Communiqués de presse

Grâce à IBM, la British Library préserve l'information du web pour les générations futures

Paris - 04 mars 2010: IBM annonce sa collaboration avec la British Library – la Bibliothèque Nationale britannique – sur un projet visant à préserver et analyser l'information en ligne avant qu'elle soit détruite.

Nouvelle technologie analytique, le prototype Bigsheets d'IBM permet d'extraire, d'annoter et d'analyser visuellement de vastes quantités d'informations sur le web, via un navigateur. Il accélère ainsi le processus d'archivage, en capturant les données avant qu'elles disparaissent (la durée de vie moyenne d'un site web étant entre 44 et 75 jours).

La British Library reçoit une copie de chaque publication produite au Royaume-Uni et en Irlande, soit quelques 150 millions de documents, qu'elle doit archiver. En parallèle, et depuis 2004, elle archive également les pages web d'une sélection de noms de domaines britanniques. Grâce à BigSheets, les futurs utilisateurs pourront accéder à un large panel d'archives Web, issues de sites historiques.
****

IBM HELPS BRITISH LIBRARY PRESERVE INFORMATION ON THE WEB FOR GENERATIONS TO COME

New Analytics Software Makes Searching Vast Amounts of Data Easy

IBM (NYSE: IBM) today announced it is working with the British Library on a project that will preserve and analyse terabytes of information on the Web before it is lost forever.

The new analytics software project, called IBM BigSheets, helps extract, annotate and visually analyse vast amounts of Web information using a Web browser. IBM's new technology prototype is helping the British Library archive and preserve massive amounts of Web pages, and then unlock the virtual door to its archives for generations to come.

IBM's new analytics technology is helping the British Library speed up the archival process before Web data is lost forever. The Web is rapidly changing with new pages created every day causing an explosion of data that is disappearing almost as quickly as it is published. Recent research estimates the average life expectancy of a Web site is just 44 – 75 days. In turn, every six months, 10 percent of Web pages on the UK domain are lost.

"IBM BigSheets does for big data what spreadsheets did for personal computing," said Rod Smith, vice president, Emerging Internet Technologies, IBM. "Within a matter of minutes, researchers, academics and students will be able to search many terabytes archived Web pages from the UK domain, analyse the results and effortlessly visualise the results of the search."

Preserving Data for Generations to Come

Each year more than six million searches are generated by the British Library online catalogue, and nearly 400,000 people visit the British Library reading rooms, looking for information. The British Library receives a copy of every physical publication produced in the UK and Ireland, amounting to more than 150 million maps, manuscripts, musical scores, newspapers and magazines that it must archive. Beyond just the physical

assets, the British Library has been archiving selected Web pages from the UK domain since 2004. With BigSheets, users of the Library in the future will be able to access vast archives of historic Web sites, and easily research and analyse their queries and visualise the results of the search.

"We estimate the UK Web space will contain over 11 million Web sites by 2011. To take on the enormous challenge of capturing this content, we need a system capable of taking the UK Web Archive to Web-scale," said Helen Hockx-Yu, Web Archiving Programme Manager, The British Library. "IBM can help us analyse the web archive containing millions of pages and unlock embedded knowledge which otherwise is difficult to discover using traditional search methods."

Whether it's someone interested in their own genealogy or a student working on a project for school, people need help making sense of this growing sea of information on the Web. For example, the 2005 election marked the first attempts by UK politicians to use the Web as a campaigning tool. With the use of Web campaigns expected to explode during the 2010 election, the 2005 collection will enable researchers studying the evolution of politics and the Web to access hugely valuable primary source material.

BigSheets: The Technical Foundation

This year, the amount of digital information is expected to reach 988 exabytes which is the equivalent to a stack of books from the Sun to Pluto and back. The Web is exploding with data and business professionals want to access that data -- both structured and unstructured -- to get better insights to their business. IBM BigSheets is an insight engine that helps businesses get insights from really large data sets easily and in a timely manner. By building on top of the Apache Hadoop framework, IBM BigSheets is able to process large amounts of data quickly and efficiently.

IBM BigSheets is a new technology prototype. Users can explore and generate new data insights using a Web application and then the IBM software publishes Web 2.0 standard data feeds which can be searchable by British Library patrons.

BigSheets is an extension of the mashup paradigm that integrates gigabytes, terabytes, or petabytes of unstructured data from Web-based repositories; collects a wide range of unstructured Web data stemming from user-defined seed URLs; extracts and enriches that data using an unstructured information management architecture; and lets the user explore and visualise this data in specific, user-defined contexts. For example, users can see search results in a pie chart and look at the data in a tag cloud.

For more information about IBM's Emerging Technology projects, visit
http://www.ibm.com/software/ebusiness/jstart/

For more information about the British Library Web Archiving Programme, visit
http://www.bl.uk/aboutus/stratpolprog/digi/webarch/index.html