UNITED STATES PATENT AND TRADEMARK OFFICE

OFFICE OF GENERAL LAW

January 26, 2010

VIA CERTIFIED MAIL
RETURN RECEIPT REQUESTED

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

RE: ***Freedom of Information Act (FOIA) Request No. F-10-00031 (Appeal)***

Dear Mr. Malamud,

This is in response to your Freedom of Information Act (FOIA) appeal received on December 24, 2009. After careful consideration, your appeal is granted in part and denied in part.

Your original FOIA request dated November 7, 2009 (FOIA Request) sought "any submissions received by your office in response to your [s]olicitation for USPTO's Data Dissemination Solution. [SS-PAPT-09-10008]." FOIA Request at 1. On December 1, 2009, the Agency's FOIA Officer responded (Agency's FOIA Response) and identified ninety-three (93) pages of documentation that were responsive to your request and released thirty-three (33) pages to you. Agency's FOIA Response at 1. Further, the FOIA Officer partially denied your request as to the remaining sixty (60) pages of documents that were determined not suitable for public disclosure based on the FOIA exemption found at 5 U.S.C. § 522(b)(4) (Exemption 4).

Your appeal challenged the Agency's decision to withhold 60 pages of documents that were responsive to your request. FOIA Appeal at 1. You further alleged that "...the United States Patent and Trademark Office may have simply done a blanket denial of all RFI responses and that [the Agency] instead should have redacted the portions 'clearly delineated' and released the rest." Id. Finally, you inferred that, at the very least, the Agency should have released a "list of respondents" to the RFI. Id.

Exemption 4 of the FOIA protects "trade secrets and commercial or financial information obtained from a person [that is] privileged or confidential." 5 U.S.C. § 522(b)(4). This exemption safeguards submitters of both commercial and financial information from competitive disadvantage that can occur due to the disclosure of this type of information. Exemption 4 contains three components that must be met in order to make a determination that certain information is exempt from a FOIA request. First, the information must be commercial or financial in nature. 5 U.S.C. § 522(b)(4). This has been interpreted to mean that the information is anything pertaining, relating or dealing with commerce. Am. Airlines, Inc. v. Nat'l Mediation Bd., 588 F.2d 863, 870 (2d Cir. 1978). The second component necessitates that the information come from a "person" which has been interpreted to include a wide range of entities. Nadler v. FDIC, 92 F.3d 93, 95 (2d Cir. 1996). The third and final requirement of Exemption 4 is met if the information is privileged or confidential. A determination of whether information is deemed privileged or confidential will depend on whether the submitter of the information is obliged to furnish the information to the Government or whether the information is voluntarily submitted. Information is deemed privileged or confidential in the former if: (1) it impairs the government's ability to collect information in the future; or (2) cause substantial harm to the competitive position of the person from whom the information was obtained. Nat'l Parks & Conservation Ass'n v. Morton, 498 F.2d 765, 770 (D.C. Cir. 1974). Information is deemed privileged or confidential in the latter if the information is not customarily made available to the public by the submitter. Critical Mass Energy Project v. NRC, 975 F.2d 871, 879 (D.C. Cir. 1992).

The information that you requested is subject to Exemption 4, because it satisfies the three-part test described above. The information is commercial due to the nature of the Request for Information (RFI) for the USPTO Data Dissemination project. The RFI sought information that could lead to potential opportunities for vendors to acquire patent and trademark data in bulk and to provide such data at no cost to the Government or the public. In exchange, vendors would be allowed to repackage (i.e., add value) and sell any resulting enhanced data sets and retain any fees collected. Therefore, the information pertains, relates and deals with commerce. The respondents to the RFI consist of different entities and thus fit within the definition of "person" as interpreted by the courts. Finally, the voluntarily submitted information not released is confidential. The Agency requested that each respondent classify the nature of the information it submitted in response to the Agency's RFI. The information that was withheld under Exemption 4 was identified as harmful to the competitive position of each respective respondent by each respective respondent and thus not customarily made available to the public by the respective respondent.

Further, the argument in your appeal that the Agency issued a blanket denial is inaccurate. In response to your FOIA request, the Agency released thirty-three (33) documents along with a description of the methodology used in determining that certain documents are exempt. This contradicts your claim that a blanket denial was issued.

However, on appeal, the Agency has identified two additional RFI responses which have been determined to be non-confidential. These documents are responsive, in part, to your request. We are releasing these documents to you in full, without redactions.

Also, your appeal raised, for the first time, a request for " ... a list of respondents." Appeal at 1. You originally requested "any submissions received by the [USPTO] in response to [USPTO] [s]olicitation for USPTO's Data Dissemination Solution. [SS-PAPT-09-10008]." This new request was not raised in your original FOIA request and thus is not an appropriate matter for appeal. See 37 C.F.R. 102.10. As such, this request is being referred back to the Agency's FOIA Officer for processing as a new FOIA request according to the Agency's FOIA regulations. 37 C.F.R. pt. 102. The FOIA Officer will update you with processing information.

Further, you also claim, due to a presumption of openness, all documentation should be released. Your reliance in your appeal on the President's Memorandum for the Heads of Executive Departments and Agencies, dated January 21, 2009, is misplaced, as that Memorandum expressly "does not create any right or benefit, substantive or procedural, enforceable at law or in equity by any party against the United States, its departments, agencies, or entities, its officers, employees, or agents, or any other person." As fully explained above, the Agency's initial decision is consistent with the provisions of FOIA, 5 U.S.C. § 552, and with various court decisions addressing FOIA claims.

Accordingly, the Agency denies your appeal in part and grants it in part. This is the final decision of the USPTO with respect to your appeal. You have the right to seek judicial review of this denial as provided in 5 U.S.C. § 552(a)(4)(B). Judicial review is available in the United States District Court for the district in which you reside or have a principal place of business, the United States District Court for the Eastern District of Virginia, or the United States District Court for the District of Columbia.

Sincerely,

William R. Covey

WILLIAM R. COVEY
Deputy General Counsel for General Law

Enclosures

VENABLE® LLP

575 SEVENTH STREET NW WASHINGTON, DC 20004
T 202.344.4000 F 202.344.8300 www.Venable.com

November 16, 2009

James R. Burdett
T (202) 344-4893
F 202.344.8300
jrburdett@venable.com

V. Anne Tugbang
Contracting Officer
United States Patent and Trademark Office
P. O. Box 1450 - Mail Stop 6
600 Dulany Street, MDE, 7th Floor
Alexandria, Virginia 22313-1450

Re: Request For Information (RFI) For USPTO's Data Dissemination Solution (last amended October 26, 2009) – SS-PAPT-09-10008

Dear Ms. Tugbang:

I write to you on behalf of the Coalition for Patent and Trademark Information Dissemination ("the Coalition"), whose members include CT Corsearch, Dialog LLC, IFI Patent Intelligence, Reed Elsevier Inc., and Thomson Reuters, regarding the "Request For Information (RFI) For USPTO's Data Dissemination Solution" (last amended October 26, 2009) – SS-PAPT-09-10008. The Coalition's members are very interested in this project, and provide herein below the Coalition's written response which:

- raises concerns about the potential anti-competitive effect of the USPTO proposal
- discusses the feasibility of this no-cost approach; and
- includes other solutions for USPTO consideration such as a consortium approach to data dissemination.

Members of the Coalition appreciate that the USPTO embraces the President's initiative to provide increased transparency of government operations and information, and seeks to make bulk data available to the public at no charge. For example, bulk data is currently available at no charge on a weekly basis from the DATA.GOV website in the case of patent application bibliographic data and patent grant bibliographic data. Members of the Coalition also appreciate that recent efforts have been focused on making virtually all public information from the USPTO accessible on the Internet, but significant challenges remain. Among the challenges which most concern members of the Coalition, however, is the proposal's potential impact on competition.

The proposed model would discriminate against all other resellers and would likely conflict with the Administration's transparency goals and objectives as it would grant an exclusive competitive advantage to a vendor(s) of patent and trademark information which would be harmful to an open competitive market in such information at competitive prices and would undermine private sector incentives to invest in value-added information products.

The USPTO's Approach Will Have A Potentially Anti-Competitive Effect

While the USPTO's FISCAL YEAR 2010 PRESIDENT'S BUDGET stands strangely silent on the matter, its FY2009 counterpart[1] and the 2008 Performance and Accountability Report[2] both proclaim that "[t]he USPTO is committed to achieving performance enhancements and cost-savings through competitive sourcing."

In support of the Department of Commerce goal to "[p]romote U.S. innovation and industrial competitiveness by protecting intellectual property...," the USPTO identified in its strategic plan objectives, initiatives, and performance measures that will enhance the degree of excellence or quality in every aspect of its patent and trademark processes—from the information the USPTO receives from applicants to the support it provides its own employees.[3] That plan is built upon four guiding principles: quality, timeliness, cost-effectiveness, and transparency.

- Quality in this regard means accurate and consistent results in examination. It presumes improved inputs, better-focused examination, improved review processes, and consistent examination results.
- Timeliness means processing, including review of applications, is completed without delay. It presumes that applicable laws, regulations, and policies work synergistically to eliminate frictions or uncertainties that lead to delays.
- Cost-Effectiveness to the USPTO implies efficiency, accountability, and a focus on results. It requires leadership and commitment to ensure that activities and processes result in value. This means using resources in the most effective manner to deliver quality, timely USPTO services. It also means being responsible stewards of the public trust.
- Transparency demands impartiality, fairness, accessibility, availability, and a public-service mentality. This includes the USPTO's continuing commitment to opening the agency to the public by providing electronic filing, patent and trademark file contents, and as much training and research material as possible via its Web site. It requires

[1] UNITED STATES PATENT AND TRADEMARK OFFICE FISCAL YEAR 2009 PRESIDENT'S BUDGET, p. 55.
[2] UNITED STATES PATENT AND TRADEMARK OFFICE PERFORMANCE AND ACCOUNTABILITY REPORT: FISCAL YEAR 2008, p. 38.
[3] UNITED STATES PATENT AND TRADEMARK OFFICE 2007–2012 STRATEGIC PLAN, pp. 12-13.

discipline on the part of the public to use public means to influence change in USPTO practices and activities.[4]

According to the strategic plan, quality begins with the fundamentals—a high-performing workforce that is properly trained and given the tools and information technology (IT) systems needed to accomplish the job.[5]

The USPTO has, indeed, developed an ambitious agenda to help position the agency to both meet the challenges it now faces, and to operate more successfully and efficiently in the 21st century. One important component of that agenda is "leveraging information technology." In that regard, the Office of the Chief Information Officer has envisioned the critical role of IT in his Strategic Information Technology Plan.[6] Equally important to that agenda is compliance with legislative and executive mandates which have an impact on those plans, including the Paperwork Reduction Act, guidance from the Office of Management and Budget and General Accountability Office, as well as other mandates that require USPTO to achieve operational efficiency and cost-effectiveness through a reliable information technology capital planning and investment control management process and a robust information technology enterprise architecture.

Paperwork Reduction Act

On May 22, 1995, the Paperwork Reduction Act[7] went into effect with important purposes, including the following:

- ***ensure the greatest possible public benefit from and maximize the utility of information created, collected, maintained, used, shared and disseminated by or for the Federal Government;***
- improve the quality and use of Federal information to strengthen decisionmaking, accountability, and openness in Government and society;
- minimize the cost to the Federal Government of the creation, collection, maintenance, use, dissemination, and disposition of information;

[4] *Id.*
[5] *Id.* at p. 12.
[6] For a practical update on that plan, *see, e.g.*, OCIO ROAD MAP TO IT MODERNIZATION: HOW THE OCIO WILL UPGRADE IT FOR BETTER SERVICE, Road Map Status Report – Advisory Committees (June 2009).
[7] Public Law 104-13, 44 U.S.C. § 3501 *et seq.*

- strengthen the partnership between the Federal Government and State, local, and tribal governments by minimizing the burden and maximizing the utility of information created, collected, maintained, used, disseminated, and retained by or for the Federal Government;
- ***provide for the dissemination of public information on a timely basis, on equitable terms, and in a manner that promotes the utility of the information to the public and makes effective use of information technology***; and
- ensure that the creation, collection, maintenance, use, dissemination, and disposition of information by or for the Federal Government is consistent with applicable laws.[8] [emphasis added]

Of particular relevance to the RFI are the following PRA references:

> (d) With respect to information dissemination, each agency shall--
>
> (1) ***ensure that the public has timely and equitable access to the agency's public information***, including ensuring such access through--
>
> (A) ***encouraging a diversity of public and private sources for information based on government public information***;
>
> (B) ***in cases in which the agency provides public information maintained in electronic format, providing timely and equitable access to the underlying data (in whole or in part);*** and
>
> (C) agency dissemination of public information in an efficient, effective, and economical manner;
>
> * * * * *
>
> (4) ***not, except where specifically authorized by statute--***
>
> (A) ***establish an exclusive, restricted, or other distribution arrangement that interferes with timely and equitable availability of public information to the public***;

[8] 44 U.S.C. § 3501.

(B) ***restrict or regulate the use, resale, or redissemination of public information by the public***;[9] [emphasis added]

In contravention to these very specific responsibilities, however, the USPTO's approach to the RFI appears not only to "... establish an exclusive, restricted, or other distribution arrangement that interferes with timely and equitable availability of public information to the public ...", but also to "... restrict or regulate the use, resale, or redissemination of public information by the public ...".

For example, in answer to Question #59 from the September 24, 2009 public meeting in Alexandria, which stated:

> 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 USPTO merely answered:

> The objective is to increase access through better distribution mechanisms.
>
> Appropriate controls will be evaluated after consideration of the information collected through this RFI.[10]

Likewise, in answer to Question #2 from the October 19, 2009 public meeting in San Francisco, which stated:

---

[9] 44 U.S.C. § 3506, Federal Agency Responsibilities.

[10] Data Dissemination QA for Amendment 1.doc (available at https://www.fbo.gov/index?s=opportunity&mode=form&id=5e22bea31baa3f05ca52f41a7fda599f&tab=core&_cview=1.

> For weekly front files, does the USPTO plan to provide the partner organization with raw data in advance of publication date in order to prepare data on time? If so:
>
> a. What is the USPTO expected time lag between data availability to the partner and data availability to the public?
>
> b. What are the terms of use for the data on the partners internal server before publication date? For example, would the partner have the capability to start working with the raw data as soon as it has been transformed? Is there any restriction on the publishing of value added PTO data?

the USPTO merely answered:

> Service level agreements for any subsequent procurement actions have not yet been finalized. Currently, weekly front files are provided to the publishing organization within USPTO a week in advance of publication date and are made available to the public by 2 a.m. on the date of publication. The intent of any subsequent procurement action is to meet or exceed the current data distribution timelines. The partner may start working with the data as soon as they receive it but may not distribute value added data prior to distribution of the non-value added data.[11]

These answers completely avoid the details of the "appropriate controls" and run the risk of establishing "... an exclusive, restricted, or other distribution arrangement that interferes with timely and equitable availability of public information to the public ...", and restricting or regulating "... the use, resale, or redissemination of public information by the public ..." in direct violation of the Paperwork Reduction Act.[12]

---

[11] Amendment_3__Data_Dissemination_QA_for_SF_Public_Meeting.doc (available at https://www.fbo.gov/index?s=opportunity&mode=form&id=5e22bea31baa3f05ca52f41a7fda599f&tab=core&_cview=1.

[12] *See also* "Synopsis for Public Data Dissemination Sole Source Contract to Google, Inc." (Solicitation Number Special Notice 10-11001) posted November 6, 2009, which notices the USPTO's intent "to award a no cost, sole source contract ... for public data dissemination services".

OMB Circulars

Among the priorities set forth by OMB Circular No. A-11,[13] one in particular is directed to the question of how an agency's budget should address workforce planning and restructuring. The answer requires agency heads to identify the training, development, leadership development, and staffing actions proposed in each budget to "[p]repare for and respond to changes driven by e-Government and competitive sourcing."[14] Nothing could be more important than this principle to the RFI.

Moreover, OMB Circular No. A-11 requires that investments in major capital assets proposed for funding in the Administration's budget should, among other things:

> Employ an acquisition strategy that ***appropriately allocates risk between the Government and the contractor, effectively uses competition***, ties contract payments to accomplishments, and takes maximum advantage of commercial technology.[15] [emphasis added]

While the "Replacement and Modernization of Information Technology Infrastructure" is a key item within the USPTO's FISCAL YEAR 2010 PRESIDENT'S BUDGET,[16] there is nothing within that document which "... appropriately allocates risk between the Government and the contractor ..." and "... effectively uses competition ...", particularly with respect to dissemination to the public of patent and trademark data.

OMB Circular No. A-76 (Revised) establishes federal policy for the competition of commercial activities. The longstanding policy of the federal government has been to rely on the private sector for needed commercial services. To ensure that the American people receive maximum value for their tax dollars, OMB Circular No. A-76 stands for the proposition that commercial activities should be subject to the forces of competition.[17]

---

[13] Circular No. A–11, PREPARATION, SUBMISSION, AND EXECUTION OF THE BUDGET, Executive Office of the President, Office of Management and Budget (August 2009).
[14] *Id.* at page 1 of Section 85.
[15] *Id.* at page 2 of Appendix J.
[16] UNITED STATES PATENT AND TRADEMARK OFFICE FISCAL YEAR 2010 PRESIDENT'S BUDGET, pp. 7-8.
[17] For a more detailed treatment of the requirements of OMB Circular No. A-76 in this regard, *see* Joseph L. Ebersole, *The 21st Century Strategic Plan: USPTO's Adjustment To The Future,* World Patent Information 25 (2003), pp. 289-302.

Finally, OMB Circular No. A-130 is directed to management of Federal information resources. It requires that agencies must, in the acquisition of information technology, "[m]ake use of adequate competition, allocate risk between government and contractor, and maximize return on investment when acquiring information technology.

The RFI and the USPTO's proposals with respect to the implementation of a system for dissemination of patent and trademark data seemingly ignore these mandates. For example, if a single vendor is selected under the hypothetical situations discussed in Questions #59 and #2 above, the potential exists for that vendor to have an unfair competitive advantage over other resellers of patent and trademark information. Not only would the selected vendor have access to the USPTO data well before such other resellers of patent and trademark information, but it would also be able to begin adding value to that data at that earlier date. Moreover, in order to work effectively with and on behalf of the USPTO, the selected vendor would need to know well in advance of the release date of new data, what new data formats might be required to disseminate such data. Other resellers of patent and trademark information would have much less time to wrestle with such new data formats when processing their own versions of value-added data, putting them at a significant competitive disadvantage to the vendor(s)..[18]

Principles to Avoid This Potentially Anti-Competitive Effect

Maximizing effective functioning of global patent systems requires vital and vigorous dissemination of value-added patent and trademark information. The most effective overall system can be achieved by a partnership between the USPTO and private sector organizations that recognizes and encourages the core strengths and responsibilities of each. Specifically, the USPTO should encourage a robust, competitive private sector information industry providing value-added patent and trademark information services while, at the same time, meeting its own requirements of providing information to the general public. Members of the Coalition believe that the principles outlined below not only provide criteria for making decisions about public versus private roles in patent and trademark information dissemination, such as under the current RFI, but also can be a source of guidance for the USPTO in policy formulation.

[18] This situation is more than a mere hypothetical. For example, when the USPTO changed the structure of its data from "b" XML to "c" XML, many large, sophisticated resellers of patent and trademark information took three to four months to modify their data sets.

USPTO Policy Should Encourage A Diversity Of Sources For Patent And Trademark Information

It seems common sense that one should not rely on a single source of information. Indeed, with respect to the patent data currently disseminated on its public web site, the USPTO acknowledges this fact with the following disclaimer:

> The fact that an invention cannot be found by searching in the Patent Full-Text Database does not mean that the invention is patentable. The USPTO's text-searchable patent database begins with patents granted since 1976. A complete patentability search must consider all prior art, including earlier patents, foreign patents and non-patent literature.

In the United States, constitutional law related to the First Amendment has emphasized that a well-informed and educated public is best arrived at in the marketplace of ideas from a multiplicity of sources. First Amendment principles underlie the encouragement of publication by private sector companies of value-added products and services based on government information and the Paperwork Reduction Act referenced above provides that federal government agencies shall insure public access to an agency's public information by "encouraging a diversity of public and private sources for information based on government public information".[19] Thus, it is built into the law. How must an agency provide information to the public? According to the statute's enforcement vehicle, the answer to that question requires agencies to:

> Take advantage of all dissemination channels, Federal and nonfederal, including State and local governments, libraries, ***and private sector entities***, in discharging agency information dissemination responsibilities[20] [emphasis added].

The concept of "a diversity of sources" has special applicability to patent and trademark information. Each area of technology — each art area — benefits from different types of search tools to achieve optimal results. There are as many types of uses of patent and trademark information as there are types of users — researchers, business intelligence analysts, marketing specialists, financial analysts, technology specialists. In short, many uses in addition to searches for registrability, patentability, infringement, validity, *etc.* And even these professional uses

---

[19] 44 U.S.C. § 3506(d)(1)(A).

[20] *See, e.g.*, OMB Circular No. A-130, *supra*, at Section 8.a.5.(d)(iii). The most recent version is available at <http://www.whitehouse.gov/omb/circulars_a130_a130trans4/> [last accessed September 18, 2009].

benefit from multiple sources each of which has a different type of value added and a resulting different perspective on the data. If there is one favored source — the USPTO vendor with advantages over all others — this diversity will be put in jeopardy. And yet, this is what will likely happen if the USPTO does not proactively take into account this principle when it makes decisions about providing access to patent and trademark data.

Perhaps the greatest advantage of a diversity of sources is that it maximizes dissemination and enables patent and trademark information to reach places where it would otherwise not be used; thus helping to realize one of the major policy goals of both the USPTO and the Obama Administration. The selection of one vendor to handle public access could affect the ability and desire for other companies to invest in the further advancement of market-specific tools, and will lower choice and functionality for all.

The model considered by the USPTO that one or more entities would invest in the data dissemination project in return for some undefined special benefit would seem to face some potentially unworkable issues. Either the benefits to the vendor would be insufficient to warrant the significant investments or they would put the vendor at an unfair competitive advantage over other resellers of patent and trademark information.

USPTO Policies Should Create An Environment For Maximizing Competition Among Private Sector Patent And Trademark Information Providers

Maximizing competition requires creating — not destroying — incentives for investment. Incentives are greatest when it is clear that the USPTO or its selected vendor(s) with an inherent advantage as the data source, are not going to directly compete with the features or functions that will be created by an investment under consideration by other private sector firms. Rational investors will be highly reluctant to risk capital where there is not certainty that fair and open competition will prevail. This means open competition among private sector companies in a marketplace, and fair competition with a patent office or its vendor(s). Fairness implies that the USPTO or its vendor(s) is not adding value it is providing free or with a monopoly advantage. If the USPTO takes steps to directly compete with private sector companies, by itself or through the assistance of its selected vendor(s), a market distortion is created and this can lead to destruction of the marketplace itself.

USPTO Policies Should Be Informed By Competition Law and Antitrust Law Principles Of Fair Competition

The impact USPTO policies and services of its vendor(s) can have on the patent and trademark information marketplace is well illustrated by doctrines from antitrust and competition law. Antitrust law in the United States and competition law in the European Community have similar concepts. Two of those concepts — "essential facility" and "predatory pricing" — are especially relevant. Where an entity is the single source of a given product or a given type of information, antitrust law in the United States refers to this as an "essential facility," and competition law includes the same concept. Although a public entity, the USPTO (or its vendor(s)), in effect, are essential facilities. As such, they should reflect on the obligations they would be under if they were a private entity with such power, be aware of the reasons why the concept emerged in antitrust law, and use this to inform their own decision making. Competition and antitrust law also address situations, referred to as predatory pricing, where market power is used to under-price products or services. The law attempts to correct the abuses that can result from this power of predation enabled by a monopoly position. These practices have been found to be illegal because of the destructive effects on a competitive marketplace. The effect on the patent and trademark information marketplace can be destructive if it involves under-priced or free services by the USPTO or its selected vendor(s). Thus, we strongly urge that the known impacts of such activities be one of the considerations applied in the USPTO's decision making surrounding the proposals in the RFI.

Solutions For USPTO Consideration

USPTO Should Ensure Maximum Reasonable Competition

Under the initiative announced in the synopsis posted November 6, 2009,[21] the USPTO intends to award a no cost, sole source contract to a single vendor, purportedly using the authority set forth in the Patent and Trademark Office Acquisition Guidelines (PTAG).[22] According to the synopsis:

> The PTAG authorized the UPSTO to acquire products and services using "maximum reasonable competition" instead of "full and open competition" when possible on all acquisitions. However, the USPTO will use competition as a principal tool in achieving results and intends to adopt means of affording

[21] *See, e.g.*, footnote 12, *supra.*
[22] 68 Federal Register 25, pp. 6120-6122 (February 6, 2003).

> competition that it determines will effectually serve the performance goals established for particular acquisitions. Under PTAG the USPTO is exempt from the requirement to meet the test of "full and open competition" as defined in FAR Part 6, and may also utilize a streamlined contract approach to solicit and award contracts to the most highly qualified contractor. In addition, the PTAG provides greater flexibility in the identification of sources and exempts the USPTO from portions of FAR part 15.[23]

However, an exemption from "full and open competition" is not an excuse for the USPTO to ignore its PTAG requirement for the USPTO to "endeavor to acquire products and services to the maximum extent possible in all acquisitions on a competitive basis."[24] The PTAG emphasizes that:

> It is the policy of the USPTO to promote competition to the maximum extent possible. Competition reduces the risk of having to rely on only one source for critical goods or services and reduces costs. USPTO intends to balance these considerations with the program benefits that can be gained from developing a reduced supplier base and building strategic alliances with its suppliers. The degree of competition sought will be influenced by knowledge of the marketplace and successful past performance records, with competition in most cases limited to a reasonable number of capable sources.
>
> Under the USPTO process, all firms will be apprised of opportunities, but only those judged to be the most viable will commit the resources to fully participate. USPTO intends to have an open interchange with industry about USPTO potential requirements and contractor capabilities long before any formal solicitation is issued. It is the policy of the USPTO to inform all firms of opportunities and seek to ensure only the most viable will need to commit resources to fully participate.[25]

The synopsis concludes by noting that:

---

[23] *See* "Synopsis for Public Data Dissemination Sole Source Contract to Google, Inc." (Solicitation Number Special Notice 10-11001) posted November 6, 2009, However, the PTAG further requires the USPTO to "endeavor to acquire products and services ***to the maximum extent possible in all acquisitions on a competitive basis***." [emphasis added]

[24] *Id.*

[25] 68 Federal Register 25 at pp. 6120-6121.

> The USPTO will continue to develop an acquisition strategy under the previously posted RFI (SS-PAPT-09-10008) using competition to the maximum extent possible for establishing a long-term no cost solution to public data dissemination and transparency of governmental data.

Therefore, the PTAG should not be used to justify sole source awards for patent and trademark dissemination services to the public.

A More Competitive, Transparent Approach Is Possible

Members of the Coalition believe a consortium of interested parties—including the vendor(s) selected to ultimately disseminate the bulk data—could be used in conjunction with the USPTO to jointly establish the data format and delivery schedules to be used for bulk data dissemination under the RFI. This would serve to ameliorate concerns of potential anti-competitive behavior by establishing the data format and any subsequent changes to that data format well in advance of its dissemination to the public. Other resellers of patent and trademark information who participate as part of the consortium of interested parties would not be disadvantaged, because they would be as knowledgeable of the data format as the selected vendor(s).

In closing, the coalition is very concerned that the USPTO proposal, if brought to fruition, will be inherently anti-competitive as it will grant the vendor(s) advantages in exchange for a no-fee contract. Such government-granted inequitable advantages in timing, branding and inside technical information, are clearly in violation of existing statutes, and would result in unfair competitive advantage over other resellers of patent and trademark information. This would result in an erosion of an open competitive market that is the driving force for private sector innovation and broad high-quality access to patent and trademark information.

We urge you to consider the Coalition's concerns and proposals.

Very truly yours,

James R. Burdett

cc: John B. Owens II
JRB/lrh
DC2DOCS1 #1074161 v1

A Data Dissemination Solution for the U.S. Patent and Trademark Office (USPTO)

November 16, 2009

Submitted in response to USPTO's RFI SS-PAPT-09-10008

Submitted to the U.S. Patent and Trademark Office
c/o V'Anne Tugbang, Contracting Officer
publicdatadissemination@uspto.gov

Company Name(s): Access Innovations, Inc. (proposed prime contractor); Perfect Search Corporation (proposed subcontractor)

Primary Point of Contact: Marjorie Hlava, President, Access Innovations, Inc.
Phone Number: (800) 998-0800
E-mail Address: mhlava@accessinn.com

Alternate Point of Contact: Jay Ven Eman
Phone Number: (800) 998-0800
E-mail Address: j_ven_eman@accessinn.com

- **Cage Code of Proposed Primary Contractor: 0FRF2**
- **NAICS Code of Proposed Primary Contractor: 511210**
- **Business Size of Proposed Primary Contractor: Woman-owned, Small Business**
- **FSS Contract Number: GS-35F-0820N**

Table of Contents

INTRODUCTION 3

TECHNICAL CAPABILITY 6

- Perfect Search Corporation Solutions 6
- Access Innovations, Inc. 6
- The Enhanced Search Interface in Search Harmony 9

CORPORATE EXPERIENCE / PAST PERFORMANCE 11

- Corporate Experience – Access Innovations, Inc. 12
- Past Performance – Access Innovations, Inc. 13
- Corporate Experience – Perfect Search Corporation 15
- Past Performance – Perfect Search Corporation 16
- References 18

INTRODUCTION

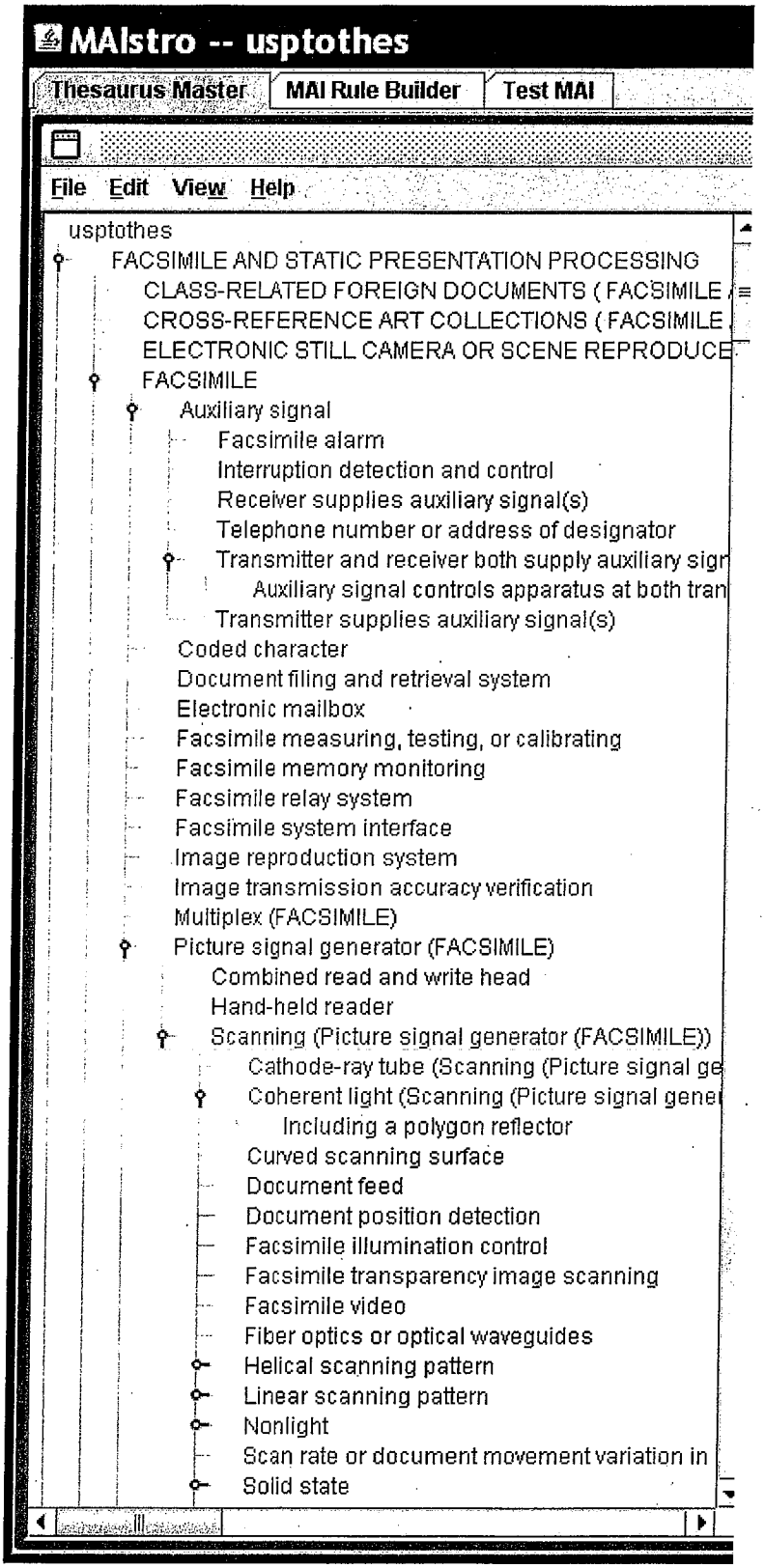
The technical solution that we offer is based on Data Harmony's Search Harmony, which combines the search capabilities of Perfect Search Corporation's products in concert with the user interface and taxonomic integration made possible with Access Innovations' Data Harmony suite of software solutions.

Search Harmony is taxonomy enhanced search that includes hierarchical browse navigation, automatic query completion using taxonomy terms and their synonyms, contextual presentation of related terms to broaden the search and narrower terms to refine the search, spell check, "more like these" sets, and recursive searches. Based on Access Innovations' MAIstro automatic indexing and thesaurus software suite (part of Access Innovations Data Harmony product line), it provides well formed data and highly accurate search results. Used on top of Perfect Search's search software, it provides the best of breed in recall, precision, and relevance searching.

Perfect Search Corporation software provides the capability to quickly index and query terabytes of data with sub-second search response. This innovative search technology is exceptionally compact, fast, and accurate, providing over 50% reduction in staff search time, using fewer servers, and resulting in low indexing latency. The search technology of Perfect Search is based upon ground-breaking technological innovation that provides the ability to quickly index and query very large data sets. Extremely scalable, Perfect Search already hosts databases of more than 1.4 billion records with sub-second search response time on a single Windows based commodity server. The search can be faceted, full text, or a combination.

The solution set would also include integration with a taxonomy or thesaurus. The U.S. Patent Classification System could be used as a basis for either the main thesaurus structure or as a supplemental classification system that would augment a more keyword-oriented thesaurus . These two combined, one as a classification and the other for conceptual indexing, would enable deep access to the patent literature. See the next page for a taxonomic view of the U.S. Patent Classification System.

We propose that Perfect Search Corporation host the USPTO databases on a dedicated server at Perfect Search Corporation headquarters. While we understand that the contract will be on a no-cost basis, we are also aware of the enormous amount of data and traffic involved; we are therefore hoping that USPTO is planning to cover the actual operating costs of hosting the data and search platform for access by their community of examiners. We also propose an implementation that would enable combined search across both patents and grants; we understand that searches on patents and grants are now necessarily separate.


MAIstro -- usptothes
Thesaurus Master
MAI Rule Builder
Test MAI
File Edit View Help
usptothes
FACSIMILE AND STATIC PRESENTATION PROCESSING
CLASS-RELATED FOREIGN DOCUMENTS (FACSIMILE
CROSS-REFERENCE ART COLLECTIONS (FACSIMILE
ELECTRONIC STILL CAMERA OR SCENE REPRODUCE
FACSIMILE
Auxiliary signal
Facsimile alarm
Interruption detection and control
Receiver supplies auxiliary signal(s)
Telephone number or address of designator
Transmitter and receiver both supply auxiliary sign
Auxiliary signal controls apparatus at both tran
Transmitter supplies auxiliary signal(s)
Coded character
Document filing and retrieval system
Electronic mailbox
Facsimile measuring, testing, or calibrating
Facsimile memory monitoring
Facsimile relay system
Facsimile system interface
Image reproduction system
Image transmission accuracy verification
Multiplex (FACSIMILE)
Picture signal generator (FACSIMILE)
Combined read and write head
Hand-held reader
Scanning (Picture signal generator (FACSIMILE))
Cathode-ray tube (Scanning (Picture signal ge
Coherent light (Scanning (Picture signal gene
Including a polygon reflector
Curved scanning surface
Document feed
Document position detection
Facsimile illumination control
Facsimile transparency image scanning
Facsimile video
Fiber optics or optical waveguides
Helical scanning pattern
Linear scanning pattern
Nonlight
Scan rate or document movement variation in
Solid state

We believe that this solution set can fully meet the USPTO's search and data accessibility needs. The access to both the grants and applications will be very fast, allow additional search features not currently available and help speed and refine the patent process. They files may be search together or separately as needed. Access by all XML elements is allowed.

Access Innovations, Inc. and Perfect Search software both have patents and patents pending. Access Innovations' patents include one for a system and method for database design and maintenance (the basis of the Data Harmony line of software); in addition, there is a patent pending for a method for identification of chemical names in text-containing documents.

TECHNICAL CAPABILITY

Perfect Search Corporation Solutions

Perfect Search provides a core search engine that handles massive datasets with relative ease, has extremely good precision and recall, and has scalability that is currently unmatched in the marketplace.

Perfect Search Corporation software provides the capability to quickly index and query terabytes of data with sub-second search response. This innovative search technology is exceptionally compact, fast, and accurate, providing over 50% reduction in staff search time, using fewer servers, and resulting in low indexing latency. The search technology of Perfect Search is based upon ground-breaking technological innovation that provides the ability to quickly index and query very large data sets. Extremely scalable, Perfect Search already hosts databases of more than 1.4 billion records with sub-second search response time on a single server. The search can be faceted, full text, or a combination.

Perfect Search technology differentiates itself from other products in the following areas.

- Scope – we can search across all fields and types of data, structured or unstructured so responses include data from all sources.
- Federation – the system can search data from multiple databases, multiple file system, multiple document repositories with a single search.
- Scalability – the search server is truly game changing in the amount of data that can be indexed and searched on a single box.
- Incremental Indexing – massive data sets can have data added easily through incremental indexing with no impact on precision of search or fragmentation of the index.

Perfect Search is a key component of the Access Innovations product Search Harmony, which takes full advantage of the thesaurus and tagging that Access Innovations provides.

Access Innovations, Inc.

MAIstro software

MAIstro is part of Access Innovations' Data Harmony software product line. The Data Harmony suite is platform-independent software written in Java to increase the effectiveness of document management personnel and of a collection's search engine.

Documented APIs (application programming interfaces) and similar connectors enable integration of Data Harmony applications with virtually any content management system. We have established a comprehensive library of connectors. These connectors extend the flexibility and capabilities of Data Harmony products.

The MAIstro user interface uniquely combines the functionality of Access Innovations' M.A.I. (Machine Aided Indexer) and Thesaurus Master to simplify the creation and management of controlled vocabularies and their use for categorization.

M.A.I. is a natural language processing tool that uses a rule-based system for interactive assignment of indexing terms to individual documents. For indexing that must be 100% accurate, editors are assisted in selecting indexing terms but retain control of the final selections. For large volumes of data, M.A.I. has provided auto-categorization at up to 92% accuracy.

The Thesaurus Master module of MAIstro is used for the creation and maintenance of a taxonomy or development of a full standards-compliant thesaurus.

Since the two modules are integrated, a rule is created automatically when a thesaurus term is added, and changes to terms in the thesaurus are immediately reflected in associated M.A.I. rules. The editor can immediately see the effect of rule changes in the Test MAI feature.

The Thesaurus Master Component of MAIstro

The language of an enterprise reflects its scope and content, calling for a specialized taxonomy to ensure efficient, effective management of that content.

Whether an organization develops this vocabulary (term list, authority file, thesaurus) in-house or obtains it whole or in part from external sources, Thesaurus Master thesaurus management software provides a means for controlling the entries, the hierarchy, and the conditions of use.

Thesaurus Master lets the user:

- Create and/or import the terms
- Fine-tune and refresh definitions easily as terms evolve
- Maintain hierarchical consistency while adding terms at any level
- Link directly to automatic indexing and documents

Advantages include the following:

- Easy import and export of taxonomy file
- Flexible integration with external software systems or display on website
- Increased consistency in indexing
- Increased productivity for human editors
- Increased relevancy of retrieval and higher recall of relevant documents

- Reduced misses and noise
- Greater satisfaction for database users

Thesaurus Master also:

- Operates independently in managing structured taxonomies
- Accommodates formal thesaurus structure of Broader Term, Narrower Term, Use and Used For references, History, Related Terms, and Scope Notes
- Enables restricted access to maintenance controls for database integrity
- Can be used in conjunction with a organization's existing systems
- Adheres to thesaurus standards ISO 2788 (monolingual), ISO 5964 (multilingual), and NISO Z39.19

The M.A.I. (Machine Aided Indexer) Component of MAIstro

Remarks from clients:

"We have found M.A.I. to be accurate and efficient, meeting and exceeding our expectations. The rule builder is flexible, powerful, and intuitively clear ... its term suggestions are right where we want them to be. We are delighted with both the package itself and the technical support and training we have received. ... We have achieved a 6.7 percent increase in productivity with the MAI system."

Scott Ryan, Development Specialist
Cambridge Scientific Abstracts

"Our staffing was going down, the work load was increasing, and we needed more efficient ways to store and move data without major staff intervention. I estimate that M.A.I. has improved our productivity by 50 percent."

Kurt Keeley, Database Manager
American Water Works Association (AWWA)

M.A.I. makes it possible for human categorizers ("indexers") to increase their indexing efficiency and consistency while adding superior descriptive data. M.A.I. promotes selection of precise and appropriate indexing / categorization terms from controlled vocabularies, authority files, or full thesauri. It presents a list of approved terms to the editor for selection, which saves time spent looking up terms manually. At the same time, M.A.I. gives total editorial control to the indexer / categorizer.

M.A.I. allows flexibility in categorization ("indexing") of documents because the editor can add or reject terms as needed. All editorial actions are gathered by the Statistics Collector, which then submits hit, miss and noise lists to the Rule Builder module for continued improvement of the rule base.

Customers have experienced up to a seven-fold increase in productivity using M.A.I. while measurably improving consistency and coverage of individual records. M.A.I. improves consistency by providing the same term in the same conditions every time, preventing editorial drift.

Indexing using M.A.I. mines the entire depth of the vocabulary applied, improving document retrieval, relevance and precision for the end user.

Additional features:

- M.A.I. can be used in conjunction with Thesaurus Master.
- M.A.I. is written in Java with APIs (application programming interfaces) for integrating with other software.
- M.A.I.'s operation depends on an automatically created rulebase that is easily modified to increase its effectiveness with specific document collections.

The Enhanced Search Interface in Search Harmony

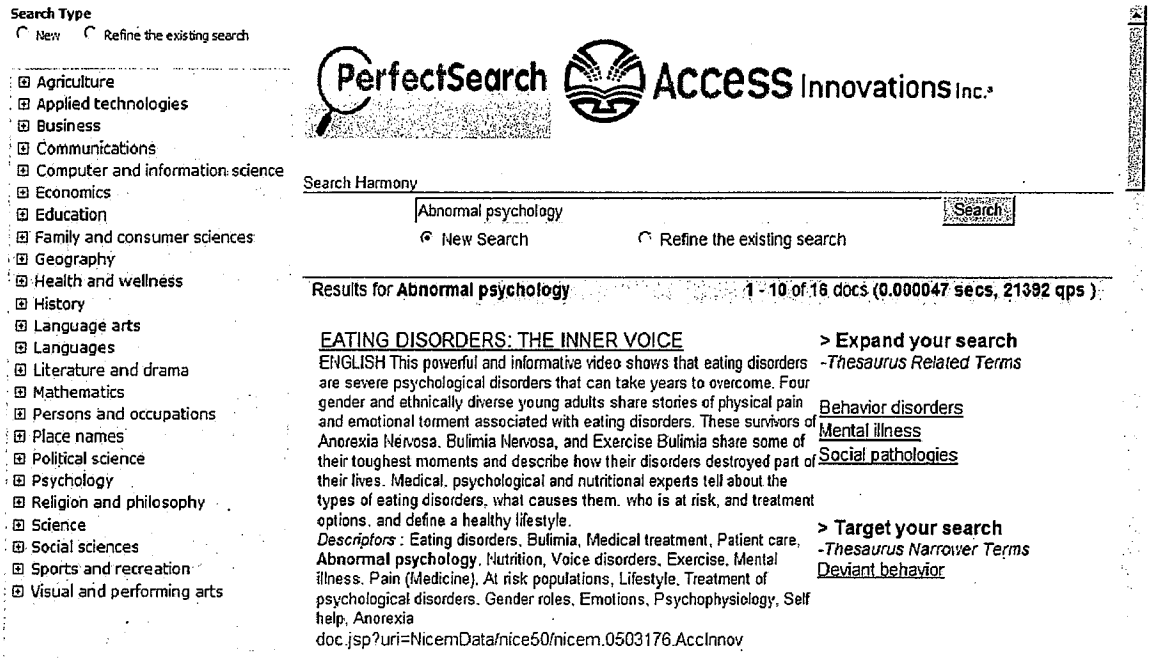
Data Harmony's Search Harmony is taxonomy enhanced search that includes hierarchical browse navigation, automatic query completion using taxonomy terms and their synonyms, contextual presentation of related terms to broaden the search and narrower terms to refine the search, spell check, "more like these" sets, and recursive searches. Based on Access Innovations' MAIstro automatic indexing and thesaurus software suite (part of Access Innovations Data Harmony product line), it provides well formed data and highly accurate search results. Used on top of search software, it provides the best of breed in recall, precision, and relevance searching.

Adding the presentation layer from Access Innovations' Data Harmony software suite over the Perfect Search Corporation search engine gives the user the ability to improve search results through taxonomy-enhanced methods such as automatic completion based on the synonyms and taxonomy terms, related terms to broaden the search, narrower terms to refine the search, browsing of taxonomies associated with document sets, spell check, and recursive sets. We are therefore proposing that an adaptation of, for example, the Defense Technical Information Center (DTIC) thesaurus, or a similar one, be incorporated into the search.

Key advantages of the joint product include the following:

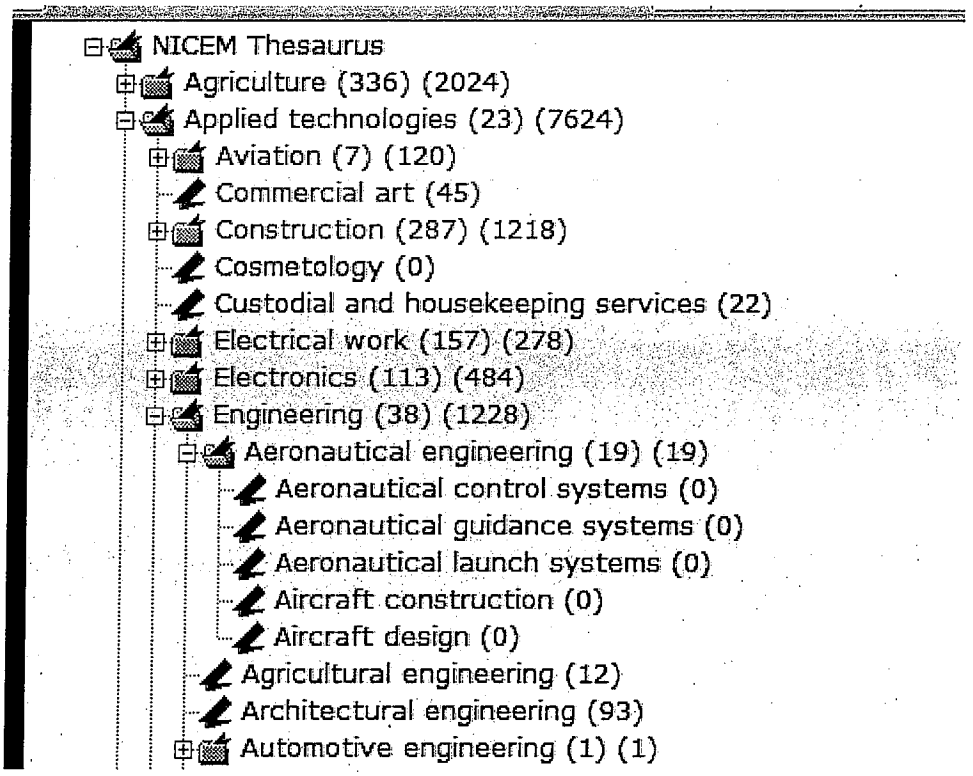
- Clean XML for the patent collection back to 1976 already
- Clear and concise search responses, helping users get to information more quickly.
- Better, more exact answers based on rule based ranking.

Refinement, expansion, and targeting of search



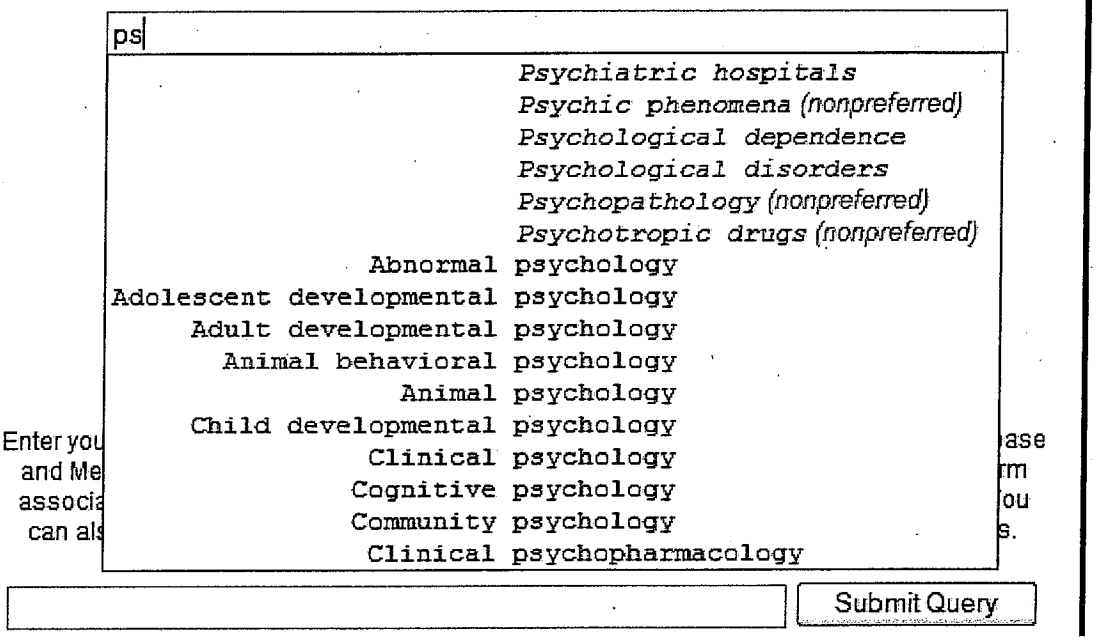
Browseable navigation (with one branch expanded)

Clicking on a term brings up a list of records that have been indexed with that term.



Autocompletion of possible terms to use in searching

Start typing the word(s) you want to search on and a list of suggested completions appears. These represent every term in the thesaurus that matches the starting letters you've just typed. Click on one of the terms in the pick list and then the *Search* button to see a list of item titles that are indexed with the chosen term.



CORPORATE EXPERIENCE / PAST PERFORMANCE

U.S. Patents Demonstration Site

In a jointly developed demonstration project, Access Innovations and Perfect Search Corporation have created a Web site containing approximately 3.5 million U.S. Patent grants. This strictly experimental demonstration site can be found at

http://patent.perfectsearchcorp.com:30015/index.html and
http://patent-app-plus.perfectsearchcorp.com/

Each url represents different approaches and different treatments of the data. The first site has all of the converted patent grants loaded and searchable. The second site has a few thousand grants loaded that have been indexed using the DTIC thesaurus. We are using the DTIC thesaurus for demonstration purposes. (Please note, it is not necessarily the best choice for indexing the entire patent database but does provide an indication of the possibilities for the USPTO.)

In first stage of this pilot project, Access Innovations converted all of the grants and applications files from 1976 to 2007. This includes sources files in ASCII and various XML DTDs that have been used over the years by the USPTO. The resulting database is

now in a single, clean XML file. The XML DTD Access Innovations designed covers, – in a single DTD, all the variations between the grants files and the applications files.

Access Innovations has also loaded the Patent Classification Codes into its Data Harmony Thesaurus Master software. As part of the demonstration site we will complete the indexing of the grants file adding the DTIC thesaurus terms and using the USPTO assigned Patent Classification Codes. The second url (http://patent-app-plus.perfectsearchcorp.com/) shows the results of indexing a few thousand patents using just the DTIC thesaurus. In the very near future, we intent to complete the indexing of the converted grants file of over 3.5 million patents using both the DTIC thesaurus and the Patent Classification Codes.

Access Innovation's has loaded the DTIC thesaurus and the Patent Classification Codes into its Data Harmony software tool. We are using the tool to automatically index and classify the 3.5 million grants file. Once this process is complete, we will be able to deploy the Data Harmony Search Interface for guided navigation. These capabilities are described starting on page 7, above. They can be explored at the following urls:

http://www.mediasleuth.com/ and
http://www.accessinn.perfectsearchcorp.com:32121/

The functionality displayed at these sites are not yet fully available at the demonstration patent sites mentioned above. These capabilities will be added as soon as the indexing is done.

The Perfect Search Corporation search software has the ability to provide Boolean logic queries with all the advantages of high precision and recall, but it can also provide guided, faceted navigation. On the two Web sites mentioned the preceding paragraph, you can experience guided navigation using a taxonomy. Once we have indexed the 3.5 million grants database, guided and faceted navigation can be done using a variety of the fields of a patent. For example, one can search using the grant number. Since the underlying database is XML, any XML element found in the data set can be used for either faceted navigation, Boolean logic queries, or as part of a power/advanced search page.

By combining Access Innovation's conversion skills (honed over more than 31 years of data conversion), its Data Harmony software, and the tremendous power, flexibility, and speed of the Perfect Search software solution, we are demonstrating the capabilities and capacity to handle the requirements of the USPTO. Access Innovations has worked with the USPTO data in the past and set the quality control system for the conversion of the patent files to image format using the original patents in the Boyers, PA collection.

Corporate Experience – Access Innovations, Inc.

Access Innovations, a small woman-owned business, has extensive experience with, and deep knowledge of, Internet technology applications, bibliographic database creation, visualization knowledge, and thesaurus/taxonomy creation and application. Incorporated in 1978, Access Innovations, Inc. is now in its 32nd year of providing information management services. Established in Albuquerque, New Mexico, the organization

maintains its corporate headquarters there. The organization's services are provided to a broad range of public and private organizations in the United States as well as in international locations.

Employing a core staff of 18 professionals, Access Innovations is small, but highly regarded in the information services and software industry. In 2005, the organization was included in KMWorld's prestigious "List of 100 Companies That Matter in Knowledge Management." In 2008, KMWorld included Access Innovations' Thesaurus Master/MAIstro software on its list of "Trend Setting Products of 2008". Key personnel are extremely active and well known in information management organizations.

Access Innovations has set up workflow analysis and production guidelines modeled on the Baldrige and Six Sigma quality management methodologies. The organization has won recognition from Quality New Mexico, an organization dedicated to promoting organizational excellence.

Throughout Access Innovations' years of service to the information industry, clients have included companies of all sizes and focuses, from Fortune 100 companies to university libraries, from corporate and technical information centers to government agencies, foreign and domestic. The organization has provided the following organizations, among others, with information management software and/or services: the American Mathematical Society; Amgen; the Getty Institute; the American Society for Information Services and Technology (ASIS&T); the American Society of Radiologic Technologists (ASRT); the American Water Works Association (AWWA); BIOSIS Zoological Abstracts; Cambridge Scientific Abstracts; Derwent –engineering and Derwent Pharma Bio (now part of Thomson Scientific); the Electric Power Research Institute (EPRI); the National Agricultural Library; the National Transportation Institute; the National Institute of Environmental Health Sciences (NIEHS); and the Weather Channel.

Past Performance – Access Innovations, Inc.

Institute of Electrical and Electronics Engineers (IEEE)

The IEEE Indexing Group is responsible for daily processing from 200 to 3,000 new research articles for the *Xplore* online delivery platform, generating the metadata needed for each article. To manage the indexing vocabulary and categorization for the wealth of electronic documents and to promote productivity in the Indexing Group, IEEE uses Access Innovations' Thesaurus Master and M.A.I., seamlessly integrated as MAIstro.

IEEE maintains its thesaurus of approximately 5,000 terms and nearly 2,000 synonyms in Thesaurus Master. The taxonomy tool is synchronized and integrated with the categorization rule base in M.A.I. The rules in M.A.I. govern application of the taxonomy terms to the stream of incoming documents as well as to IEEE's vast store of legacy documents. The rules are initially generated programmatically, and IEEE editors can fine-tune rules to capture new expressions of taxonomy concepts in incoming documents.

In the IEEE indexing workflow, content streams into the IEEE's Digital Asset Management System to start processing. Programmatic checks filter records to pass through material with certain metadata entries automatically applied. For automatic subject indexing by taxonomy categories, selected metadata are processed by M.A.I., including the publication title, abstract and first paragraph of XML articles. PDF documents are preprocessed to extract text and then indexed.

Of the 180,000 articles auto-indexed annually, the editorial team performs a quality review on about 15,000-20,000, or 10 to 200 in a day. At this point, they may catch the odd error such as the taxonomy term "Gold" as a chemical element being suggested by the phrase "the gold standard." Discovery of a categorization error or missing a significant concept prompts an editor to fine-tune a rule in M.A.I., adding or modifying a condition for more precise application of indexing terms.

Indexed documents stream forward to publication on IEEE's Xplore database, where the subject metatags on the documents enable precision in search and retrieval by topic. In addition to subject metadata, subscribers can access content by searching authors, terminology from additional taxonomies, and other metadata.

International Foundation of Employee Benefit Plans (IFEBP)

Since 1986, Access Innovations has been producing the Employee Benefits INFOSOURCE database, which now contains more than 60,000 article summaries from more than 350 periodicals and newsletters. The database is used by the Information Division to answer questions posed by the Foundation membership concerning any aspect of the employee benefits field. Access Innovations editors do abstracting and indexing of journal articles, including texts on highly technical financial and legal matters. Editors use in-house machine aided indexing software to obtain lists of controlled vocabulary terms suitable for categorizing each article, and manually select the terms, for extremely high indexing accuracy and completeness.

National Information Center for Educational Media (NICEM) Database

In April of 1984, Access Innovations purchased the NICEM database from the University of Southern California. Today, the comprehensive NICEM database of over 440,000 bibliographic records, representing over 640,000 items, is available directly from NICEM as the Film and Video Finder Online or from licensed partners A-V Online from Ovid and The Library Corporation. CD-ROM versions of the database include the A-V Online on CD-Rom from SilverPlatter and the NICEM A-V MARC CD-ROM from The Library Corporation.

Access Innovations provides all the abstracting and indexing to maintain, enlarge, and update the database. Formats cataloged include videotape, videodisc, various audio

formats, filmstrip, CD-ROM, and software, as well as slides, transparencies, streaming audio and video, and Web-based audio and audiovisual materials.

The NICEM database is widely regarded as the world's most comprehensive audiovisual database and a crucial reference tool for librarians, media specialists, training directors, teachers, university faculty, and researchers.

Revolution Health Group (RHG)

Access Innovations assisted Revolution Health Group LLC (RHG) in updating and expanding their online consumer health database, RevolutionHealth.com. RHG is a leading consumer-centric health organization founded to transform the manner in which people approach their overall health and wellness. RevolutionHealth.com is their free comprehensive health and medical information website.

The data normalization required using special proprietary conversion and metadata enhancement techniques. As Randy Yanoshak, RHG's content systems product manager, has stated, "The challenges Revolution Health faced were substantial as we tried to integrate medical content from more than a dozen editorial partners into a nascent portal, and Access Innovations' depth of knowledge in data transformation and taxonomy development were essential to our successful launch. Each partner offered distinct nuances in their data, and Access Innovations flexibly and expertly adapted the Revolution Health schema and their own processes to ensure a smooth integration."

Corporate Experience – Perfect Search Corporation

Perfect Search Corporation was incorporated in January of 2007 with the idea of building game changing, breakthrough technology in the area of search. The founders of the company had been wrestling with a puzzling phenomenon: Hardware was doubling in performance every 18 months, according to Moore's law. All performance gains in search had been the result of running on faster hardware, rather than better software. Why had there not been any breakthrough in search software performance?

They set out in 2005 to find an answer to increase search performance similar to the performance gains that had been seen in hardware. As they worked together and came up with ideas they started to realize that these ideas could be implemented in software and after 2 years formed Perfect Search to bring those ideas and technology to market.

The main engineering team at Perfect Search has a cumulative 75+ years in the search industry. They are credited with writing the original search program for WordPerfect and Novell. They were also involved with Folio Corporation (which later sold to Fast Search & Transfer (FAST), which is currently owned by Microsoft).

Examples of Perfect Search technology in action include the following:

- World Vital Records, who provide search on 1.4 billion documents of both structured and unstructured data, running on a single commodity Windows server. All query responses come back sub-second. See them at http://worldvitalrecords.com

- OneGreatFamily, a genealogy company with 200 million records that is looking to drastically reduce the latency of data being input to when it is available for search.
  In implementation.

- I.TV, a company that provides a TV listing service for iPhones that needed to join data from 5 databases in and provide sub second response to queries from their customers.
  In implementation

- NICEM database of 500,000 educational titles for search
  http://www.accessinn.com:8081/PerfectSearch/navtree/index.html

- Demonstration based on USPTO data
  http://patent.perfectsearchcorp.com:30015/index.html
  http://patent-app-plus.perfectsearchcorp.com/

Past Performance – Perfect Search Corporation

The Problem

Paul Allen knew the challenge he was facing. After all, he was one of the founders of the world's largest genealogy companies, MyFamily.com. MyFamily (now The Generations Network) has over 6 billion genealogical records and requires thousands of query servers to handle customer searches. These thousands of servers cost millions to purchase and millions per year to operate with energy and maintenance costs. Paul was now launching World Vital Records, a new genealogy company that was rapidly acquiring hundreds of millions of genealogical records. His goal was to become the second largest genealogical company. The question was how could he provide sub-second search to his customers without building the costly and massive infrastructure that MyFamily had built to support the query load from customers?

The Choice

Paul Allen had the following choices:

- Build his own proprietary search engine as MyFamily had done
- Build his own search platform using open source solutions such as Lucene
- Buy a solution from an Enterprise Search Vender

As a new start-up, World Vital Records was constrained with limited resources. Paul wanted to use the majority of his resources towards data acquisition, new social networking features, and marketing, not towards his search platform. Building his own proprietary search engine would take multiple man-years to develop and was not feasible.

As World Vital Records acquired more data, solutions from the current Enterprise Search Vendors became too cost prohibitive. World Vital Records had over 800 million genealogical records with a mix of both structured and unstructured data in over 10,000 files. Solutions from existing enterprise search vendors ranged into the millions of dollars, well beyond the reach of World Vital Record's limited budget.

With a limited budget and some talented developers, World Vital Records decided to use Lucene, a free, open source search engine, as their initial search platform, providing exact and near-exact search. Lucene worked great with low traffic and the initial small data sets, but as the data and the traffic grew, the Lucene platform started to have its own performance and financial costs.

As the data sets grew in size, indexing times became a bigger and bigger hassle. It would take over 880 hours of processing time to index 40 gigabytes of data. The Lucene system could handle about 1 query per second per server. To meet the traffic demands and to keep their index stored in cache, they partitioned their index across 6 servers and utilized a Collation Server to distribute the query demand. In spite of their load balancing attempts customer query volume frequently peaked and query response Times slowed to unacceptable levels and would sometimes time out during a customer query. The CPUs were often maxed at 100% utilization trying to process this volume of traffic and query load.

To continue to grow with Lucene, World Vital Records would need to continue to add servers to handle additional data and additional queries. Paul Allen was continuing to aggressively add genealogical data, with plans to add 50% more records to the 800 million existing records.

The Solution

About this time, World Vital Records accepted a proposal from Perfect Search Corporation to test Perfect Search's new search engine in a parallel system. The requirements were to replace Lucene, to match existing business rules, to incorporate exact and near-exact search, to match or improve results, to perform on fewer servers, and to provide query results back to World Vital Records in the same format as expected from Lucene.

The Result

Perfect Search was able to reduce the server requirement of World Vital Records from 7 servers to 1 server, while handling a 60% growth in data. Today, more than 1.3 billion records exist on 1 server, with an additional server for redundancy. The same 40 gb of

data that took over 880 hours to index on Lucene is now indexed in about 8 hours by Perfect Search. Wait times were reduced so query response times were sub-second, even with an increase of traffic and data size. CPU utilization seldom exceeds 10% at peak query loads. Perfect Search's solution delivered the following benefits to world Vital Records:

- Reducing indexing processing time to 1/100 of the Lucene times
- Reducing query servers from 7 to 1 server
- Reducing query times to sub-second
- Allowing for continued dramatic data growth without significant server expansion
- Allowing World Vital Records to compete with the market leaders at a fraction of the server capitalization and maintenance costs.

This system has been in production since Aug 2008 with no downtime.

References

MAIstro customer:
Adam D. Philippidis
Manager, Indexing & Database Production
Institute of Electrical and Electronics Engineers (IEEE)
PO Box 1331
445 Hoes Lane, Piscataway NJ, 08855-1331 (USA)
Voice mail: 732-562-6840
Fax: +1-732-562-1182
aphilipp@ieee.org

MAIstro customer:
Grant Slade
American WaterWorks Association
(see his article at http://www.asis.org/Bulletin/Aug-05/slade.html)
6666 W. Quincy Avenue
Denver, CO 80235
Phone: 303-347-6170
Fax: 303-795-1440
gslade@awwa.org