PROTOCOLOS DE CONVERSION PARA EL INTERCAMBIO DE INFORMACION ENTRE BASES DE DATOS BIBLIOGRAFICAS HETEROGENEAS.

JOSE ABASOLO PRIETO
Grupo de Investigación SINBAD
Depto. de Ingeniería de Sistemas
Universidad de los Andes
Apartado aéreo 4976
Bogotá-Colombia.

RESUMEN:

Se presenta una solución al problema de cómo brindar transparencia hacia la heterogeneidad a los usuarios de una Base de Datos distribuida heterogénea con información bibliográfica. La solución resuelve el problema de traducción entre diferentes representaciones de datos adoptando un formato puente a través del cual debe pasar toda conversión.

1. DESCRIPCION GENERAL.

El intercambio de información entre Bases de Datos bibliográficas automatizadas que no siguen los mismos estándares de clasificación y que son manejadas por diversos paquetes de 'software' crea la necesidad de conversión de datos entre los diferentes formatos y estándares utilizados.

Una alternativa de solución al problema de la heterogeneidad sería desarrollar convertidores entre cada par de Bases de Datos. Esta solución se descarta por su complejidad: a cada nueva Base de Datos que se quiera integrar al sistema habría que desarrollarle convertidores con cada una de las Bases ya existentes.

La alternativa correcta consiste en adoptar un **Modelo Global** (formato y estándar de clasificación) que sirva de puente entre los otros. Bajo este esquema solo se requiere un convertidor entre el Modelo Global y cada Base de Datos perteneciente al sistema [CERI 84].

El presente artículo describe la situación actual y la proyección futura del proyecto **"Protocolos de Traducción bibliográfica desde/hacia el Formato Común de Comunicación Colombiano (FCCC)"** [ABAS 87] en el cual se han desarrollado protocolos de conversión que utilizan un Modelo Global y permiten el intercambio de información entre Bases de Datos bibliográficas automatizadas en microcomputadores IBM PC XT/AT y manejadas por uno cualquiera de los siguientes tres paquetes de 'software': ISIS, SCIMATE ó SCIB. Dichos protocolos han sido desarrollados en el marco de un convenio entre la Universidad de los Andes y el Fondo Colombiano de Investigaciones Científicas y Proyectos Especiales "Francisco José de Caldas" (Colciencias).

Una característica importante de los protocolos es que los programas que los soportan constan de un gran núcleo, **independiente de los paquetes de 'software'** que se estén utilizando, más unos componentes muy sencillos específicos a cada uno de esos paquetes. La integración al sistema de intercambio de Bases de Datos bibliográficas manejadas por paquetes distintos a los tres mencionados antes - por ejemplo NOTIS, LIBRUNAM, etc.- requerirá únicamente del desarrollo de los componentes que tengan en cuenta los detalles específicos del paquete en cuestión. Dichos componentes representan menos del 10% del sistema total, y generalmente son muy sencillos de desarrollar.

Otro aspecto importante de los protocolos es que son **independientes del tipo de información manejada**: en lugar de registros bibliográficos se podrían manejar historias clínicas, información de catastro, inventario de recursos naturales, etc., siempre y cuando se disponga de un Modelo Global para cada uno de esos tipos de información. Este aspecto asegura un gran campo de aplicación del proyecto "FCCC".

Este artículo se ha organizado de la siguiente manera:

En la sección 2 se describen las características principales, relacionadas con el problema de intercambio, de los tres paquetes de 'software' contemplados inicialmente en el proyecto "FCCC": SCIB, SCIMATE e ISIS.

La sección 3 presenta el FCCC (Formato Común de Comunicación Colombiano), utilizado como Modelo y formato puente para facilitar las conversiones.

En la sección 4 se describe un formato de entrada/salida a/de los convertidores, llamado Formato Interno, que fué necesario definir para poder independizar esos convertidores de los paquetes de 'software' utilizados.

La sección 5 explica en forma general el procedimiento que debe seguirse para poder integrar una nueva Base de Datos al sistema.

En la sección 6 se presentan los protocolos actuales de intercambio de información utilizando diskettes como sistema de comunicación.

En las secciones 7 y 8 se presentan las extensiones previstas para los protocolos así como algunas conclusiones.

2.CARACTERISTICAS DE LOS MANEJADORES DE LAS BASES DE DATOS LOCALES.

La clasificación, estructuración y manipulación de los datos de una Base bibliográfica dependen del paquete de 'software' escogido para manejarla. En las siguientes secciones se describen los aspectos clasificación y estructuración para los tres paquetes incluidos actualmente en el proyecto "FCCC".

2.1. SCIB.

Todas las Bases de Datos bibliográficas administradas por el paquete SCIB (Sistema Colombiano de Información Bibliográfica) utilizan como estándar de clasificación el formato MARC (Machine Readable Cataloging) diseñado por la Biblioteca del Congreso de los Estados Unidos, con algunas modificaciones. Bajo dicho forma-

to los elementos de información se agrupan en campos de longitud fija ó variable. Cada campo tiene una etiqueta que identifica el tipo de elemento de información que aparece en dicho campo. Un campo puede ser repetible y contener a su vez subcampos. Cada subcampo es identificado por un conjunto de caracteres especiales, y puede ser repetible.

La etiqueta de un campo se complementa con dos indicadores que en algunos casos sirven para especificar en más detalle el tipo de información contenida en el campo, y en otros casos se utilizan para facilitar procesos administrativos tales como producción de fichas catalográficas, indización, etc.

Ejemplo: en SCIB, el título de una obra se codifica en el campo con etiqueta 245. Dicho campo puede contener dos subcampos que corresponden a subtítulo y mención de autoría. El subcampo subtítulo se identifica con el caracter especial ':' mientras que el subcampo mención de autoría se identifica con el caracter '/' si es la primera vez que aparece y con ';' para los demás. Un ejemplo de cómo se codificaría un campo título es:

24510EDGAR WALLACE : THE MAN WHO MADE HIS NAME

El 10 que viene después de la etiqueta 245 representa los valores que toman los indicadores en este caso específico. No hay menciones de autoría asociadas a este título.

SCIB permite importar un archivo de texto para integrarlo a una Base de Datos local cuando dicho archivo de texto se ajusta a un formato específico que refleja el modelo SCIB. Igualmente una Base de Datos SCIB puede exportar un archivo de texto en ese formato.

2.2 SCIMATE.

Las Bases de Datos manejadas por SCIMATE no están obligadas a seguir un estándar de clasificación específico. El formato al que se deben ajustar es el siguiente:

Para una misma Base de Datos se pueden definir varias máscaras ('templates') para los registros. Cada registro incluye un identificador de la máscara que utiliza. Un registro es una agrupación de campos no repetibles de longitud variable. El máximo número de campos en un registro es 20 y la longitud máxima de un registro es 1894 caracteres. Cada campo tiene una etiqueta que identifica el tipo de información que contiene dicho campo. Un campo puede contener subcampos identificados por un conjunto de caracteres especiales, pero esta subdivisión es transparente para SCIMATE.

Ejemplo:

En una Base de Datos particular se pueden tener registros bibliográficos correspondientes a monografías, seriadas o audiovisuales. Los datos que describen una monografía no son los mismos que describen una seriada o un audiovisual. Se pueden tener tres máscaras : una para monografías, una para seriadas y otra para audiovisuales.

La máscara para monografías podría tener los siguientes campos: título con etiqueta **TI**; autor personal con etiqueta **AP**; autor institucional con etiqueta **AI**; editor con etiqueta **ED**; y fecha de publicación con etiqueta **FP**. El campo autor personal puede contener dos subcampos: en el uno se coloca el apellido y en el otro el nombre. El subcampo apellido se identifica porque aparece al comienzo del campo. El subcampo nombre se precede por una coma y un espacio. Ejemplo:

AP GARCIA MARQUEZ, GABRIEL

Para SCIMATE es transparente el hecho de que después de la coma venga el nombre: todo lo que venga en el campo con etiqueta AP es el autor personal, sin importarle cómo se subdivide la información al interior.

SCIMATE permite importar/exportar archivos de texto hacia/desde una Base de Datos SCIMATE, siempre y cuando los archivos de texto se ajusten a un formato específico que refleja el modelo SCIMATE.

2.3 ISIS.
Las Bases de Datos manejadas por ISIS tampoco están obligadas a seguir un estándar de clasificación específico. El formato que impone es el siguiente:

Un registro es una agrupación de campos repetibles de longitud variable pero con un máximo definido. Cada campo tiene una etiqueta que identifica el tipo de información que contiene dicho campo. Un campo puede contener subcampos repetibles identificados por un caracter especial y una letra. Los subcampos son manejados por ISIS. Adicionalmente, un campo o subcampo puede contener otros subcampos repetibles identificados por unos caracteres especiales, pero estos últimos subcampos son transparentes para ISIS: solo las personas responsables de la Base de Datos conocerán de su existencia.

Ejemplo:

En una Base de Datos particular el título y sus menciones de autoría se van a codificar en el campo con etiqueta **10**. Cada mención de autoría será un subcampo con identificador ^a. Un ejemplo de codificación sería:

10Asterix in Switzarland^atext by Goscinny^adrawings by Underzo^atranslated by Anthea Bell and Dereck Hockridge

Si además del título hubiera subtítulo, éste se podría poner después del título, separado por dos puntos (:). En este caso habría dos subcampos -título y subtítulo- que no serían manejados como tales por ISIS, pero que serían interpretados como subcampos por el responsable de la Base de Datos.

ISIS permite importar/exportar archivos en formato ISO 2709 hacia/desde una Base de Datos ISIS.

3. MODELO GLOBAL: EL FCCC.

Como Modelo Global se escogió una adaptación del CCF de Unesco [SIMM 84].

El CCF -The Common Communication Format- es un formato que se ajusta a los estándares ISO y fué diseñado, entre otras razones, para permitir el intercambio de registros bibliográficos entre grupos de Bibliotecas. Los estándares de clasificación que establece son diferentes a los de MARC.

En CCF un registro bibliográfico consta de información de control, un directorio, y un conjunto de campos formados por subcampos. Campos y subcampos tienen identificadores que indican la información que contienen. Unos y otros pueden ser repetibles y los textos que contienen son de longitud variable. Un campo puede tener indicadores que especifican en más detalle el tipo de información contenida. Un registro puede dividirse en segmentos, y se pueden establecer encadenamientos entre dichos segmentos ó entre campos de un segmento. Los valores posibles de algunos subcampos tales como nivel bibliográfico, idioma, rol, tipo de material, etc., han sido codificados.

La adaptación del CCF al caso colombiano, que fué bautizada con el nombre de Formato Común de Comunicación Colombiano -FCCC-, simplificó el formato de los registros bibliográficos, eliminando directorios, por considerar que en un formato de transporte es importante minimizar el volumen de datos transportado y que la información de directorios es relevante para el almacenamiento eficiente de datos pero no para el transporte. También se suprimieron algunos campos y/o subcampos y se añadieron otros. Se eliminaron los indicadores, la segmentación y los encadenamientos. Las tablas de códigos para ciertos subcampos también fueron modificadas.

Un registro FCCC tiene la siguiente estructura:

- Separador de registro
- Nivel bibliográfico
- Etiqueta de campo
- Identificador de subcampo
- Texto de subcampo
- [Separador de subcampo]*
- [Identificador de subcampo
- [Texto de subcampo
- [Separador de campo]*
- [Etiqueta de campo
- [Identificador de subcampo
- [Texto de subcampo
- [Separador de subcampo]*
- [Identificador de subcampo
- [Texto de subcampo

Para ilustrar los conceptos de campo y subcampo en el FCCC tomemos como ejemplo el título de una obra y sus menciones de autoría: Se codifican en el campo con etiqueta 200. El título propiamente dicho es el subcampo con identificador @A. Las menciones de autoría, repetibles, constituyen cada una un subcampo con identificador @B. Un ejemplo de esta codificación sería:

200@AAsterix in Switzerland@Btext by Goscinny@Bdrawings by Underzo@Btranslated by Anthea Bell and Dereck Hockridge

Dado que el FCCC debe poder cambiar para adaptarse a las nuevas necesidades de intercambio de información, se desarrolló un módulo que permite definirlo y editarlo dinámicamente. La responsabilidad de la definición y de las modificaciones ha sido centralizada en Colciencias, quién será la encargada de distribuir las versiones actualizadas del formato a los responsables de las Bases de Datos que participen en el programa de intercambio.

4. FORMATO INTERNO.

La solución adoptada para independizar los programas convertidores de los paquetes de 'software' utilizados para manejar las Bases de Datos locales, fué introducir un formato interno que incluyera las características importantes de los formatos utilizados por SCIB, ISIS y SCIMATE: cualquier registro producido por alguno de esos paquetes es pasado a formato interno antes de entrar al convertidor hacia FCCC; a la inversa, todo registro en FCCC es convertido a formato interno antes de pasarse al formato del paquete manejador de la Base de Datos que lo va a recibir.

En formato interno un registro bibliográfico se estructura de la siguiente manera:

- Separador de registro
- Máscara (vacio si no hay máscara)
- Separador de etiqueta
- Etiqueta
- Separador de subcampo
- Identificador de subcampo (vacio si no hay subcampo)
- Separador de texto
- Texto de subcampo

Para cada uno de los posibles paquetes manejadores de las Bases de Datos locales que van a intercambiar información es necesario desarrollar programas especiales que conviertan del formato de exportación/importación de cada uno de ellos hacia/desde el formato interno. Esta conversión es puramente sintáctica.

5. PASOS INICIALES PARA QUE UNA BASE DE DATOS SE INTEGRE AL SISTEMA DE INTERCAMBIO.

Para que una Base de Datos pueda integrarse al sistema de intercambio es necesario que su responsable defina cómo y dónde se codifican en su Base de Datos cada uno de los elementos de información del Modelo Global (FCCC). Este proceso de definición lo llamaremos **especificación de equivalencias.**

5.1 ESPECIFICACION DE EQUIVALENCIAS Y EJEMPLO.

5.1.1 ESPECIFICACION DE EQUIVALENCIAS.

La especificación de equivalencias se hace interactivamente, utilizando un módulo que para cada elemento del FCCC pide la siguiente información:

```
MASCARA:          ETIQUETA:       SUBCAMPO:
DELIMITADORES INICIALES(POSICION):
DELIMITADORES FINALES(LONGITUD):
VALOR:
DIRECCION DE LA CONVERSION:
-------------------------------------------------------
condición:
ETIQUETA:         SUBCAMPO:
POSICION(DELIMITADORES INICIALES):
LONGITUD(DELIMITADORES FINALES):
VALOR:
```

Si el paquete local es SCIMATE, que permite múltiples máscaras para una misma Base de Datos, hay que dar el nombre de la **máscara** donde se encuentra el elemento de información. Si el elemento se puede encontrar en varias máscaras, se dará una equivalencia por cada una de ellas. Si el paquete local no maneja múltiples máscaras, MASCARA se deja en blanco.

A continuación se debe dar la **etiqueta** del campo donde se encuentra el elemento. Si se puede dar en **más de** un campo pero **éstos** se excluyen mutuamente, se colocan las etiqueta de esos campos, separados por '|'. Si el elemento se puede encontrar en varios campos no excluyentes, se da una equivalencia por campo.

Si el paquete local es ISIS, que permite el manejo explícito de subcampos, y el elemento de información se encuentra al interior de un subcampo, se coloca el identificador de dicho **subcampo**. Si no se está en ISIS ó el elemento no se encuentra en un subcampo, SUBCAMPO se deja en blanco.

Luego se especifica entre qué conjuntos de caracteres especiales (**delimitadores iniciales y finales**) se puede encontrar el elemento en cuestión. En lugar de caracteres especiales se puede especificar una posición fija y una longitud.

En caso de que el elemento esté codificado en la Base de Datos local, se debe especificar qué código (**valor**) le corresponde. Si son varios, se debe especificar una equivalencia para cada uno.

En **Dirección de la Conversión** se indica si la equivalencia es válida únicamente en la dirección formato local->formato puente, si es válida solo en la dirección formato puente->formato local, ó si es válida en ambas direcciones.

En algunos casos un dato que se encuentra en un campo, entre ciertos delimitadores, corresponde a un elemento del FCCC si y solo si en algún otro campo (que puede ser el mismo) aparece(n) cierto(s) valor(es). Esto se especificará en forma de **condición**, indicando la localización y valores posibles de los datos condicionantes.

Una vez especificadas las equivalencias, éstas serán compiladas y guardadas como tablas específicas de esa Base de Datos. Dichas tablas se utilizarán posteriormente cada vez que se vaya a enviar ó recibir información por intercambio. Las equivalencias pueden ser modificadas y recompiladas en cualquier momento.

5.1.2 EJEMPLO.

Para el elemento de información "fecha de copyright normalizada" del FCCC, en una Base de Datos manejada por SCIB la especificación de equivalencia será:

```
MASCARA:        ETIQUETA:008     SUBCAMPO:
DELIMITADORES INICIALES(POSICION):6
DELIMITADORES FINALES(LONGITUD):4
VALOR:
DIRECCION DE LA CONVERSION:3 (formato local <->formato puente)
-----------------------------------------------------------------
condición:
ETIQUETA:008     SUBCAMPO:
POSICION(DELIMITADORES INICIALES):1
LONGITUD(DELIMITADORES FINALES):1
VALOR:c
```

En la equivalencia se está diciendo que la fecha de copyright normalizada, cuando aparece, se encuentra en el campo con etiqueta 008 en la posición fija 6, y tiene una longitud de 4 caracteres. Sinembargo, es posible que lo que se encuentre allí no sea la fecha de copyright normalizada sino otra fecha. Para cerciorarse de de ésto hay que mirar la posición 1 del mismo campo: si allí hay una c, entonces la fecha que está en la posición 6 es efectivamente la deseada. La equivalencia debe operar en ambas direcciones.

5.2 ESPECIFICACIONES ADICIONALES.

Dado que el formato puente maneja tablas de códigos para idioma del documento y rol de un autor, el responsable de la Base de Datos que se desea integrar al sistema de intercambio deberá indicar, utilizando unos módulos del sistema, la equivalencia entre los códigos que él maneja y los códigos del formato puente.

Además de especificar equivalencias, el responsable de cada Base de Datos manejada por SCIMATE ó ISIS deberá hacer unas definiciones adicionales:

- Para las Bases de Datos en SCIMATE se deberá indicar el orden en que pueden aparecer los campos al interior de cada máscara . Si un registro muy largo se puede partir en dos ó más, hay que indicar el campo que se replica para hacer el encadenamiento lógico entre esos registros;

- Para las Bases de Datos en ISIS se deberán especificar los identificadores de los campos con modo de inversión 2 ó 3.

Para SCIB se deben especificar los valores por defecto que se deben dar a los indicadores que acompañan a cada etiqueta cuando se vaya a recibir información.

6. PROTOCOLO ACTUAL DE INTERCAMBIO.

A falta de una red de comunicaciones de gran alcance que interconecte las Bases de Datos que participan en el proyecto FCCC, el intercambio de información entre Bases de Datos se está haciendo actualmente a través de diskettes.

El protocolo que se sigue es el siguiente:

A) SOLICITUD Y ENVIO DE INFORMACION (figura 1):

1. El responsable de una Base de Datos solicita información a su homólogo en otra Base de Datos. Esta solicitud puede ser verbal ó escrita.

2. El responsable de la Base de Datos a quien se solicita la información hace la consulta utilizando el lenguaje del paquete local (ISIS, SCIMATE ó SCIB). El conjunto de registros producidos se deja en un archivo tipo exportación (por ejemplo utilizando la opción de exportar archivos en formato ISO 2709 de ISIS).

3. Se utiliza un módulo específico al paquete local para convertir el archivo del formato de exportación al formato interno estándar.

4. Se utilizan los módulos que convierten el archivo en formato interno a otro archivo en el formato puente (FCCC). Para hacer la conversión, los módulos consultarán las tablas de equivalencias de la Base de Datos de la cual se extrajo la información. Como en un mismo computador puede haber varias Bases de Datos, cada una con sus tablas de equivalencias propia, los módulos preguntarán el nombre de la Base de Datos específica.

5. Una vez convertido al formato puente, el archivo será copiado a diskettes.

6. Los diskettes son enviados al responsable que hizo la solicitud de registros bibliográficos.

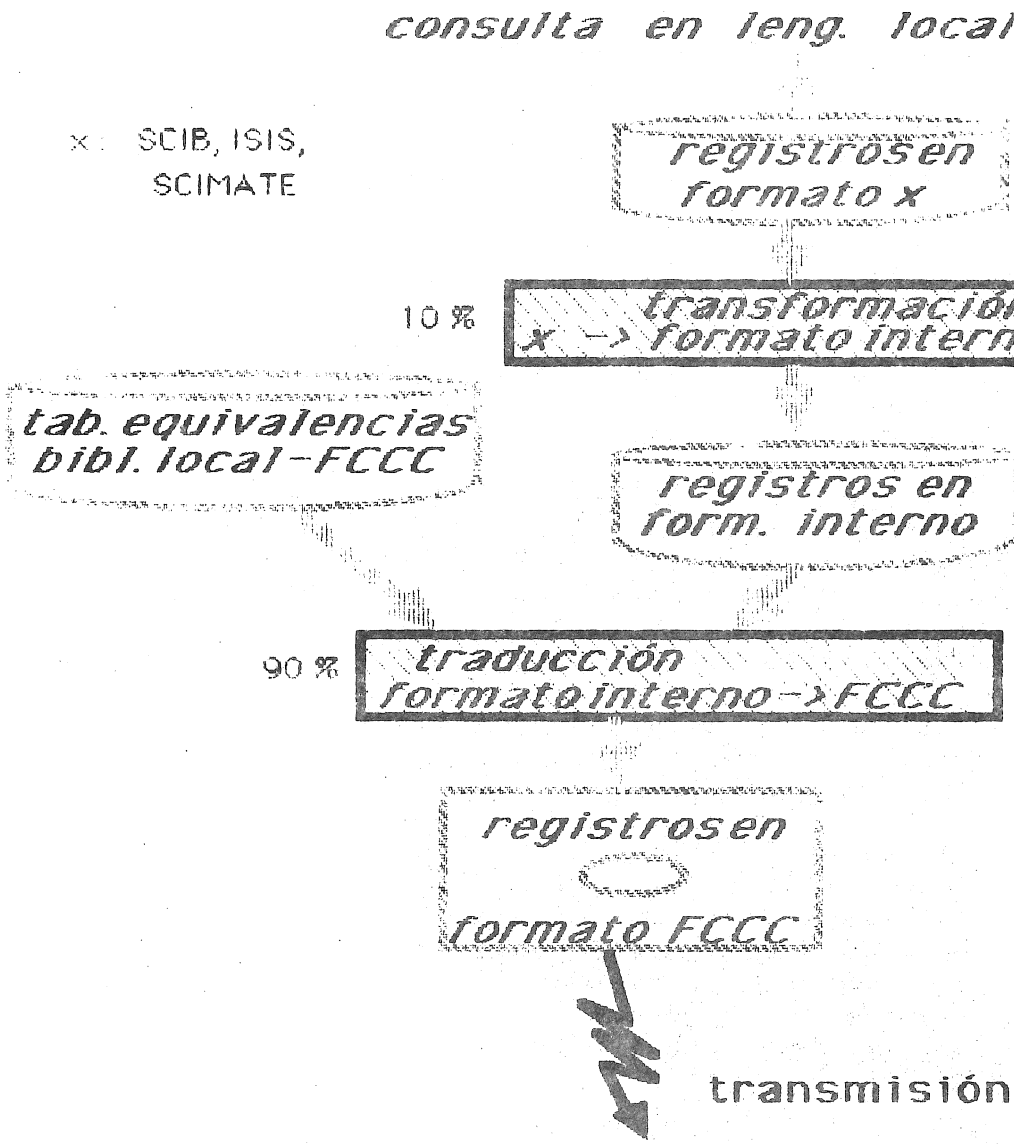

FIG. 1 : SOLICITUD Y ENVIO DE INFORMACION

B) RECEPCION Y EVENTUAL INTEGRACION DE LA INFORMACION (figura 2):

7. El responsable que hizo la solicitud copia en disco duro el archivo que recibió en diskettes.

8. Se utilizan los módulos que convierten el archivo en formato puente (FCCC) a otro archivo en formato interno estándar. Para hacer la conversión los módulos consultarán las tablas de equivalencia de la Base de Datos local a la cual, eventualmente, se integrarían los registros recibidos.

9. Se utiliza un módulo específico al paquete local (ISIS, SCIB ó SCIMATE) para convertir el archivo en formato interno a otro archivo en el formato de importación del paquete local.

10. El responsable que hizo la solicitud examina el archivo recibido, ya en su formato local, lo edita y, eventualmente, lo integra parcial o totalmente a su Base de Datos.

7. DESARROLLOS FUTUROS.

Se está empezando a trabajar en convertidores FCCC<->CCF que permitan intercambiar información bibliográfica con otros paises que utilicen el CCF como formato de transporte. También se planea desarrollar versiones de los programas del proyecto "FCCC" para máquinas diferentes de IBM PC XT/AT.

Una vez instalada la red pública de transmisión de datos colombiana -COLDAPAQ- se piensa montar un servicio de consulta en línea a un conjunto de Bases de Datos Distribuidas Heterogéneas que ofrezca transparencia tanto a la heterogeneidad de las Bases de Datos participantes como a la localización física de los datos. La información bibliográfica estará en una de esas Bases de Datos Distribuidas. Otras Bases podrán contener información estadística, de salud, factual, etc. Cada consulta será enrutada automáticamente a la Base de Datos correspondiente.

Para poder montar tal sistema de consulta en línea será necesario definir un lenguaje de consulta global, así como resolver problemas ligados a la confidencialidad de los datos, tarificación de los servicios, etc.

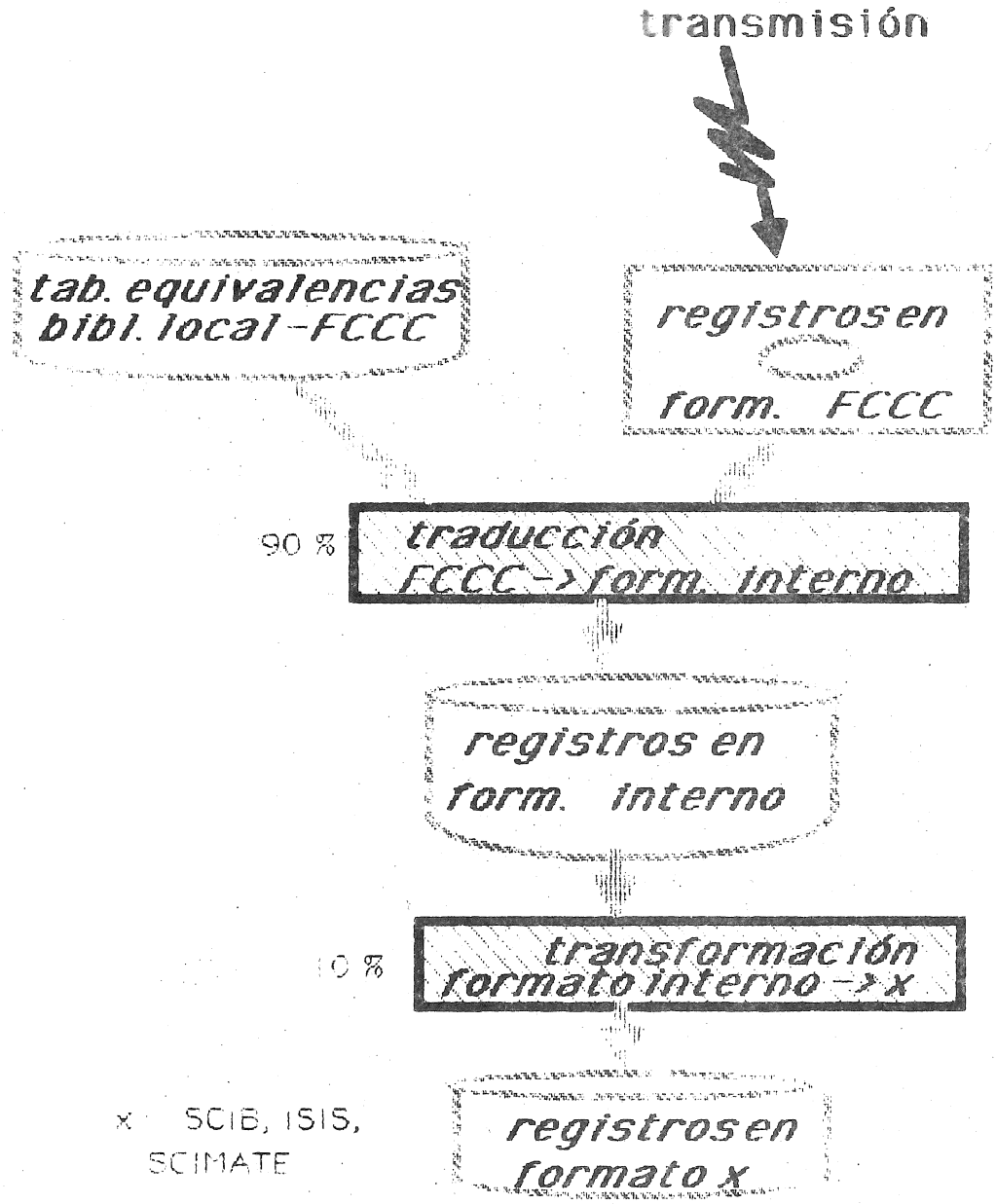

FIG. 2 : RECEPCION Y EVENTUAL INTEGRACION DE LA INFORMACION

8. CONCLUSIONES.

Los protocolos presentados en este artículo resuelven, para el caso específico de la información bibliográfica, el problema de cómo brindar transparencia hacia la heterogeneidad a los usuarios de una Base de Datos Distribuida heterogénea.

Esos protocolos podrían utilizarse también en la integración de Bases de Datos no bibliográficas, siempre y cuando exista un Modelo Global para el tipo de información que se esté integrando.

En futuros desarrollos se espera resolver el problema de cómo brindar transparencia a la localización física de los datos, integrando , por tipo de información, muchas Bases de Datos automatizadas que existen en el país, utilizando como sistema de comunicación la red pública de transmisión de datos colombiana COLDAPAC.

9. BIBLIOGRAFIA.

[ABAS 87] Abásolo, José; Franky, Ma. Consuelo; Muñoz Adriana
"Protocolos de traducción bibliográfica desde/hacia el Formato Común de Comunicación Colombiano (FCCC)"
Informe final preliminar proyecto 1204-03-098-85
Colciencias-Universidad de los Andes. Junio 1987.

[CERI 84] Ceri, Stefano; Pelagatti, Giuseppe
"Distributed Databases: Principles and Systems"
McGraw-Hill, 1984

[SIMM 84] Simmons, Peter; Hopkinson, Alan
"CCF: The Common Communication Format"
United Nations Educational, Scientific and Cultural Organization, 1984.