JORGE ANTONIO MEJÍA ESCOBAR
FRANCISCO JAVIER ÁLVAREZ
JOHN ALBEIRO SÁNCHEZ
Universidad de Antioquia, Medellín, Colombia
E-Mail: Jamejia@catios.udea.edu.co, falvarez@cesde.edu.co, jsanchez@cesde.edu.co

Presentación del programa Cratilo, su manejo y sus usos

CRATILO es un programa de lingüística computacional que sirve para poner al servicio de los intérpretes de texto varias herramientas lexicográficas. Trabaja sobre textos que se encuentran en los formatos ***.txt** y ***.rtf**. Sin embargo, es recomendable trabajar en formato **rtf** pues ello da la posibilidad de conservar el aspecto que tiene un texto normalmente en el procesador, incluyendo los saltos de página y las características tipográficas. Se recomienda que el texto magnético sea *facsimilar* respecto de la edición crítica (si la hay) o, en su defecto, de la edición príncipe, puesto que el programa genera, como su producto más acabado, una *concordancia universal* ordenada alfabéticamente por *lemas* y sub - ordenada por *formas gráficas*, en la cual cada aparición de una palabra (ocurrencia) es documentada con su dirección en el texto expresada en página, línea y columna y ella puede imprimirse para ser utilizada por fuera de ambientes computacionales. Para lograr el efecto buscado se requiere que haya saltos de línea fijos al final de cada renglón, lo que implica digitar un **[enter]** (o la señal de fin de párrafo: ¶). Igualmente se debe digitar un salto de página manual al final de cada página, lo cual se hace en MSWord pulsando **Ctrl + Enter**. Si el texto procesado es facsímil de uno de los textos sobre papel, la concordancia podrá ser utilizada para rastrear el contexto más amplio directamente sobre el libro. La última operación que se debe hacer antes del procesamiento es añadir el número de la página inicial y los saltos de página legibles para CRATILO. El número de la página inicial se debe colocar entre paréntesis angulares dobles en el primer renglón del texto (por ejemplo **<<9>>** si ese es el número de la primera página de texto). Igualmente debe ponerse al final del último renglón de cada página la señal que el programa reconoce como fin de página, a saber, **<<p>>**. Esto último puede hacerse programando una macro basada en los saltos de página manuales, que es recomendable que no desaparezcan, pues no generan ninguna distorsión para el programa.
Una vez que el texto está disponible en uno de los dos formatos mencionados se procede a grabar una copia de él en la carpeta **Textos** que se encuentra como sub-carpeta de la carpeta "Cratilo".

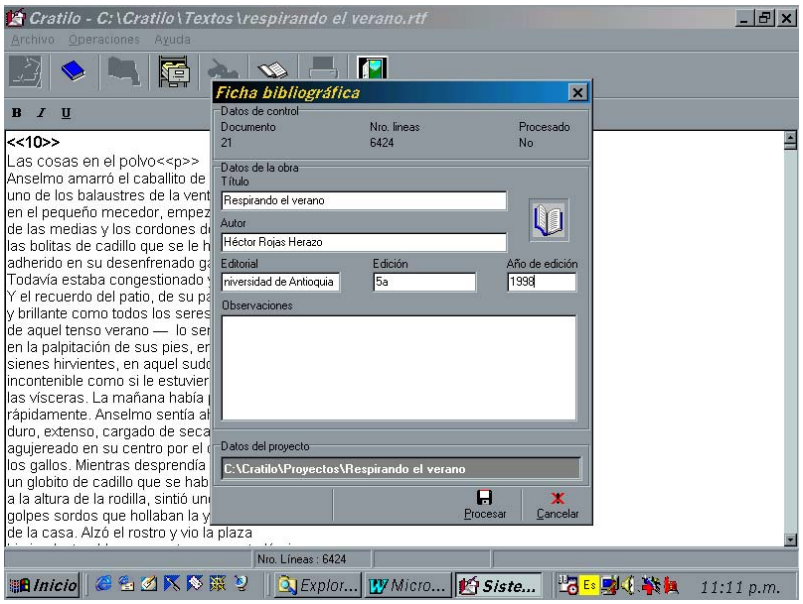

Figura 1: Ambiente básico y ficha bibliográfica

Luego viene la conversión del *texto* en *proyecto*. En términos prácticos consiste en convertir el texto de procesador de palabra, con su formato secuencial en una base de datos en la cual cada *ocurrencia* (aparición de una palabra) es un registro con varios campos:

1. La palabra, tal como aparece (su nombre es "forma gráfica").
2. El lema, que es de ordinario la palabra bajo la cual entraría en un diccionario.

3. La clasificación gramatical (si es flexión nominal, verbal, invariable etc.).
4. El lugar donde aparece, o dirección, expresado en página, línea y columna.
5. El microcontexto: la palabra se ubica en medio de un párrafo artificial formado por unos 4 renglones de texto (si se hace la consulta por pantalla), o en medio de una secuencia de unos 80 caracteres (si se imprime la concordancia).

El requisito que hay que cumplir para pasar de *texto* a *proyecto* es la elaboración de una breve ficha bibliográfica del texto que va a ser procesado (parte central de la figura 1) Incluye Autor, nombre del texto, Editorial, Edición y Año. Además hay un campo que sirve para consignar observaciones de diferente índole. Esta ficha debe llenarse en forma completa para que sea aceptada y se inicie el procesamiento.

La velocidad del proceso es bastante buena: un texto como La Marquesa de Yolombó, con sus 135.000 palabras (*ocurrencias*), se procesa en unos 4 minutos en un Pentium III de 450 Mhz.
Una vez realizado ese procesamiento se tiene a disposición un primer análisis, al cual se accede por medio del tercer icono (de izquierda a derecha). Una vez activado aparece una pantalla cuyo cuerpo principal son tres columnas: **Palabra**, **Frecuencia** y **Longitud**. Inicialmente aparece **Palabra** en color amarillo y las otras dos en negro. Quiere decir que el orden de aparición de los datos procedentes del texto es el de palabra, que es el orden alfabético, o mejor el ASCII, en el cual los números están primero y luego las letras. En esa forma se ven las palabras ordenadas alfabéticamente. Si se pulsa el botón izquierdo del ratón sobre **Frecuencia**, el color del título de la columna pasará de negro a amarillo y se verán ordenadas las palabras según su frecuencia, de mayor a menor. Lo mismo sucederá con **Longitud** (en número de caracteres).
Cuando la columna activada es **Palabra** y se hace un “click” con el botón izquierdo del ratón sobre la casilla de la línea inferior de la pantalla llamada **Índice inverso** el orden alfabético cambiará de derecho a inverso y tendremos a nuestra disposición el *índice inverso o reverso* o *index a tergo* con base en la primera columna de la izquierda (**Palabra**). De ese modo podremos consultar rápidamente las “familias” de palabras que resultan de su agrupación por terminaciones comunes. Esta característica permite a Cratilo convertirse en *rimario* e igualmente permite al investigador examinar con facilidad asuntos como el manejo de la temporalidad en las formas verbales, los adverbios de modo en “mente”, y todo lo que pueda seguirse por terminaciones comunes.

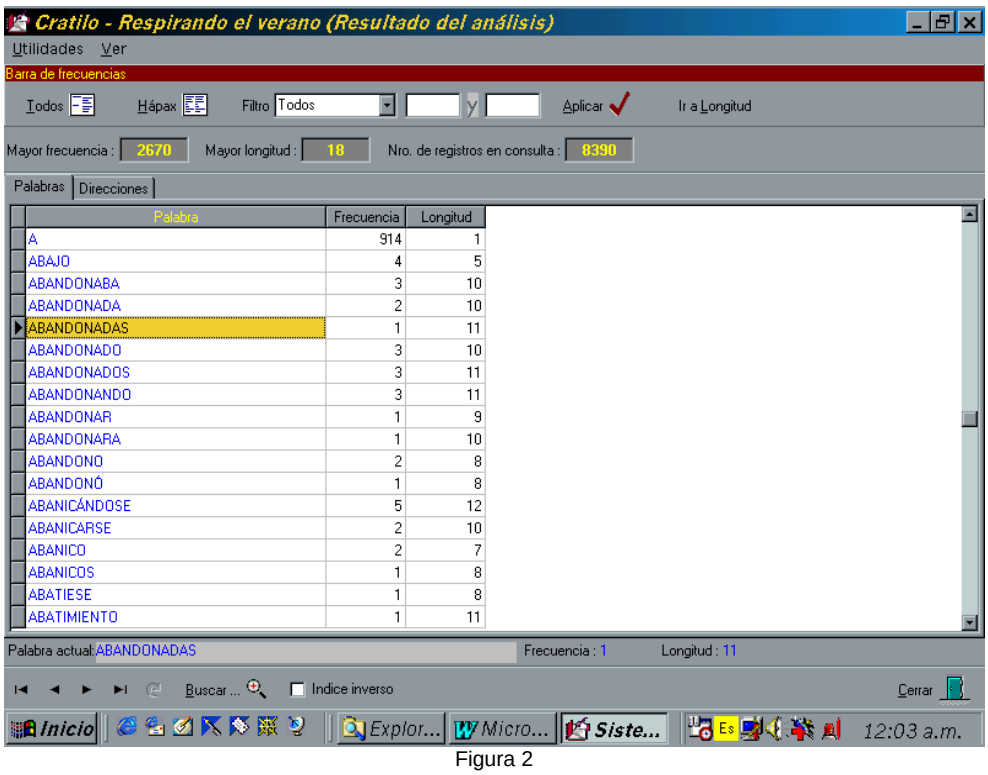

Figura 2

Con las flechas del teclado, o con el ratón, es posible desplazarse de arriba hacia abajo y viceversa sobre las columnas. Cuando una palabra es señalada se puede ver un marcador negro, a modo de

puntero, en el extremo izquierdo del renglón (En la figura 2 se encuentra en **ABANDONADAS**). En esa forma podemos seleccionar una palabra cualquiera y veremos cuántas veces está en el texto y qué longitud tiene. Es posible llegar a una palabra cualquiera por medio del botón **Buscar**, que se encuentra en el último renglón de la pantalla, al lado de la lupa. Cuando se oprime, aparece una caja interactiva que permite digitar la palabra buscada y va inmediatamente a localizarla con el puntero. Si la pantalla no cambia, y el puntero no se mueve de donde se encuentra, ello significa que la palabra no está en el texto procesado.
En la parte superior, inmediatamente encima de las columnas, se encuentran dos pestañas. La primera, que inicialmente aparece activada (sobresale en falso relieve en la figura 2), se llama **Palabras** y la segunda **Direcciones**.
Una vez localizada una palabra se puede saber en qué lugares de la obra se encuentra, pulsando el botón izquierdo del ratón sobre la pestaña **Direcciones**.
Si queremos examinar los contextos de esa palabra podemos pulsar ahora el botón **Contexto** que se encuentra en el renglón superior de la pantalla de direcciones. Aparecerá una caja con el contexto de la *ocurrencia* donde está el puntero y la dirección correspondiente (en la figura 3 se ve que corresponde a la tercera *ocurrencia* de **abuela**). Por medio de las flechas inferiores de la caja es posible *navegar* por todos los contextos de todas las *ocurrencias* de una misma palabra. Para cambiar de palabra es menester cerrar la caja de contextos y luego activar nuevamente la pestaña **Palabras**.

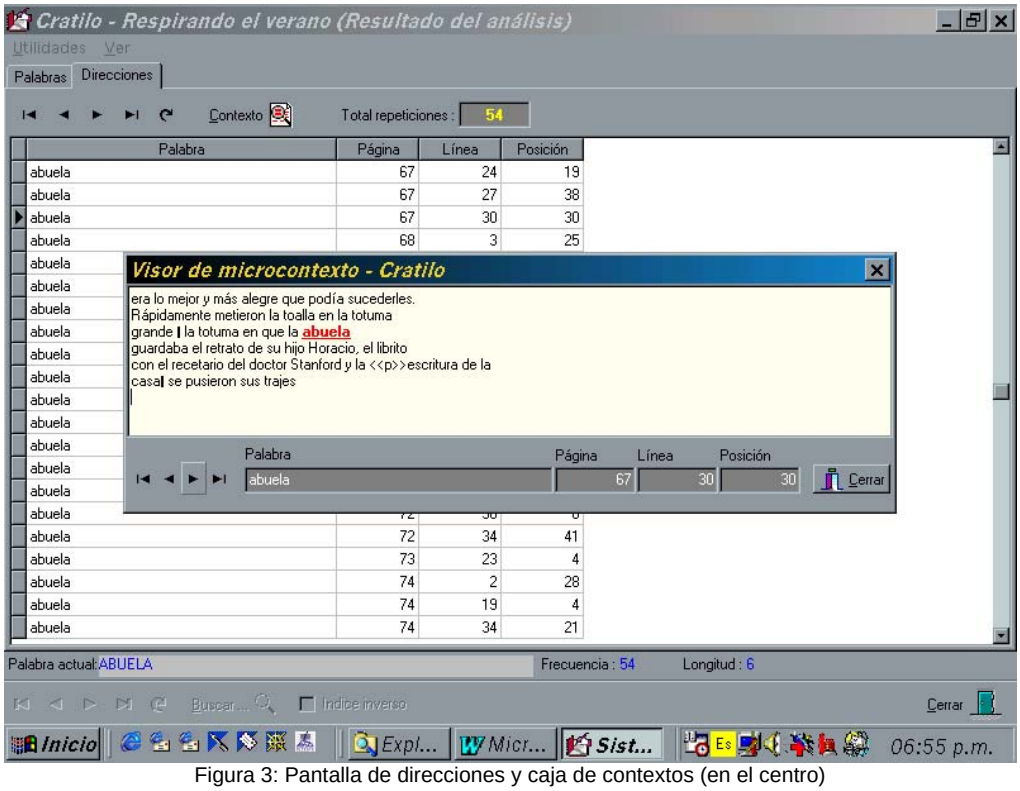

Figura 3: Pantalla de direcciones y caja de contextos (en el centro)

Por medio del icono inferior derecho cerramos este ambiente y volvemos desde los análisis al ambiente básico, que puede distinguirse porque el texto aparece como fondo y encima están los ocho iconos de las funciones del programa (figura 1). El quinto de ellos, de izquierda a derecha, pasa al proceso de *lematización*.
Pero antes de hacerlo es muy importante anotar lo siguiente: una vez que tenemos un texto convertido en proyecto el acceso a él será siempre a través del cuarto icono del ambiente de trabajo básico (el archivador). El proyecto se selecciona en el listado que aparece, y cuando el puntero lo señala se va al renglón inferior de la pantalla y se oprime el icono **Activar proyecto**, que en realidad corresponde a la orden de cargar la base de datos de un texto determinado. Una vez *activado* podemos cambiar de proyecto por medio del segundo icono básico, que representa un libro cerrado, y al ser oprimido nos lleva a la pantalla de presentación del programa o *splash*. También es importante anotar que un mismo texto puede ser procesado en varias formas según las necesidades del usuario, con el fin de hacer análisis contrastivos. Ello requiere manipulaciones con el diccionario y es un tópico más avanzado de uso.

Por *lematización* se entiende, como ya habíamos dicho, la asignación a cada *ocurrencia*, de la palabra bajo la cual figuraría en un diccionario, que es el *lema*, además de una clasificación gramatical y de una señal adicional que dice si bajo la misma forma escrita puede haber otra u otras palabras diferentes, caso en el cual se dice que esta forma es *ambigua*. Existen en el programa un modo manual y un modo automático de *lematizar*. En el modo manual se va avanzando secuencialmente en el texto, de la primera palabra hasta la última. Así, en "Cien años de soledad", comenzaríamos por *Muchos* luego *años*, luego *después* y terminaríamos con *sobre*, *la* y *tierra*.
El modo automático exige que haya palabras grabadas en el diccionario y esto se logra a medida que *lematizamos*, pues Cratilo *aprende* cada palabra que el operario *lematiza*. También es posible arrancar desde un comienzo con un diccionario básico registrado en el programa. En modo automático existe la variante *Lematización autónoma* que permite al programa trabajar con base en el diccionario residente excluyendo las palabras que implican decisión, que son las *ambiguas* y las *vicarias* (aquellas palabras que reemplazan a otras o que manifiestan deixis). En el caso de las *vicarias* el operario debe decir cuál es la palabra reemplazada (por ejemplo por un pronombre), pues será guardada en un campo especial para consultar luego apariciones implícitas, y en el caso de las *ambiguas* el operador debe deshacer la ambigüedad eligiendo el valor que se aplica en cada caso particular). El modo automático también permite escoger cuáles tipos de palabras (según su clasificación gramatical) van a ser *lematizados* por el operador en cada sesión de trabajo. Esta característica permite dedicarse a un mismo tipo de palabras para concentrarse en ellas y facilitar el trabajo. A medida que avancemos tendremos un diccionario de máquina cada vez más rico.
El diccionario trabaja con una clasificación gramatical básica y abreviada, inspirada en la de Roberto Busa que opera *por defecto* o en ausencia de modificaciones, e incluye los siguientes clasificadores:

0. Sin lematizar
1. Vicarias o deícticas: especialmente los pronombres
2. Vicarias abiertas, cuando lo reemplazado es tan numeroso que no puede ser delimitado en el campo disponible
3. Vicarias indefinidas, cuando lo reemplazado no puede ser definido.
4. Flexiones nominales: nombres, adjetivos y artículos definidos e indefinidos
5. Flexiones verbales: se exceptúan los infinitivos precedidos de artículos y los participios que no hacen parte de tiempos compuestos.
6. Invariables: adverbios, preposiciones y conjunciones.

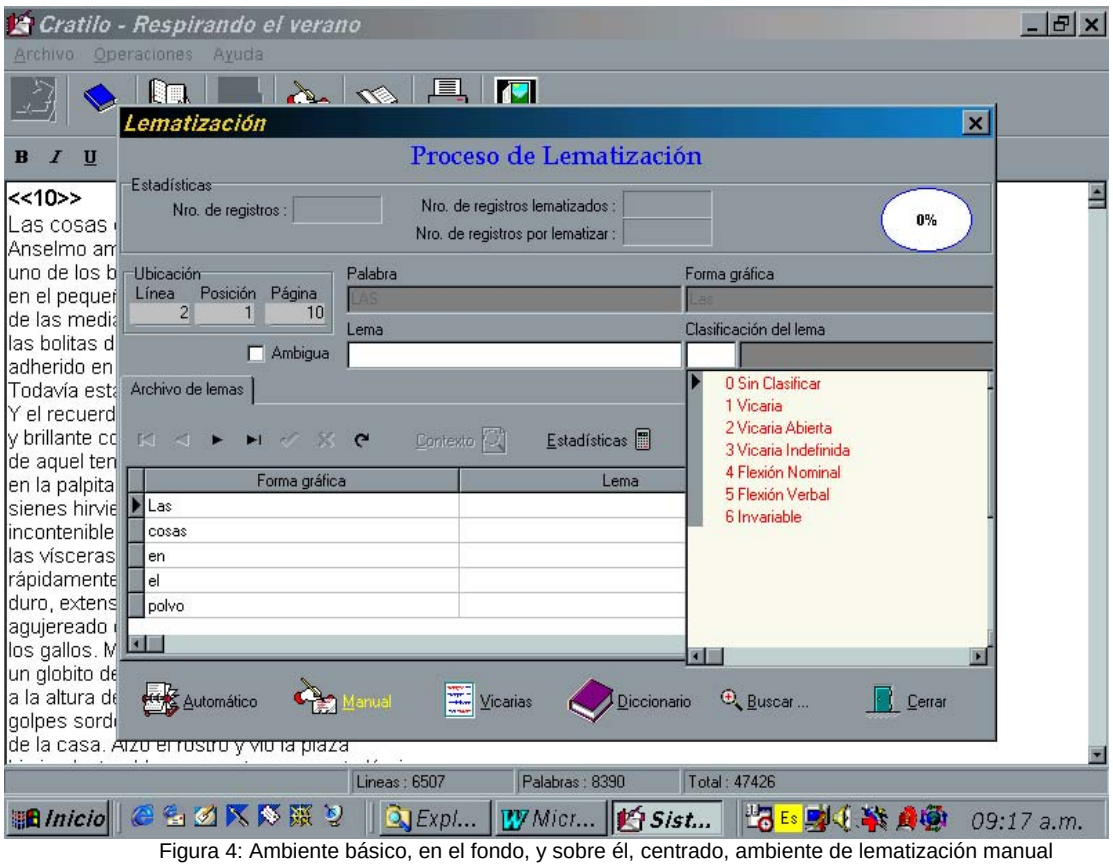

Figura 4: Ambiente básico, en el fondo, y sobre él, centrado, ambiente de lematización manual

No obstante, existe la posibilidad de introducir hasta 93 clasificaciones más según los propósitos del investigador. Su adición se hace en el modo **Diccionario** (véase el cuarto botón de acceso en el renglón inferior del ambiente de lematización, figura 4) a partir del uso de la pestaña **Clasificadores**.
Es importante tener en cuenta que se debe conservar rigurosamente la relación entre el texto *lematizado* y la versión correspondiente del diccionario pues si se cambia el diccionario (al modificar o introducir valores diferentes) podría aparecer por pantalla o impresora una clasificación incoherente de las palabras de un texto.
El proceso de lematización se realiza acumulativamente por sesiones de trabajo que pueden programarse con muchas variaciones (por clasificación gramatical, por orden de aparición, sin interacción, con interacción, manual, por rango de las ocurrencias, etc.). Una vez completado se tiene el producto más acabado de Cratilo, que es la **concordancia universal por lemas**, que permite ver la pluriforme diversidad del texto desde el orden clasificador que han introducido los *lemas* y luego las *formas gráficas*. Los lemas comandan el ordenamiento y las formas gráficas subordenan dentro de cada lema. Este producto sólo es accesible por impresora y no por pantalla, pero puede ser impreso sobre disco y luego consultado como texto. Este procedimiento se maneja a partir del séptimo icono del ambiente básico de trabajo.

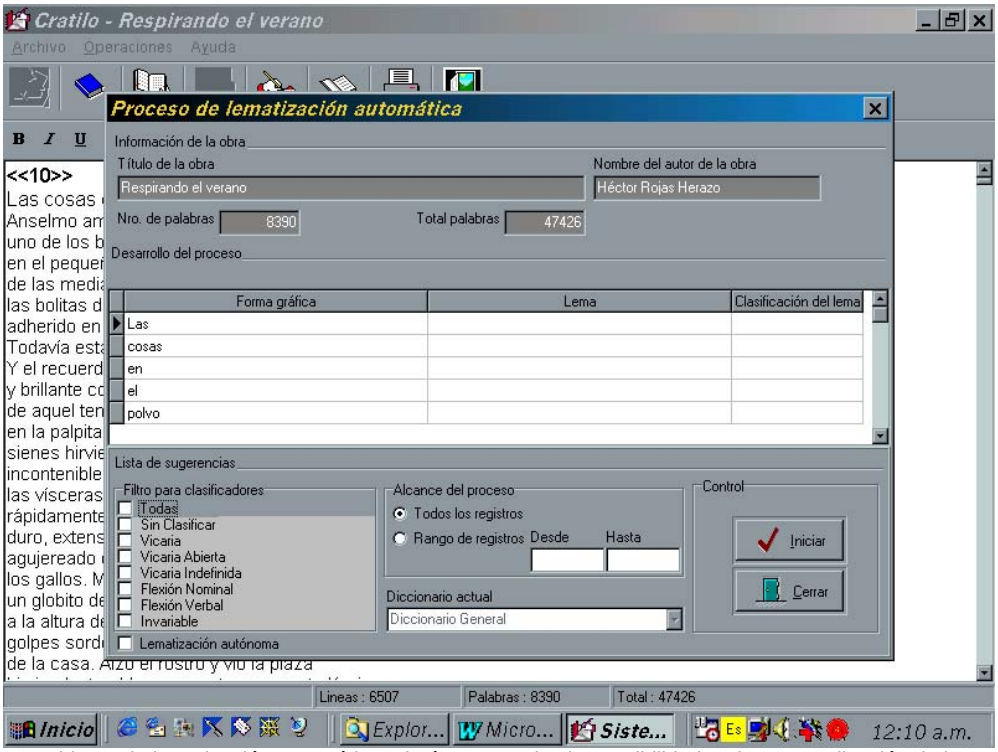

Figura 5: Ambiente de lematización automática, obsérvense todas las posibilidades de personalización de las sesiones

A partir del uso del séptimo icono se pueden imprimir otros subproductos como el listado de formas gráficas con sus frecuencias en orden alfabético, el índice reverso, el índice reverso de lemas, las estadísticas (20 frecuencias mayores de formas gráficas, 20 palabras más largas, todas las frecuencias de los caracteres, número de ocurrencias, líneas y formas gráficas), el diccionario (para hacer control de calidad sobre papel y luego transcribirlo), las palabras vicarias y las vicariadas. Todos estos subproductos pueden imprimirse sobre disco en formato ***.QR** o en formato ***.txt**, que tiene la ventaja de poder ser convertido a los procesadores de texto y hojas de cálculo y por eso se recomienda para manejos por fuera de Cratilo. El formato ***.QR** (*quick report*) permite tener a disposición el archivo ya elaborado y simplemente abrirlo desde el icono de impresora antes mencionado.