Enabling Data Access with Automatic Mapping

Jess Kozman
Technical Integration
MetaCarta
Perth, Western Australia
jkozman@metacarta.com

Jonathan Williams, Richard Daniel Ellitt
Information Management
Schlumberger
Aberdeen, Scotland, U.K; Perth, Western Australia
Jwilliams2@slb.com, REllitt@slb.com

Geo-processed attributes are pervasive in natural resource geotechnical data, information and knowledge elements, and users expect integrated architectures for enterprise search to be map-enabled. Integrated architectures similar to those already deployed at oil and gas companies are now being piloted in the mineral extraction segment of the natural resource industry. These deployments apply best practices and lessons learned from the last decade of technology advancement in oil and gas data management. Recent pilot projects have demonstrated significant value from integration of enterprise keyword and spatial search technology, Natural Language Processing (NLP) for automatic mapping (geo-tagging) and geo-spatial web services that combine map views of structured and unstructured data. While enterprise-level resource and asset management tools have been recently more widely adopted and deployed [1], the mining industry is still less standardized in the use of geospatial data than oil and gas when measured on industry capability maturity models [2]. This project shows the ability of the mining industry to adopt integrated architectures that are indicators of higher levels of maturity from functioning examples in other natural resource sectors, such as oil and gas exploration. Progression from resource intensive thick client applications to role-based and web-enabled access and the incorporation of data validation methodologies [3] are both indicators of a move from Level II (Aware) to Level III (Systematic) on the spatial data management maturity model. This pilot project demonstrates the value of applying such previously validated and field tested integrated architectures.

Keywords - Integrated architectures; Natural resource information systems; Automatic mapping

Introduction

The pilot project is part of a larger strategy to discover globally, access regionally, and manage locally all of the data, information and knowledge elements utilized by a mineral exploration division. This includes geochemical and geophysical, documents (both internal and external as well as those stored in structured and unstructured data repositories), GIS map data, and geo-referenced image mosaics. The initial stage involved validating a technology for spatial searches to enable streamlined, intelligent access to a collection of scanned documents by secured users, through scheduled automated crawls for geo-tagging, and following corporate security guidelines. This stage also included administrator training, including defining procedures for managing the document collections, procedures for maintaining the hardware appliance used for generating the spatial index, customizing the User Search Interface (Fig. 1), and developing and implementing support roles and responsibilities. Functionality testing was run on a subset of documents representative of the enterprise collections that would need to be addressed by the Exploration Data Access (EDA) solution. The goal was to expand indexing of legacy documents related to mining and energy prospects, enabling geoscientists to instantly discover and retrieve stored reports and other data relevant to their specific areas of interest.

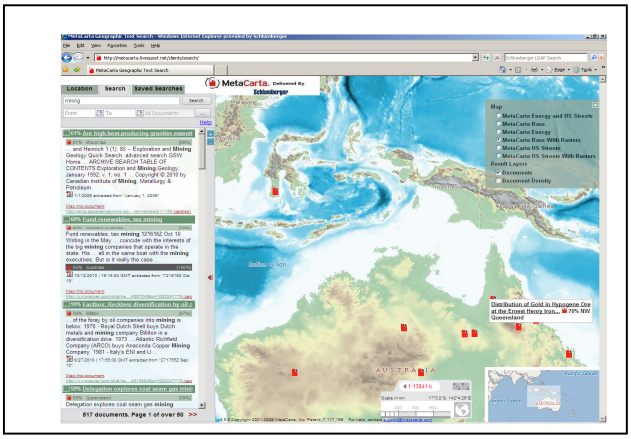

Figure 1. User Search Interface (USI) for unstructured data showing automated mapping of document content.

Implementation

The next stages will focus on broadening the user base, with a goal of having access and use by all corporate geoscientists by the end of 2010. This will be accomplished by defining, prioritizing and publicizing the automatic mapping of additional document collections, developing a custom gazetteer with geographic place names specific to the Australian mineral industry, and integrating with existing GIS map layers such as land rights. A proof of concept enhanced USI (Fig. 2) will be rolled out to a User Reference group for input and feedback. This USI will be similar to those deployed at major global oil and gas operators, for incorporating both structured and unstructured data search results, and supporting both keyword and spatial search [4]. This stage will be supported by a testing and development hardware appliance, connectors to existing electronic document management systems (EDMS), geo-spatial web services and SQL data stores, and a feature rich enhancement of the User Interface. This stage will align the Spatial Discovery project with the larger Exploration Data Access (EDA) initiative and enterprise search strategies based on integrated architectures from other resource industries.

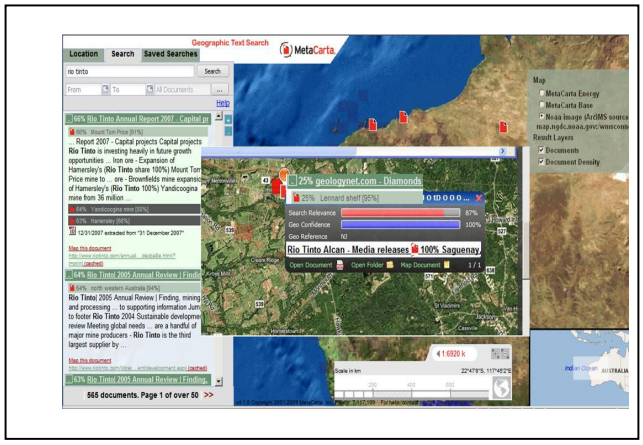

Figure 2. Mock-up example of the proposed enhanced user interface.

An essential part of this stage is the management and updating of a Customized Gazetteer (Fig. 3) to work with the NLP engine and automatic mapping software, which identifies geographic place names in text from multiple formats of unstructured documents and categorizes the index by location types such as country, region, populated place, mines, Unique Well Identifiers (UWI), camps, or concession identifiers. The index also allows sorting of search results by relevance based on natural language context, and Geo-Confidence, or the relative certainty that a text string represents a particular place on a map.

Future improvements to the system will include increasing the confidence in automatic mapping by correctly identifying ambiguous text strings such as "WA" in locations and street addresses from context. This will give documents referencing Asia Pacific regions a higher probability of "WA" referring to "Western Australia" instead of the default assignment to "Washington", the state in the United States. The natural language processing engine can be trained using a GeoData Model (GDM) to understand such distinctions from the context of the document, and can utilize international naming standards such as the ISO 3166-2 list of postal abbreviations for political subdivisions such as states, provinces, and territories [5] or other geo-coding standards such as the U.S. Federal Information Processing Standard No. 10 (FIPS) [6]. The capabilities of the natural language processing engine to use grammatical and proximity context become more important for the correct map location of documents when a populated place such as "Belmont, WA" exists frequently in company documents because of the location of a data center in Western Australia, for example, but could be confused with the city of Belmont, Washington, in the United States without contextual clues.

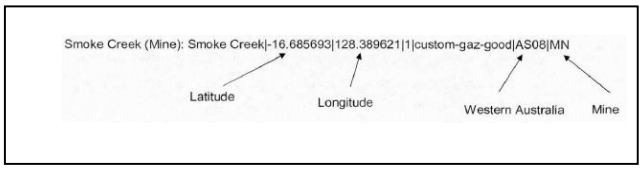

Figure 3. An example of the Geo-Tagging software output, illustrating spatial meta-data passed from an enterprise automatic mapping system to an enterprise keyword search engine (KSE).

The NLP engine is made more robust by an understanding of relative text strings such as "30 km NW of Darwin" and support for foreign language grammar and special characters such as those in French and Spanish. The current NPL engine also has the ability to locate and index date text strings in documents so that documents can be located temporally as well as spatially. Next stages of the deployment will include improvements to the current USI such as automatic refresh of map views and document counts based on selection option context, support for the creation of "electronic data room" collections in EDMS deployments, URL mapping at directory levels above a selected document, and the capture of backup configurations to preserve snapshots of the index for version control of dynamic document collections such as websites and news feeds. The proof of concept USI already includes some innovative uses of user interface controls, such as user-selectable opacities for map layers, the ability to "lock" map refreshes during repeated pans, and utilities for determining geoid centers of polygonal features. Further results of the pilot show that there is the potential to replace the connectors currently in use, enabling an enterprise keyword search engine (KSE) to perform internal content crawls and ingest additional document types and to pass managed properties to the automatic mapping engine to enhance the search experience. The performance of remote crawling versus having search

appliances physically located in data centers is also being evaluated against the constraints of limiting the content crawled from individual documents. The pilot project is designed to validate the ability of the automatic mapping tool to share an index with enterprise KSE's in an integrated architecture like those deployed to oil and gas exploration geotechnical users [7], and to use Application Programming Interfaces (API's) to provide the results of document ingestion and SQL-based structured data searches to both geo-spatial web services and map-based "mash-ups" of search results (Fig. 4).

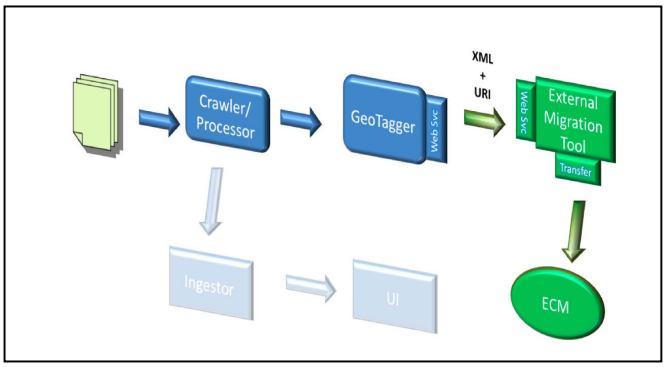

Figure 4. Proposed conceptual integrated architecture of the Spatial Discovery pilot project.

Results

An initial pilot project had previously improved information retrieval efficiency, prospect identification, and decision making, by making available over 88% of legacy documents, and accomplishing results at one-tenth the cost of using a manual metadata-cataloging approach, taking 3 weeks rather than 1 year. The goals of the successful proof of concept stage were; to demonstrate that the automatic mapping tool could ingest text provided by the keyword search ingestion pipe, without having to duplicate the crawl of source documents; to use metadata from keyword search for document categorization such as product type, related people, or related companies; and to provide a metadata list of place names, confidence and feature types back to the search engine. The resulting demonstrated functionality moves towards providing "Enterprise Search with Maps". The completed EDA project is sponsored by the head of exploration and will remove the current "prejudice of place" from global search results for approximately 250 geotechnical personnel for legacy data and information, in some cases dating back to 1960. The solution supports a corporate shift in focus from regional activity focused on projects and prospects with a 24 to 36 month timeline to move to global access that will no longer be biased toward locations with first world infrastructure, and eliminate the need for exploration personnel to take physical copies of large datasets into areas with high geopolitical risk. The corporate Infrastructure Services and Technology (IS&T) group is the main solution provider in the project with the ongoing responsibility for capacity, networking and security standards management. The deployed solution will support search from global to prospect scales, and roles including senior management, geoscience, administrative, data and information managers, research and business development. The focus is on a single window for data discovery that is fast and consistent, with components and roles for connected search and discover solutions. The entire solution will be compatible with the integrated architecture used for the broader context of a discovery user interface and data layer for mineral exploration (Fig. 5). Deployment strategies for data stewardship and governance that will move mineral exploration divisions into higher levels of data management maturity are also being adapted from current best practices in oil and gas exploration [8].

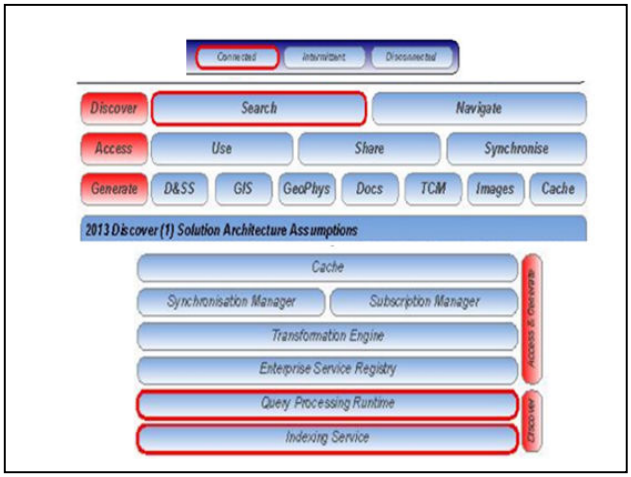

Figure 5. Connected search and discovery integrated architectures in an exploration data access context.

Future Work

Future work identified during the Proof of Concept includes strategies for indexing documents already ingested prior to establishing the keyword search pipe, merging licensing models for the keyword and spatial search engines, and adding full Boolean search capability to the spatial keyword functions. Users are supplied with a larger search result from the keyword search, while spatial search returns only documents with spatial content that can be placed on a map. Conversely, keyword results will receive place name metadata for searching, but will be limited in map capabilities. Separate collections of documents do not need to be built for the spatial search engine, the single crawler reduces load on the file repository, and additional connector framework development is not required. The next stage will validate a security model managing document security tokens in the ingestion pipe using Access Control List (ACL) based security [9]. The baseline integrated architecture was validated during the Proof of Concept phase, with the enterprise KSE passing text from crawled documents

individually to the enterprise automatic mapping engine. The enterprise SSE then extracts metadata and processes text using the NLP engine looking for geographic references. The enterprise SSE (ESSE) passes back managed properties for locations, rating of confidence in location, and feature type (such as mining area, populated place, or hydrographical feature), and the GDM and Custom Gazetteer provide a database of place names, coordinates and features. The system is combined with an existing ESSE component licensed on production for 1 million geo documents, to be used with the automatic mapping stream processor (Fig. 6).

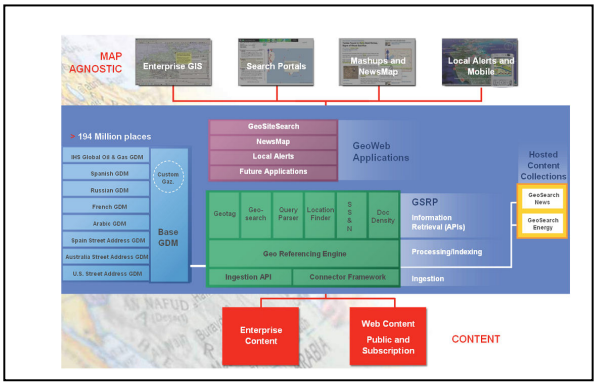

Figure 6. Integrated architecture for geo-spatial search.

Results

Geo-confidence results are analyzed to evaluate the impact of misread characters from digital copies of documents [10] produced through optical character recognition, (such as landscape-oriented numeral "2" being read as state abbreviation "NJ" in the United States or scanned images of pages removed from ring binders having the rectangular holes read as capital letter "I") and ambiguous character strings such as "tx" being an abbreviation for "transmission" in field notes for electromagnetic surveys as well as a the postal abbreviation for Texas. The accuracy of the OCR system can have a significant impact on the relevancy of automatic mapping results, and the success of the user search experience. Again, experience from oil and gas data management practitioners is helpful in understanding the impact of data quality on user acceptance of data management tools [11]. Another indicator of advancing data management maturity is partnering with technology providers to determine best practices. Recent technology partnerships in the mining industry have included providers of geo-spatial web services that incorporate the idea of large amounts of static or base layer data (land boundaries, geo-referenced images and grids) overlain by dynamic operational data such as geophysical and geochemical interpretations. Other development strategies adapted from oil and gas may include launching search in context from analytic applications to create Level IV (Dynamic) observational characteristics [12], conforming to public Open Geospatial Consortium (OGC) standards, using the "shopping cart" concept of commercial web services, and arranging spatial metadata and taxonomies along the lines of ISO content categories. There is also a vision in the mining industry for using Level V (Optimized) expert systems from oil and gas such as real time remote monitoring of spatially enabled sensors, predictive analytics for preventing downtime [13], and artificial intelligence for logistics and supply chain management [14].

Conclusions

The pilot project team identified several achievements from the Proof of Concept phase (Fig. 7). Documents ingested by the keyword search engine that had place name references were successfully located on the user map view.

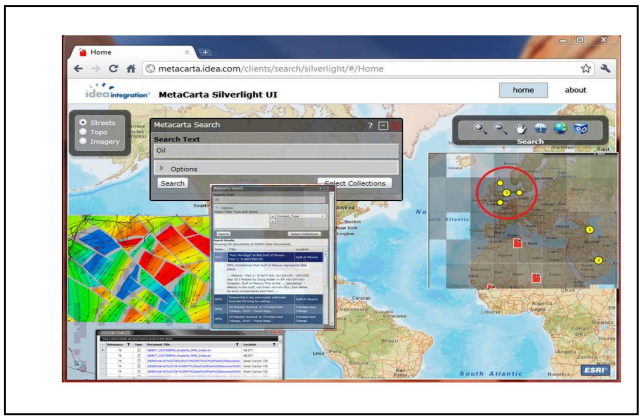

Figure 7. Proof of Concept: Providing structured and unstructured results in a single interface.

Categories passed from the keyword search such as source or company names were able to be searched in the spatial search engine as document metadata. Also, feature types and place names with location confidences were provided, appearing on the spatial search page as managed properties. The system will be enhanced in the deployment phase with security implemented by passing ACL's associated with each document through the ingestion pipeline, and processing for replicated security in the enterprise SSE. Improved presentation of returned managed properties will allow them to be used as a refined list. Search categories can be selectable from an enhanced user interface to allow, for example, selection of a product type for search refinement. This will complement the current Boolean search parameters available in the map view. The ability to merge search results from structured and unstructured data will place mining exploration GIS systems [15] on a similar Data Management Maturity level with existing oil and gas implementations [16]. The enhanced User Interface also presents the density of search results, the directory location of located documents, and the file type of the document. The map view allows a more Australasia centric map experience by removing the arbitrary "seam" at the International Data Line (Longitude

180 degrees) so the region can be centered on a map. The concept of "Enterprise Search with Maps" will be driven as part of the architecture of the Exploration Data Access project, in order to meet the business needs of exploration users. Use cases (Fig. 8) and data lifecycle information is being gathered and analyzed for comparison with existing similar studies of oil and gas exploration systems, for benchmarking and planning purposes.

Figure 8. Use case analysis of a geographic knowledge portal.

Final levels of integration may be impacted by decisions on future versions of geo-spatial web services. Next steps include a cost and benefit analysis of enterprise license usage during crawl and display processes, impact of licenses for each active geo-tagged document versus the use of managed properties, direct indexing of spatial databases and geotechnical repositories using the keyword search engine, and security implementation. A third party application is being used to scan and categorize documents discovered with automatic mapping in order to extract and protect potentially sensitive information.

The deployed solution will provide a holistic search interface that allows geotechnical users to answer essential questions about both the structured and unstructured data in their enterprise, improving efficient access to mission critical data and reducing the risk of geotechnical decisions. This proof of concept project has already validated the value of both adopting and adapting best practices and lessons learned from oil and gas data management implementations that have higher levels of Data Management Maturity.

REFERENCES

[1] R. Roberts, "Software firms do the hard sell," Mining & Technology Australia, issue 1, pp. 74-75, 2010.

[2] M. Dougherty. J. Kozman, J. Maskell, T. Ripley, "Enabling a sustainable spatial data managemnt strategy," unpublished, Westheimer Energy Consultants Ltd., September 2010, http://www.westheimerenergy.com/internal-page/white-paper-enabling-a-sustainable-spatial-data-management-strategy, accessed 28-Sep-2010.

[3] N. Quin, "Data system enhances efficiency," Mining Weekly, September 03, 2010, pp. 50-51, Creamer Media at www.miningweekly.com.

[4] T. Donovan, "The synergy of structured and unstructured data in a geographic environment," Proc. Schlumberger Information Solutions Global Forum 2010, London, U.K.

[5] International Standards Organization, "ISO 3166-2:2007 Codes for the representation of names of countries and their subdivisions - Part 2: Country subdivision code," ISO 1998, accessed 12-Sep-2010, http://www.iso.org/iso/country_codes/background_on_iso_3166/iso_3166-2.htm.

[6] Earth Sciences Research Institute (ESRI), "New geocoding solutions provide many options, ESRI ArcNews Online, Spring 2009, http://www.esri.com/news/arcnews/spring09articles/new-geocoding.html, accessed 30-Sep-2010.

[7] S. Byrd and J. Williams, "Optimizing migration of unstructured data with automated geographic metadata creation," Proc. 14th International Conference on Petroleum Data Integration,Information and Data Management, May 18-20, 2010, Houston, Texas, USA.

[8] F. Kunzinger, P. Haines and S. Schneider, "Delivering a data governance srategy that meets business objectives," 14th International Conference on Petroleum Data Integration,Information and Data Management, May 18-20, 2010, Houston, Texas, USA.

[9] L. Guoqing, L. Chenhui, Y. Wenyang and X. Jibo, "Security accessing model for web service based geo-spatial data sharing application," Proc. 3rd ISDE Digital Earth Summit, 12-14 June, 2010, Nessebar, Bulgaria.

[10] A. Gevinson, "Research results: utility of large-scale digitizing efforts," Council on Library and Information Resources, Washington D.C., USA, publication 147, The Idea of Order: Transforming Research Collections for 21st Century Scholarship. June 2010, http://www.clir.org/pubs/reports/pub147/sumGevinson.pdf, accessed 04-Sep-2010.

[11] R. Radhay, "Facilitating data quality improvement in the oil and gas sector," Proc. SPE Asia Pacific Oil and Gas Conference and Exhibition, 20-22 October 2008, Perth, Australia, Paper #116415, doi: 10.2118/116415-MS.

[12] E. Abecassis, "Google search vs. cataloguing," Proc. SMi E&P Information and Data Management Conference, London, UK, February 2009.

[13] K. Greiner, J. Kozman and C. Piovesan, "Intelligent platforms to manage offshore assets", Proc. 2010 ISA Automation Week, 4-7 October, 2010, Houston, Texas, USA.

[14] M. Brace, "Innovate to succeed," Earthmatters CSRIO Exploration and Mining Magazine, issue 19, Mar/Apr 2009, pp 10-11.

[15] C. Burns, "Handling and managing exploration geodata: stories of uranium in Africa and gold in Nevada," Directions Magazine, Wednesday, June 24th 2009, Directions Media, Glencoe, Illinois, USA, http://www.directionsmag.com/articles/handling-and-managing-exploration-geodata-stories-of-uranium-in-africa-and-/122525, accessed 24-Sep-2010.

[16] Flare Solutions, "Cracking information management," Digital Energy Journal, February/March 2007, pp. 11-12, accessed 14-Sep-2010, http://www.digitalenergyjournal.com/issues/DEJjanfeb07lowres.pdf.