

## **Búsquedas y patrones de uso en Metalib: un análisis a través de los archivos de log**

Gaspar Olmedo (acbic@cica.es). CSIC. Unidad de Coordinación de Bibliotecas *VI Jornadas de Expania, Asociación de Usuarios de Ex Libris en España. Pamplona, 18 y 19 de junio de 2009.*

Se ha utilizado el análisis de los archivos de log del servidor web (www\_server) y del servidor de búsquedas (search\_server) de Metalib para obtener datos sobre el comportamiento de los usuarios en las sesiones y sobre el resultado de las búsquedas en función de los recursos utilizados, aspectos que sobre los que las estadísticas estándares de la aplicación proporcionan poca información. Los resultados muestran un uso de la aplicación bastante sencillo por parte de los usuarios, demuestran que el número de sesiones se encuentra muy sobrevalorado en las estadísticas estándar y proporcionan un método para valorar de forma comparativa la utilidad de los recursos electrónicos incluidos en Metalib.

Log files of Metalib web and search servers have been analyzed to obtain data about user behaviour during sessions and about response of e-resources used in meta-search queries. Log files were used because standard Metalib statistics give very little information about those aspects. Results show that users work with Metalib in a rather simple way, also that number of sessions is highly overestimated in standard statistics due to the presence of many time-out sessions. Graphic analysis of search data can be used to evaluate e-resources.

### Introducción y métodos

La Red de Bibliotecas del Consejo Superior de Investigaciones Científicas (CSIC) instaló Metalib como portal de recursos electrónicos y metabuscador en el año 2006, desde entonces se ha hecho un seguimiento de su uso a través de las estadísticas que ofrece la aplicación, pero éstas ofrecen muy poca o ninguna información sobre dos puntos concretos que son muy relevantes: el comportamiento de los usuarios durante las sesiones y el grado de utilidad de los recursos cuando se utiliza Metalib como metabuscador. Los archivos de log del servidor web (www\_server) contienen información que permite aislar cada sesión y reconstruirla con bastante detalle y la información de los logs del servidor de búsquedas (search\_server) permite analizar el resultado de cada una de ellas. A partir de estos archivos de log se han analizado algunos aspectos concretos de las sesiones de usuario y de las metabúsquedas.

Se han utilizado los archivos de log del servidor web (www\_server) y del servidor de búsquedas (search\_server) de la instalación de Metalib en el CSIC correspondientes al año 2008 completo. Se filtraron, trataron y cargaron en una base de datos para su análisis posterior. En el caso del www\_server, para cada registro se tienen los datos de fecha, hora, identificador de sesión y función, que permiten realizar una reconstrucción bastante detallada del transcurso de cada sesión (ver figura 1 y vídeos anexos). Se han obtenido 1.439.443 registros correspondientes a 311.039 sesiones. A partir del archivo de log del servidor de búsqueda se obtuvo una base de datos en la que en cada registro figura el nombre e identificador del recurso en el que se hizo la búsqueda, la cadena de búsqueda, el número de documentos obtenido como resultado y si la búsqueda produjo error o no (figura 2). Se obtuvo información sobre 263.937 búsquedas.

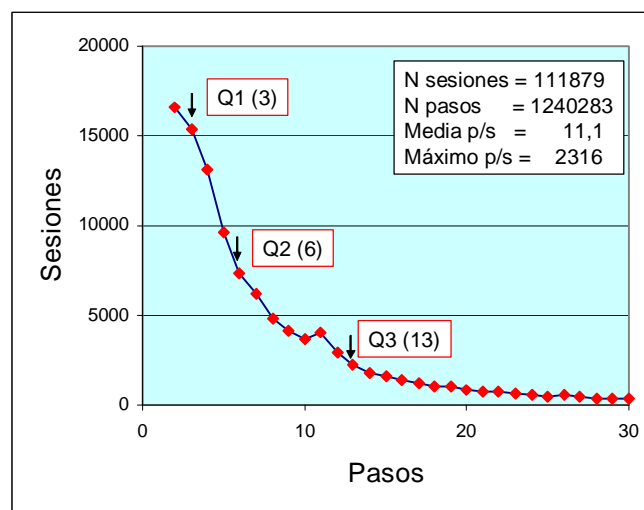
Para las sesiones se ha analizado el número de pasos que compone cada sesión, como indicador de su complejidad, la presencia de estas tres



## Resultados: sesiones

Lo primero que llama la atención al analizar el número de pasos de las sesiones es la existencia de 199.160 sesiones (el 64% del total) que tienen un solo paso. Mirando estas sesiones con detalle, se puede ver que corresponden a sesiones generadas por el efecto del cierre automático de sesión establecido por la aplicación (time-out) por motivos de seguridad. Sólo un tercio de las sesiones corresponden, pues, a sesiones reales de usuario. La distribución de los pasos de las sesiones con más de un paso y sus indicadores estadísticos pueden verse en la figura 3.

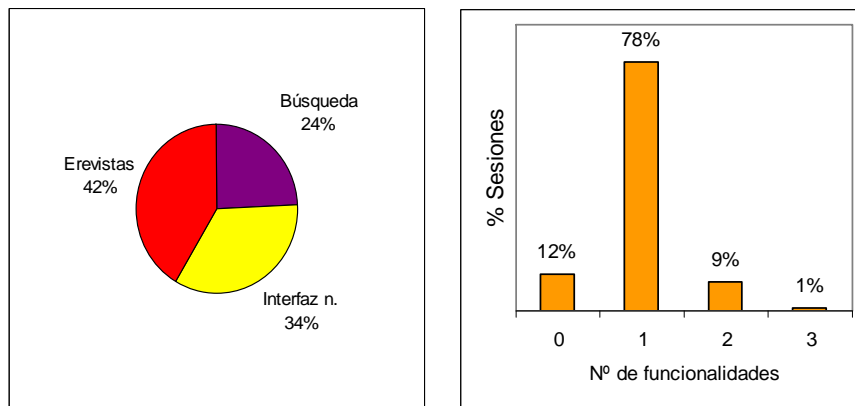
### Pasos por sesión (sin sesiones de 1 paso)



La mayoría de las sesiones tienen pocos pasos, son muy simples. La cuarta parte tiene 2 o 3 pasos, se trata de sesiones fallidas o de sesiones en las que el usuario busca de forma muy directa un recurso o título de revista para enlazar a su interfaz nativa. Nunca o casi nunca contienen metabúsquedas. Las sesiones de 6 pasos (mediana de la distribución, punto que delimita el 50% de los casos) pueden ya contener una búsqueda y algún pase de página, visualización de registro completo y enlace a texto completo. Los 13 pasos marcan el tercer cuartil (75% de los casos), y éstas son ya sesiones de recorrido más complejo que pueden contener una o varias búsquedas, varias visualizaciones de registros, enlaces a texto completo, uso de las opciones privadas, etc. (Se pueden ver ejemplos de sesiones extraídas de los logs con diferentes número de pasos en los vídeos adjuntos).

En las gráficas de la figura 4 puede verse el porcentaje de sesiones en las que está presente al menos un caso de cada uno de los tipos de funcionalidad que se analizan (metabúsqueda, entendida como búsqueda en uno o varios recursos lanzada desde Metalib, enlace a la interfaz nativa de un recurso y búsqueda de e-revistas), y la distribución de las sesiones en función del número de funcionalidades que contienen de forma simultánea.

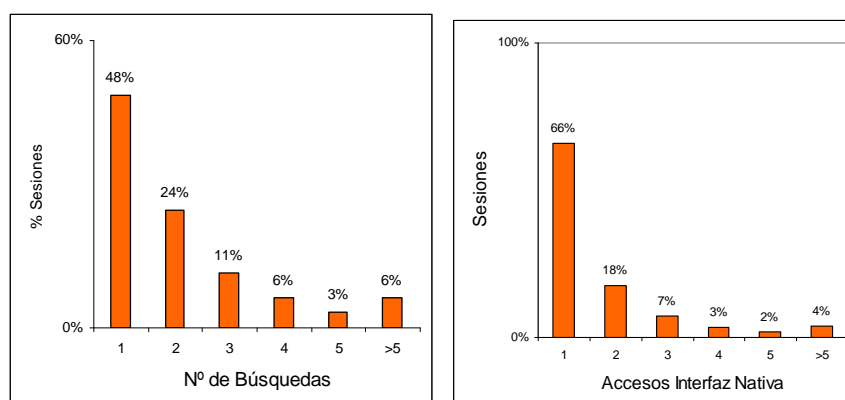
### Presencia y simultaneidad de funcionalidades en las sesiones



Como se puede ver la funcionalidad que más aparece es la búsqueda de e-revistas, una funcionalidad que en realidad no corresponde a Metalib, sino a SFX, en segundo lugar se sitúa el enlace a la interfaz nativa. La metabúsqueda se puede encontrar aproximadamente en una de cada cuatro sesiones. Hay pocas sesiones (9%) en la que se hayan utilizado dos funcionalidades de forma simultánea, y las que contienen las tres son ya realmente escasas (1%).

Cuando la metabúsqueda está presente, se trata casi siempre (82% de 59.072 acciones de búsqueda) de una búsqueda rápida (sobre conjuntos predefinidos o quick-sets), en el 13% de los casos se utiliza la opción de "Metabúsqueda" y sólo en el 5% se usa la búsqueda en un solo recurso. En cuanto al número de búsquedas que se realizan en cada sesión, la media es de 2,2 (calculada sobre las sesiones que contienen al menos una búsqueda). La media de accesos a interfaz nativa para las sesiones que contienen al menos un acceso es de 1,8

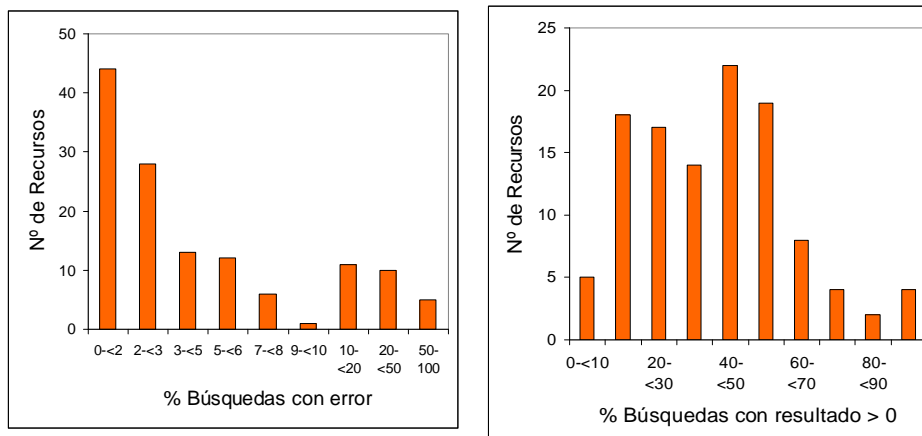
### Búsquedas y accesos a i. nativa por sesión



## Resultados: Búsquedas

El porcentaje global de búsquedas que devuelven error es del 5,1% (nº de búsquedas = 263.937 ; nº de recursos =130) y el porcentaje global de búsquedas no erróneas que devuelven al menos un documento como resultado es del 53% (nº de búsquedas = 250.396 ; nº de recursos=114). En las distribuciones se puede ver que la mayor parte de los recursos tienen unas tasas de error bajas. También se puede ver que la mayoría de los recursos tienen tasas de acierto inferiores al 50%, sin embargo el porcentaje global sube por el hecho de que algunos recursos muy utilizados (p.ej. WOK/SCI, Catálogos CSIC, Scopus, Pubmed) tienen tasas de acierto medias o altas, en todo caso superiores al 50%.

### Distribución de los recursos en función de la tasa de error y de obtención de resultados en las búsquedas

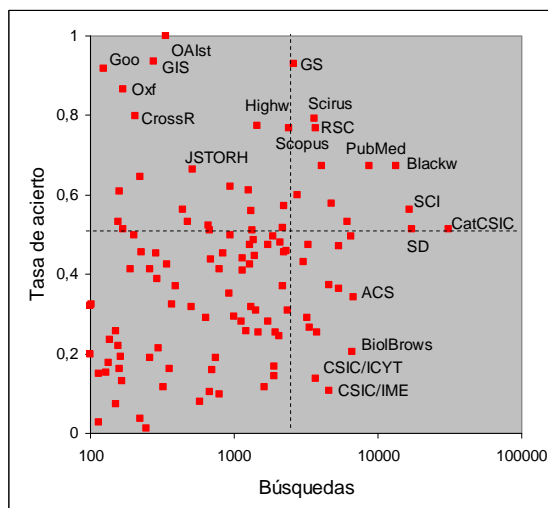


Aunque las tasas de error y acierto sean las expuestas en el párrafo anterior, la percepción del usuario es otra debido a que a mayor parte de las búsquedas se lanzan de forma paralela a varios recursos. La agrupación de las búsquedas tiene como efecto que, aunque en muchas de las metabúsquedas exista algún recurso que devuelva error (el 22%, de 36.128 búsquedas agrupadas), es muy raro que toda la búsqueda en conjunto resulte errónea (sólo el 1% de los casos). De forma similar, cuando se agrupan las búsquedas, el porcentaje en las que existe al menos un recurso que devuelve algún resultado sube hasta el 78% (sobre 35.574 búsquedas agrupadas).

Un gráfico de la distribución de los recursos electrónicos en función de su uso (número de búsquedas) y tasa acierto (figura 7) puede ser útil para identificar patrones que ayuden a la gestión de Metalib. Por ejemplo, se podrá pensar en potenciar, incluyéndolos en grupos de búsqueda rápida, los recursos del cuadrante superior izquierdo, poco usados pero con altas de acierto, en retirar de éstos grupos los del cuadrante inferior derecho, muy usados pero con bajas

tasas de acierto, o en retirar por completo los de la esquina inferior izquierda, poco usados y con bajas tasas de acierto.

### Recursos: Búsquedas vs.tasa de acierto



Una representación similar, pero utilizando las tasas de error en lugar de las de acierto, puede servir para identificar recursos cuyo uso sea potencialmente problemático por sus altas tasas de error (p. ej., en nuestro caso OAlster, con una alta tasa de acierto pero con una tasa de error del 48%)

### Conclusiones

- El número real de sesiones viene a ser de 1/3 del que reflejan la estadísticas, si se descartan las sesiones automáticas de expiración de sesión.
- La mayor parte de las sesiones reflejan un uso bastante sencillo y directo de Metalib:
  - Sesiones de pocos pasos
  - Con una sola funcionalidad
  - Los enlaces tienen mayor peso que las búsquedas
  - La búsqueda rápida es con mucho la más utilizada
- La tasa global de error en las búsquedas (5%), aunque no es alta no es tampoco despreciable.
- La tasa global de búsquedas con resultado positivo es relativamente baja (53%)
- Las búsquedas agrupadas (federadas) tienen como efecto disminuir la tasa de error percibida por el usuario hasta el 1% y aumentar la tasa de acierto hasta el 78%
- La distribución de las tasas de acierto y error con respecto al número de búsquedas puede proporcionar un método para analizar la utilidad de los recursos