en la figura 4.1. La entrada en el agente serían dos tipos de *templates*, los que describen documentos, bien sean artículos o *papers* y los que describen personas. La lectura de éstos se realiza mediante un módulo denominado **find_files**, que haría las funciones de sensores para detectar el entorno. Una vez que CitEc ha llevado a cabo su trabajo de enlaces de referencias, procede a modificar el entorno añadiendo nueva información a través de *templates* sobre citas.

En este sentido el *input* de CitEc serán los documentos a texto completo existentes en RePEc, mientras que el *output* será una serie de información que podrá ser utilizada por servicios como WoPEc o IDEAS para presentar al usuario final un nuevo valor añadido. CitEc no está planteado como un servicio destinado a usuarios finales, sino como un apoyo al resto de servicios existentes. No se pretende crear un servidor web donde se ofrezca la información generada por el proyecto sino que se trabajará de forma autónoma auxiliando a otros servicios. Serán ellos los que decidan si hacen uso de los datos que se generen o no y cómo van a llevar a cabo la integración de los mismos. Esta

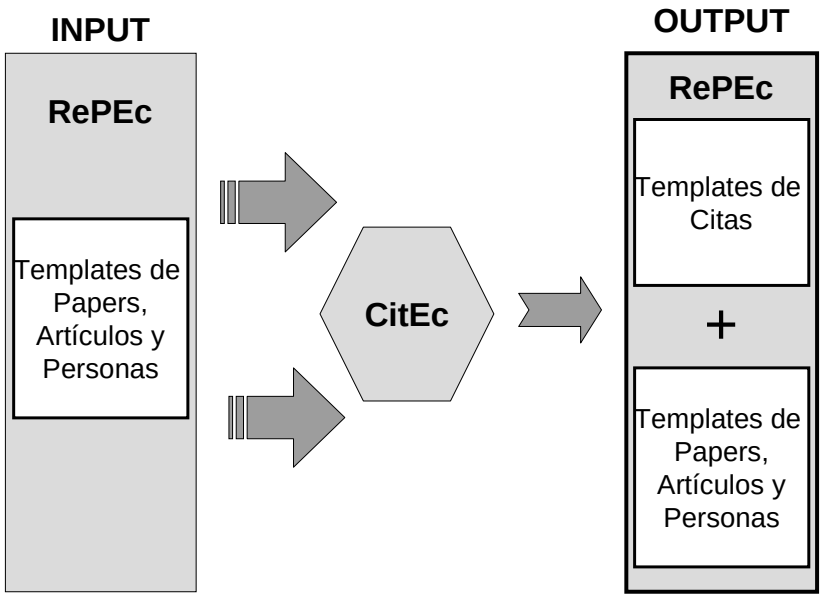


Figura 4.1: CitEc y su entorno

forma de actuar tiene la ventaja de que nos aseguramos una mayor difusión del trabajo realizado al utilizar la masa de clientes que ya vienen utilizando los servicios tradicionales. En el capítulo anterior mencionamos los problemas de OpCit para difundir sus resultados dado que los usuarios tienden a utilizar el interface antiguo, mucho más conocido que el nuevo. Por otro lado, sería contraproducente crear un nuevo servicio que compita con los existentes, el usuario podría verse confundido sin saber a ciencia cierta donde acudir para buscar la información que necesita.

De esta forma, la información que produzca CitEc deberá ser compatible con el resto de datos en RePEc de forma que pueda ser utilizada por los servicios. Para ello el resultado estará redactado en formato ReDIF y se distribuirá a través de un archivo especial llamado **cit**. En el caso de ReDIF se creará un *template* nuevo denominado **ReDIF-Cite 1.0** y cuyos contenidos son expresados más adelante. Adicionalmente se estudiará la posibilidad de utilizar el formato AMF.

Como ya se señaló en la introducción de este trabajo, el proceso de enlace de referencias se podría dividir en tres fases o niveles: la **recolección** de los documentos sobre los que se va a trabajar, el **análisis** de los mismos para encontrar las referencias y aislar los elementos que las componen y finalmente el **enlace** de esas referencias con aquellos documentos que estén disponibles en formato electrónico. Una representación gráfica de este proceso aparece en la figura 4.2. Seguidamente se analizará cómo lleva a término CitEc cada una de estas tareas.

El proceso de recolección tiene una doble vertiente: recolección de la información bibliográfica y recolección de textos propiamente dichos. El punto de partida del proceso son los datos distribuidos a través de RePEc. A través de ellos se sabrá si un documento se encuentra disponible en la red o no. En caso afirmativo nos proporcionarán la dirección para acceder al mismo. Hay que notar que solamente se trabajará con documentos en formato PDF o PostScript, no obstante ésta no es una limitación grave ya que, como se verá a continuación, más del 95% de los documentos en RePEc están en alguno de estos dos formatos.

En un segundo nivel, el proceso de análisis toma la información sobre los documentos y los descarga a nuestro sistema. Posteriormente, los convierte a formato ASCII para poder llevar a cabo los trabajos de análisis y finalmente identifica las referencias con todos sus elementos constitutivos: autor, título, fecha, etc.

Esta información sobre citas es pasada al tercer y último nivel, el de enlace, quien se encarga de producir los datos que después serán utilizados por el resto de servicios de RePEc.

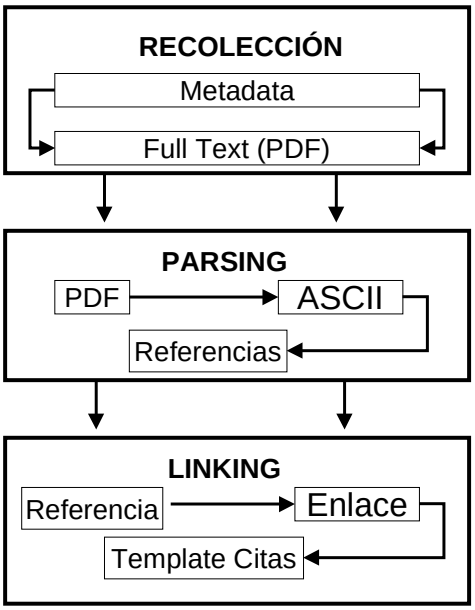


Figura 4.2: Proceso de extracción de referencias

4.2 La base de datos

Internamente todo el flujo de datos generado se canaliza a través de una base de datos relacional. RePEc fue diseñada teniendo en mente una estructura relacional, como se puede deducir de lo visto en la segunda sección de este trabajo. Por ello resulta muy fácil trasladar esta estructura a una base de datos. Como sistema de gestión de ésta se ha elegido el software MySQL en su versión para linux.

Entre las razones por las que se ha seleccionado como sistema de gestión MySQL y no otro están:

- Es gratuito. Dado que el proyecto carece de financiación se ha de recurrir a herramientas que se distribuyan gratuitamente.
- Es uno de los sistemas más populares y robustos para linux/unix.
- Se puede acceder a el desde una variedad de lenguajes de programación tales como **perl** o **php**.
- Dispone de abundante documentación.
- Existen multitud de módulos **perl** que facilitan las tareas de acceso y manipulación de los datos.

En la figura 4.3 se detalla el esquema entidad-relación de la base de datos. Toda su estructura se articula en torno a la tabla **Documento**. Seguidamente se ofrece el diseño lógico de esta estructura con la enumeración de todas las tablas existentes.

DOCUMENTO(tclave:VARCHAR(200), titulo:VARCHAR(200),
docid:VARCHAR(60), año:YEAR(4), control:CHAR(0),
acodigo:CHAR(9), scodigo:CHAR(16))
Clave Primaria: docid
Clave Ajena: acodigo hace referencia a ARCHIVO
Clave Ajena: scodigo hace referencia a FUENTE

Como se aprecia en el esquema ésta es la tabla central del sistema. Almacena información sobre cada uno de los documentos de los cuales se dispone del texto completo en RePEc, tanto si están accesibles de forma gratuita como si están restringidos. Se prescinde de todos aquellos documentos de los que no se dispone del texto completo. Los campos que componen esta tabla son:

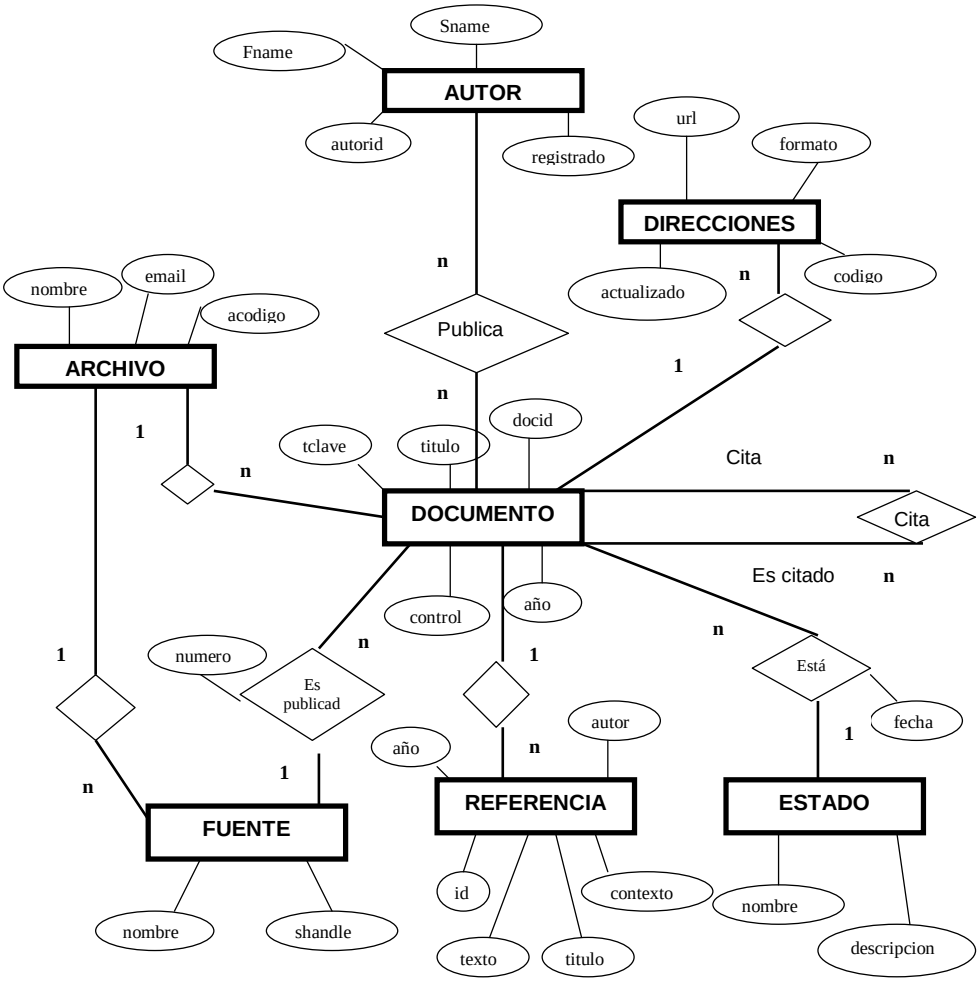


Figura 4.3: Esquema entidad-relación para CitEc

- Título. Almacenará los títulos de los documentos tal como aparezcan en la fuente.
- Tclave. Contendrá el título clave. Esta es una versión normalizada del título a la que se le han eliminado todos los signos de puntuación, se ha pasado mayúsculas a minúsculas y se han eliminado los artículos. Su objetivo es obtener una versión del título más normalizada en la que se reduzcan posibles errores de trascripción desde la fuente a la referencia bibliográfica. Se utilizará para comprobar si dos títulos pertenecen al mismo documento.
- Docid. En este atributo se almacenará el *handle* del documento, que como ya se ha indicado sirve a la vez para identificar de forma unívoca al documento dentro de RePEc y por otro nos ofrece puntos de enlace con el archivo al que pertenece y la serie en la que ha sido publicado.
- Año. Fecha de publicación del documento. Con este campo podemos tener problemas ya que en el caso de los *working papers* no es un campo obligatorio según la versión actual de ReDIF. No obstante la mayoría de documentos lo incluyen e incluso se puede extraer de la numeración de los mismos. Habitualmente esta numeración se compone del año de publicación, en formato largo o abreviado, seguido de un número correlativo que se renueva cada año. Por ello se han incluido una serie de funciones para que CitEc sea capaz de extraer esos datos a partir del número.
- Control. Este es un campo booleano que podrá tomar los valores 0 y 1. En principio está pensado para auxiliar en tareas como por ejemplo el borrado de registros.
- Acodigo. Este es un campo que contiene la clave ajena que permite realizar un enlace con la tabla ARCHIVO.
- Scodigo. Este es un campo que contiene la clave ajena que permite realizar un enlace con la tabla FUENTE.

AUTOR(autorid:VARCHAR(50)), registrado:CHAR(0),
Fname:VARCHAR(50), Sname:VARCHAR(50)
Clave Primaria: autorid

Esta tabla contiene la información relativa a los autores. Su clave es el campo autorid. El contenido de este campo es de dos tipos, en caso de que el autor esté registrado en HoPEc, su identificador será el *handle* de persona que se le haya asignado, en caso contrario será una cadena compuesta por las letras de su

nombre completo traducidas a minúsculas, eliminados los signos de puntuación los acentos y los espacios.

Los campos Fname and Sname almacenan los nombres y apellidos respectivamente de los autores.

PUBLICA(autorid:VARCHAR(50),docid:VARCHAR(60))
Clave Primaria: autorid,docid
Clave Ajena: autorid hace referencia a AUTOR
Clave Ajena: docid hace referencia a DOCUMENTO

Esta tabla está destinada a representar la relación entre las tablas DOCUMENTO y AUTOR.

DIRECCIONES(url:VARCHAR(200), docid:VARCHAR(60),
formato:VARCHAR(50), actualizado:TIMESTAMP(12),
codigo:CHAR(4))
Clave Primaria: url
Clave Ajena: docid hace referencia a DOCUMENTO

Tabla en la que se almacenarán las direcciones electrónicas de los documentos. El caso ideal sería que cada documento estuviera disponible en un sólo fichero. Lamentablemente esto sólo se cumple en ocasiones. Es habitual que el autor ofrezca varias versiones del mismo documento, por ejemplo una en PDF y otra en Microsoft Word. Incluso se puede complicar si ofrece versiones comprimidas de esos formatos y versiones sin comprimir. En el caso del archivo RePEc:wpa (Washington University at Saint Louis) se ofrecen no menos de siete ficheros por cada documento. Otra opción que utilizan algunos autores aunque afortunadamente en menor medida es dividir el documento en varios ficheros, por ejemplo uno para cada capítulo y otro para la bibliografía, etc.

En esta tabla se almacenarán por cada documento todas las direcciones existentes en su ficha bibliográfica. Será trabajo del software, como se verá después, determinar cual de esos ficheros contiene la versión correcta o al menos la utilizable por CitEc.

Por cada dirección almacenada se tendrá disponible la fecha en la que se actualizó el fichero en el servidor. Esto servirá para detectar futuras versiones de

este documento. También se guardará en el campo código el código devuelto por el servidor HTTP cuando se accedió al mismo. En caso de que el fichero haya sido procesado satisfactoriamente, es decir, que su estado sea **ready**, este campo se pondrá a **111**.

REFERENCIA(id:VARCHAR(50), docid:VARCHAR(60),
autor:VARCHAR(150), titulo:VARCHAR(200),
año:YEAR(4),texto:TEXT,contexto:TEXT)
Clave Primaria: id
Clave Ajena: docid hace referencia a DOCUMENTO

Ésta es una tabla muy importante ya que almacenará información sobre cada una de las referencias que se hayan identificado. Por cada referencia se salva el texto completo de la referencia así como el contexto en el que ha aparecido. Además se guardará el resultado de su *parsing*: lista de autores, título y año de publicación. Dado que cada referencia puede ser citada varias veces en un documento, es posible también que existan varios contextos. Si este es el caso los diferentes contextos se separan en el campo por el carácter "|".

Finalmente nos encontramos con dos claves, la ajena que enlaza con la tabla DOCUMENTO y la primaria que es un código compuesto de: el docid seguido de ":" y un número correlativo que se renueva con cada documento.

Es importante matizar que solamente se guardarán las referencias que hayan sido analizadas correctamente. Será necesario crear algoritmos que permitan identificar cuándo una referencia es correcta o no. Todo ello se analizará más adelante.

ESTADO(nombre:VARCHAR(50)), descripcion:VARCHAR(250)
Clave Primaria: nombre

Otro punto importante es el estado de proceso de un documento en un determinado momento. Desde que comienza el procesado de un determinado documento hasta que se llevan a cabo los enlaces de sus referencias con los documentos citados, este puede pasar por una serie de estados que son:

1. Undownloaded. Estado inicial, solamente se ha añadido la información al sistema pero el documento no ha sido procesado.

2. Downloaded. El documento ha sido descargado al disco duro.

3. DownError. Se ha producido un error al descargar el fichero. En este caso el último campo, que se ha llamado dígito de control, contiene el código devuelto por el servidor HTTP. El sistema periódicamente comprobará si el error continua existiendo. Si hubiera sido solucionado procedería a la descarga del documento y continuaría así el proceso. En caso contrario, tras intentar la descarga un número de veces determinado por la configuración de CitEc, enviaría un mensaje al administrador del archivo comunicando el error.

4. Restricted. Es necesario una subscripción o algún tipo de pago para acceder al texto del documento. No es un documento gratuito por lo que no se puede procesar.

5. BadDocument. El url no va al texto del documento sino a una página de resúmenes intermedia. Esta es una práctica de algunos archivos pero que no está admitida en ReDIF. Cuando sucede esto el documento no puede ser procesado.

6. IncompatibleFormat. El documento ha sido descargado pero se ha encontrado que no está ni en PDF ni en PostScript, por lo tanto no se puede procesar.

7. Converted. El documento ha sido convertido a ASCII de forma satisfactoria.

8. NoReferences. El documento ha sido convertido a ASCII satisfactoriamente pero no se ha encontrado una sección de referencias, bien porque no existe o porque ha sido imposible detectarla.

9. NoEnglish. El documento ha sido convertido pero aparentemente no está en inglés. Para comprobar este punto se examina si en el contenido del documento aparecen las palabras más comunes del inglés. Documentos en otros idiomas son descartados.

10. BinaryFile. Se ha producido un error al convertir el documento a ASCII, el resultado es un fichero binario.

11. Ready. Una vez convertido el documento a ASCII se ha encontrado una sección de referencias que nos permitirá seguir procesándolo.

12. PsToTextError. Se ha producido un error indeterminado al convertirlo a ASCII.

13. UnableToReadAsciiFile. Ha sido imposible leer el fichero texto generado por el programa de conversión.

14. WrongNumberOfReferences. El documento ha sido procesado pero se ha encontrado un número de referencias que no se encuentra entre los límites establecidos en la configuración del sistema. Ello nos lleva a inducir que se ha producido un error y no se han analizado las referencias correctamente.

15. Done. El documento ha sido analizado correctamente y está listo para pasar al último nivel.

16. Linked. Las referencias del documento han sido enlazadas y por lo tanto el proceso del mismo ha terminado.

Cada estado lleva asociada la fecha en que se ha establecido.

Una representación gráfica de estos estados se puede ver en la figura 4.4.

ESTA(docid:VARCHAR(60), nombre:VARCHAR(50),
fecha:TIMESTAMP(12))
Clave Primaria: docid
Valor No Nulo: nombre
Valor No Nulo: fecha
Clave Ajena: docid hace referencia a DOCUMENTO
Clave Ajena: nombre hace referencia a ESTADO

Esta tabla establece la relación entre un documento y el estado en que se encuentra. Está compuesta por las claves ajenas de las tablas que relaciona más un campo fecha asociado a la relación que almacenará la fecha, en formato AAAAMMDDHHMM, en que se ha asociado un estado al documento. Destacar que existe una restricción de valor no nulo sobre los atributos "nombre" del estado y "fecha".

FUENTE(nombre:VARCHAR(150), scodigo:CHAR(16), acodigo:CHAR(9))
Clave Primaria: scodigo
Clave Ajena: acodigo hace referencia a ARCHIVO

En esta tabla se guardará la información sobre las series de documentos que existen en RePEc. Por definición, según está especificado en ReDIF, todo documento debe pertenecer a una única serie. No pueden existir los documentos

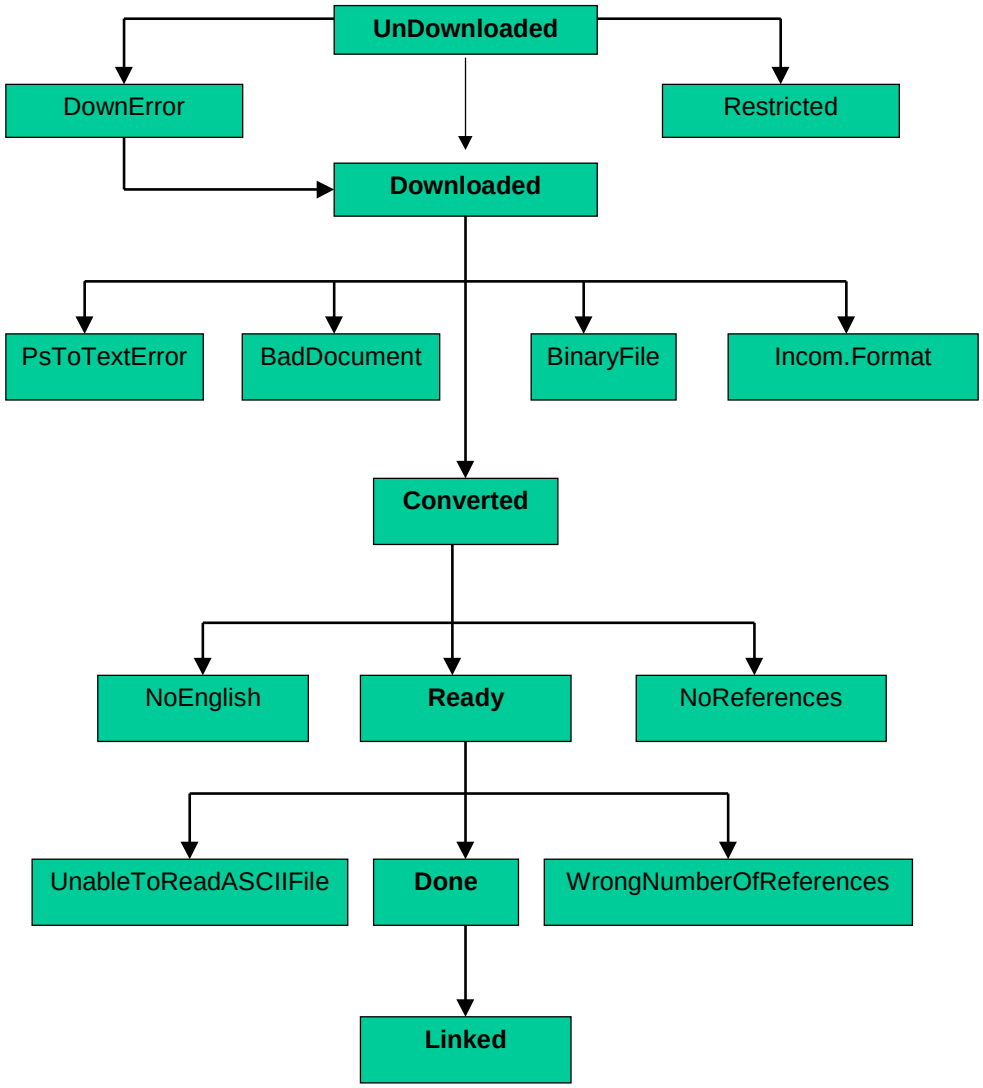


Figura 4.4: Estados de un documento

de forma aislada. La información sobre series que se necesita a nuestro nivel es bastante limitada, así es que simplemente se recoge de cada una su nombre y su handle.

El caso del nombre de las series de documentos de trabajo es particularmente complicado. Mientras en el caso de las revistas las series equivalen a revistas y cada una suele tener un título suficientemente detallado para distinguirla de las demás, las series de *working papers* acostumbran a llamarse de la misma forma: *working papers* o *discussion papers* o documentos de trabajo, etc. Con tan poca variedad de títulos es necesario acudir, siguiendo las técnicas bibliotecarias y documentales, a adjuntar el nombre de la entidad editora al título de la misma. De esta forma ya podremos identificarlas de forma unívoca. Ambas partes se separan por "/". Así por ejemplo *Universitat de València / Quaderns de Treball*.

ESPUBLICADO(docid:VARCHAR(60), scodigo:CHAR(16)
numero:VARCHAR(10))
Clave Primaria: docid
Valor No Nulo: shandle
Clave Ajena: docid hace referencia a DOCUMENTO
Clave Ajena: shandle hace referencia a FUENTE

Con esta tabla se establece la relación entre un documento y la serie en la que se ha publicado. Además de las claves ajenas, la relación tiene un atributo que es el número con que ha sido publicado el documento. En el caso de los artículos de revistas esto incluye el conjunto de los datos fuentes, esto es: volumen y número.

CITA(cita:VARCHAR(60), escitado:VARCHAR(60))
Clave Primaria: cita,escitado
Clave Ajena: cita hace referencia a DOCUMENTO
Clave Ajena: escitado hace referencia a DOCUMENTO

Aquí se incluye las relaciones de cita y es citado entre documentos disponibles a través del sistema.

ARCHIVO(acodigo:CHAR(9), nombre:VARCHAR(80), email:VARCHAR(80))
Clave Primaria: acodigo

Esta tabla contiene información de los archivos, es decir, instituciones y departamentos, que han proporcionado datos a RePEc. Cada archivo contiene una o más series de documentos y toda serie pertenece por definición a un archivo.

Esta tabla no es estrictamente necesaria para nuestro trabajo pero se incluye porque puede ser interesante saber qué archivos contienen documentos que no se pueden descargar, cuáles contienen los documentos más citados, etc. En definitiva se intenta producir información que pueda ser de interés para los archivos colaboradores.

La tabla ARCHIVO contiene tres atributos: el código o handle del archivo, el nombre del mismo y la dirección de correo electrónico del responsable.

4.3 El servidor

Para el diseño y prueba de CitEc se ha utilizado un servidor cedido por el Instituto de Investigaciones Económicas de la Hitosubhasy University de Tokyo (Japón). Su dirección es: netec.ier.hit-u.ac.jp. A grandes rasgos sus características técnicas son: 10 procesadores, 128 Gb de memoria y 300 Gb de disco duro para nuestro proyecto. El sistema operativo es linux Red Hat.

Para CitEc es muy importante disponer de espacio en disco suficiente ya que ha de descargar un elevado número de documentos electrónicos. Los 300 Gb son por el momento suficientes.

4.4 Software adicional requerido

El sistema se apoya en una serie de módulos **perl** y programas adicionales para su funcionamiento. Todos comparten la característica de tratarse de software de acceso libre que es distribuido bajo la licencia GNU. Seguidamente se enumeran y describen cada uno de ellos.

1. Módulos **perl**
   (a) Disponibles en una instalación **perl** estándar

- Getopt::Long. Escrito por Johan Vromans <`jvromans@squirrel.nl`> permite procesar de forma sencilla la línea de comandos de un programa. Se utiliza para leer las opciones con las que se llaman los distintos módulos que componen CitEc.
- File::Find. Utilizado para buscar ficheros en el sistema que cumplan una determinada condición.
- File::Copy. Utilizado para copiar o mover ficheros en el sistema

(b) Módulos proporcionados por RePEc:

i. ReDIF-Perl. Es un paquete de Ivan Kurmanov <`ivan@tm.minsk.by`> que contiene una serie de módulos y programas destinados a facilitar la administración de los archivos así como el manejo de registros en formato ReDIF. De él se utilizan por el momento:

- rech. Es un programa para la detección y corrección de errores en los *templates ReDIF*. Comprueba que la sintaxis de cada *template* es la correcta, que aparecen todos los campos obligatorios, que su formato es el adecuado, etc. Si detecta un problema abre un editor de texto (**emacs**) y coloca el cursor en el lugar donde está el problema indicando una breve descripción del mismo.
- ReDIF::init. Es un módulo que se encarga de inicializar una serie de variables de obligado cumplimiento según el *Guildford Protocol*: entre otras el *path* al directorio en el sistema donde se guarda el mirror de RePEc. Igualmente detecta donde están ubicados los ficheros de configuración del programa, ficheros de especificación de formatos, etc.
- rr.pm. Es el módulo que se encarga de leer los *templates* desde el disco, identificar los correctos y eliminar los que contienen errores, etc. Una vez analizados los pasa al programa desde el que ha sido llamado en una estructura de datos que se puede acceder en forma de *hash de hashes*. Véase (19) para una descripción detallada de esta estructura.

(c) Módulos generados por CitEc

- stopwords.pm. Básicamente contiene una lista configurable de las palabras más utilizadas en inglés.
- citec.pm. Contiene todas las funciones que comparten dos o más módulos de CitEc. Entre ellas las funciones de conexión a la base de datos MySQL, de grabación de datos en los ficheros log, etc. Igualmente contiene una serie de variables que el usuario no debería tocar.

(d) Otros módulos disponibles en CPAN[1] (Comprensive Perl Archives Network).

[1] `http://www.cpan.org`

- DB_File.pm. Es un módulo escrito por Paul Marquess <Paul.Marquess@btInternet.com> que permite a programas **perl** hacer uso de las facilidades proporcionadas por Berkeley DB. En CitEc se utiliza como paso intermedio para transferir datos de las tablas que contienen información generada por el sistema (ESTA y DIRECCIONES) en cada actualización. Como se verá, con cada actualización se borran todos los datos de la base y se genera de nuevo, a partir de los *templates* de documentos y personas que existan en ese momento. La mayor parte de los datos se extraen de los diferentes *templates,* pero hay datos que no se salvan a disco porque sólo tienen sentido para la gestión del sistema. Por ejemplo el estado de proceso en que se encuentra cada documento. Esta información necesita también transferirse en cada actualización. Para conseguirlo, se salva temporalmente a una base Berkeley DB y seguidamente se va recuperando a medida que se van insertando documentos, según las necesidades.
- Lingua-EN-Nameparse. Es un módulo escrito por Kim Ryan <kimaryan@ozemail.com.au> para analizar nombres de personas. Toma como entrada una cadena que supuestamente contiene un nombre en formato libre e intenta determinar qué parte contiene el nombre y cuál los apellidos. En CitEc se utiliza para normalizar los nombres de los autores ya que esta información, según la especificación ReDIF, puede ir en formato libre.
- Libwww-perl. Es una colección de módulos **perl** escrita por Gisle Aas. Su objetivo es proporcionar una serie de funciones que permitan al usuario crear fácilmente clientes web. Entre los módulos **perl** que incluye en CitEc se utilizan solamente: LWP::UserAgent y LWP::Request, para descargar el texto completo de los documentos.
- File-MMagic. Es un módulo para determinar el formato de un fichero. En CitEc se utiliza para analizar los documentos descargados y determinar si están comprimidos o si están en un formato aceptado (PostScript o PDF).
- String::Approx. Es un módulo escrito por Jarkko Hietaniemi <jhi@iki.fi> para calcular la distancia de edición entre dos cadenas. La distancia de edición se define como el número de inserciones, borrados o modificaciones que hay que realizar para convertir una cadena **A** en una cadena **B**. En CitEc se utiliza, junto con otros indicadores, para determinar si dos títulos son lo suficientemente parecidos como para inferir que se refieren al mismo documento.
- DBI. DBI son las siglas de Database Independent Interface, un módulo escrito por Tim Bunce. Es un módulo que permite acceder

desde programas **perl** a una gran variedad de bases de datos. En CitEc se utiliza para todas las transacciones con la base MySQL desde cualquier módulo.

2. Software adicional

   (a) Programas de conversión de formatos:

   - Pstotext[2]. Es un programa escrito por Andrew Birrel como parte del proyecto *Virtual Paper* y es distribuido gratuitamente. Básicamente es un conversor de ficheros PostScript a ASCII, aunque como veremos también puede convertir documentos en PDF si bien con menor efectividad. Se puede consultar en (17) una comparación de herramientas disponibles para la conversión de PDF o PostScript a ASCII.

   (b) Programas de mirror:

   i. w3mir[3]. Es un programa muy potente para construir mirrors de uno o varios servidores web o incluso para copiar sistemas de archivos completos como por ejemplo desde un CDROM. La utilización que hacemos de él se limita a copiar la información bibliográfica en formato ReDIF desde los archivos que se distribuyen a través del web.

   ii. mirror.pl[4]. Es un programa escrito en **perl** por Lee McLoughlin <lmjm@icparc.ic.ac.uk> con el objeto de duplicar una jerarquía de directorios entre dos máquinas. Evita copiar innecesariamente documentos entre ambas máquinas al comparar las fechas y tamaños de los ficheros origen y destino antes de proceder a su descarga. Básicamente lo que hace es un mirror de servidores FTP. Aunque es un programa muy antiguo, la última versión es de 1998, es muy potente y mucho más rápido que el anterior. Igual que w3mir en RePEc se utiliza para copiar los archivos que se distribuyen en servidores FTP.

4.5 El sistema de ficheros

Todos los ficheros generados por el sistema se guardan en una estructura de directorios como la que sigue. Asumimos que **CitEc** es el directorio raíz.

- **CitEc/bin** Directorio que alberga los módulos ejecutables. Entre otros: **find_files** y **get_papers**.

[2] http://research.compaq.com/SRC/virtualpaper/pstotext.html
[3] `http://www.math.uio.no/~janl/w3mir/`
[4] `http://sunsite.org.uk/packages/mirror/`

- **CitEc/lib** Destinado a recoger los módulos **perl** que necesitemos. Por el momento el único existente es **citec.pm**
- **CitEc/Papers** Contiene una estructura jerárquica de subdirectorios que comienza con uno por cada archivo existente en RePEc, seguidamente cada archivo se subdivide en series y éstas en números. Cada documento particular se incluye en una de estas carpetas. Aquí se guardarán los ficheros en los formatos originales así como el documento convertido a ASCII y una serie de fichero auxiliares:
  - **CitEc/Papers/<arc>/<series>/<numero>/nombre.txt** Contiene el documento convertido a ASCII y procesado, de tal forma que se le han eliminado por ejemplo saltos de página, determinadas marcas, etc.
  - **CitEc/Papers/<arc>/<series>/<numero>/nombre.raw** Contiene el documento en formato ASCII tal cual ha sido generado por **pstotext**.
  - **CitEc/Papers/<arc>/<series>/<numero>/nombre.cit** Almacena la parte del documento susceptible de contener las referencias bibliográficas.
  - **CitEc/Papers/<arc>/<series>/<numero>/nombre.bod** Almacena el texto del documento propiamente dicho, sin referencias ni preliminares.
  - **CitEc/Papers/<arc>/<series>/<numero>/nombre.hed** Contiene información de control generada por el programa que creó el fichero PDF o PostScript original
  - **CitEc/Papers/<arc>/<series>/<numero>/nombre.abs** Contiene el abstract del documento. Si CitEc ha sido incapaz de encontrarlo almacena las primeras frases del cuerpo del documento.

Los tres directorios principales podrían estar diseminados por el sistema de archivos, no es necesario que sigan la estructura mostrada. La ubicación de cada uno de ellos es necesario pasársela a cada módulo a través de un fichero de configuración. Incluso se podrían permitir distintas ubicaciones si se quisiera por ejemplo tener bases de datos separadas para distintos servicios.

Capítulo 5

CitEc : Funcionamiento

5.1 Recolección

Ya se ha dicho que el entorno de trabajo de CitEc es el conjunto de datos que forman RePEc. En este sentido es imprescindible disponer físicamente en nuestro sistema de todos esos datos. Esto se consigue creando un servicio RePEc, que llevará por nombre CitEc y como código de identificación **cit**. Recordamos que técnicamente un servicio es un espacio en el disco duro de una máquina donde se mantiene una copia o *mirror*, total o parcial de los datos disponibles en RePEc. En nuestro caso se ha elegido la opción de crear un *mirror* parcial con objeto de ahorrar espacio en disco. Obviaremos copiar los datos de aquellos archivos que con seguridad no contienen documentos a texto completo. El más importante de ellos es el CEPR.

Siguiendo las instrucciones del *Protocolo de Gildford*, el destino de los datos será un directorio en nuestra máquina accesible en la dirección `ftp://netec.ier.hit-u.ac.jp/pub/RePEc/remo`. Para llevar a cabo el proceso de *mirror* de los datos se ha utilizado *software* distribuido por RePEc: en concreto, *remi* un programa escrito en **perl** por Sune Karlson y disponible en `ftp://netec.mcc.ac.uk/NetEc/RePEc/all/soft/RePEc/remi`.

Se ha añadido una entrada al **cron** de **netec.ier.hit-u.ac.jp** para que diariamente se ejecute el *mirror* y se comprueben las modificaciones que se han producido en los diferentes archivos.

Con esto se ha finalizado la parte más sencilla. Lo que se tiene en este momento en nuestra máquina es una representación del entorno de datos de RePEc, es decir, las descripciones bibliográficas de los documentos en formato ReDIF almacenadas en una serie de ficheros con extensión ***.rdf**. Solamente algunos de los documentos referenciados estarán disponibles a texto completo en la red. Sabremos cuáles son porque su descripción incluirá como mínimo un campo

File-URL con la dirección electrónica donde se encuentran.

Una vez que se ha reproducido en nuestra máquina el entorno de datos de RePEc se puede pasar a lo que sería la recolección de los metadatos. De esta tarea se encarga el módulo **find_files**. Este *script* recorre todos los ficheros almacenados en el directorio *remo* de RePEc y guarda los datos que encuentra en la base relacional.

Uso:
find_files -c [-rhm]

-c <fichero> Es obligatorio especificar un fichero de configuración donde se recogen una serie de variables como: localización y nombre de la base de datos, directorio para almacenar el texto de los documentos, etc.

-r <directorio> Directorio del que se va a leer la información a añadir al sistema. Por defecto lee todo el directorio *remo*.

-n 1|0 Permite determinar si queremos reutilizar la información generada por el sistema (tablas ESTA y DIRECCIONES) o si se borran todas las tablas. Por defecto está seleccionado 1, si reutilizar.

-h Mensaje de ayuda.

El algoritmo que sigue **find_files** se puede resumir en los siguientes pasos:

1. Si existe una base de datos con el nombre especificado la borra y genera una nueva con las tablas definidas anteriormente. Borrando cada vez los datos existentes nos aseguramos que la base es una representación exacta de RePEc y se simplifican las tareas de actualización que de otra forma podrían ser complicadas, por ejemplo para trazar la vida de un documento que cambia de archivo o simplemente para determinar qué documentos son nuevos y cuáles han sido modificados.

   La información que guardamos en la base de datos es de dos tipos:

   (a) Aquella que está disponible en RePEc y puede ser recuperada en cualquier momento a partir de los *templates* creados por los archivos. A este grupo pertenecen todas las tablas excepto ESTA y DIRECCIONES. Una vez borrada la base, los datos se recuperan de los *templates* de documentos, personas y citas.

   (b) Aquella que es generada por el propio sistema para su gestión y que no puede ser recuperada de ningún sitio. Las tablas ESTA y DIRECCIONES no se pueden recuperar de ningún sitio por lo cual el sistema hace una copia de seguridad de los datos destinados a ser recuperados si se necesitan en la forma que se especifica más adelante. Las copias tienen la forma de bases de datos Berkeley.

2. Procede a rellenar las tabla invariable ESTADO.

3. Lee e inserta en las tablas correspondientes la información sobre series y archivos. Esta información está disponible en el archivo **RePEc:all**. En este punto se modifican las tablas ARCHIVO y FUENTE.

4. Lee el archivo **per** que contiene información sobre autores y que debe estar disponible en el directorio: `ftp://netec.ier.hit-u.ac.jp/pub/RePEc/remo/per`

5. Por cada *template* encontrado, inserta el nombre y apellidos en la tabla AUTOR. Además pone el campo REGISTRADO a 1 e inserta el *handle* en el campo AUTORID. Seguidamente añade una linea a la tabla PUBLICA consistente del *handle* de la persona y el código del documento, de forma que se lleva a cabo la asociación entre autores registrados y sus documentos.

6. Lee la información sobre *papers* y artículos disponibles en el directorio `ftp://netec.ier.hit-u.ac.jp/pub/RePEc/remo`. Si en algún fichero se encuentran *templates* que ya han sido actualizados se prescinde de ellos (archivos, series o personas).

7. Por cada documento se procede a determinar los datos que no aparecen expresamente en los *templates*. Así el año de publicación, que no es un campo obligatorio, estará ausente en muchos casos. Por ello se deberá incluir algún conocimiento para que el agente sea capaz de inferirlo a partir de otros datos. De esta tarea se encarga la siguiente función:

   (a) Si aparece el campo **revision-date**, que tiene un formato preestablecido de AAAA-MM-DD se toman las primeras cuatro cifras de la fecha. Si hay más de una fecha de revisión se comparan para tomar la más actual.

   (b) En caso contrario se comprueba si aparece **creation-date** y de existir se sigue el mismo proceso anterior.

   (c) Se comprueba si la fecha obtenida es correcta.

   (d) Si no es o no existe, se procede a analizar el *handle*.

      i. En el caso de un artículo se analiza si el año aparece en el handle como lo requiere ReDIF. Si es así, se toma. De lo contrario se comprueba si existe un campo **Year**, optativo según ReDIF. Si existe, se toma.

      ii. En el caso de los *papers* las series se suelen numerar con una combinación de año (en diferentes formatos) más un número correlativo que empieza en 1 cada año. Para inferir la fecha se hace un análisis de la parte final del handle, destinada al número del documento, para ver si responde a alguno de los siguientes formatos:

AA[signo]NN
NN[signo]AA
NNAA
AANN
AAAA[signo]NN
En lenguaje **perl**, las expresiones que se intenta encontrar son:

(e) Se comprueba si la fecha que se tiene es correcta, esto es, son cuatro dígitos de la siguiente forma: [12][90][0-9][0-9]

(f) A no ser que la fecha sea incorrecta o esté vacía, se devuelve el año inferido

Con esta función se ha podido recuperar buena parte de la información no disponible expresamente en los *templates* como se aprecia en la figura 5.1.

En esta muestra de más de 35000 documentos se ve cómo sólo la mitad de los documentos incluyen expresamente la fecha de publicación. También destaca la escasa utilización del campo **Revision-Date** destinado a indicar sucesivas revisiones de un mismo documento. Utilizando la función descrita se ha conseguido reducir el porcentaje de documentos sin fecha a sólo un 7%. Para ello se han utilizado el *handle* de los *working papers* en el 17% de los casos, el *handle* de los artículos en el 82% y el campo **Year** en sólo el 1%. Destaca el caso de los artículos, donde lo habitual es especificar la fecha de publicación insertada en el *handle* en lugar de utilizar los campos adecuados.

8. Se procede a determinar el título clave.

9. Otros datos que no aparecen expresamente en la información bibliográfica son los apellidos y nombres de los autores. ReDIF en este sentido es muy ambiguo. Se limita a sugerir que la forma del nombre sea la invertida, pero no requiere nada al respecto. Esto hace que cada archivo ofrezca diferentes formatos: en forma directa, invertida, con uno o dos apellidos, con iniciales, con títulos, etc. Además la cuestión se complica con nombres de distintas nacionalidades.

   En definitiva la cadena de caracteres que se ha introducido en el campo **Author-Name** de ReDIF debe ser analizada para acomodarse a los cam-

	Antes	**Después**
Documentos	35618	35618
Revision-Date	679	679 (2%)
Creation-Date	1712	32595
Sin Fecha	17827 (50%)	2344 (7%)
Fecha Correcta	17791	33274

Figura 5.1: Efectividad para encontrar fechas

pos establecidos. Para este propósito se ha recurrido a **Lingua:EN:NameParse**. Este módulo toma una cadena que presuntamente contiene el nombre de una persona y lo divide en los elementos que lo componen. El problema que se plantea es que solamente es fiable con nombres anglosajones, con el resto se ha tenido que definir alguna excepción. La más importante es sustituir los caracteres acentuados por los respectivos caracteres sin acentuar ya que no reconoce los acentos.

En un primer momento se intenta hacer el análisis utilizando el módulo. Si falla se han creado una serie de condiciones que solucionan los principales problemas que se han planteado. Básicamente, cuando un nombre sigue la forma:

palabra palabra palabra palabra

se ha considerado que el nombre son las dos primeras palabras y los apellidos las siguientes. Si el nombre es:

palabra inicial inicial palabra

el nombre estará compuesto por los tres primeros elementos y el apellido por el último.

10. Una vez que se dispone de todos los datos se procede a cargarlos en la base de datos. En el caso de los autores primero se comprueba si están registrados. En caso afirmativo no se hace nada pues esta información presumimos que ya la tenemos. En caso contrario se procede a crear un código de identificación para el autor y a insertar una línea en las tablas AUTOR y PUBLICA.

11. Se inserta el resto de información en la tabla DOCUMENTO: título clave, título, año, código del archivo y código de la serie.

12. Si no existe se crea un subdirectorio en **CitEc/Papers** para almacenar los ficheros que se descarguen más adelante.

13. Por cada campo **File-URL** del *template* se procede a comprobar en la copia de la tabla DIRECCIONES si ya se había recuperado ese fichero. En caso afirmativo se recuperan los datos anteriores en cuanto a formato, fecha de actualización y código devuelto por el servidor. En caso negativo se incluye la dirección y se pone a NULL el resto de campos.

14. Finalmente se recupera, si existe, la información de la tabla ESTA. Aquí simplemente se almacena el estado en que se encuentra cada documento en un momento determinado.

Una vez cargada la información, el siguiente paso es descargar a nuestro sistema el texto de los documentos y convertirlo a ASCII para ser analizado. De ello se encarga el módulo **get_papers**.

Uso:
get_papers -c <fichero> -[mhas]

-c <fichero> Es obligatorio especificar un fichero de configuración donde se recogen una serie de variables como: localización y nombre de la base de datos, directorio para almacenar el texto de los documentos, etc.

-m Número máximo de documentos a procesar. Por defecto sólo se procesan cien.

-a Descargar solamente documentos del archivo solicitado.

-s Descargar solamente documentos cuyo estado sea el especificado.

-h Mensaje de ayuda.

El módulo funciona de la siguiente manera:

1. En función de las opciones indicadas en la línea de comandos procede a consultar la base de datos para determinar qué documentos debe descargar.

2. Por cada documento seleccionado, se procede a descargar el primer fichero que lo componga. Si la descarga es satisfactoria se continua, de lo contrario se cambia el estado del documento y se pasa al siguiente fichero. Si no hay más ficheros se pasa al siguiente documento.

3. Se intenta determinar el formato del fichero que se ha descargado. Si está comprimido se descomprime y se vuelve a analizar el formato del fichero resultante. Si es PDF o PostScript continuamos, de lo contrario se cambia el estado del documento a *incompatibleformat* y se pasa al siguiente fichero.

4. Seguidamente se procede a convertir el fichero a formato ASCII utilizando el programa **pstotext**. Se comprueba si el resultado es efectivamente un fichero de texto o si se ha producido algún error. El tipo de error puede variar entre *pstotexterror* o *binaryfile*.

5. Si todo ha funcionado correctamente se procede a abrir el fichero ASCII y leer el contenido. Este contenido se comprueba que está en el orden correcto de impresión y no en sentido inverso. Si este es el caso se le da la vuelta.

6. Se maquilla el texto, por ejemplo borrando los números de páginas o arreglando las palabras truncadas al final de cada linea.

7. Se comprueba si el documento está en inglés buscando en el texto las palabras más comunes de este idioma. Si no está se pone el estado a *nonenglish* y se pasa a otro documento.

8. Se comprueba si el texto contiene una sección de bibliografía. Para ello se buscan en el mismo los términos que habitualmente suelen encabezar dicha sección. Si no aparece se cambia el estado del documento a *noreferences* y se pasa al siguiente fichero.

9. Finalmente se reescribe el fichero ASCII con los cambios que hemos realizado.

10. Si se ha llegado a este punto es porque el documento ha sido convertido satisfactoriamente y se dispone de un fichero texto correcto con una sección de referencias bibliográficas. Por lo tanto se procede a cambiar el estado del documento a *ready* y se pasa al siguiente documento. No es necesario seguir analizando más ficheros de este documento, si los hubiera, dado que el objetivo ya está conseguido.

5.2 Parsing

Este paso del proceso implica analizar la parte del documento que presumiblemente contiene las referencias, para identificar cada una de ellas, así como los diferentes elementos de que constan.

Sería de gran ayuda para llevar a cabo este proceso el conocer cuáles son los hábitos de los investigadores a la hora de confeccionar bibliografías. Aunque esto puede variar de una subdisciplina a otra o incluso por hábitos personales, se ha llevado a cabo el siguiente estudio.

Se han seleccionado al azar 36 documentos a los cuales se les ha aplicado el módulo de extracción de referencias. Del total de referencias existentes en esos documentos se ha eliminado aquellas que no representan ni artículos ni documentos de trabajo. En total han quedado 399 referencias.

Seguidamente se ha analizado el formato de las referencias con objeto de determinar las pautas de trabajo de los investigadores. Es claro que con una muestra tan pequeña no se puede obtener resultados concluyentes pero nos servirá para tener una visión global del proceso.

Lo primero que se ha analizado es el tipo de documento que los economistas citan. En el 82% de las referencias se han citado artículos, mientras en el 18% restante se han citado documentos de trabajo. En conclusión en nuestra muestra los investigadores prefieren citar artículos más que *papers*.

Un previsible problema que se planteará en la identificación de las series donde aparecen los *papers* será la falta de uniformidad a la hora de escribir los títulos de las mismas. Si bien en el caso de las revistas el título de la misma debe ser suficiente para identificarla, en el caso de los documentos de trabajo es

A D T R F	77%
A T R F D	16%
A T R F	3%
A T R A F	3%
A D T R	1%

Figura 5.2: A: Autor, D: Fecha, T: Título, R: Revista, F: Datos Fuente

mucho más difícil, ya que todas las series tienen títulos similares. Por ello es necesario acompañar el título con el nombre de la institución de la que procede. En ocasiones el nombre de la institución tampoco es suficiente y se deberá incluir el nombre de la institución más genérica. Por ejemplo en el caso de un departamento de Economía de una universidad, se deberá mencionar la universidad, el departamento y el título de la serie.

De las referencias que han citado un *working paper* se ha analizado si el autor ha incluido los datos necesarios para identificar la serie de la que procede. De esta forma se tiene que sólo en el 68% de los casos sería posible identificarla. En el resto, una cantidad bastante significativa, no sería posible ni siquiera para un humano identificar de qué serie se trata por falta de información.

En algunas disciplinas como por ejemplo la Física, es habitual referenciar los artículos citando el título abreviado de la revista. En esta disciplina se va mucho más lejos al suprimir el título del artículo, de forma que la referencia se limita a los autores, título abreviado de la revista y datos fuente. Se ha analizado cuál es la situación en Economía y se ve que lo habitual es citar el título completo. En el 93% de los casos el título estaba completo mientras el 7% restante estaba abreviado.

Otro aspecto que se ha estudiado es el formato utilizado para escribir el nombre de los autores. Aquí se pueden dar varias posibilidades: formato directo, inverso o una mezcla de ambos. Este último se utiliza cuando hay dos o más autores. El primero suele estar en formato inverso y el resto en formato directo. De las referencias analizadas el 12% estaban completamente en formato directo aunque hubiera más de un autor. El caso contrario, es decir, completamente en formato inverso, se encuentra en el 49% de los casos y una mezcla de los dos en el 38% restante.

Dentro de una referencia se pueden identificar una serie de elementos que pueden variar en número. Los básicos sin los cuales sería imposible identificar el documento que se está citando son: autor/es, título, año, revista o serie y datos fuente. La colocación de estos elementos en la referencia puede variar como se ve en la figura 5.2.

La inmensa mayoría de citas siguen el estilo de finido por el Manual de estilo de la Universidad de Chicago (1).

5.2.1 El módulo process_papers

Teniendo en cuenta estos resultados se ha creado el módulo **process_papers**. Su objeto es aislar cada una de las referencias que aparezcan en la sección de bibliografía del documento e identificar los elementos que la componen. De todos los posibles elementos que forman una referencia solamente se identifican los más importantes: título de la obra, fecha de publicación y autores. Aunque la situación ideal sería aquella en la que se identificaran todos y cada uno de los elementos, para nuestros objetivos veremos como estos tres son suficientes.

El módulo se usa de la siguiente forma:

process_papers -c <fichero> [m|a|h]

-c Fichero de configuración como en el caso anterior.

-m Número máximo de documentos a procesar. Por defecto 100.

-a Procesar documentos solamente del archivo especificado.

-h Mensaje de ayuda.

El algoritmo utilizado podría resumirse en lo siguiente:

1. Selecciona de la base de datos los documentos a procesar. Esto dependerá de la linea de comandos. Como se indica más arriba el módulo puede ejecutarse sobre los documentos de un archivo cuyo estado sea **ready** o sobre un determinado número de documentos. Si no se especifica nada se toman los 100 primeros documentos cuyo estado sea **ready**.
2. Por cada documento seleccionado:
   (a) Busca el subdirectorio bajo **Papers** donde deben estar los ficheros en formato ASCII conteniendo el documento, tal y como han sido generados por **get_papers**.
   (b) Si no lo encuentra pasa al siguiente documento.
   (c) Abre el fichero ASCII y pone el texto del documento en una variable.
   (d) Intenta determinar si existe una sección de bibliografía mediante la búsqueda de un patrón de texto que contenga las palabras: References o Bibliography seguidas de espacios y algún *tag* que marque un cambio de fuente.
   (e) Divide el texto en dos partes una, desde el comienzo hasta la posición donde se ha encontrado el patrón anterior, conteniendo el cuerpo del documento y otra desde esta posición hasta el final conteniendo la sección de bibliografía.
   (f) Borra de la bibliografía las marcas de fuentes.

(g) Intenta borrar también cualquier información que se acompañe al final del documento y que no forme parte de la bibliografía como por ejemplo índices, agradecimientos, figuras, etc.

(h) Graba el resultado a ficheros en disco. Así se crean: nombre.cit, nombre.bod y nombre.full.

(i) Comienza el trabajo sobre las citas.

(j) Quita los saltos de página que existan reemplazándolos por una nueva línea de **perl**.

(k) Cuenta las líneas de bibliografía.

(l) Intenta determinar como van enmarcadas las referencias. Es decir qué formato ha elegido el autor para establecer la clave de cada una. Se pueden dar tres casos, que vayan enmarcadas por corchetes, por paréntesis o simplemente incluyan un número correlativo. Para determinar que esquema siguen cuenta los paréntesis o corchetes abiertos que hay, así como los números precedidos de un salto de linea y pone en relación el resultado con el número de lineas.

(m) Si ha deducido que van enmarcadas por un número, entonces espera encontrar en cada clave una cadena de números. El inicio de cada referencia vendrá marcado entonces por el patrón: nueva línea, cero o más espacios, uno o más dígitos.

(n) Si ha deducido que se trata de paréntesis o corchetes, el contenido de la clave será una combinación de caracteres y números. El inicio de cada referencia seguirá el esquema: nueva linea, cero o más espacios, paréntesis o corchete abierto.

(o) Intenta determinar el formato en que se ha escrito el año para incluir estos datos en el patrón utilizado para marcar el comienzo y final de cada referencia.

(p) Una vez determinados el patrón de comienzo y el del final, se hace un análisis del texto para aislar las referencias.

(q) Por cada cita encontrada y mientras no se alcance un valor de error:

   i. Comprueba que la longitud de la referencia se encuentra entre los valores admitidos. Si el algoritmo de aislamiento anterior ha fallado podríamos encontrarnos con un texto demasiado grande o demasiado corto. Los valores de aceptación son modificables en el fichero de configuración.

   ii. Se eliminan los saltos de línea, tabuladores y espacios múltiples.

   iii. Intenta determinar cual es el título de la referencia.

      A. Si aparece una cadena entrecomillada la toma como título.

      B. De lo contrario comprueba si aparece un año en la primera parte de la referencia, en cuyo caso busca un patrón de texto compuesto de: año + palabra + palabra + todo lo que no sea

un signo de puntuación. Así todo lo que sigue al año lo toma como título.

C. Si por el contrario el año aparece en la parte final de la referencia el algoritmo se complica: por cada secuencia de tres palabras seguidas de cualquier cosa que no sea un signo de puntuación, busca la más larga y que no contenga una de las palabras consideradas vacías en el título de los documentos.

iv. Intenta determinar los autores del documento. Para ello calcula la posición en que comienza el título y deduce que la información sobre autores comprende desde esa posición hasta el comienzo de la referencia.

v. Elimina del autor los identificadores, años y signos de puntuación no habituales en este campo.

vi. Determina cual es el identificador de la referencia. Para ello busca al comienzo de la referencia cualquier texto entre corchetes o paréntesis. Si no encuentra nada pasa a buscar al comienzo uno o más dígitos seguidos de punto. Si esto también falla, por última instancia, iguala el identificador al autor del documento.

vii. Determina el año de publicación buscando una cadena de cuatro dígitos que comience por 19 o 20.

viii. Finalmente intenta determinar en qué lugar del texto se ha citado el documento, es decir, lo que hemos llamado más arriba contexto de la cita.

(r) Cuenta las referencias que ha encontrado y si el número está entre los límites máximo y mínimo permitidos se sigue adelante, de lo contrario, se informa que ha habido un error y se pasa al siguiente documento.

(s) Por cada referencia identificada se comprueba si ha encontrado satisfactoriamente el título y autores:

i. Un título se considera erróneo si esta vacío, comienza por " and " o por una inicial o por una palabra más inicial.

ii. Un autor es incorrecto cuando su longitud es superior al 25% de la longitud de la referencia.

(t) Si no ha habido errores se procede a guardar en el fichero de *log* la información que hemos encontrado: autor, título, año y posiciones.

(u) Se inserta en la tabla REFERENCIAS una nueva línea conteniendo la información.

(v) Se cambia el estado del documento indicando que ya está procesado.

Finalmente escribe en pantalla un resumen de los datos sobre los documentos que se han procesado: número, referencias detectadas, referencias correctas e incorrectas, etc.

5.3 Enlace o linking

El ultimo nivel del proceso es el que propiamente se encarga de realizar el enlace de las referencias. Como ya hemos mencionado solamente enlazamos documentos que estén disponibles en RePEc.

Además del enlace de documentos le corresponde a este nivel el generar los *templates* de citas que posteriormente serán utilizados por los servicios para proporcionar un valor añadido más a sus aplicaciones. En este sentido, este nivel se encargará de la distribución y difusión de resultados.

Uno de los objetivos de este trabajo es conseguir que toda la información sobre citas que se genere sea perfectamente integrada en el resto de RePEc para que sirva de apoyo a los servicios existentes y en definitiva contribuya a un mayor uso de los documentos. La forma de integrar nuestros resultados en la arquitectura distribuida de RePEc es crear un *template* donde se almacene toda la información relativa a citas y referencias que hemos extraído de los documentos. Para ello se ha diseñado el *template* ReDIF-Cite que se insertará en la especificación ReDIF en un futuro próximo. Este nuevo *template* está formado por los siguientes elementos:

- **Template-Type**: Es el campo con el que se debe iniciar el *template*. Siempre contendrá la cadena: **ReDIF-Cite 1.0**
- **Handle**: (Obligatorio) Es el identificador del *template*. Está formado por la cadena: **RePEc:cit**: seguida de los códigos de archivo, serie y número del documento que se está describiendo. Así el identificador de las citas del documento **RePEc:sur:surrec:9604** será: **RePEc:cit:sursurrec9604**.
- **Document-Handle**: (Obligatorio) Almacenará el handle del documento al cual pertenecen las referencias. Aunque este dato se podría inferir a partir del código enunciando anteriormente, se ha considerado oportuno por el momento, incluirlo de forma explícita.
- **Cite cluster**. Existirá un *cluster* de citas por cada referencia que se haya identificado. Al menos debe existir un *cluster* por cada documento no existiendo un límite máximo. El *cluster* estará formado por los siguientes campos:
  1. **Cite-Text**: (Obligatorio) Es la clave del cluster y por lo tanto es obligatorio que aparezca en primer lugar. Contiene el texto de la referencia tal y como ha sido extraído del documento por el módulo **process_papers**.
  2. **Cite-Title**: (Optativo) Es la cadena que presumiblemente contiene el título del documento referenciado.

3. **Cite-Author**: (Optativo) Contiene los posibles autores.
4. **Cite-Year**: (Optativo) Contiene el año de publicación del documento referenciado.
5. **Cite-Context**: (Optativo) Contiene la frase o frases con las cuales el autor del documento se ha referido a la obra citada. Si la cita se ha producido en varias ocasiones, cada contexto estará separado del resto por una barra vertical.
6. **Cite-Handle**: (Optativo) Este es un campo que solamente se completará si el documento referenciado está disponible a través de RePEc, en cuyo caso contendrá el *handle* del mismo.

En las secciones siguientes analizaremos detalladamente cómo se realiza el enlace de documentos y como se distribuyen los resultados.

5.3.1 El módulo link_papers

El objeto de este módulo es doble, en primer lugar genera un *template* de citas conteniendo la información que se ha extraído de cada documento. En segundo lugar intenta determinar si las referencias del documento analizado se refieren a algún documento disponible en RePEc. Si es ese el caso se añade una linea a la tabla CITA y se actualiza el campo handle de la tabla REFERENCIA con los datos del documento citado.

Se usa de la siguiente forma:

link_papers -c <fichero> [m|a|s|h]

-c Fichero de configuración como en el caso anterior.
-m Número máximo de documentos a procesar. Por defecto 100.
-a Procesar documentos solamente del archivo especificado.
-s Procesar documentos que se encuentren solamente en el estado especificado.
-h Mensaje de ayuda.

El algoritmo utilizado podría resumirse en lo siguiente:

1. Lee opciones desde la linea de comandos y selecciona los documentos a procesar.
2. Por cada documento a procesar, selecciona la lista de referencias identificadas en el mismo.
3. Por cada referencia de la lista:
   (a) Obtiene el título clave de la misma.

(b) Selecciona de la tabla DOCUMENTO aquellos registros que contengan en su título clave las mismas palabras que el título clave de la referencia.

(c) Para comprobar si algún documento de la lista obtenida es verdaderamente el documento citado, se realizan dos comprobaciones. En primer lugar se descarta aquel cuyo año de publicación no coincide y en segundo lugar se calcula la distancia de edición de los dos títulos clave. Si el valor absoluto obtenido es inferior a 8 se infiere que estamos ante el mismo documento, de lo contrario se rechaza la asociación.

(d) Si se ha encontrado una relación se procede a añadir una fila a la tabla CITA con los handles de los documentos citante y citado.

4. Se genera un *template* conteniendo la información sobre citas y se salva a disco. Por cada documento procesado se generará un fichero en el archivo **RePEc:cit** que llevará el nombre: <**archivo**><**series**><**número**>**.rdf**. De esta forma la información sobre citas del *template* cuyo handle fuera: **RePEc:sur:surrec:9604** se almacenará en un fichero: `ftp://netec.ier.hit-u.ac.jp/pub/RePEc/cit/cites/sursurrec9604.rdf`

5. Finalmente se cambia el estado del documento a **linked** y finaliza el proceso.

5.4 Distribución de resultados, el programa get_data.pl

Además de la distribución de datos utilizando *templates* de citas se ofrece un canal alternativo para que los servicios consulten directamente la base de datos MySQL a través de un *interfaz* web. El objetivo es diversificar la oferta de accesos a la información para que aquellos servicios que no dispongan de la infraestructura o del tiempo para procesar e incluir en sus sistemas los nuevos *templates*, puedan acceder a los datos de CitEc de la forma más rápida y cómoda posible. A diferencia de los *templates* en los que existirá inevitablemente un retraso entre la actualización del archivo **RePEc:cit** y su repercusión en los servicios, la consulta a través de esta *interfaz* se realiza en tiempo real.

Para ello se ha desarrollado un programa **CGI** en **perl** que funciona en el servidor del proyecto en la dirección `http://netec.ier.hit-u.ac.jp/adnetec-cgi-bin/get\_data.pl`. Este programa recibe alguna de las ordenes especificadas más adelante, ejecuta la búsqueda en la base de datos y devuelve una página HTML con los resultados. La posibilidad de devolver el resultado en formato XML queda abierta para una futura ampliación del proyecto.

El programa admite peticiones tanto utilizando el método POST como GET del protocolo HTTP. Las opciones que admite son tres:

1. **Handle**: Es el identificador del documento del cual se quiere obtener la información.
2. **Acción**: Es el tipo de información que se quiere obtener del documento. Las posibilidades son las descritas en la arquitectura de enlaces de referencias de (6):
   (a) **GetMyData**: Obtiene la información bibliográfica del documento especificado. Solamente devolverá el título, autores y año de publicación.
   (b) **GetMyCite**: Obtiene los identificadores de las obras que han citado al documento.
   (c) **GetRefList**: Obtiene una lista de las referencias del documento con enlaces si la obra referenciada se encontrara en RePEc.
   (d) **All**: Incluye todas las opciones anteriores.
3. **Format**: Esta opción solo se aplica a **GetRefList** e indica la cantidad de información que se proporcionará de cada referencia. Ofrece dos posibilidades:
   (a) **Sort**: Solamente se ofrece el contenido del campo **Cite-Text** del *template* de citas, esto es, el texto de la referencia tal cual fue identificado por **process_papers**.
   (b) **Long**: Incluye toda la información existente sobre la referencia, es decir, el contenido completo del *cluster* de esa referencia tal cual se ha especificado en la definición del *template* de citas.

Con objeto de integrar los datos en la mayor medida posible en los servicios que los requieran, la página HTML que devuelva el programa no incluirá ninguna alusión al proyecto CitEc ni ninguna información que induzca al usuario a abandonar el servicio que está utilizando. En la medida de lo posible los resultados deberían encauzarse hacia una nueva ventana del navegador del usuario que se debería cerrar en el momento que acabe de consultar las referencias.

Esta implementación es la que se ha seguido en el servicio WoPEc. En la figura 5.3 puede verse la descripción de un documento en la que se han insertado los enlaces (marcados con una flecha) para acceder a la información de CitEc. Esta página está disponible en la dirección: `http://netec.mcc.ac.uk/WoPEc/data/Papers/crecrefwp123.html`

Cuando se pica sobre el enlace marcado como "This paper's references" se abre una ventana nueva como la de la figura 5.4 con la lista de referencias del

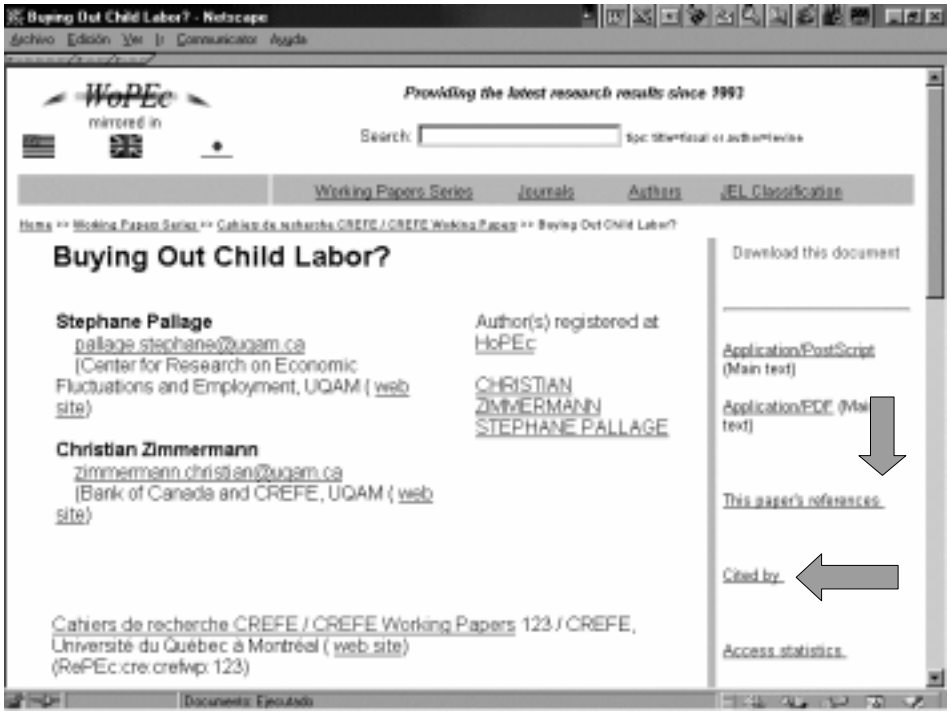


Figura 5.3: Ejemplo de implementación en WoPEc

documento. Igualmente si se solicitan las citas se abre una ventana como la de la figura 5.5. Ambas disponen de un icono en la parte inferior para cerrarla una vez que se ha consultado la información requerida. Alternativamente el usuario puede continuar navegando en este nivel de referencias o citas picando en los enlaces que se ofrecen. Las distintas ventanas que se vayan abriendo le darán una idea de los distintos niveles por los que ha pasado.

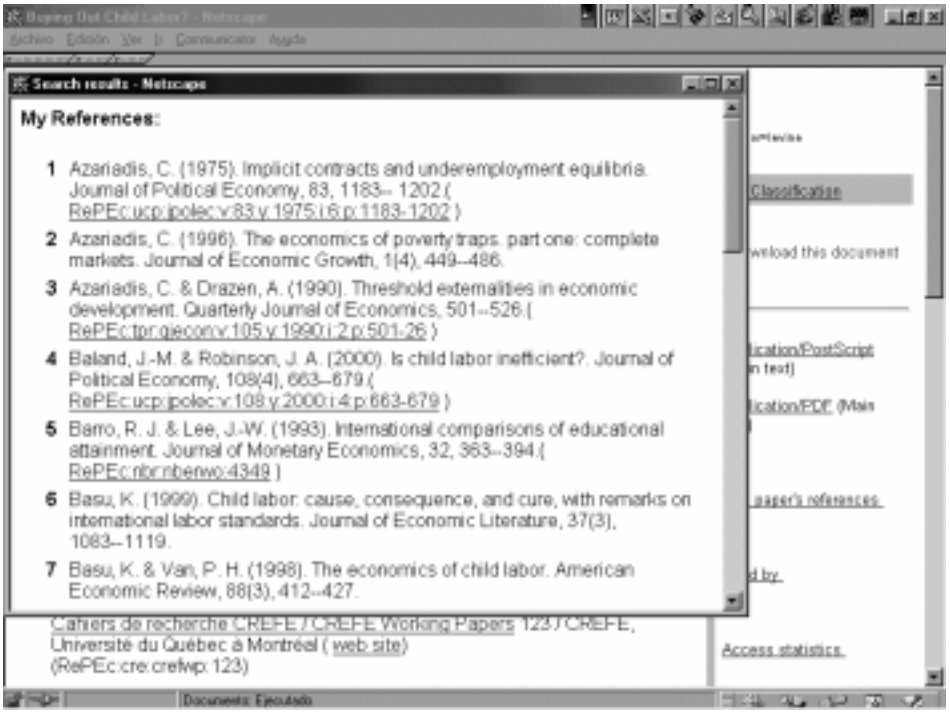


Figura 5.4: Listado de referencias

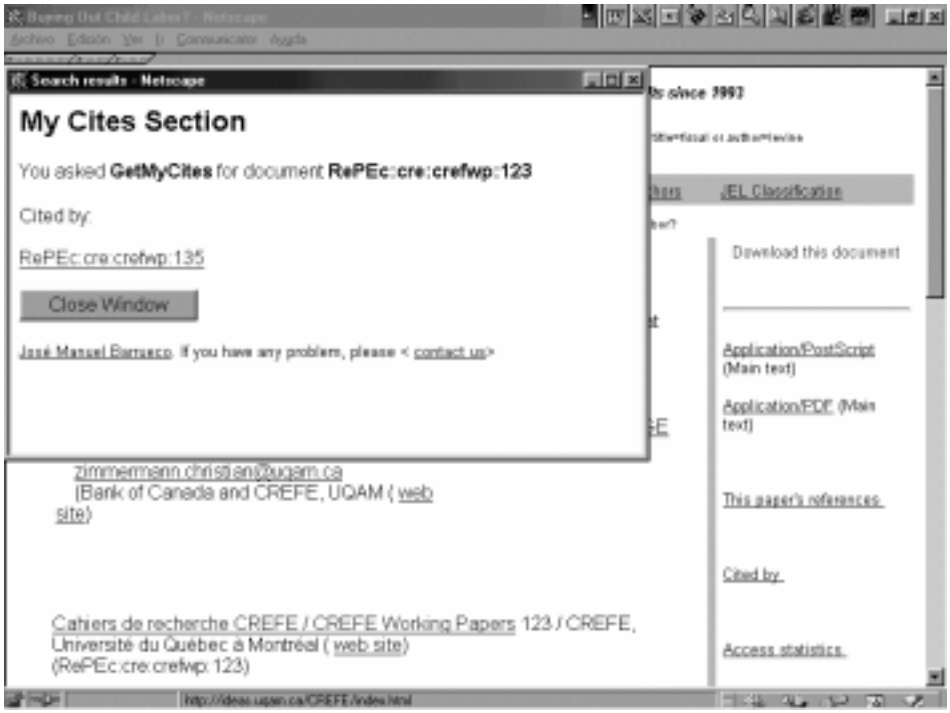


Figura 5.5: Listado de citas

Capítulo 6

Análisis de Resultados

En el presente capítulo se hace un análisis de los resultados obtenidos en el procesamiento inicial de todos los documentos disponibles en RePEc. El análisis se ha dividido en función de los distintos módulos de CitEc que intervienen en el proceso.

El estudio que se presenta corresponde a la situación de RePEc tal y como estaba a mediados de Diciembre de 2001. El procesado inicial de todos los documentos ha llevado más de un mes, por lo que en este momento aún no se han hecho actualizaciones. No obstante, la actualización es sencilla una vez procesada la masa inicial de documentos. El volumen de nuevos documentos oscila entre los 100 y 200 cada semana.

6.1 Resultado de find_files

El proceso inicial de lectura de toda la información, es decir, del entorno RePEc, se inició a las 04.00 horas del día 20 de Diciembre de 2001 y acabó a las 15.17 horas del mismo día. Se detectaron 58175 *templates* con información sobre documentos electrónicos. Los *templates* se distribuyeron entre 1106 series procedentes de 210 archivos de diferentes instituciones. El número de *templates* por archivo varía considerablemente desde aquellos que solamente ofrecen 1 o 2 documentos hasta archivos que ofrecen información procedente de varias instituciones. Es el caso, por ejemplo de **RePEc:hhs** para los documentos de centros suecos.

Un aspecto a estudiar es la calidad de los *templates*. En ellos es fundamental la existencia de la fecha de creación de los documentos. Como se ha descrito en la sección anterior este dato es utilizado para detectar si un documento citado se encuentra o no en RePEc. Por ello, si no se proporciona deberá ser inferido por

	Frecuencia	*Porcentaje*
Fecha de creación indicada por el archivo	28746	49%
Fecha de revisión indicada por el archivo	613	1%
Sin fecha indicada por el archivo	28816	50%
De ellos:		
No inferida por el sistema	3351	12%
Inferida por el sistema	25465	88%
De ellos:		
A partir del handle de los artículos	21849	86%
A partir del handle de los papers	3482	14%
A partir del campo Year	89	0%

Figura 6.1: Extracción de las fechas

el sistema siempre que sea posible. En los datos analizados solamente apareció explícitamente en la mitad, concretamente en 28746 (49%) documentos. Hay que resaltar también la escasa frecuencia de aparición del campo **Revision-Date**, utilizado para indicar sucesivas puestas al día de un mismo documento. Solamente 613 (1%) documentos la incluyeron. Esto podría llevar a afirmar que, en contra de lo que se presupone de las prepublicaciones (documentos de transición hacia publicaciones formales como artículos o libros), en muchos casos, al menos en Economía, la distribución de estos documentos constituye un fin en sí misma. No obstante sería necesario un estudio más exhaustivo para demostrar esa hipótesis.

A los 28816 *templates* sin fecha se les aplicó el algoritmo de búsqueda de fechas descrito. Los resultados pueden verse en la figura 6.1. Este algoritmo fue capaz de inferirla a partir del *handle* de los documentos en 25376 casos (3482 en *Papers* y 21894 en artículos). Es de destacar el hecho de que la práctica totalidad de los artículos incluyen la fecha como parte integrante del *handle*, sin duda porque los administradores consideran redundante incluirla dos veces: en el *handle* obligatorio y en **Creation-Date** opcional. No obstante esto mejoraría la lectura no automática de los datos. En otros 89 casos el sistema obtuvo la fecha a partir del campo **Year**.

El algoritmo ha sido capaz de inferir la fecha en 25465 documentos. Así, solamente 3351 documentos carecen de fecha. Para ellos se podrá obtener su lista de referencias pero será imposible determinar si han sido citados o no.

Finalmente, se detectaron 36294 autores. De ellos 2976 están registrados en HoPEc, con lo cual tenemos una total garantía de que la asociación con sus documentos publicados es correcta, independientemente de la forma del nombre que hayan utilizado para firmar.

Error	*Frecuencia*	*Porcentaje*
restricted	20878	36%
baddocument	6387	10%
downerror	2353	4%
available	28557	50%

Figura 6.2: Documentos no disponibles

6.2 Resultado de get_papers

Como vimos en la sección anterior, el módulo **get_papers** tiene una doble función: descargar los ficheros que componen cada documento y convertirlos a formato ASCII para que puedan ser procesados por el siguiente nivel o módulo.

El proceso de descarga de los documentos llevó mucho más tiempo que el anterior. Para agilizar la descarga se lanzaron al mismo tiempo varios procesos sobre archivos diferentes. Así en algunos momentos estuvieron funcionando hasta seis procesos diferentes de descarga y conversión. En total la descarga se inició el día 22 de Diciembre y finalizó el 4 de Enero.

El aspecto más importante de este paso es discernir qué documentos están realmente accesibles y por lo tanto podrán ser analizados y cuáles no. De los 58175 documentos que ofrecen inicialmente información sobre su disponibilidad electrónica, 28557 (el 50%) están realmente disponibles.

A medida que se van descargando, a cada documento se le asigna un código de estado. Según el gráfico 4.4 los estados que indican que un documento no se puede procesar son: *restricted, downerror* y *baddocument*. En la tabla 6.2 se muestra la proporción de documentos no disponibles por cada uno de estos motivos, junto con la proporción de documentos que sí lo están.

Destacar el elevado número de documentos que no son de acceso gratuito (*restricted*). La mayor parte de ellos son artículos procedentes de editoriales comerciales o del proyecto JSTOR. JSTOR es un proyecto comercial para escanear y poner en Internet los números antiguos de las revistas más importantes de cada disciplina.

En la figura 6.3 se observa la evolución del número de documentos disponibles frente al total de documentos a texto completo. Se puede ver cómo hasta comienzos de los años noventa no existen documentos disponibles ya que todos los fondos pertenecen a JSTOR o a editoriales que ponen en Internet sus artículos de forma retrospectiva. Es la llegada de Internet la que facilita la distribución gratuita de documentos. Lo que más llama la atención es el escaso número de documentos electrónicos añadidos durante el año 2001. Frente a los 11.000 del año anterior, éste no se han alcanzado ni siquiera los 6.000. Esto puede deberse al hecho de que los datos han sido tomados antes de finalizar el

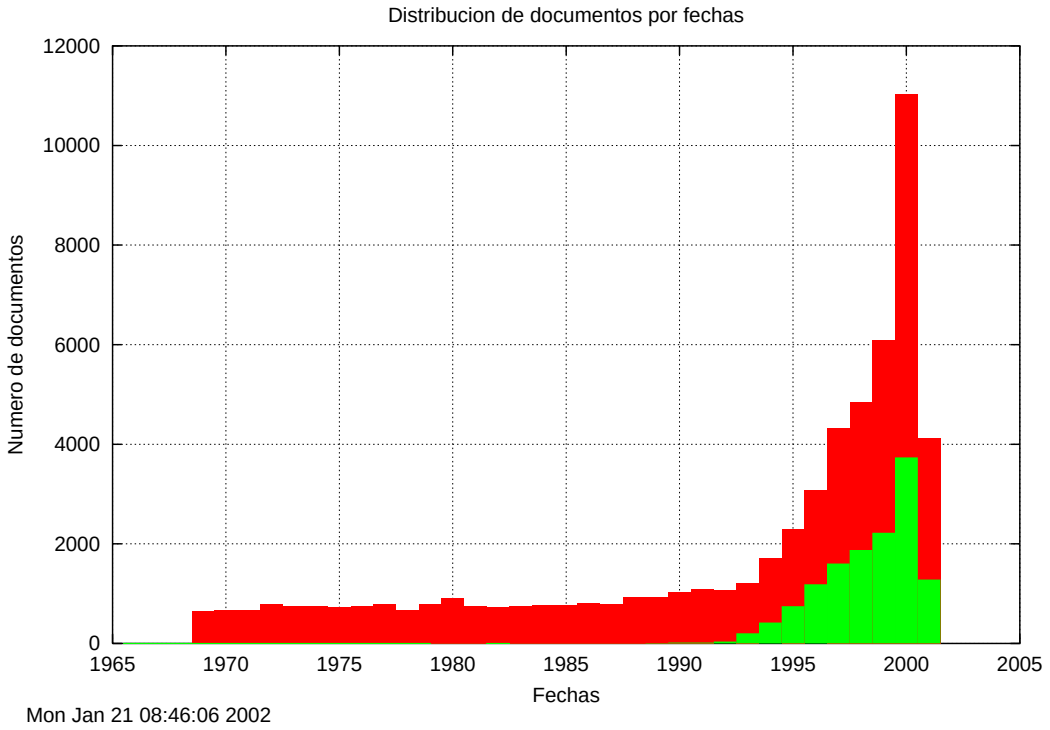


Figura 6.3: Distribución de documentos por años

año y hay archivos que acostumbran a actualizar su información solamente a finales de año. No obstante, aunque la tendencia es al alza, será difícil igualar la marca de 2000.

Posteriormente, se procedió a la conversión de los 28557 documentos a formato ASCII. De acuerdo al gráfico 4.4 en este paso se podrían dar los siguientes errores: *incompatibleformat* cuando el fichero no sea de tipo PDF o PostScript, *pstotexterror* cuando el programa de conversión falle, *nonenglish* cuando el documento esté en una lengua diferente del inglés, y *noreferences* cuando se haya convertido a texto pero haya sido imposible encontrar una sección de referencias. En la práctica este último error está relacionado con el segundo, ya que en un elevado porcentaje de casos CitEc no encuentra referencias porque éstas están al final del documento y el proceso de conversión ha fallado antes de llegar a ellas. El resultado es un fichero de texto correcto pero incompleto.

De los 28557 documentos que han llegado a este punto, 15522 lo han pasado satisfactoriamente. En la tabla 6.4 puede verse la distribución de errores.

Hay que destacar aquí el elevado número de documentos que no ha sido posible convertir satisfactoriamente. Un punto a tener en cuenta en sucesivas actualizaciones del sistema será probar nuevos programas de conversión de PDF a ASCII, que vayan apareciendo.

Error	*Frecuencia*	*Porcentaje*
pstotexterror	6931	24%
nonenglish	963	3%
noreferences	3476	12%
incompatibleformat	1665	6%
convertidos	15522	55%

Figura 6.4: Errores en la conversión a ASCII

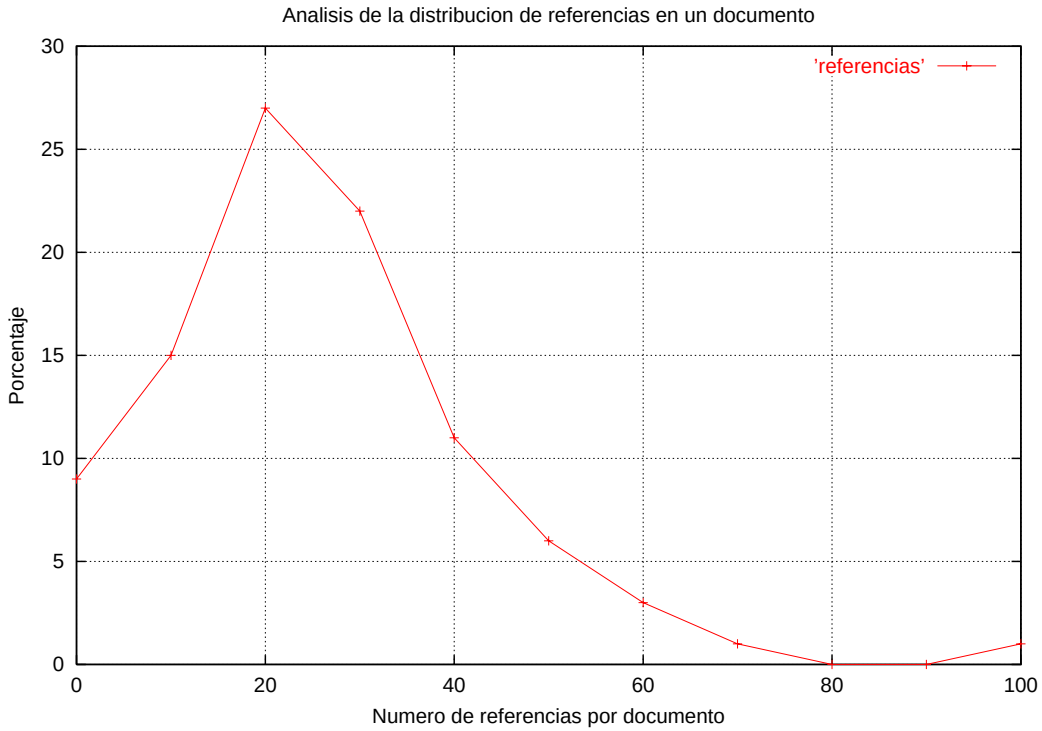


Figura 6.5: Porcentaje de referencias por cada documento

6.3 Resultados de process_papers

process_papers es el módulo encargado de analizar los ficheros ASCII producidos por el módulo anterior. Su objeto es aislar cada una de las referencias que aparecen en el documento e identificar los elementos que las componen (autor/es, título, año, contexto).

15522 documentos fueron procesados para localizar sus referencias bibliográficas e identificar cada uno de los elementos que las componen. En este paso el único error que se puede producir para que un documento sea excluido del análisis, en su totalidad, es que el sistema encuentre un número de referencias que no esté entre los límites especificados. Dichos límites son configurables y se han fijado entre 1 y 70.

Error	*Frecuencia*	*Porcentaje*
wrongnumberofreferences	2081	13%
analizados	13446	87%

Figura 6.6: Documentos excluidos al identificar las referencias

	Frecuencia	*Porcentaje*
Total referencias identificadas	293069	
Referencias correctas	283932	97%
Referencias incorrectas	9137	3%
De ellas:		
Errores en el título	7921	87%
Errores en el autor	1216	13%

Figura 6.7: Causas de exclusión de referencias

Estos límites se han establecido después de analizar la distribución de referencias que se puede ver en el gráfico 6.5. Como se puede apreciar, casi el 90% de los documentos quedan englobados en estos límites. A ellos hay que añadir un 9% de documentos en el que el sistema, a pesar de haber identificado que existe una sección de bibliografía, ha sido incapaz de aislar convenientemente las referencias. No ha sido capaz de identificar dónde comienza y acaba cada una. El mismo problema se produce en el extremo opuesto, el 1% restante ha encontrato más de 80 referencias, lo que también se interpreta como error. De esta forma el análisis queda como se ve en la tabla 6.6

A partir de este momento los errores que se pueden producir son en la localización de los elementos que componen cada referencia. Hay tres elementos clave que deben ser encontrados en todas ellas: el año de publicación, el autor o autores y el título del documento. Si el proceso falla en identificar alguno, la referencia es descartada. Los resultados pueden verse en la tabla 6.7. En total se identificaron 293069 referencias en 13039 documentos. Ello nos da una media de 22.5 referencias por cada uno, que se ajusta a la distribución vista en la figura 6.5. De ellas CitEc fue capaz de identificar correctamente 283932 (97%) y falló solamente en 9137 (3%) casos. Entre las causas de error tenemos fallos en identificar el título en 7921 referencias y en identificar el autor en 1216 casos.

De las 283932 referencias identificadas, 52352 (el 18.4%) representaban a un documento cuyo texto electrónico está disponible en RePEc y por lo tanto se puede establecer un enlace entre ambos.

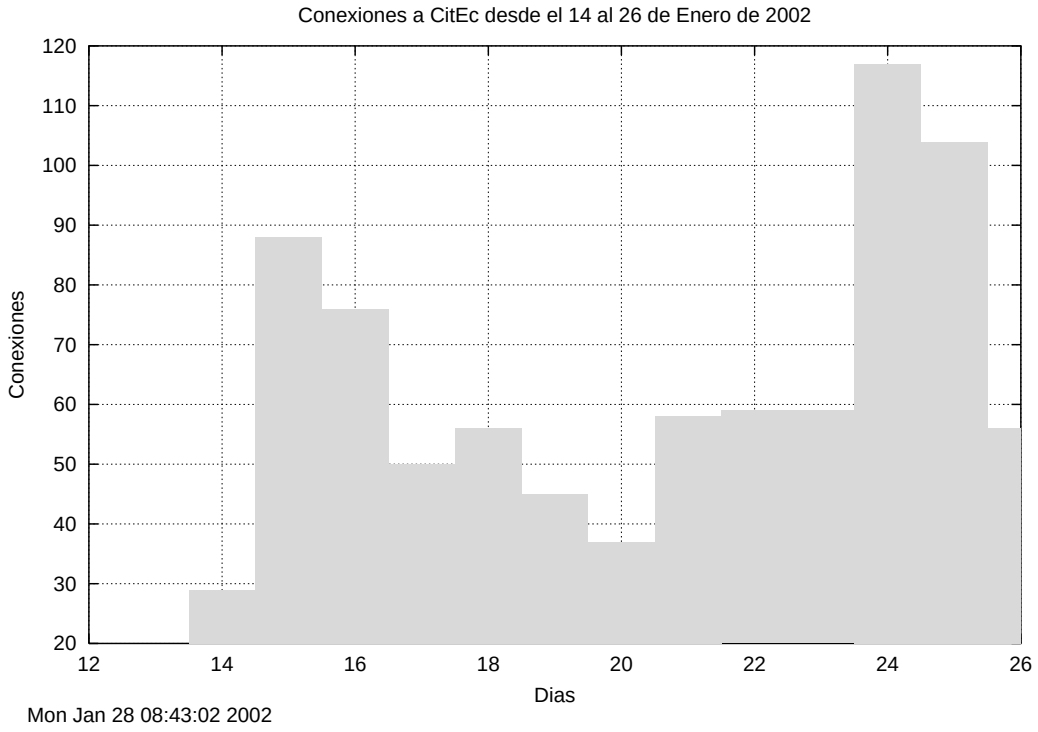


Figura 6.8: Conexiones a CitEc

6.4 Resultados de get_data.pl

El paso final que se ha de evaluar es la consulta de los datos extraídos. Como se ha indicado en el capítulo anterior la consulta se puede realiza a través de una pasarela CGI llamada **get_data.pl**. Aunque estrictamente hablando este módulo no forma parte del agente se incluye aquí para dar una idea de la utilización que se está realizando de este trabajo.

Todas las consultas que se han realizado a CitEc están registradas en su fichero de *logs*. Para este análisis se ha extraído de este fichero los datos de los días 14 a 26 de Enero de 2002. Durante estos 12 días se han recibido 834 peticiones, lo que nos da una media de 69.5 peticiones por día. En la figura 6.8 se puede ver gráficamente la evolución por cada día. Excluyendo los días 19 y 20 de Enero que fueron fin de semana, se aprecia que se mantiene una tendencia estable desde el primer momento. Las únicas excepciones son los días 24 y 25 de Enero que hay el doble de conexiones. Ello está motivado porque esos días CitEc se anunció públicamente entre los coordinadores de RePEc.

La opción más solicitada ha sido la petición de referencias de un documento. Se ha solicitado en 549 casos, es decir, el 66% de los mismos. En segundo lugar está la solicitud de citas con 204 peticiones, el 24%. Y finalmente una

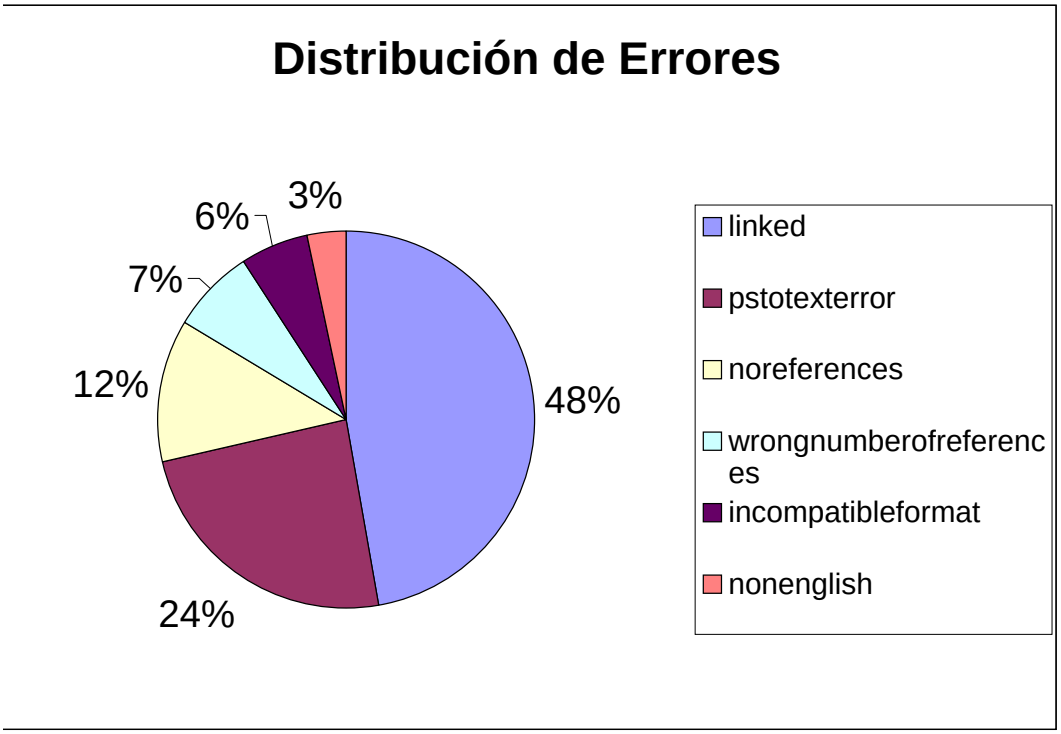


Figura 6.9: Distribución total de documentos según errores

combinación de ambas en 81 casos, 10% de las peticiones.

Los archivos más consultados han sido **RePEc:wop** y **RePEc:dgr** con 104 y 101 consultas respectivamente. Les sigue **RePEc:wpa** con 95. En total se ha consultado información de 89 archivos diferentes, lo que da una media de 9.4 peticiones por archivo.

Las peticiones se han recibido desde 562 direcciones IP diferentes. Ello nos da una media de 1,5 peticiones por máquina. Las máquinas con mayores peticiones son: 132.66.16.12 y 147.210.108.196 con 15 y 16 peticiones respectivamente.

6.5 Resultados globales

Como conclusión de este capítulo se ofrece un resumen de los datos más importantes.

De los 28557 documentos que se han analizado se ha concluido el proceso satisfactoriamente en el 48% de los casos. En el resto se han producido los errores que se ven en el gráfico 6.9. Destacar que el error más importante se ha producido en la conversión de los ficheros PDF o PostScript a ASCII.

Archivo	ind	NV/AV	linke	incom	noref	nonen	pstot	wrong
RePEc:san	0.91	0 / 12	11 (91)			0	1 (8)	
RePEc:boe	0.88	0 / 84	74 (88)			0	3 (3)	7 (8)
RePEc:ukc	0.86	0 / 67	58 (86)		2 (2)	0	5 (7)	2 (2)
RePEc:bef	0.83	1 / 11	10 (90)			0	1 (9)	
RePEc:upj	0.83	0 / 60	50 (83)	1 (1)	3 (5)	0	2 (3)	4 (6)
RePEc:hwe	0.78	2 / 59	48 (81)		6 (10)	0	0	5 (8)
RePEc:ise	0.76	0 / 13	10 (76)			0	2 (15)	1 (7)
RePEc:dal	0.75	0 / 12	9 (75)		1 (8)	0	0	2 (16)
RePEc:vir	0.74	0 / 27	20 (74)			2 (7)	3 (11)	2 (7)
RePEc:els	0.73	2 / 39	30 (76)		2 (5)	0	3 (7)	4 (10)
RePEc:irs	0.72	0 / 18	13 (72)		2 (11)	0	2 (11)	1 (5)
RePEc:aal	0.71	0 / 82	59 (71)		11 (13)	0	0	12 (14)
RePEc:lec	0.70	0 / 65	46 (70)		7 (10)	0	2 (3)	10 (15)
RePEc:anu	0.69	4 / 19	16 (84)	1 (5)	1 (5)	0	1 (5)	
RePEc:ham	0.69	1 / 22	16 (72)	5 (22)		0	0	1 (4)
RePEc:sbu	0.68	0 / 16	11 (68)		1 (6)	0	0	4 (25)
RePEc:cri	0.68	0 / 16	11 (68)		2 (12)	0	1 (6)	2 (12)
RePEc:tor	0.66	10 / 135	96 (71)	3 (2)	9 (6)	4 (2)	11 (8)	12 (8)
RePEc:cre	0.66	0 / 135	90 (66)	3 (2)	22 (16)	1	6 (4)	13 (9)
RePEc:bon	0.66	5 / 158	112 (69)	5 (3)	12 (7)	4 (2)	17 (10)	7 (4)

Figura 6.10: Archivos con mayor número de documentos enlazados

De los documentos analizados se han encontrado 283932 referencias. De ellas 52352 son a documentos existentes en RePEc.

Finalmente, se ha definido un *indicador de éxito* que trata de evaluar la calidad de los datos aportados por los archivos en función del mayor o menor número de documentos que han superado todos los módulos. Este indicador se define como:

$$ind = \frac{documentoslinked}{totaldedocumentos}$$

Es decir, número de documentos enlazados partido por el total de documentos que ofrece el archivo. Esto da un rango de resultados entre 1, cuando todos los documentos han sido enlazados, y 0 cuando ninguno de los documentos en el archivo han sido enlazados. En la figura 6.10 se detallan los 20 archivos que han alcanzado una mejor puntuación. Se indica el número de documentos disponibles y no disponibles. El total de documentos es la suma de ambas cantidades. También se especifican, cuantos han sido enlazados correctamente y cuantos han fallado en alguno de los errores indicados. Las cifras entre paréntesis indican porcentajes.

Los archivos resaltados pertenecen a instituciones que colaboran en la gestión de RePEc. No son archivos normales en el sentido de que están estrechamente vinculados a la iniciativa RePEc, colaboran en el desarrollo de software y es-

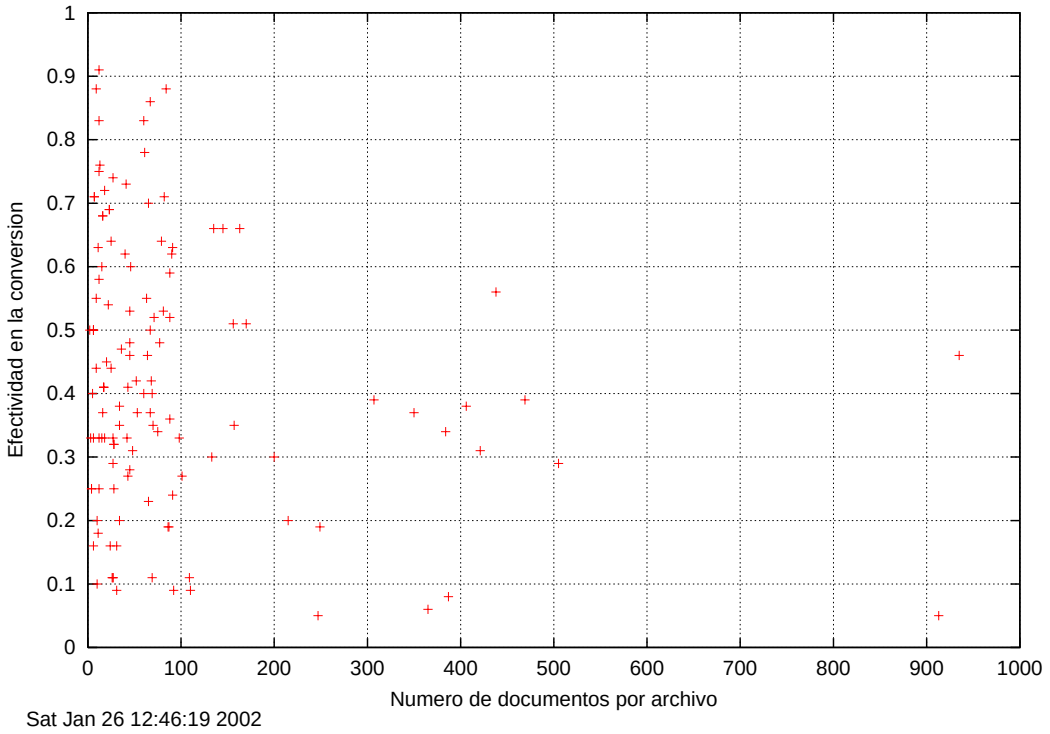


Figura 6.11: Distribución de indicadores según numero de documentos por archivo

pecificaciones, etc. Por tanto, sería de esperar que estas instituciones obtuvieran una puntuación más elevada en el *ranking*.

Destacar que los archivos que ocupan las primeras posiciones son relativamente pequeños. En la figura 6.11 se aprecia la relación entre el número de documentos por archivo y su indicador de éxito. Se puede apreciar que entre ambos no existe una correlación definida. Pensamos que este hecho no se debe a que se produzcan necesariamente más errores cuanto más grande sea el archivo. En concreto pensamos que se debe al software con el que se crean los ficheros PDF. Cuando un archivo utiliza un programa que genera PDFs convertibles a texto, todos sus documentos, independientemente del número, pueden ser convertidos. Cuando utiliza un software que no produce buenos PDFs, los resultados son negativos aunque contenga un reducido número de documentos.

Capítulo 7

Conclusiones

A la vista de los datos descritos en el capítulo anterior se puede concluir que el objetivo de este trabajo tal como se planteó en la introducción, es decir, **experimentar la posibilidad de llevar a cabo un enlace de citas entre los documentos disponibles en RePEc**, se ha alcanzado satisfactoriamente.

A lo largo de este trabajo se ha descrito un sistema que hace posible la extracción automática de información sobre citas de los documentos existentes en una biblioteca digital. Se ha presentado un sistema que difiere de iniciativas similares en el sentido que se ha trabajado en un entorno distribuido, con documentos procedentes de instituciones de todo el mundo y con una gran variedad de formatos.

Se ha diseñado un agente capaz de determinar cuando ese entorno cambia y actuar en consecuencia. Su actuación se concreta en la recuperación del texto de los nuevos documentos desde cualquier parte de Internet, extracción de sus referencias bibliográficas cuando existen y establecimiento de enlaces con los documentos citados que se encuentran también en nuestra biblioteca. Finalmente también es capaz de modificar el entorno al codificar la información obtenida en unos *templates* especiales y ponerla a disposición de aquellos servicios RePEc que deseen utilizarla para mejorar sus aplicaciones. Por el momento los resultados de CitEc están siendo implementados en tres servicios RePEc:

- WoPEc (Inglaterra)
  `http://netec.mcc.ac.uk/WoPEc/data/Papers/crecrefwp123.html`
- Socionet (Rusia)
  `http://socionet.ru/RuPEc/xml/cre/paper-crefwp/crecrefwp123.xml`
- Biblioteca de Ciéncies Socials, Universitat de València
  `http://www.uv.es/bibsoc/GM/data/Papers/upfupfgen521.html`

En el capítulo anterior se ha demostrado hasta qué punto este sistema se ha mostrado efectivo. No obstante quedan algunos puntos que inciden directamente en la calidad de los resultados pero que quedan fuera del alcance de este trabajo. Por ejemplo, la colaboración de los administradores de archivos es un punto crucial para el buen funcionamiento del sistema. Lo fundamental es que proporcionen unas buenas descripciones de sus documentos. Entre otras cosas, deben asegurarse que las direcciones de los textos de los documentos son correctas y deben proporcionar información sobre las fechas de creación de los documentos.

Como se ha demostrado a lo largo del trabajo, la investigación en enlace de referencias entre documentos científicos es una de las áreas de mayor expansión dentro de las bibliotecas digitales. El interés es tal que los resultados de este proyecto han sido aceptados para ser presentados en varias conferencias internacionales: (ISKO 7(3) e ICEIS 2002(4)).

7.1 Trabajos futuros

El trabajo descrito en este proyecto no es un fin en si mismo sino que deja las puertas abiertas a nuevos desarrollos. Lo que sigue a continuación es una relación de propuestas de mejora del sistema CitEc.

- El módulo con mayor índice de errores es, sin duda, **process_papers**, en el cual ha fallado la conversión de PDF/PS a ASCII. En este sentido son necesarias mejores herramientas de conversión.
- Enlaces a documentos externos a RePEc. El 18.4% de referencias enlazadas que hemos visto en los resultados se refiere exclusivamente a documentos disponibles en RePEc. Esta tasa se podría incrementar sustancialmente si se extendiera el ámbito a cualquier documento disponible en formato electrónico. Aunque todo lo que sea salir de RePEc sería chocar con barreras económicas y financieras ya que se tendría que recurrir a la industria editorial, el primer candidato a probar estos enlaces sería CrossRef.
- Análisis de los documentos que están restringidos. Una vez que tenemos un sistema funcionando podríamos utilizarlo para convencer a las editoriales para que nos faciliten una copia de los documentos restringidos con objeto de analizarlos. De ello se podrían beneficiar tanto CitEc como las propias editoriales. De la misma forma que nos han suministrado los metadatos, sería posible obtener el texto completo de los documentos.
- Mejora de los algoritmos de búsqueda de referencias. Cualquier intento de enlazar documentos externos a RePEc pasa por una mejora de los

algoritmos de identificación de los elementos de una referencia. Por el momento CitEc no es capaz de encontrar los títulos de las revistas. Esta es una mejora necesaria. Sin ella no se puede pensar en utilizar CrossRef o cualquier otra fuente externa.

- Análisis de nuevas versiones de un documento. Al menos la mitad de los documentos disponibles en RePEc son trabajos en curso. Es de esperar que cambien a lo largo del tiempo. En estos momentos CitEc no tiene en cuenta esta característica. Simplemente descarga y analiza los documentos a medida que sus metadatos es incluida en el archivo que los distribuye. Una vez analizado, el proceso acaba. El sistema debería ser capaz de monitorizar cada documento y alertar cuando detecte que se han producido cambios en la lista de referencias con objeto de actualizar la base de datos. Con ello se aumentaría la reactividad del agente.

- Extraer los metadatos del documento para compararlos con los *templates.* Todos los proyectos similares de enlaces de referencias tratan de encontrar la propia información bibliográfica de cada documento. En nuestro caso no es necesario ya que disponemos de metadatos de calidad. No obstante sería interesante incluir esta función en CitEc con objeto de calibrar el grado de exactitud de la información que proporcionan los archivos.

- Establecer un sistema de correcciones para que los autores o usuarios puedan corregir los errores del sistema. Dado que el sistema nunca será fiable al 100% sería de gran ayuda la colaboración tanto de usuarios como de los propios autores a la hora de corregir errores.

- Estudio bibliométrico de las citas. Una vez finalizada la operación de enlace, CitEc cuenta con una extensa base de datos de citas sobre la disciplina que podría ser utilizada para la elaboración de estudios bibliométricos, mapas conceptuales de la disciplina, etc.

Referencias

[1] *The Chicago Manual of Style*. The University of Chicago Press, 1993.

[2] Arms, W., Bergmark, D., and Lagoze, C. An architecture for reference linking. Tech. Rep. CSTR 2000-1820, Cornell Digital Library Research Group, 2000.

[3] Barrueco, J. M., and Julian, V. Reference linking in economics: the citec project. In *ISKO 2002 International Conference* (2002).

[4] Barrueco, J. M., and Krichel, T. Automatic extraction of citation data in a large digital library. In *4th International Conference on Enterprise Information Systems. New Developments in Digital Libraries* (2002).

[5] Bergmark, D. Automatic extraction of reference linking information from onlin documents. Tech. Rep. CSTR 2000-1821, Cornell Digital Library Research Group, 2000.

[6] Bergmark, D., and Lagoze, C. Reference linking the web's scholarly papers. Tech. rep., Cornell Digital Library Research Group, 2000.

[7] Cameron, R. D. A universal citation database as a catalyst for reform in scholarly communication. *First Monday 2*, 4 (1997).

[8] Claivaz, J.-B., Meur, J.-Y. L., and Robinson, N. From fulltext documents to structured citations: Cern's automated solution. *High Energy Physics Libraries Webzine*, 5 (2001). http://library.cern.ch/HEPLW/5/papers/2.

[9] de Sompel, H. V., and Lagoze, C. The santa fe convention of the open archives initiative. *D-Lib Magazine 6*, 2 (2000).

[10] Garfield, E. *Science*, 122 (1955), 108–111. http://www.garfield.library.upenn.edu/papers/science_v122(3159)p108y1955.html.

[11] Krcihel, T. Guildford protocol. ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/

[12] Krichel, T. Redif (research documents information format). ftp://netec.mcc.ac.uk/pub/NetEc/RePEc/all/root/docu/redif_1.html.

[13] Krichel, T., and Warner, S. Design of a metadata framework to support scholarly communication. In *International Conference*

on Dublin Core and Metadata Applications (October 24-26 2001). http://openlib.org/home/krichel/papers/kanda.a4.pdf.

[14] Lawrence, S., Bollacker, K., and Giles, C. L. Indexing and retrieval of scientific literature. In *Eighth International Conference on Information and Knowledge Management, CIKM99* (1999), pp. 139–146.

[15] Maltrás, B. *Los indicadores bibliométricos en el estudio de la ciencia: fundamentos conceptuales y aplicación en política científica*. PhD thesis, Universidad de Salamanca, Facultad de Filosofía, 1996.

[16] Piñero, J. M. L., and Terrada, M. L. Los indicadores bibliométricos y la evaluación de la actividad medico-científica (ii): la comunicación científica en las distintas áreas de las ciencias médicas. *Medicina clínica*, 98 (1992), 101–106.

[17] Robinson, N. A comparison of utilities for converting from postscript or portable document format to text. CERN-OPEN-2001-065.

[18] Small, H. The synthesis of specialty narratives from co-citation cluster. *Journal of the American Society for Information Science 37*, 3 (1986), 97–110.

[19] Wall, L., Christiansen, T., and Schwartz, R. L. *Programming Perl*. O'Reilly, 1997.