UNITED STATES PATENT AND TRADEMARK OFFICE

OFFICE OF THE GENERAL COUNSEL

December 01, 2009

Mr. Carl Malamud
Public.Resource.Org
1005 Gravenstein Hwy, North
Sebastopol, CA 95472

Re: Freedom of Information Act (FOIA) Request No. F-10-00031

Dear Mr. Malamud:

The United States Patent and Trademark Office (USPTO) FOIA Office is in receipt of your letter dated November 7, 2009, requesting, under the provisions of the Freedom of Information Act (FOIA), 5 U.S.C. § 552, a copy of: "any submissions received by your office in response to your Solicitation for USPTO's Data Dissemination Solution. [55-PAPT -09-10008]"

The USPTO identified 93 pages of documents that are responsive to your request. 33 pages of the documents are releasable. A copy of this material is enclosed.

The remaining 60 pages of responsive documents are not suitable for public disclosure through the Freedom of Information Act pursuant to 5 U.S.C. § (b)(4). They are copies of the proposals voluntarily submitted to the USPTO in response to its Request for Information wherein the confidentiality of these submissions was assured.

Since this commercial information was submitted voluntarily, and it was not the type of information customarily disclosed by these submitters to the public, the information is prohibited from disclosure through the Freedom of Information Act under the Exemption (b)(4), 5 U.S.C. § 552(b)(4), which protects "trade secrets and commercial or financial information obtained from a person and privileged or confidential." See Critical Mass Energy Project v. NRC, 975 F.2d 871, 872-73 (D.C. Cir. 1992) (en banc); National Parks & Conservation Association v. Kleppe, 547 F.2d 673, 682-83 (D.C. Cir. 1976); Professional Review Organization of Florida v. HHS, 607 F. Supp. 423, 425-26 (D.D.C. 1985).

Since the processing costs of this request did not exceed $20.00, applicable fees are hereby waived. See 37 C.F.R. § 102.11(d)(4). Accordingly, your request for a fee waiver is moot, and its merit was not evaluated.

This Exemption (b)(4) withholding determination constitutes a partial denial of your request for records under the FOIA. The undersigned is the denying official. You have the right to appeal this initial decision to the General Counsel, United States Patent and Trademark Office, P.O. Box 1450, Alexandria, VA 22313-1450. An appeal must be received within 30 calendar days from the date of this letter. <u>See</u> 37 C.F.R. § 102.10(a). The appeal must be in writing. You must include a copy of your original request, this letter, and a statement of the reasons why the information should be made available and why this initial denial is in error. Both the letter and the envelope must be clearly marked "Freedom of Information Appeal."

Sincerely,

Robert Fawcett
FOIA Officer

Enclosure

Tugbang, Vanne (Viola)

From:	David LeDuc [dleduc@SIIA.net]
Sent:	Monday, November 16, 2009 1:50 PM
To:	Tugbang, Vanne (Viola); Public Data Dissemination
Cc:	David LeDuc; Mark Bohannon
Subject:	SS-PAPT-09-10008 - Comments Submitted by SIIA
Attachments:	SS-PAPT-09-10008_SIIA_20091116doc.pdf

Hi V. Anne,

Thanks for taking the time to talk with me on Friday about this solicitation, the upcoming RFP process and about the recent announcement of the sole source contract with Google. Please find enclosed comments on behalf of the Software & Information Industry Association. Please associate these comments not only with the solicitation, but also the recently announced sole source contract. If you have any questions or would like to discuss, do not hesitate to contact me.

Thanks again.

Best regards,

David LeDuc
Senior Director, Public Policy
Software & Information Industry Association
www.siia.net
office: 202-789-4443
mobile: 703-220-5943

Software & Information Industry Association
1090 Vermont Ave NW Sixth Floor
Washington, DC 20005-4095

SIIA

November 16, 2009

Mr. John B. Owens II
Chief Information Officer
U.S. Patent and Trademark Office
P.O. Box 1450 – Mail Stop 6
600 Dulaney Street, MDE, 7th Floor
Alexandria, VA 22313-1450

RE: USPTO Data Dissemination Solution Solicitation (SS-PAPT-09-10008)

On behalf of the Software & Information Industry Association (SIIA), thank you for the opportunity to submit comments to the Request for Information (RFI) regarding the U.S. Patent and Trademark Office (USPTO) Data Dissemination Solution Solicitation # SS-PAPT-09-10008.

SIIA is the principal trade association of the software and digital information industries, representing nearly 500 leading technology companies that provide the backbone of the Internet through the development of cutting edge software applications and digital information services. SIIA members include many companies that have long played a critical role in promoting and enhancing public access to government information, providing many information products and services based in whole or in part on government information.

Government information is a critical national asset. SIIA strongly supports the President's commitment to harnessing new technologies to rapidly disclose information and engage citizens, particularly policies and practices that lead to a diversity of sources for the public to access this information. SIIA also supports the USPTO taking efforts to embrace the President's goal to increase transparency of government information "by making data available directly to the public." In your efforts to establish a data dissemination solution that accomplishes these objectives, both in the long term solicitation process and the short-term sole source contract for data dissemination, I urge you to uphold the critical guidelines established by the Paperwork Reduction Act (PRA).

The PRA established three critical requirements to ensure that the public has timely and equitable access to Government information: (1) require that when agencies maintain information in electronic format, they provide timely and equitable access to the underlying data, (2) encourage a diversity of sources, including both public and private sources, for government information, and (3) require agencies to avoid, unless specifically authorized by statute, establishing an exclusive, restricted, or other distribution arrangement that interferes with timely and equitable availability of public information. [1]

[1] 44 USC § 3506(d)

Having reviewed the RFI released on September 4, 2009, as well as subsequent information provided at the Public Meeting held on September 24th, SIIA is concerned that the proposed Data Dissemination Solution could have an adverse effect on improving public access to valuable USPTO information.

SIIA is pleased that the Solicitation identifies as a primary requirement for the selected vendor, or group of vendors "to make the data that is provided by the USPTO available to the public on a no charge basis." However, in providing as an incentive for a vendor to "maintain, repackage (add value), distribute, and sell any resulting enhanced data sets and retain any fees collected," it is a likely outcome that the selected vendor(s) would seek to delay access to the bulk data by competing vendors, or to discriminate by some other means, in order to maintain an advantage and recoup the investment costs of providing the necessary IT infrastructure services required by the USPTO.

As you know, the current system that provides for public access to this data enables a wide range of vendors to compete on a level playing field to provide access to value-added products and services based on the data. This current system is consistent with information policies, including the requirements of the PRA. To ensure continuation of the same level of competition in value-added products and services, it is critical that the level playing field for access to this information is not diminished.

Therefore, SIIA urges you to make it a requirement of any contract that access to the bulk data be provided in a timely manner, without discrimination based on the recipients' intended usage of the data, and that there not be any other inherent competitive advantages. For instance, we urge that the data be made available without either unnecessary delay or any association with the vendor(s)' brand(s) which would otherwise put the vendor at a significant competitive advantage for any added value services based on USPTO data that it might provide.

A failure to require equitable and timely access, essentially a level playing field for access to the underlying USPTO data in bulk would provide an exclusive distribution arrangement, virtually ensuring that there would not be fair competition in the market to add value and disseminate this information. Such a result would be to the detriment of public access to USPTO public information, therefore proving contrary to the President's laudable objectives to maximize openness and access to Government information. Further, providing an exclusive arrangement without a level playing field could also be inconsistent with antitrust laws, in that it would effectively disadvantage other vendors from competing in the sale of digital value-added information products and services based on USPTO information.

Thank you again for the opportunity to provide feedback on this proposal to make USPTO patent and trademark data more easily accessible to the public. As you continue to consider alternatives to accomplish this important initiative, we welcome the opportunity to work with you to help craft an approach that will meet the best interest of the USPTO and citizen access to this information. If you would like to discuss, please do not hesitate to contact me at 202-789-4443.

Sincerely yours,

Ken Wasch

Ken Wasch
President

Public Data Dissemination RFI Questions and Answers

No.	Question	Answer
Questions Received Through 2:00 pm, September 22, 2009		
1	USPTO delivers various data in limited quantities to the public today. Will USPTO continue to perform this function once data is delivered to a contracted partner? Will USPTO dissemination also be at no charge to the public?	The goal is to enter into a partnership where the awardee(s) remove the stress of data delivery from the USPTO infrastructure by delivering the data to the public. However, USPTO will continue to sell packaged data as long as there is a demand for that service.
2	Will third party partners (awardees) be required to enhance the data they receive from USPTO before providing it to the public for free? If yes, what are the requirements?	No. Awardees must pass the data that USPTO provides to them on to the public, as is, at no charge. They may also enhance the data or delivery mechanism and charge for the enhancements.
3	If the data that USPTO provides to awardees is passed on to the public unaltered, will USPTO continue to be accountable for the accuracy of the data?	Yes. Legal responsibility for the accuracy of the unaltered data remains with USPTO.
4	This RFI seems focused on PAIR data. Does USPTO plan to disseminate other, existing data sets, such as patent full-text, via this same mechanism? If so, will they also be distributed to the public at no charge?	Yes, USPTO intends to use this mechanism to disseminate all public data, including existing fee-based data sets via this mechanism, at no charge to the public. However, USPTO will continue to sell packaged data as the agency does today for as long as there is a demand for the packaged data.
5	Does USPTO plan to scan or convert paper files into electronic form? Does USPTO expect awardees to do that?	USPTO has no plans to convert existing paper-based files to electronic form at this time. USPTO encourages awardees to include plans to perform this function if they believe that doing so provides a benefit to the awardee and/or the public.
6	What kind of technical support would awardees be expected to provide?	Data expertise on the data that is provided by USPTO would be provided directly to customers by the USPTO. Data that has "added value", as well as technical assistance with delivery mechanisms would be part of the awardees' responsibility.
7	What is the historical and projected percent growth per year of the data sets?	Attachment 1, Current Data Sets provides the size and growth data for each current and prospective data set. USPTO encourages offerors to propose, in their responses to the RFI, methods for reformatting the data to reduce storage and bandwidth requirements for subsequent distribution.

No.	Question	Answer
8	Does the USPTO expect to add additional data sets in the future? How would these be handled in the contract?	USPTO intends to make all of the existing data sets available under this arrangement and we do not currently anticipate the introduction of new data sets. Handling of any new data would have to be addressed through subsequent contracting actions.
9	Is the vendor expected to process, validate, correct, or improve, any of the data provided by the government for free dissemination?	No. The data passed on to the public does not need to be modified in any way but must be provided to the public at no charge. However, USPTO encourages offerors to add value to the data and distribute the enhanced data for profit.
10	Has/will the USPTO considered opening this opportunity under an 8(a) set-aside?	The USPTO will consider setting this opportunity aside if it is determined that at least 2 -8(a) vendors are capable. At this time an acquisition strategy has not been made.
11	The RFI indicates that responses will only be accepted in Microsoft Word. Is there any reason that responses cannot be accepted in other standard industry formats such as OpenOffice or PDF? Has Microsoft furnished any monetary or in-kind compensation as part of the requirements for use of Microsoft Word, Microsoft Visio, and Microsoft Excel for submissions?	The Microsoft Office suite of products is USPTO's standard tool set. Accepting electronic submissions in other formats may inhibit the agency's ability to read the files. Responses may be provided in PDF format as well. Microsoft has not provided compensation of any kind relative to this RFI.
12	It was unclear if all bids will be made public after the redaction of any unmarked proprietary information. Knowing what bids have been submitted will allow the public a greater ability to understand the tradeoffs made by the U.S. Patent and Trademark Office in evaluating the submissions and for any subsequent decisions to issue an RFP.	Information submitted as part of the RFI is for market research and planning purposes of the USPTO only. The office does not intend to make the submissions public any responses to the RFI that are marked proprietary or confidential
13	Given the fundamental and long-lasting repercussions of this initiative, has any thought been made to holding a West Coast or Midwest open meeting in addition to the one being held inside the Washington Beltway? Likewise, will audio and video from the meeting be made available to the public on the Internet?	USPTO will consider alternatives for this request.
14	As this initiative may fundamentally alter your distribution strategies, will the Under Secretary and other members of the U.S. PTO senior management team also be participating in this process?	USPTO senior management will participate in the process.

No.	Question	Answer
15	Can USPTO provide more details of the 2 petabytes of data: a. Details of the data formats and volumes for each data set that is available in electronic format? b. Details of the data sets that are not currently in electronic format	Yes. Attachment 1 to the RFI contains a broader description of the data being considered within the scope of this effort.
16	Can USPTO provide an overview of its current data dissemination platforms and operations?	Current bulk data products are distributed via download, magnetic tape, or optical disc, depending on factors such as size and frequency. Some products are provided as an annual subscription. While some data sets are readily produced as part of current data processing operations, others would have to be periodically extracted from internal databases.
17	What is the current approach to timing for making bulk data available to the public and to commercial vendors? What are the goals and objectives of the PTO RFI as it relates to these data sets? (status quo, versus enabling data mining, versus additional system enhancements)	Attachment 2 to the RFI represents the current plan for making USPTO data available to the public through data.gov. The goal of this effort is to enhance the accessibility and usability of USPTO data in an accelerated manner. There are two objectives: 1) to offer no-charge access to current bulk data sets; 2) to offer no-charge access to data that is not currently available in bulk (e.g., PAIR)
18	What are USPTO's initial expectations for Service Level Agreements, Quality Control, and Accessibility? a. Requirements for uptime / peak access b. Scheduled maintenance	Certain data sets must continue to be provided at weekly intervals (Tuesdays for Patent Grants, Thursdays for Patent Applications). Other data should be distributed at least as frequently as currently done: distributed daily (Trademark Applications, and Trademark Trial and Appeal Board data, as well as Trademark and Patent Assignment data); or bimonthly for classification data. Attachment 1 to the RFI contains the issue frequency of current data sets. Service level agreements and requirements to include quality and accessibility have not yet been finalized. These requirements will be incorporated into any subsequent procurement actions.
19	Should data exchange with other Patent Offices, including the delivery of data prior to the official publication date, be covered by any proposed Data Dissemination solution?	No data is disseminated prior to the official publication date. The USPTO exchanges data with a limited number of other Patent Offices (< 10), and would consider using a 3rd-party solution for that delivery.

No.	Question	Answer
20	Would all components of a data dissemination solution need to be located in the United States?	USPTO believes that the systems used to transform existing data sets for high volume dissemination need to be located in proximity to the existing data, most likely within the USPTO data center, because of the limited capability to move the data. The systems used to disseminate the data could be located anywhere.
21	What if any restrictions would be put in place on the internal or commercial use of data?	No restrictions other than timely and equitable delivery of the unaltered data provided by USPTO to the public at no cost.
22	Would a successful candidate be allowed to apply their brand to the public data dissemination platforms?	Yes. However, the raw data provided by USPTO to the successful candidate(s) must be labeled as such so that its authenticity is clear.
23	Would all users and commercial vendors have equal access to the data?	Published data must be delivered to all members of the public, including commercial vendors, within specified time frames in a manner that permits everyone to access the data at the same time without regard to geographic location. The requirements for universal delivery will be included in any subsequent procurement actions.
24	Would a successful candidate have any advantage over public or would there be equal access to the data and equal rights for its use, for all users and commercial vendors?	See the answer to question 23.
25	Can USPTO provide details of the budget allocated for Patent and Trademark Data Dissemination in FY 2009, and the budget proposed for FY 2010?	This information is not applicable to the market research being conducted at this time.
26	USPTO provides bulk data sets to commercial Patent and Trademark information companies and other organizations at marginal cost. Will USPTO continue the marginal cost policy if Data Dissemination operations are outsourced?	Yes. USPTO will continue to sell packaged data as long as there is a demand for that service.
27	What revenues did USPTO derive from Patent and Trademark Information Dissemination in FY 2008 and 2009?	This information is not applicable to the market research being conducted at this time.
28	Would USPTO be willing to provide facilities on the Campus free of charge for a Data Dissemination contractor/'partner?	Yes. The system(s) that is used to repackage the existing data must be hosted at the USPTO facility. The system(s) that are used to disseminate the data to the public may be hosted anywhere.

No.	Question	Answer
29	Liability – would a successful candidate be required to hold and save the Government, its officers, agents, and employees, harmless from liability of any nature or kind, including costs and expenses, for, or on account of infringement of any patent or copyright or any other unauthorized disclosure or use of any confidential secret, or proprietary data, process, product or invention, whether or not patentable, in the performance of an RFI related contract?	Awardees must pass the data that USPTO provides to them on to the public, as is, at no charge. USPTO will retain liability for the accuracy of this data provided that it is unaltered prior to dissemination. The awardee will be solely accountable to it's customers for any data that is altered, repackaged, or modified in any way.
30	Can the US Patent and Trademark Office make available the Road Map and Transformation Plan that was referenced in the FY2010 President's Budget Request?	This information is not applicable to the market research being conducted at this time.
31	Are there budgetary materials, perhaps a CIO's budget, which more clearly defines IT spending on various projects? While some of that was broken out in the FY2009 and FY2010 President's Budget Requests and the 2008 Performance and Accountability Report, more detail would be very helpful.	This information is not applicable to the market research being conducted at this time.
32	How many people work in IT at the USPTO and is this information possibly available broken down by type of position?	This information is not applicable to the market research being conducted at this time.
33	How does the USPTO distinguish, if at all, between data sets of a "bulk" format and data sets of a "machine-readable" format?	The term "bulk data" is used to represent a product that has been assembled as a collection and designed for data processing. "Machine-readable" data includes bulk data, but may also represent data that was designed for display but can also be read by a machine – for example, through Web-based "bots". The USPTO currently restricts machine readability on its public PAIR system through the use of a CAPTCHA mechanism that distinguishes between human- and machine-originating queries.
34	Are such data sets compliant with the USPTO Information Quality Guidelines?	Yes – all data sets are addressed under the USPTO Information Quality Guidelines.
35	Can the USPTO provide more details regarding the data format(s) used in and volume of each data set that is currently available in electronic format?	Yes. Attachment 1 to the RFI contains a broader description of the data being considered within the scope of this effort.

No.	Question	Answer
36	What, if any, data sets are not currently in electronic format?	The data represented in Attachment 1 are all available in electronic format. The bulk of non-electronic data are the patent application file contents from 2002 and earlier. Other non-electronic data would be the data in Attachment 1 that does not fall within the identified date ranges. USPTO encourages awardees to convert paper or microfilm based records into electronic format if they believe that doing so provides a benefit to the awardee and/or the public.
37	In what non-electronic format are such data sets maintained?	Not applicable – see answer above.
38	Would the Selected Vendor(s) be allowed to exclusively obtain and market these data sets?	Not applicable – see answer above.
39	The RFI appears to say that data production functions would remain with the USPTO, but the entire dissemination process would be moved to one or more Selected Vendor(s). Is this an accurate understanding of the RFI?	Yes.
40	In what form and with what frequency would the USPTO provide new data sets to the Selected Vendor(s)?	The intent is to meet or exceed the current data distribution timelines. Attachment 1 to the RFI contains the issue frequency of current data sets.
41	In what form and with what frequency would the USPTO provide changes to the data to the Selected Vendor(s)?	Ideally, changes to the data will be extracted in the existing format from the production systems and transferred to the dissemination systems in real time or near real time. Changes occur daily.
42	What formats are "desired by the Intellectual Property (IP) community"?	PDF is a commonly-requested format that is not currently offered. Custom extracts of well-formed XML data are also desirable.
43	Is there an expectation that a Selected Vendor would standardize and make consistent bulk and machine-readable data formats that are now inconsistent across data sets?	Standardizing the current data is desirable as a component of preparing data sets for distribution in bulk. USPTO encourages awardees to address this in their responses.
44	Will a Selected Vendor be required to use currently used data formats (i.e., the Daily Application "C" file in XML for trademarks) and supply adequate advance notice to current recipients of data of any changes?	See the answer to question 43. USPTO will retain responsibility for notifying recipients of changes to the data formats as it relates to the distribution of the raw data to the public.
45	What are the USPTO's current demands from customers and related IT systems requirements to deliver the content? The RFI suggests that the USPTO does not have the necessary resources to accomplish its objectives. What, if any, budgetary estimates have led the USPTO to that conclusion?	Current requests include unlimited data mining and electronic download of all data sets. USPTO's current infrastructure is incapable of supporting that volume. The budget estimates question is irrelevant to the market research being conducted under this RFI.

No.	Question	Answer
46	What does the USPTO define as a "value-add"? For example, would offering the data sets on a high-speed connected server that allows the download in 1 hour rather than hours/days from the USPTO server be considered a value-add, or would value-add be defined as only data enhancements? Can the USPTO provide other examples of what the USPTO would consider to be value-added and what it would not?	At this point, USPTO is defining "value added" as data enhancements. Examples may include custom extracts of the data, custom combinations of the data to meet specific needs, or mash ups of data from multiple data sources. USPTO will consider other types of distributions enhancements in any subsequent procurement actions.
47	Would any restrictions be put on the Selected Vendor(s) in terms of when they could make the data available in its value added form (for which they would charge a fee) and/or when they can make the data available in a patent research platform for which they would charge a fee?	No restrictions other than the fundamental requirement for timely and equitable delivery of the unaltered data provided by USPTO to the public at no cost.
48	The RFI states that the Selected Vendor(s) must make bulk data available to the public at no charge. Would pricing of the value-added services have any government restrictions placed on it?	None are envisioned at this time
49	What advantages accrue to the benefit of Selected Vendor(s) in providing the services described in the RFI? Clearly, there are significant investments in staff/infrastructure, so there needs to be a way to recover this cost and run a margin.	Selected Vendor(s) may add value and market their value-added products.
50	What are the USPTO's initial expectations for Service Level Agreements and Quality Control?	See question 18.
51	Would the contract contain an "Authorization and Consent" clause under the Federal Acquisition Regulations? Or, would a Selected Vendor be required to hold the USPTO, its officers, agents, and employees, harmless from liability of any nature or kind, including costs and expenses, for, or on account of infringement of any patent or copyright or any other unauthorized disclosure or use of any confidential secret, or proprietary data, process, product or invention, whether or not patentable, in the performance of the contract?	See question 29.
52	Would updating the content with changes be a required activity that the Selected Vendor(s) would have to implement and offer for free (or would the Selected Vendor(s) simply be responsible for making the data provided by the USPTO available to the public in its original form and without comment or consideration for the actual content of the data)?	The vendor would be expected to maintain updates consistent with current USPTO practices, at a minimum. Withdrawn and corrected information is routinely made available to the public under current dissemination practices.

No.	Question	Answer
53	What specific activities does the USPTO see as being included in moving the entire dissemination process to one or more Selected Vendor(s)?	Availability of timely, free, downloadable bulk data at the customer's convenience is the essential activity.
54	How quickly after the data is made available to the Selected Vendor(s) would it have to be made available to the public?	The intent is to meet or exceed the current data distribution timelines. Attachment 1 to the RFI contains the issue frequency of current data sets. Specific time frame requirements for redistributing the data will be included in any subsequent procurement actions.
55	What customer support requirements would the USPTO envision the Selected Vendor(s) providing (e.g., only support for the technical downloading of the raw government data)?	Vendor support is expected to be limited to supporting the data delivery mechanisms. Support for the raw data would continue to be provided by the USPTO.
56	What customer support requirements does the USPTO envision they will maintain for this data? For example, Selected Vendor(s) will not be able to comment on the accuracy of the data. Will the USPTO maintain a resource to answer questions about the data that they provide?	The USPTO will provide support for the raw data content.
57	How will the USPTO react to the reporting of data errors either from the Selected Vendor(s) or from members of the general public? Will the USPTO maintain full responsibility for communicating to the public about data errors and plans for fixing the same?	Awardees must pass the data that USPTO provides to them on to the public, as is, at no charge. USPTO will retain liability for the accuracy of this data provided that it is unaltered prior to dissemination. The awardee will be solely accountable to its customers for any data that is altered, repackaged, or modified in any way.
58	Would parties other than a Selected Vendor have any restrictions on their ability to use, package or distribute data (e.g., in the form of search reports or via web services)?	None are envisioned at this time.
59	Is there an expectation that a Selected Vendor would make applicable data available via distribution mechanisms similar to those mechanisms employed by the USPTO (e.g., HTTP)? Would controls exist to ensure a Selected Vendor could not make significant changes to distribution mechanism and procedures without advance notice? Would controls exist to ensure a Selected Vendor does not delay data dissemination to other providers or create a discernable competitive gap by making data either more complete or more current from their own systems?	The objective is to increase access through better distribution mechanisms. Appropriate controls will be evaluated after consideration of the information collected through this RFI.

No.	Question	Answer
60	Would the USPTO be responsible for initially keying in new records, such that information pertaining to new records would be channeled to a Selected Vendor's infrastructure at the USPTO?	The USPTO will manage the source data.
61	Would the USPTO be responsible for electronically updating existing records, such that information pertaining to updated records would be channeled to a Selected Vendor's infrastructure at the USPTO?	The USPTO will manage the source data.
62	Would the USPTO be responsible for performing corrections to existing records and adding cross-indexing to records, such that all corrections would be channeled to a Selected Vendor's infrastructure at the USPTO?	The USPTO will manage the source data.
63	Is there an expectation that a Selected Vendor would make available in bulk or machine-readable format all existing information that is not currently in bulk or machine-readable format (e.g., all TDR information for trademarks)?	The Vendor would be expected to provide unaltered data for all data sets identified in Attachment 1 to this RFI.
64	Would the USPTO consider alternatives to the proposed in-house infrastructure solution, such as an infrastructure hosted completely offsite?	Yes.
65	Would the USPTO continue to maintain public interfaces for searchability (e.g., those available at the USPTO.GOV), independent of any systems that may be maintained by a Selected Vendor?	Yes.
66	Is there an expectation that a Selected Vendor would maintain all of the nearly two petabytes of data in the original formats currently maintained by the USPTO (e.g., as archives)? Is there an expectation that a Selected Vendor would make such data available in bulk or machine-readable formats? Or, might there be an understanding that this information could be maintained in the current format?	Information may be maintained and distributed via the most effective method for timely and equitable distribution, provided its content is unaltered from the authoritative data source.
67	Would there be any responsibilities for regularly contributing data to either the Trademark Official Gazette or the Patent Official Gazette?	No.
68	Would there be an option to implement an infrastructure solution only for trademarks, and not involving patents?	Not at this time.

No.	Question	Answer
69	The Obama Administration cites small businesses as important engines of economic recovery. Yet the ambitious size and complexity of the proposed endeavor precludes all but the largest Information Technology organizations from bidding successfully. Thomson, Reed, Google, Microsoft, IBM, and EDS readily come to mind. How does the Administration foresee small businesses and minority-owned businesses participating meaningfully in this development opportunity?	At this time an acquisition strategy has not been made. Teaming arrangements and subcontracting possibilities will be considered.
70	Much of the material submitted to the Patent Office is a matter of corporate confidentiality and even national security at some time in its life-cycle. What limits will Homeland Security place on vendors that have large proportions of foreign nationals in their employ?	Only published data will be available for dissemination. It is anticipated that any applicable IT or other security provisions will be included in any resulting contract.
71	The RFI contains a tacit admission that USPTO's Information Technology infrastructure is inadequate to the task of implementing the envisioned system effectively and is willing to surrender that responsibility. Yet it is axiomatic that the management and oversight of an Information Technology project requires substantially more expertise than the implementation tasks. How can the public and Intellectual Property community credibly trust that USPTO has the wisdom and capacity to oversee and manage this endeavor to a successful completion?	This comment is not specifically relevant to the market research currently being conducted.
72	The proposed scheme franchises Governmental Functions to the private sector in return for 'free' systems development for internal USPTO processes and some level of 'free' public access to the resulting database. The vendor, in turn, will profit by selling add-on features and data repackaging to its captive public audience. What USPTO mechanisms are in place to prevent the vendor from setting the 'free' baseline so low that the public will have to pay for everything beyond the trivial and inadequate data access the USPTO presently provides?	The requirement is for timely, equitable, and no-cost access to USPTO data in bulk.
73	The RFI depicts a hypothetical outsourcing proposal that simultaneously conserves agency resources, promotes free public access to public data, and provides adequate profit margins to the private sector. Can USPTO point to a precedent for a completed project of this nature and magnitude that optimally balances these competing interests?	Although we cannot point to a project of identical nature and magnitude, the USPTO is aware of other Federal projects whose aim was to provide free access to public data at no cost to the Government while providing commercial opportunities for the private sector. The Government Printing Office recently announced a contract opportunity for the digitization of documents at no cost to the government.

No.	Question	Answer
74	What development time-frame does USPTO envision for procurement, implementation, and operational life? From the ambitious scope, this looks to be a five-year project at very least. Does USPTO intend to suspend needful improvements to its existing prepublication platforms in deference to this new enterprise?	An implementation timeframe has not been established. Responders to the RFI are encouraged to provide thoughts on the timeframe for implementing their proposed solutions.
75	Federal Information Technology procurements have a sad history of being overweight at birth, obese during midlife, and sclerotic in old age. Some have been obsolete even before completion – FBI's systems come readily to mind. Surely USPTO recognizes the problem since some of its own systems are in a state of legacy paralysis. They are so ponderous and baroque that only the incumbent contractor can successfully bid on maintenance. What novel design approach does USPTO intend to employ to insure that this effort does not go similarly awry?	This comment is not specifically relevant to the market research currently being conducted.
76	The RFI mentions mostly prepublication and assignee data but is silent regarding the Application and Grant Red Book and Yellow Book data which has been routinely packaged and sold in bulk by USPTO for over 30 years. This is likewise public data. Is a similar scheme anticipated by USPTO for its dissemination? Most of its users would readily state that it is priced too high relative to other Patent Authorities such as EPO, JPO, WIPO and SIPO. Yet these users do not uniformly embrace 'free' distribution for fear data quality will suffer. Has USPTO studied the impact such a change would have on these ongoing activities?	The Red Book and Yellow Book data is included in this RFI, and is described in Attachment 1. In accordance with the President's Open Government Initiative, the intent of this RFI is to make data more accessible.
77	Is the data all electronic?	Yes.
78	If there is physical data, is there an expectation to convert the data to electronic within the scope of this project	Not applicable under this RFI.
79	Can you clarify what you mean by machine readable format?	See answer to number 33.
80	The RFI describes an estimated 2 petabytes of data sets maintained by USPTO. Could more detail about each of these data sets be provided such as: a. Format(s) b. Number of records/pages/documents c. Estimated size of each data set	Yes. Attachment 1 to the RFI contains a broader description of the data being considered within the scope of this effort.
81	Could sample documents of each data set be provided?	USPTO will consider providing example data sets with any subsequent procurement action.

No.	Question	Answer
82	Is it a requirement that any potential solution be housed at USPTO offices? Could a vendor supplied location be used instead?	USPTO believes that the systems used to transform existing data sets for high volume dissemination need to be located in proximity to the existing data, most likely within the USPTO data center, because of the limited capability to move the data. The systems used to disseminate the data could be located anywhere.
83	Is the hardware and software to be mandated by the USPTO or does the vendor have the ability to provide that as part of a proposed response?	Responses should propose the tools that will be part of the solution.
84	How does USPTO envision that the data to be disseminated at no charge be handled? Would dissemination via the web and / or FTP be sufficient for the USPTO?	Responses should describe any appropriate dissemination methods that ensure timely and equitable access to data.
Questions Received at the September 24, 2009 Meeting		
85	If the goal is dissemination of information, would not this goal be better met by offering an incentive, to disseminators that is, raising the price of the information (not free) to the public?	The goal is to continue to disseminate the data for free as required by law and to meet the access goals defined through data.gov.
86	System is owned by OCIO – would OCIO consider code rights to be open sourced?	USPTO will not rule out use of open source code at this stage of market research. We are moving to more open source. However, security concerns may limit where we can use open source.
87	How do the plans for data.gov as specified in the handouts relate to this RFI?	Attachment 2 to the RFI represents the current plan to make data available to the public through data.gov. We hope to accelerate that timeline by partnering with external service providers as defined in the RFI.
88	Is USPTO willing to consider access to the source data via API?	Yes, however the solution will have to overcome any security concerns that may be introduced through this method.
89	Is the vendor expected to perform the data transformation?	No, that task would be performed by USPTO personnel. The vendor would design the system and USPTO and the vendor would jointly deploy it.
90	On whose site would the data be transformed?	See answer to question 82.
91	How big is USPTO's current Internet pipe and will USPTO allow a separate "pipe" between USPTO and the vendor for transmission of the data?	USPTO currently employs a single OC3 line and a T3 line used primarily for backup. Yes, USPTO will allow a separate pipe between USPTO and the vendor.
92	Will there be an Authorization of Consent to protect the vendors?	As the USPTO is conducting market research and exchanging information during this RFI, neither the acquisition strategy nor the solicitation clauses have been selected. However, the USPTO anticipates that all required FAR clauses, including FAR §52.227-1 Authorization and Consent, if applicable, will be included.

No.	Question	Answer
93	Will USPTO consider splitting the effort? For example, one procurement for improving the infrastructure (cost paid by USPTO), and one procurement for hosting the data (the vendor)?	Not at this time.
94	What are USPTO expectations for the timeliness of data updates?	See answer to question 18. Specific requirements have not yet been established.
95	How much of the data is comprised of patent products, and how much is trademark?	The amount of trademark data is very small compared to the amount of patent data. See Attachment 1 for more details.
96	How will the data be monitored to ensure it is kept current?	The system will continuously monitor the data for changes.
97	Who will maintain the data as it changes?	The USPTO will maintain the data.
98	The data sets listed in Attachment 1 do not add up to 2 petabytes. Why the difference?	USPTO's total data storage of 2 petabytes includes duplication of data in multiple systems.
99	Where will the data be hosted?	Can be anywhere, preferably outside the USPTO so it does not impact our production operations. Bulk data will be check-summed so we can verify authenticity.
100	What sort of internal infrastructure changes are needed – at a high level?	We hope to separate the dissemination infrastructure from the internal prosecution systems and reduce the load on our existing infrastructure.
101	How does this initiative relate to Director Kappos' desire to separate the Trademark systems?	The two initiatives are independent.
102	Will PTO continue to host search systems for the public?	Yes.
103	Would you like to hear about search system improvements in the RFI response?	Changes to our search systems are outside the scope of this RFI but we always welcome suggestions.
Questions Received After the September 24, 2009 Meeting		
104	The RFI anticipates that a vendor who would operate what is now known as PAIR for the benefit of the USPTO would receive the rights to commercial distribution of the data in PAIR. Is it the position of the USPTO that all of the data included in PAIR today is eligible for repackaging and redistribution? What limits would you propose to place on the third party commercial vendor of PAIR data?	*The USPTO will continue to operate PAIR. The vendor will extract data from PAIR and distribute it in bulk. The extracted data, which includes all data contained in Public PAIR, will be freely distributed to the public. The vendor may enhance the data and market the enhanced data set.*

No.	Question	Answer
105	Within PAIR, search reports may depict third party search systems, search strategies used by USPTO examiners and STIC personnel, and in some instances copyrighted material (database records) retrieved from those systems. Under the current PAIR system, it would be inconvenient at best to aggregate this data together and turn it into a product for commercial distribution, since PAIR can only be searched using a single patent number at a time. The next generation PAIR product may provide different search capabilities, and may also provide a means for customers who should be searching third party systems to instead substitute some new product created from PAIR instead. Will the USPTO identify and remove copyrighted materials from documents in PAIR before granting commercialization rights to the PAIR vendor? Alternately, will the USPTO restrict the types of PAIR data that can be commercialized by its third party vendor(s), specifically by not allowing the sublicensing of copyrighted data?	*USPTO will not identify and remove copyrighted materials. We do not know, at this time, what limits might be placed on copyrighted material.*
106	Some USPTO data products are apparently already being created out of the PAIR data. One such example is the Daily Assignments File, which provides information on inventor rights assignments, reassignments and other important events relevant to US patents. a. Can you confirm that the source of the Daily Assignments File is in fact certain documents currently posted within PAIR? b. Do you expect that the Daily Assignments File would continue to be offered by the USPTO, or would its distribution rights transfer to the third party who had won the rights to commercialize PAIR data? c. Would the USPTO stipulate any type of pricing protection for current subscribers to the Daily Assignments File should responsibility for its creation and distribution transfer to a third party?	*Although assignments are intended to be distributed as part of this effort, they are not currently part of the Public PAIR data. The vendor must freely distribute the data provided by USPTO, including the Daily Assignment File, but may enhance and market the enhanced data set.*
107	The solicitation describes a data dissemination problem in terms of the data for distribution, and the potential public users of the data. What is the estimate of the number of users that will be interested in this data?	*The number of users varies by specific data set. Overall, there are currently about 50 users. However, that number is likely to rise significantly once the data sets are offered to the public free of charge. The number of potential users for Public Pair bulk data is unknown at this time.*

No.	Question	Answer
108	For a number of years the USPTO has maintained an active website to assist the public with regard to information services. What is the estimate, in terms of document download requests, of the volume of requests that would be estimated to take place during a one week time period, provided that all of the data was made available over the Internet? Your estimate of the volume of requests over a one month and one year time period would also be helpful.	*See Attachment 1 for the Issue Frequency of each data set. Question 107 addresses the potential number of users.*
109	As a follow-on to question number 108, please also specify what the anticipated demand would be for "repackaged" data, as specified in the solicitation.	*The demand for "repackaged" data is unknown at this time, as it would depend upon the particular value added by the vendor.*
110	It has been stated that the data in question is approximately 2 petabytes in size. Please specify the estimated data footprint (summary information) for this data, or, please specify the estimated number of short (1-2 page) documents and long documents that are represented by this number.	Attachment 1 details the data sets and provides estimates of their sizes. USPTO's total data storage of 2 petabytes includes duplication of data in multiple systems.
111	The solicitation mentioned that the data resides in formats that are not machine readable. Please specify these formats.	*Attachment 1 details the formats for each data set.*
112	Please specify all of the formats that the data must be rendered to in order to, as stated in the solicitation, fulfill the desires of the IP community.	*See the answer to question 42.*
113	Any additional comments regarding how the data is segmented, or should be segmented, and what the size estimates are for these volumes of data would be helpful.	*Not at this time.*

MARK
LOGIC

Mark Logic Corporation
999 Skyway Road
Suite 200
San Carlos, CA 94070
+1 650 655 2300 Phone
+1 650 655 2310 Fax
www.marklogic.com

October 15, 2009

V'Anne Tugbang
Procurement Officer
PTO

Re: Mark Logic Data Dissemination RFI Response

Ms. Tugbang:

Mark Logic Corporation is pleased to provide PTO this RFI response to the Data Dissemination RFI. This RFI response includes a description of the Mark Logic Server as it relates to Data Dissemination. Please refer to the RFI document adjunct to this letter.

(1) Company name **Mark Logic Corporation**
(2) Primary Point of Contact, **Paul Norcini**
(3) Phone Number and Email Address, **703.403.4181, paul.norcini@marklogic.com**
(4) Cage Code, **3HZ37**
(5) NAICS Code, **511210**
(6) Business Size, **150 people**
7) Federal Supply Schedule (FSS) Contract Number and SIN, **Not Applicable**

Please review this RFI response and let me know if you have any questions. You can reach me at 703.403.4181.

Sincerely,

Paul Norcini
Civilian Accounts
Mark Logic Corporation

MARK LOGIC

Mark Logic RFI Response
Patent and Trade Mark Office
Data Dissemination Solution

October 15, 2009

Mark Logic Corporation
999 Skyway Road
Suite 200
San Carlos, CA 94070

Proprietary

Table of Contents

Introduction

Mark Logic Corporation is pleased to provide this response to the PTO Data Dissemination Solution Request for Information. Mark Logic Corporation is a products vendor supplying an XML Content Server to ingest, search, analyze and render information. The MarkLogic Server will serve as a platform to store the PTO online content natively as XML, index all aspects of the content and make it searchable, provide analytics on the content and render all or parts of the content to users, other systems or third party applications. The Mark Logic server is an ACID compliant XML database with a search engine sharing the same kernel. The advantages of having the database and search engine as one kernel means that with one atomic write, the data is stored and the search indexes are created. There is no delay in when the information is stored in the repository and when it is available to be searched on the web.

The Mark Logic server is a repository that stores the XML as-is or converts content to XML and automatically builds a search index of every word and metadata field. The content is fully searchable once ingested into the ML repository using XQuery. Features of the Mark Logic repository are:

- Schema agnostic - Can ingest multiple data types without having to define a schema (i.e., NIEM, DDMS, XBRL, KML, GML, PDF, Word, etc.)

- Highly scalable - largest repository is 2 Petabytes for the DNI. A typical 2 CPU server can ingest and provide sub second search across 1 terabyte of content

- Render subsections of content - ML does not only return links to the entire record, but can return any section of the record including elements.

- Built in real-time alerting

- A specialized geospatial index for fast searches across lat/long content (i.e., point, radius, bounding box, polygon) and integration with mapping programs

- SOA based application

- Immediate access to the content upon ingest. There is no delay between when the content can be searched and when it is ingested.

Company Background

Mark Logic Corporation is a Silicon Valley software company providing an XML Server COTS product, ideal for processing (i.e., storing, indexing, searching, analyzing and rendering) content. Mark Logic has approximately 150 commercial and government customers. Mark Logic has three main divisions, including Government, Information and Media, and Enterprise Accounts, selling products in the US and Europe. In our Government sector, Mark Logic has customers in the Intelligence Community, the Department of Defense, and the Civilian Sector. In delivering solutions to this Government customer base, Mark Logic has gained unique experience and insight in the large scale use of XML to solve search, information sharing, discovery, and analysis challenges.

Technical Solution

Mark Logic proposes use of the MarkLogic XML Server as an enterprise repository of content that can address the needs of the PTO. Mark Logic provides the performance and scalability to address the increasing volume of information for data dissemination, while providing the flexibility of managing multiple data schemas and semi structured information in XML.

This section provides an overview of the Mark Logic XML Server. It discusses the core capabilities of the product, and those capabilities are relevant to the needs of the PTO Data Dissemination program.

Mark Logic offers the industry's leading Extensible Markup Language (XML) Server. The MarkLogic XML Server is designed from the ground up to load, query, manipulate and render content in any format. As such, it provides a new solution for integrating multiple sources of information with multiple formats and schemas, serving as a repository for information. It has the ability to handle fully structured, semi structured, and unstructured data, and it is frequently used as a platform for information access and sharing solutions. Figure 1 shows a high level functional depiction of the Mark Logic XML Server.

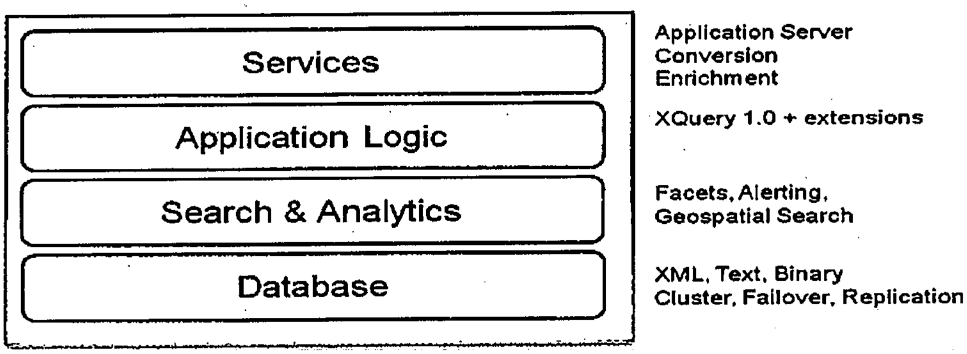


Figure 1 – Mark Logic XML Server Functional Overview

Mark Logic's XML Server has delivered solutions for information sharing; knowledge management; information discover via metadata catalogs; dynamic custom publishing based on XML; open source intelligence; and, search data analytics, among others. Today MarkLogic is being used by Federal customers including the Intelligence Community (IC) agencies, Department of Defense, and Civilian agencies. These customers chose Mark Logic because of the following key attributes:

- Scalability: The ability to manage large, multi-terabyte repositories of XML. Today Mark Logic is deployed in clusters supporting hundreds of terabytes of data
- Performance: March Logic has the unique ability to query XML data in large repositories with sub-second response time
- Agility and Flexibility: Mark Logic provides a powerful query language based on XQuery, the W3C standard for managing XML.
- Lower Total Cost of Ownership (TCO): Mark Logic was designed for 64 bit platforms, and is capable of scaling up to 1 TB on a single quad core server. This gives Mark Logic the ability to manage millions of documents in a small hardware footprint.

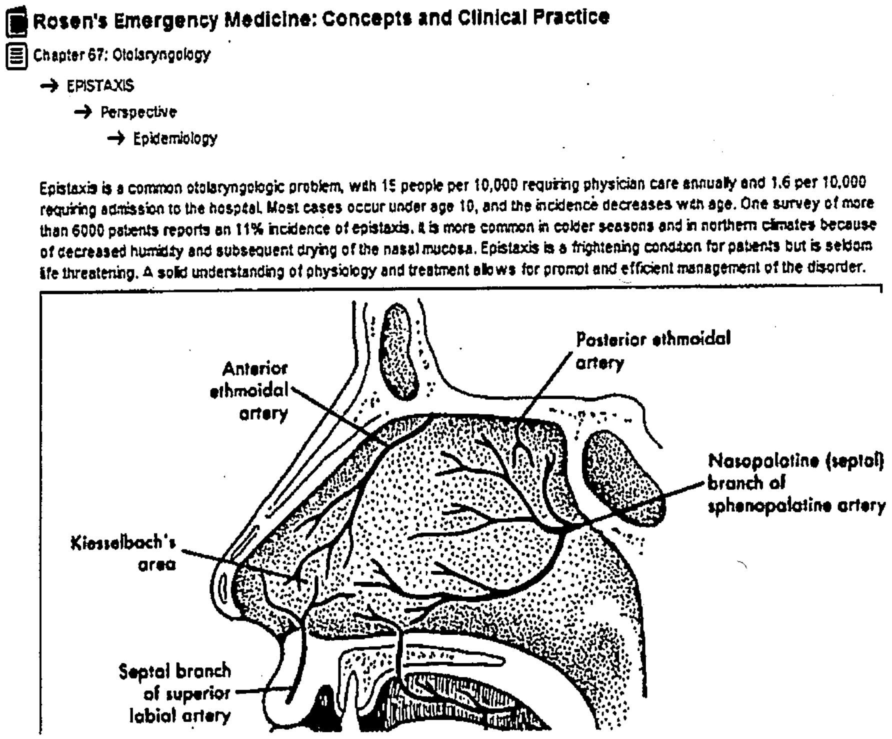
The following sub-sections illustrate some of the unique features of the Mark Logic XML Server.

Transactional Repository – MarkLogic Server is a high performance transactional repository for XML data. As a native XML repository, it has the ability to load XML data as is without any shredding or manipulation. Once in the repository, Mark Logic offers ACID transactional capabilities on this XML data.

Universal indexing - The universal index within MarkLogic Server is automatically populated with both the full-text and XML structure within XML data. These indexes are built in real time when content is ingested or updated --- within the context of the transaction. This single view of information assets leads to faster configuration, lower maintenance costs and increased agility. Other systems often require three or more indexes to achieve the same functionality, drastically increasing the storage and maintenance requirements. This universal index of structure and data enables fine-grained query and retrieval of information. This is significant for the NARA enterprise, because Mark Logic provides a high performance solution for XML data independent of how complex the schema may be.

Schema Agnostic Storage and Indexing – The Mark Logic XML Server is schema aware, but also schema agnostic. MarkLogic does not require upfront schema definition for XML data to deliver functionality and performance. This feature enables the ingestion of arbitrary sources of data conforming to different schemas, without the need to pre-configure the server. Once in the Server, this data can be quickly transformed to a single schema, in this way enabling flexible, all source, information sharing.

Fast, scalable full text and XML search – MarkLogic Server has a complete full text and search capability including keyword, phrase, Boolean expression, wildcard, proximity, thesauri, spell-checking, and highlighting. MarkLogic Server delivers millisecond search and query response times against multi-terabyte content bases. Unlike a search engine that will respond to a simple keyword query with a list of links to documents that contain the keywords, MarkLogic Server pinpoints and returns the specific information sought at the level of granularity required. MarkLogic Server goes straight to the details, providing the exact context required for each query. Figure 2 below shows an example medical library search application, in which the most relevant paragraph is displayed in context together with relevant figures and the breadcrumbs of where in the manual was the information found.



Rosen's Emergency Medicine: Concepts and Clinical Practice

Chapter 67: Otolaryngology

→ EPISTAXIS

→ Perspective

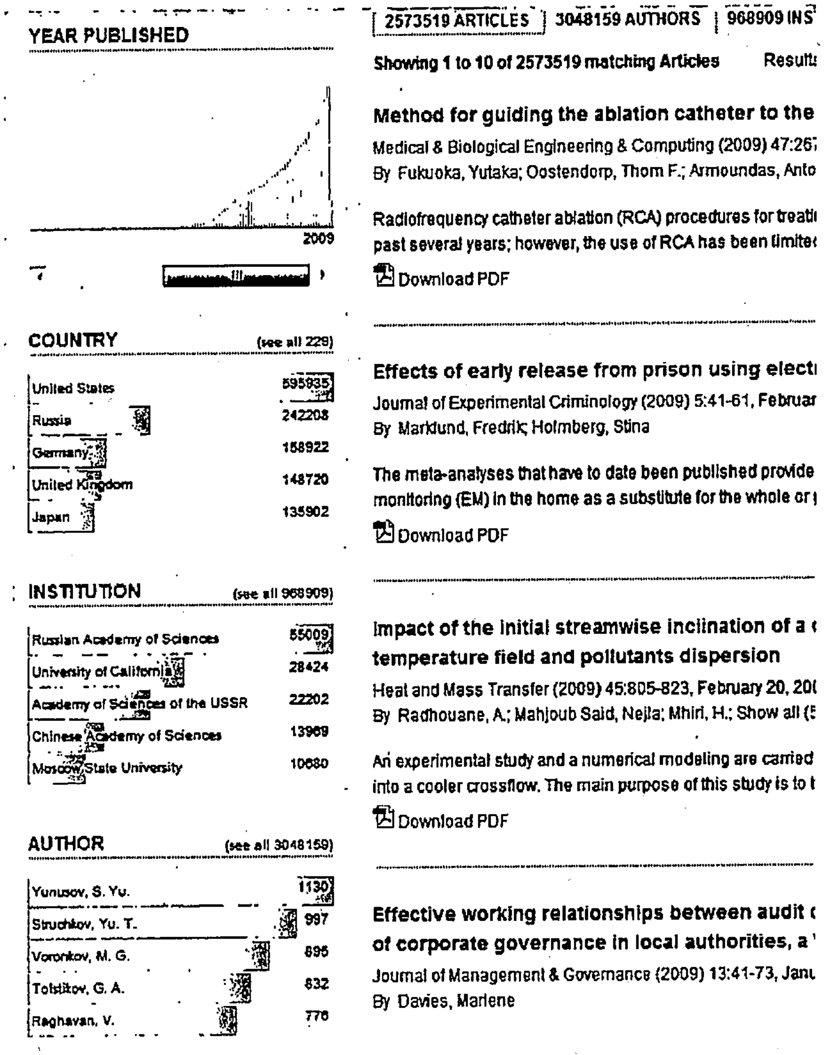
→ Epidemiology

Epistaxis is a common otolaryngologic problem, with 15 people per 10,000 requiring physician care annually and 1.6 per 10,000 requiring admission to the hospital. Most cases occur under age 10, and the incidence decreases with age. One survey of more than 6000 patients reports an 11% incidence of epistaxis. It is more common in colder seasons and in northern climates because of decreased humidity and subsequent drying of the nasal mucosa. Epistaxis is a frightening condition for patients but is seldom life threatening. A solid understanding of physiology and treatment allows for promot and efficient management of the disorder.

Posterior ethmoidal artery
Anterior ethmoidal artery
Nasopalatine (septal) branch of sphenopalatine artery
Kiesselbach's area
Septal branch of superior labial artery

Figure 2- Example of granular search result with full context preserved.

Data Analytics and Faceted Navigation– The MarkLogic XML Server includes a number of built in analytical features that greatly enhance the search and query capabilities. First, MarkLogic provides classification based on content, structure, or both. The classifier is unique in its class, and it provides the ability to automatically categorize documents. MarkLogic also supports lexicons for any word and value in the documents. These are high performance indexes of every word occurrence. MarkLogic combines these lexicons with fast, on-the-fly frequency counts of words, element values, co-occurring pairs of values, and value ranges. These unique capabilities enable faceted navigation for discovery of information, and provide insight into content and search results. Figure 3 below shows an example of a faceted navigation application that leverages Mark Logic fast analytics to return counts on each of the relevant facets.



COUNTRY (see all 229)

United States	595935
Russia	242208
Germany	158922
United Kingdom	148720
Japan	135902

INSTITUTION (see all 968909)

Russian Academy of Sciences	55009
University of California	28424
Academy of Sciences of the USSR	22202
Chinese Academy of Sciences	13969
Moscow State University	10680

AUTHOR (see all 3048159)

Yunusov, S. Yu.	1130
Struchkov, Yu. T.	997
Voronkov, M. G.	895
Tolstikov, G. A.	832
Raghavan, V.	776

2573519 ARTICLES | 3048159 AUTHORS | 968909 INS

Showing 1 to 10 of 2573519 matching Articles Result

Method for guiding the ablation catheter to the
Medical & Biological Engineering & Computing (2009) 47:26
By Fukuoka, Yutaka; Oostendorp, Thom F.; Armoundas, Anto

Radiofrequency catheter ablation (RCA) procedures for treati
past several years; however, the use of RCA has been limite

Download PDF

Effects of early release from prison using elect
Journal of Experimental Criminology (2009) 5:41-61, Februar
By Marklund, Fredrik; Holmberg, Stina

The meta-analyses that have to date been published provide
monitoring (EM) in the home as a substitute for the whole or

Download PDF

Impact of the initial streamwise inclination of a
temperature field and pollutants dispersion
Heat and Mass Transfer (2009) 45:805-823, February 20, 20
By Radhouane, A.; Mahjoub Said, Nejla; Mhiri, H.; Show all (

An experimental study and a numerical modeling are carried
into a cooler crossflow. The main purpose of this study is to

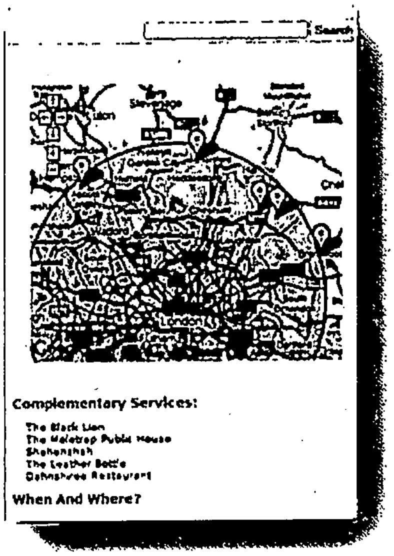
Download PDF

Effective working relationships between audit
of corporate governance in local authorities, a
Journal of Management & Governance (2009) 13:41-73, Jan
By Davies, Marlene

Figure 3 - An example of faceted navigation using Mark Logic

Content Processing Framework – MarkLogic Server enables companies to create powerful, custom ingestion processing pipelines (trigger-based sequences of content processing steps) comprised of native XQuery statements and web services-enabled external applications. This unique capability enables plug and play architecture, wherein the technology and tools can be layered on top of MarkLogic to enable an intelligent enterprise. Common uses of this flexible ingestion processing capability include transformation of content prior to ingestion, and integration with third party tools like Entity Extractors.

Geospatial - Increasingly, information needs to be delivered within a geospatial context. MarkLogic Server includes geospatial support, which provides fast search, retrieval and analysis of content marked up with geospatial data. By using the integrated full-text search and geospatial query, organizations can create high-performance location-based services, which fully leverage the value of their content, delivering it to users with greater context, based on their physical location. Figure 4 shows a location based directory application combining full text search and a point-radius geo query to find places of interest.



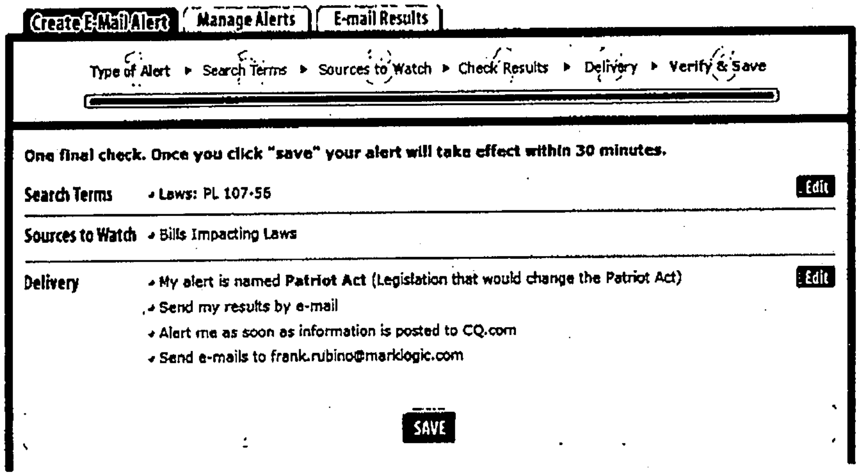
Complementary Services:

The Black Lion
The Moletrap Public House
Shahenshah
The Leather Bottle
Dalmshree Restaurant

When And Where?

Figure 4 – Example of Geo-enabled information discovery application

Large-scale alerting - The more information organizations gather, the harder it is for employees and customers to find what they are looking for. MarkLogic Server includes large-scale alerting (sometimes called triggers or profiles) functionality, which is designed to perform well across two dimensions: large numbers of alerts and extremely large amounts of content. Additionally, alerts can be defined using a wide range of factors including key word, structure, entity, geospatial information — all in any combination, which means users can immediately know about any new relevant information they seek has been added to the XML repository. Figure 5 below shows the alerting definition screen for Congressional Quarterly. This allows users to save a query of interest, and indicate how the alert will be delivered, in this case email.


Create E-Mail Alert
Manage Alerts
E-mail Results
Type of Alert ▸ Search Terms ▸ Sources to Watch ▸ Check Results ▸ Delivery ▸ Verify & Save

One final check. Once you click "save" your alert will take effect within 30 minutes.

Search Terms
- Laws: PL 107-56

Sources to Watch
- Bills Impacting Laws

Delivery
- My alert is named **Patriot Act** (Legislation that would change the Patriot Act)
- Send my results by e-mail
- Alert me as soon as information is posted to CQ.com
- Send e-mails to frank.rubino@marklogic.com

Figure 5 - Example interface for saving alert preferences to be delivered via email

High availability - MarkLogic Server is architected to support your most mission critical applications. MarkLogic natively supports and scales in a cluster, providing a basis for not only scalability, but also fault tolerance. The architecture delivers superior scalability while also providing failover, hot backup and other high-availability features. Database style journaling and transactional updates mean you can rely on MarkLogic Server to reliably store and deliver your high-value content to your users.

Figure 6 illustrates a typical Mark Logic cluster deployment. Servers in a cluster can be designated as data managers and/or query evaluators, and can be deployed in the same server or separate clustered servers

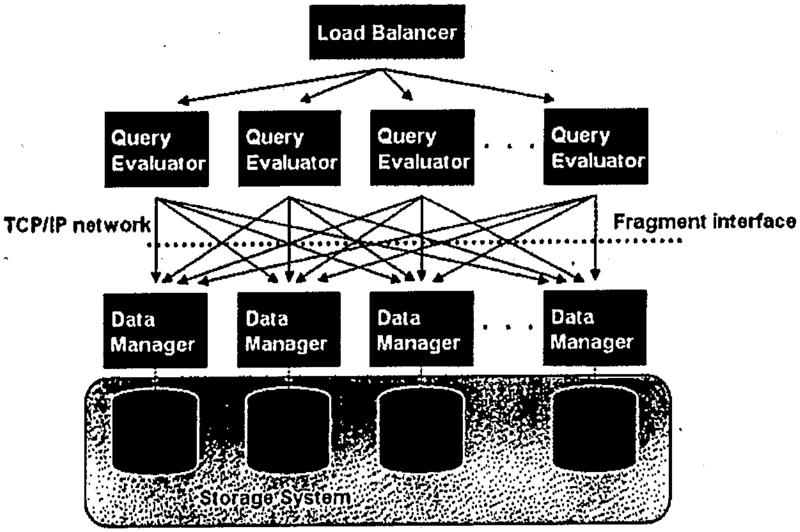


Figure 6 – Mark Logic Deployment Architecture

Auditing - MarkLogic Server makes it easy to monitor system activity by providing auditing functionality. Organizations can audit events such as document update, system shutdown, modifications of permissions, and user authentication to a log file. They can also filter the events they want to log — by user, by role, by outcome (success/failure), by event, and/or by document in order to speed analysis and understanding.

Open Standards –MarkLogic Server is based on open-standards that allows for easy integration with a wide range of products for content processing. In addition to providing native support for XML, XPath, and XQuery, it also offers Application Programming Interfaces (APIs) in Java/J2EE and .Net, and Web Services. This supports exposing Mark Logic functionality as a Service that can be used by a variety of applications.

Automatic content conversion – Mark Logic supports other document formats beyond XML. In addition to loading XML content "as is", MarkLogic Server automatically converts common document formats including Microsoft Office, Portable Document Format (PDF), and Hyper Text Markup Language (HTML) into well-formed XML. This eliminates the detailed analysis and costly effort required to "shred" or "chunk" documents into a relational database.

Past Performance

This section includes some representative Mark Logic customer references. Additional references are available upon request.

DNI Open Source Center:

The Open Source Center is dedicated exclusively to the exploitation and dissemination of all valuable unclassified open source information that it acquires on a daily basis by crawling the world's web sites and translating, converting and storing that information. The original project was expected to grow to 160 TB of content and 2 PetaBytes of related binary content, but it is likely to exceed 200 TB of content and this growth must have only minimal impact on performance. MarkLogic Server is central to the architecture in its role as the XML hub that stores and provides access to this vast store of information. Mark Logic provides the storage, search, rendering and integration to visualization products like geospatial, links, timeline, etc.

National Security Agency E-Space:

E-Space is an all-source information sharing application, ingesting 20+ different data types/schemas into a single Mark Logic repository. Prior to Mark Logic the E-Space project was experiencing unacceptably slow query performance. Analysts needed to query for, and subsequently manipulate, SIGINT data from 20+ disparate data sources. The system issued federated queries to the different data sources and received on average an 80MB – 200MB XML stream containing 100K records with up to 200 fields. These documents needed to be transformed and loaded into an Oracle relational database, something that took on average 12 minutes to complete for each query. To solve this problem Mark Logic was introduced as a high-performance XML transformation engine, and the time it took to load the query results into Oracle was reduced to six seconds for a +100X increase in speed. In the second phase of the project, Mark Logic replaced Oracle as the XML storage and search repository.

DNI NCTC Terrorist Database:

The NCTC's Terrorist Identities Datamart Environment (TIDE) is the U.S. Government's central data base on known or suspected international terrorists. The database contains all source highly classified information provided by members of the Intelligence Community (IC) such as CIA, DIA, FBI, NSA, and many others. An unclassified extract from TIDE is provided to the FBI's Terrorist Screening Center, which is used to compile various watch lists such as the TSA's No-Fly list, State Department's Visa and Passport Database, Homeland Security's Boarder System, and FBI's NCIC (National Crime and Information Center) for state and local law enforcement. MarkLogic Server is used to integrate the content and data that comes in many formats from these many sources, and then makes that information accessible to users and other applications through native search capabilities and through integration with other XML tools. MarkLogic is central to an architecture that marks a major step forward from the pre-9/11 status of multiple, disconnected, and incomplete watchlists throughout the government.

Army Battle Command Knowledge System:

BCKS is a lessons learned data sharing application using the DoD XML metadata standard called DDMS. The US Army is engaged in two unconventional wars in Afghanistan and Iraq fighting an enemy that does not use standard war fighting methods and is constantly adapting their guerilla-like methods. U.S. Soldiers needed to share and find information about the enemy's tactics in near real-time to keep up, and they could not rely on traditional Army doctrine to provide this because the development and review cycle lasted anywhere from six months to two years. The Battle Command Knowledge System was created to address these needs by allowing soldiers to share their expertise and experience through user-generated content, subject matter experts (SMEs) to classify the content, and soldiers to discover this information and have only the most relevant content delivered to them via web-based UIs and on various devices including PDAs.

DISA NCES:

The NCES Enterprise Catalog project for the DoD's NCES organization is used to explore how DDMS XML could be used to identify arbitrary collections of documents, and improve categorization of new documents. The customer wanted to improve ingestion methods for new documents, including automatic ingestion via atom and RSS feeds. The pilot delivered these capabilities, while also enabling this content to be available to users through semantic discovery in a customized web application and via web services through other applications such as Google Maps on Intelink.

US Army Training and Doctrine Command (TRADOC):

Mark Logic provides a custom publishing application for the Army which allows dynamic document creation of various document formats into a single new custom field manual. The U.S. Army produces hundreds of paper-based field manuals that contain training and doctrine. It is difficult for soldiers to find the information they need in all of these manuals, and when they do, they often tear the relevant pages out and create their own "custom manual". Mark Logic put together a system that allows the TRADOC doctrine writers to create content for specific subject areas, and to have these smaller pieces be saved in MarkLogic Server as doctrinal objects. Soldiers and Commanders can now conduct a targeted search, retrieve granular, relevant results as doctrine objects, and reassemble those objects dynamically to build custom manuals called "battlebooks" that contain approved doctrine.

Additional Federal customers include:

- Library of Congress
- FAA
- NARA
- DIA
- CIA
- NGA
- PTO
- Department of State
- Multiple Air Force Programs
- Multiple DoD Joint Agencies