Towards an Automatic Metadata Management Framework for Smart Oil Fields

Charalampos Chelmis[1], Jing Zhao[1], Vikram Sorathia[2], Suchindra Agarwal[1], Viktor Prasanna[2]
[1]Department of Computer Science, University of Southern California, CA, USA.
[2]Ming Hsieh Department of Electrical Engineering, University of Southern California, CA, USA.

This paper was prepared for presentation at the SPE Western North American Regional Meeting held in Bakersfield, California, USA, 19–23 March 2012.

Abstract

Vast volumes of data are continuously generated in smart oilfields from swarms of sensors. On one hand, increasing amounts of such data are stored in large data repositories and accessed over high-speed networks; On the other hand, captured data is further processed by different users in various analysis, prediction and domain-specific procedures that result in even larger volumes of derived datasets.

The decision making process in smart oilfields relies on accurate historical, real-time or predicted datasets. However, the difficulty in searching for the right data mainly lies in the fact that data is stored in large repositories carrying no metadata to describe them. The origin or context in which the data was generated cannot be traced back, thus any meaning associated with the data is lost. Integrated views of data are required to make important decisions efficiently and effectively, but are difficult to produce; since data is being generated and stored in the repository may have different formats and schemata pertaining to different vendor products.

In this paper, we present an approach based on Semantic Web Technologies that enables automatic annotation of input data with missing metadata, with terms from a domain ontology, which constantly evolves supervised by domain experts.
We provide an intuitive user interface for annotation of datasets originating from the seismic image processing workflow. Our datasets contain models and different versions of images obtained from such models, generated as part of the oil exploration process in the oil industry. Our system is capable of annotating models and images with missing metadata, preparing them for integration by mapping such annotations. Our technique is abstract and may be used to annotate any datasets with missing metadata, derived from original datasets.

The broader significance of this work is in the context of knowledge capturing, preservation and management for smart oilfields. Specifically our work focuses on extracting domain knowledge into collaboratively curated ontologies and using this information to assist domain experts in seamless data integration.

Introduction

Oil and gas organizations are in continuous pressure to investigate and employ innovative techniques to extract hydrocarbons from depleting reservoirs. Equipment failures, uncoordinated maintenance and other such unplanned interruptions in production may significantly increase cost of downtime [1]. With involvement of multiple vendors, partners, service companies, and contractors; their effective coordination becomes an important priority. Additionally, reporting requirements and compliance to standards provide additional push towards integration and inter-operation across disciplines, tools and data sets. Availability of relevant data plays a key role in managing the oil field. Multi-disciplinary teams of scientists, engineers, operators and managers use various datasets captured during exploration, drilling and production stages to perform modeling, simulation, interpretation, analysis, testing and decision making activities. Exploration and production (E&P) life cycle includes data intensive activities like seismic data acquisition, geologic interpretation, modeling, reservoir analysis, drilling

target selection, drilling, well logging and analysis, production and well monitoring that links the oil field measurements to oil field management decisions [2]. Illustrating the data intensive nature of one of these activities, a study by Chevron reported the case of its Kern River field with 9,000 active wells that records 1,000,000 data points on daily basis [19]. While large fraction of Chevron's data is in structured form, the rest is hidden in Microsoft Excel and Microsoft Access files on individual users' desktop that accounts for 70% of the knowledge value. For instance, engineers and scientists at Kern River field utilize more than nine datasets and eleven tools to complete the design process for a recompletion workflow. In was observed that significant time is lost in locating, accessing, transferring, transforming and using the required data at each stage. This problem is enhanced as duplicate records of various versions these datasets are stored in the network folders. Similarly, Aveva reported that its offshore platforms may contain up to 30,000 tagged items with 40 to 50 individual data fields each and require nearly 60,000 documents [1]. Such activities rely on SCADA systems, hydrocarbon accounting systems, systems of records, automated workflow systems and other domain-specific systems that are supplied by various vendors. Integration of underlying systems and realization of integrated optimization (IO) is therefore increasingly becoming a key requirement for Oil and Gas organizations.

The vision of Smart oil field is a step in this direction for improving efficiency of oil field operation by proper management of data. The *i-Field* program of Chevron [3], Shell *Smart Fields*, Integrated Operation for the High-North (*IOHN*) [4], the *Field of the Future* program of BP, Integrated Production Management (*IPM*) [5], and *UTCS* of ExxonMobil [6] are key efforts in this direction. In addition to efforts from major oil and gas organizations, service organizations like Baker Hughes has devised novel approaches for capturing, encoding, and provisioning of actionable knowledge from experts deployed in the field [7]. As part of the data management effort, it is important to adopt effective record keeping and data curation strategies that have been extensively studied and addressed in other data-intensive disciplines [20].

Among data intensive processes typically performed by oil and gas organizations, seismic processing and interpretation workflows have their prominent share. Seismic imaging is extensively employed in exploration, appraisal, development and production stages of a reservoir [8]. Several techniques are used by interpreters, processors and analysts that include application of various advanced computational algorithms [9]. The interactive geophysical interpretation workflow involves highly interactive and iterative process [10]. This results in heavy computational and storage requirements [11]. The problem also goes beyond management of large number of seismic volumes and velocity models, and intermediate data files created in the process [12].

Data management problems for data intensive processes, like seismic image processing in E&P, boil down to the challenge of effective approaches that ensure provisioning of right information, at right time, to right person in the right format. To this end, effective techniques have been proposed that demonstrated reduced time spent on search [2]. Another approach can be to enforce standards, conventions and best practices that can reduce unmanaged file handling. One such effort included introduction of standards for storage in LAN and Role based access control that significantly reduce data volumes, access time, and other associated overheads [6]. By designing a Data Services System (DSS), Saudi Aramco [13] reported effective management of well log data in a continuously changing environment.

All these approaches affirmed the role of an effective data curation strategy that may include record keeping, and retrieval using manual or automated workflows. A good curation strategy should be able to meet the needs of all involved domain specific processes [14]. Realization of integration efforts releases datasets locked up in silos that give rise to the update propagation problem. Therefore, the data curation strategy must be able to handle such scenarios that may require adding intelligent capabilities [1]. From end user point of view, the curation strategy should be able to support advance indexing, search and map based display capability based on spatial parameters [12].

Semantic web technologies are increasingly being identified as key enabling technology for integrated asset management and smart oil field applications [24]. Among the proposed data curation approaches also, several proposals explored the semantic web technology at varying levels. Semantic web techniques were used to carry out annotations for images to achieve enhanced search capabilities [23], [22]. A visual annotation framework based on common-sensical and linguistic relationships was proposed to facilitate semantic media retrieval [21]. The E&P organizations are also starting to explore possibilities to employ semantic web technology for their integration effort. For instance, Integrated Production Management Architecture (IPMA) uses Ontology for management of data to reduce search time and to facilitate exchange among participating workflows [5]. The Integrated Operations in the High North (IOHN) project developed set of ontologies to drive integration effort [4]. Baker Hughes started extending their technical domain taxonomy based knowledge management system to next level with development of Ontology [25]. This ontology is based on send control vocabulary to classify metadata and provide advance search, filtering and navigation capability for unstructured information sources. They also proposed gatekeeper stage that requires review from community as well as Subject Matter Experts [7].

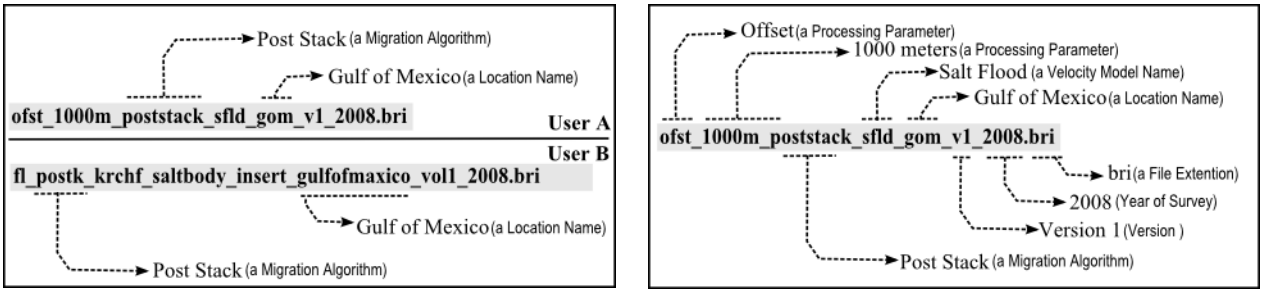
However, development of suitable ontologies from the knowledge hidden in large volumes of structured and unstructured datasets and more critically the expertise of the professional and tacit knowledge that is not externalized in any form is the

key challenge. As 80-90% business data is in unstructured form, in order exploit these rich sources of knowledge, the natural language expressions, must be converted to structured data. Alternatively, they can be semantically enriched to enable extraction of metadata.

While semantic annotation based approach can equally be useful in solving the data curation problem for the oil and gas organizations, adoption has been slow due to several reasons. The successful semantic annotation approach reported heavy reliance on existing taxonomies and ontologies, however, E&P lifecycle involves many domain specific concepts, and in absence of a comprehensive single E&P ontology that covers all involved domains, such annotation approach cannot be realized. Additionally, oil and gas organizations utilize large number of vendor products, and tools developed in-house making the coverage problem more complex. We envision a huge potential for utilizing an ontology driven approach for data curation, with due recognition of the challenges identified in realizing such vision. We argue that proper selection of enabling techniques and their appropriate application in carefully designed information management workflows can address the identified challenges and realize required data curation capabilities

Motivating Use Case

To address the data curation problem in smart oil field domain, we target the challenge of unstructured data management. One of the major contributors of unstructured data is seismic imaging domain that is increasingly being used throughout the E&P life cycle. We focus on investigating data management issues related to seismic imaging – a highly data intensive processes involving various interpretation and processing techniques [9]. Seismic volumes are created iteratively using velocity models with application of appropriate processing techniques that are selected based on the geological structure. A typical oil and gas organization handles few hundred terabytes of seismic datasets as part of seismic interpretation and characterization workflows [12]. The life cycle of seismic datasets involves loading, storing, processing, referencing and visualization processes. Interpretation experts, characterization experts and modeling experts employ various tools and techniques that enable highly interactive data processing and visualization capabilities. In doing so, they generate huge amounts of derived datasets, which are lacking proper metadata that can establish provenance indicating all historical transformations the data has undergone right from the raw field data up to the final, finished product. To track how a particular file was generated, there are several techniques and recommended best practices in place. Some of these are required as part of data reporting guidelines or metadata standards, however, scientists find the compliance to such standards tedious, ending up using their own file naming conventions based on individual preference and style. Figure 1(a), represents file name created by two different interpreters who use different terms for the same concept. For instance Gulf of Mexico is referred to as "GOM" by Use A, whereas User B used the word "gulfofmaxico". Among efforts towards addressing this issue, file naming convention is being commonly included in mandatory requirements and therefore covered in reporting standards [15], [16].



(a) (b)

Figure 1. (a) Terms in a Seismic Volume File Name (b) File Naming Convention Example

Figure 1(b) represents a file name of a seismic image volume. Here, the interpreter who carried out the image interpretation used various terms. Processing parameters (like 1000-meter offset), migration algorithm name, place name, model name etc. related terms are included in the file name. Various types of velocity models generated by geoscientists that are applicable for given geological structure are tried in this process, and therefore, such model names are also included in the file name. Processing parameters provide hints about how the image was loaded and processed in the interpretation system. The seismic survey and project related information are also included in file names or folder names. While general information on the project, the seismic survey and data source etc. is known to everyone involved in this process; the file derivation information, associated model files, processing parameters etc. for a specific volume is only known to the interpreter who generated it.

The volume name example also gives some hints about the file naming conventions followed by its interpreter. A typical file name of a seismic volume contains processing parameters, velocity model name, migration algorithm name, year of survey, version information, project name, location name, and pre and post-processing parameters selected while loading the processing the volume. While generating the volumes, the interpreters typically follow this "template". The existence of such

templates provides unique opportunities for a controlled ontology development, since all terms used in the file names belong to one of these categories. With help of this template of file naming convention and known terms, it also becomes easier to detect missing information in the file name. An interpreter may choose not to include the project name, location name, or survey number in all the derived volumes; however, it is easy to infer, once association among the derived file is established. For instance, a file name could be missing the location name or the survey year, but such information can be easily derived from the source or "seed" files that have full entry.

Based on these observations, we can establish the following key characteristics of seismic file names:

- **They do not include natural language expressions:** Seismic file names only include keywords known in seismic image processing and interpretation domain. File names do not include lengthy descriptions in natural language expression, thereby preventing the effort required for natural language processing of free form text.
- **All keywords contribute to metadata:** All keywords selected by the interpreter provide technical details and specification of the given file using terms well recognized in the seismic image processing and interpretation domain. Therefore, each user-supplied term contribute to the metadata.
- **Some keywords can provide hints to missing metadata:** User may skip capturing detailed context in the filename; however, it is easy to establish the missing information based on who created it, part of which project and other similar information.
- **File names provide hints to workflow:** Terms used in the name provide not only derivation history but also may help identify the workflow by which the current dataset was derived.

Problem Definition

Petroleum engineers and geoscientists involved in seismic image processing play the role of both producers and consumers of large volumes of seismic data sets that are generated or utilized on their workstations. In absence of formal metadata, any attempt in solving the "looking for data" problem may significantly benefit from file naming convention followed by them.

Our example in the motivation section serves as a specific case of the more generic data curation problem that we address in this paper. Given input data, for example filenames, we would like to discover the metadata from the given data. In our example, we would like to discover the different processes that the volume and model files went through and annotate the discovered processes with corresponding filenames. Thus, data annotation task can be expressed as follows:

Given a set of input data $\mathbf{S_d}$ with missing metadata and a domain ontology $\mathcal{O}$, automatically identify concepts of the domain ontology in the input data and annotate the input data with such terms. For concepts that do not currently exist in the domain ontology, automatic annotation is not possible. User supervision is required in this case in order to accomplish the annotation of such terms. Instead of just asking users to manually annotate every individual files with unknown concepts, we exploit such opportunities to capture their background knowledge and expertise about the domain by assisting them in updating the domain ontology. We therefore address the problem of data annotation as a twofold problem:

1. **Automatic annotation of data with missing metadata:** Given a set of input data $\mathbf{S_d}$ with missing metadata and a domain ontology $\mathcal{O}$, automatically identify concepts of the domain ontology in the input data and annotate the input data with such terms.
2. **User assisted ontology maintenance:** Given domain ontology $\mathcal{O}$ and a set of input data that failed to be automatically annotated $\mathbf{S_u}$, assist domain experts in enriching the ontology $\mathcal{O}$ with new concepts, capable of capturing the semantics of unknown terms in the input data.

There is a closed loop between the two problems stated above, because annotation cannot be performed without proper domain ontology in place, while on the other hand, unknown terms (those that are not expressed in ontology $\mathcal{O}$) identified during the annotation process can drive ontology evolution, thus enabling automatic annotation of similar terms in the future. In our proposal, we assume that an initial version of ontology already exists before the annotation process can begin. With the progression of the annotation process, whenever portion of the input data is not associated to domain concepts due to the fact that such concepts do not currently exist in the domain ontology, users assist in their annotation by intuitively defining new concepts and relations in the domain ontology. If an initial ontology is not available to begin with, our technique can assist domain experts in bootstrapping the ontology during the annotation process.

Here we would like to put emphasis on the nature of the ontology $\mathcal{O}$**.** The role of this ontology is not to act as a domain ontology that captures knowledge about seismic interpretation domain. We propose bootstrapping and interactive evolution of ontology that captures the knowledge about file naming convention for a given organization. Therefore, unlike relatively static nature of concepts in domain knowledge, the file naming conventions terms are constantly updated, and therefore, building a file naming convention ontology can be a constantly moving target due to following reasons:

- **Evolving with new projects:** New projects introduce new location, service companies, vendor products and workflows that may results in new keywords in file names.
- **Evolving with new people joining:** New professionals introduce new terms and brings with them a varied level of preference in capturing key information in file names (as discussed with example in Figure 2.).
- **Evolving with new vendor products, scientific techniques:** New vendor products and new tools developed in-house introduces new terms, resulting in support for newer techniques and possibly newer file extensions.
- **Evolving with new data curation policy standards:** With introduction of new regulatory requirements, new keywords can be introduced in the ontology to ensure compliance to metadata standards.

For such an evolving domain, we summarize the data curation requirements as follows:

- **Include all evolving concepts in search and retrieval:** Newly added concepts must be included in advance search and retrieval.
- **Automatically generate missing data**: It should be able to generating missing data based on captured knowledge.
- **Discover and establish relationship among derived datasets:** In addition to derivation history, it is also important to link derived datasets that are associated with specific workflows, decision process, project or equipment etc.
- **Transform metadata for compliance:** To ensure the reporting requirements and metadata standards, the system should be able to generate and maintain metadata according to different schema and content standards.

The file naming ontology is expected to play a key role in addressing these requirements. However, following knowledge representation challenges must be addressed in order to be useful:

- **Ontology coverage:** Coverage of ontology can be a key issue due to constantly evolving nature of organization. It can be bootstrapped by the domain expert, however, coverage can also be achieved by involvement of all the producers and consumers of datasets.
- **Ontology update and maintenance:** Ontology cannot be updated by user as they are not skilled in semantic web techniques:
- **Selection of terms:** Vocabulary is fixed for a domain, but how user will use it in expressing the parameters for a given file is completely personal to the individual and may evolve over time.

Proposed Approach

In this section we present our approach, which is based on semantic web, linguistic processing, and machine learning technologies. Ontology is used for indexing, search and retrieval process. Availability of a comprehensive ontology at design time however is not feasible for constantly evolving domains. As a solution, constant evolution of ontology can be achieved on the runtime with help of user intervention. This requirement assumes familiarity with semantic web techniques, and user's continuous commitment towards updating the ontology, which can be an unreasonable assumption.

Our goal is to achieve this task, without any additional knowledge or effort required by the end user. We argue that this can be achieved by intelligently processing user-supplied keywords in file names that provide hint for concepts in ontology. We propose a method to appropriately classify user-supplied keywords in ontology where a semi- supervised named entity identification approach [17] can be employed. Linguistic processing techniques may further help in addressing variations of these terms. For the ontology, we focus on instantiating a **File Naming Convention *(FNC)* Ontology** based on knowledge in seismic imaging domain that can be further extended by the users on the runtime. Concepts captured from textbook references [9], act as the initial source of domain knowledge that enables bootstrapping of FNC ontology. The source of data can be file names stored in data directories residing on personal desktops or shared locations. We create an instance of each file encountered in such directories in data repository that acts as data catalog or digital library. We assume that users follow some template - an informal file naming convention that can be utilized to our benefit. Every file name is expected to contain a finite number of slots can users can fill with some known values. Users may select different words or abbreviations to represent same concepts, completely omit some of them, or coin new variations or new terms. We denote terms that are *not* defined in the File Naming Convention ontology as **Unknowns**.I If we are unable to determine the value of a slot, we temporarily assign it a **Null** value. Unknown and Null instances are later reviewed and updated by end users, maintaining a constantly evolving FNC Ontology that is updated as new projects, users, vendor products and scientific workflows are reflected in the file name. File naming convention ontology can be mapped to work with metadata content standards. For each file annotated in the process, we create a unique instance in the FNC ontology, that helps in handling multiple versions of files and multiple copies of files stored in different directories.

A. Automated Annotation Workflow

First, we explain the automated annotation workflow that completes the annotation based on the concepts defined in FNC Ontology- without any user intervention. Figure 2(a) depicts the steps involved in this workflow. We consider a set of input data $\boldsymbol{S_d}$ *= {r1, r2, ..., rn}* to be a set of **n** records $\mathbf{r_i}$. Each record contains **k** number of attributes that describes it, hence a

record $\mathbf{r_i}$ is defined as $\mathbf{r_i}$ = *{a1, a2, ..., ak}*. Different records may have different numbers of attributes. The purpose is to annotate each record $\mathbf{r_i}$ in the set $\boldsymbol{S_d}$ automatically by associating $\mathbf{r_i}$'s attributes to concepts in the FNC ontology. .

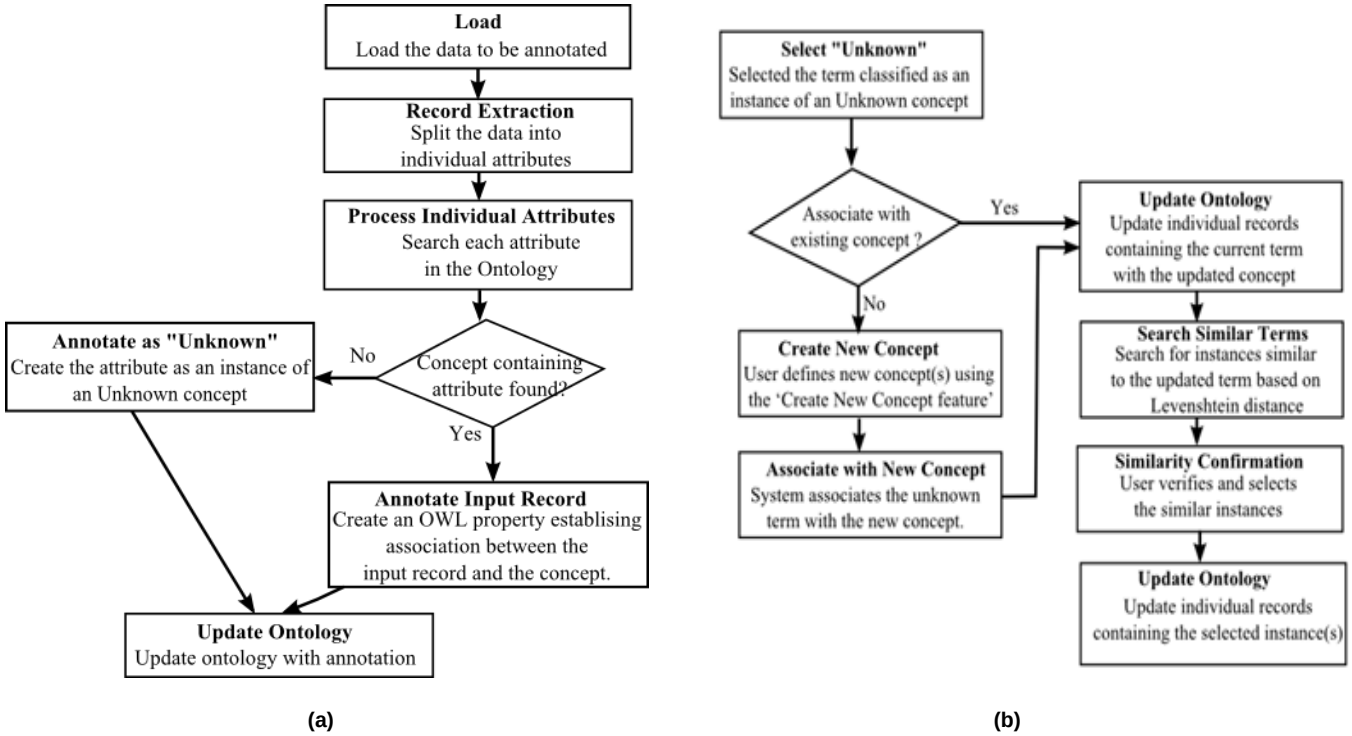


Figure 2. (a) Automated Annotation Workflow (b) User Assisted Ontology Maintenance Workflow

The input set of data $\boldsymbol{S_d}$ may be structured, semistructured, or completely unstructured. Hence, we begin by performing a preprocessing of the input data. During this step, we perform blocking of the input data, splitting them into individual records, consisting of attributes and storing each record and its set of attributes. For instance, the sample filename in Figure 1 becomes a single record $\mathbf{f_1}$ consisting of the set of attributes $\mathbf{S_{f1}}$ = *{ofst, 1000m, poststack, sfld, gom, v1, 2008}*. We process each record individually, examining each of its attributes in term. For each attribute, we probe the FNC ontology and we map the attribute to a corresponding instance of a domain concept, if we are able to find such a concept. We use exact matching to map attributes to domain concepts and we store such mappings as triples in a triplestore using Owl property *OwlProperty : hasConcept*. It is always possible that some of the input data will not match any of the domain concepts in the ontology. We treat such data as unknown concepts and we annotate them as **Unknown**.

B. User Assisted Curation Workflow

We provide an intuitive interface to assist domain experts, who we do not assume to have any prior knowledge and/or expertise in Semantic Web Technologies, in capturing domain knowledge, currently unavailable in the FNC ontology. Using our intuitive interface, domain experts are able to define new concepts for the instances of the **Unknown** class or associate such instances with existing classes in the ontology. The ontology is automatically updated to reflect these changes. By defining new classes or by associating unknown concepts to existing classes in the domain ontology, domain experts assist in disambiguating unknown concepts in the input data. The annotation component is then able to associate previously unknown terms to concepts in the domain ontology, thus establishing mappings and creating annotations. We describe this workflow in Figure 2(b).

We demonstrate our approach with an instructive example. Figure 3 presents a sample ontology, which we call FileNamingConvention (FNC) ontology. This ontology captures file-naming conventions in terms of abbreviations that users choose for their filenames, by denoting them as instances of classes, which abstract different abbreviation schemes. Figure 3 shows the partial snapshot of FNC ontology, where different types of **Migration Algorithm Names** are represented. The same ontology also included a few number of instances for some of the defined concepts. For example, "*gom*" is denoted as instance of the concept **Gulf of Mexico**, which in turn is a *Place Name* (not shown in Figure 3).

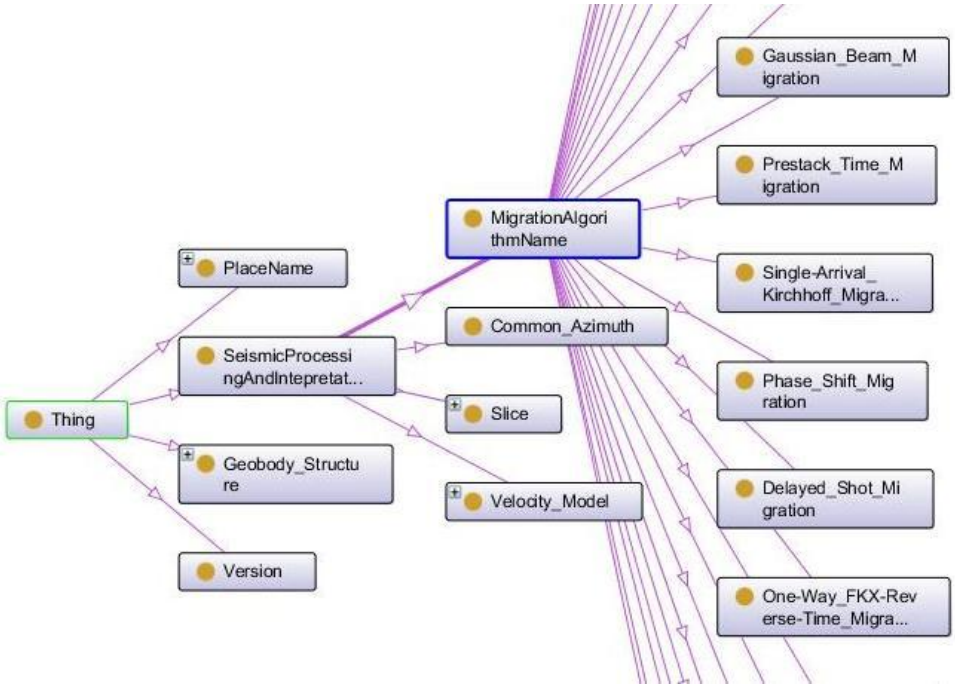
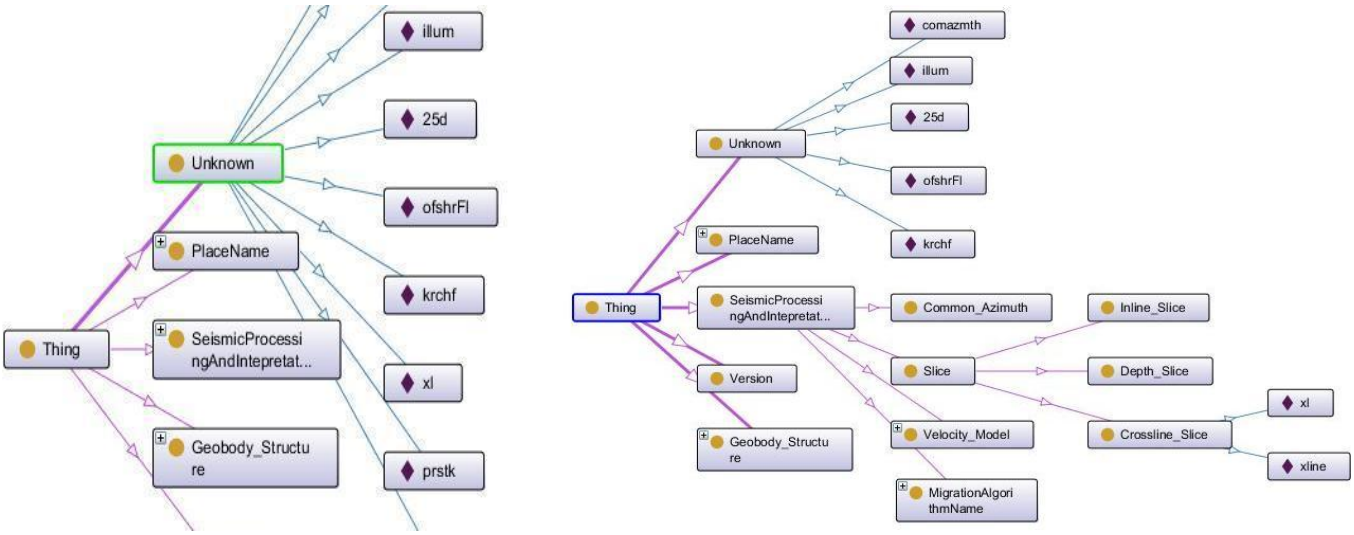


Figure 3. File Naming Convention Ontology

Let us assume for simplicity that we only have one record to annotate: **f1** (see Figure 1a), consisting of the set of attributes $\mathbf{S'_{f1}}$= *{gom, sfld, poststack, saltmute, xl}*. By examining FileNamingConvention ontology we discover that "*gom*", "*sfld*", and "*poststack*" have exact matches in the domain ontology, while "*xl*" is **Unknown.** Figure 4(a) shows partial snapshot of the newly established annotations as reflected in the updated version of FileNamingConvention ontology. Since "*xl*" is unknown, there is no annotation for this attribute of ***f1***. Using our intuitive interface, a domain expert specifies "*xl*" as belonging to the **Cross Line Slice** concept in FileNamingConvention, thereby updating the ontology, as shown in Figure 4(b).



(a) **(b)**

Figure 4. (a) Unknownterms encounted in Automated Annotation Process (b) Unknown Term "xl" Updated in FNC Ontology

Using approximate string matching based on Levenshtein distance [18] we provide domain experts with suggestions of other **Unknown** terms that lexically match. For instance, we recommend terms like "*x-l*", "*x_line*" and "*x-line*", which are similar to the "*xline*", to be included under the same concept **Cross Line Slice**. If domain experts agree with the recommendation, we update the domain ontology to reflect the newly acquired knowledge about the multiple representation of **Cross Line Slice.** We then search for all instances having "*xl*", "*xl*"," *x line*", or "*x-line*" in their attribute sets in order to annotate them with

Cross Line Slice.

Prototype System

We have applied our technique for data curation on filenames that have been created with unknown file-naming convention(s). Our system, **S**emantic Assistant to Manual **C**uration of **D**ata (SSCD), supports automatic annotation of seismic imaging filenames, while at the same time captures knowledge by assisting domain experts in ontology creation and evolution. In this case, filenames are orphan, meaning that they do not carry any associated metadata nor their content is accessible. We have utilized extracted knowledge to automatically recover missing linkage between seismic images and their ancestral velocity models, when no provenance information is recorded [26].
As shown in Figure 5, the system can be divided into three major components, Data Preprocessing, Automated Annotation and Ontology Curation. Data Preprocessing consists of Record Extraction and Record Cleaning. Record Extraction is responsible of splitting input data into a set of records, each of which having a set of attributes. Record Cleaning performs a set of verifications on produced records to ensure that their attributes have appropriate data types as well as that they conform to specific rules, like for instance that they do not have trailing white spaces.

We assume that input data are provided in a column separated file. Users are able to specify the location of input data, their type and number of columns to be annotated. If only one column is available, then we assume that it contains filenames to be annotated. If more than one column is provided, we assume that extra columns maintain metadata about filenames (file creation dates and associated project names). Having some background knowledge on the type of metadata associated to the filenames can assist in their annotation.

Once, preprocessing is complete, the Automatic Data Annotation component annotates all filenames using concepts from the domain ontology. To interact with the domain ontology, Automatic Data Annotation communicates with Ontology Curator. The curator acts as a mediator between the ontology and the other components in the system, converting input queries into SPARQL queries, which are then issued to a SPARQL endpoint. The Automatic Data Annotation component issues search queries searching for matching concepts in the FNC ontology exposed by SPARQL endpoint. If such a match is discovered, it issues an update command to the Ontology Curator, which in turn converts it into a
SPARQL update and then calls the SPARQL endpoint so as to store the annotation. Whenever annotation is not possible, Automatic Data Annotation issues an update command to store the record as **Unknown**.

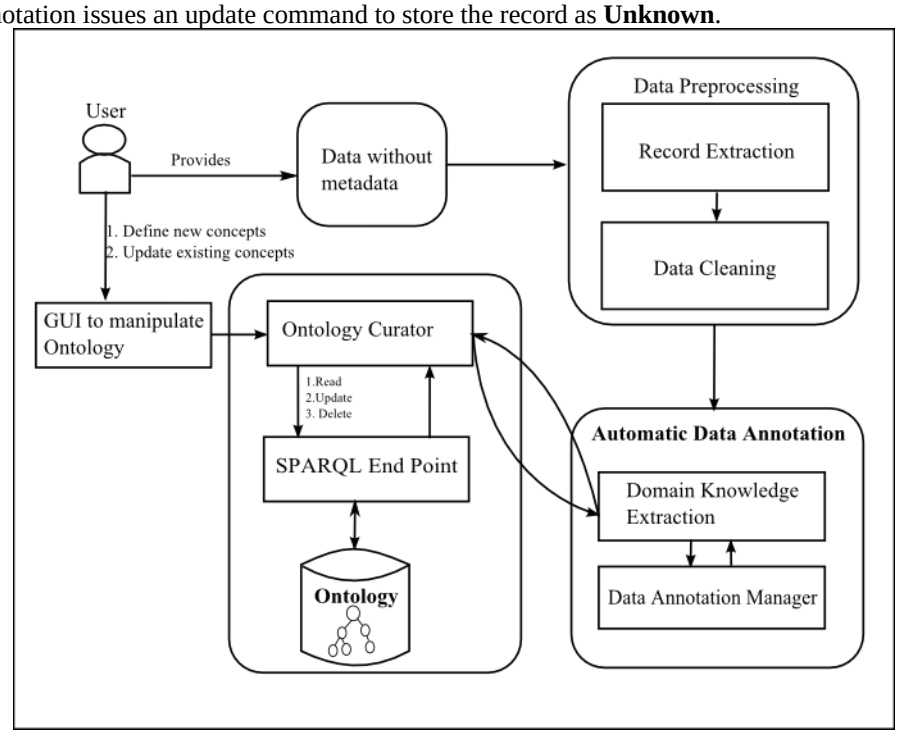


Figure 5. Semi-Automatic Semantic Assistant to Manual Curation of Data (SSCD) System Architecture

Upon completion, Automatic Data Annotation summarizes annotation statistics and prompts domain experts for unknown terms. Domain experts are then able to define new concepts or update existing concepts, resulting in updating the domain ontology. In such cases, domain experts are recommended to inspect similar terms (in terms of lexical similarity) to decide if such terms are also referring to the same concept. If this is the case, the domain ontology is updated accordingly. After

updating the ontology, Automatic Data Annotation component is responsible of inspecting a new round of annotation of input data based on the new knowledge, if required.

Results and Discussion

In this section, we illustrate the functionality of our system. We demonstrate a typical interaction workflow, where we process volume seismic imaging filenames. Based on expected file types, the annotation system can select specific relationships from FNC Ontology that can be utilized for instance creation. For volume data (Figure 6 shows a partial list of filenames), there are three columns; column one denotes file names to be annotated, column two indicates the date on which each file was generated, and column three denotes the project each file is associated with.

ofst_1000m_poststack_sfld_gom_v1_2008.bri 9/10/2008 projectgom
Kirchhoff_prestack_xline_v1.bri 9/9/2008 projectgom
krchf_prstk_xl_v1_gom_2008.bri 9/9/2008 projectgom
comazmth_xline.bri 6/21/2008 projectofl
Dpwtr_laminate_slt_Kir_ofshrFl.bri 6/21/2008 projectofl
1way_shot_prof_offshore_Florida.bri 6/21/2008 projectofl
2wy_bp_25D_w_optimum_imaging.bri 3/12/2004 projectbp
full_2way_mig_BP_2004.bri 3/13/2004 projectbp
krchf_mgr_BP_tomo_2004.bri 3/13/2004 projectbp
full_wavefield_mgr_bp_tomogrph_2009.bri 2/19/2009 projectbp

Figure 6. Input Records Example

A. Automated Annotation capability

After providing path to the input file, user can initiate the annotation process by clicking on Annotate File button. The Automated Annotation process takes over and completes the first phase of annotation. Once completed, Automated Annotation passes control back to the user and opens a new interface giving a detailed report on the annotation. In this report, the user can see the number of records processed, the target ontology, and the current total number of records for the given input type in the ontology.

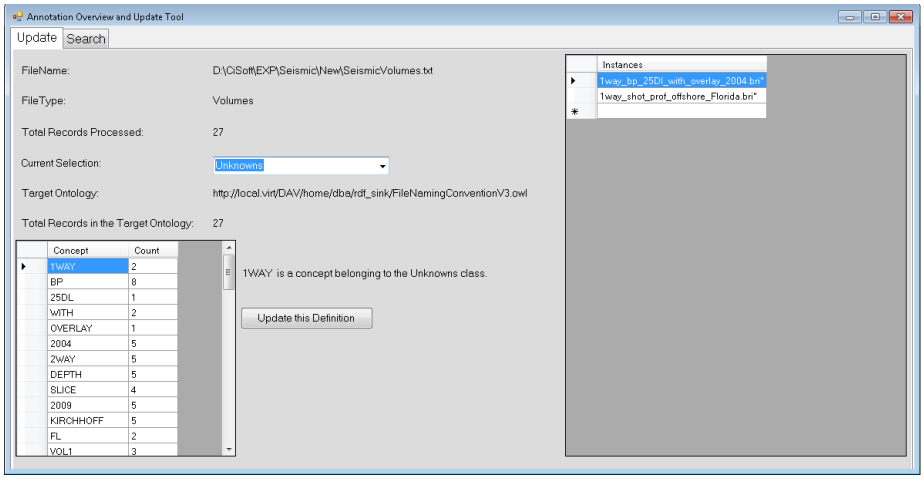


Figure 7. Unknown Terms in Annotated Files

It is important to note that executing the same annotation repeatedly on the same input files does not create new records every time as there can be only one instance of a particular file name in the same repository. This feature in important in identifying and eliminating duplicate copies of the file in local desktop as well in the folder shared over the network. The interface also provides the user with an option of filtering the annotated records based on the concepts in the ontology. For example in Figure 7, user has selected to filter based on Unknown. The tool also provides list of unknown terms along with the frequency of the term encountered during the annotation process. By selecting the unknown concept, the user is displayed the list of instances that includes the selected unknown term. In Figure 7, user selects the unknown term *"1way"* and system returns two instances of files that included *"1way"* term in the file name.

To understand the annotation process, let us consider one entry from the input file depicted in Figure 6. The record

krchf_prstk_xl_v1_gom_2008.bri 9/9/2008 projectgom is first broken down into a filename *(krchf_prstk_xl_v1_gom_2008)*, a date (*9/9/2008)*, a project name (*projectgom)*, and a file type (*Volume)* that was detected based on file extension "*bri"*. The filename is processed for identification of user-supplied terms in FNC Ontology. This steps results in identification of term "*prstk"* as an instance of **Pre-stack Time Migration**, which is a *MigrationAlgorithmName*. Similarly, "*gom"* is identified as an instance of **Gulf of Mexico** that is a *PlaceName* and "*xl"* is identified as instance of **Cross line Slice** which is a kind of **Slice** used in seismic image processing. The system was not able to recognize the terms *KRCHF*, *2008*, and *V1* and therefore, they are classified as **Unknowns**. As a seismic volume file, the filename is expected to have *VersionNames*, *ProjectNames*, *ModelNames*, *ImagingAlgorithmNames*, *ProcessingVendorNames*, *PostProcessingNames*, *Association*. Such attributes were not recognized during the annotation process and hence, they are temporarily set to **Null** value. Table 1 summarizes the annotation process outcome.

Annotation Property from FNC Ontology	**Value detected and assigned automatically**
MigrationAlgorithmName	Pre-stack Time Migration
PlaceName	Gulf of Mexico
Slice	Crossline Slice
Association	Null
PostProcessingNames	Null
ProcessingVendorNames	Null
VersionNames	Null
ProjectNames	projectgom
ModelNames	Null
ImagingAlgorithmNames	Null
FileType	Volume
Unknowns	KRCHF
Unknowns	2008
Unknowns	V1

Table 1. Initial Annotation Generated for Volume File Name "krchf_prstk_xl_v1_gom_2008.bri"

B. Interactive Ontology Update Capability

The interface provides an option of filtering annotated records with concepts in the ontology. For example in Figure 7, user has focused on records that were classified as **Unknowns**. A domain expert might decide to change the definition for the attributes in the **Unknowns** class. This is accomplished through activating the "Update Definition" function.

As shown in Figure 8, the user decides that currently unknown attribute *gulfofmaxico* appearing in two file instances, is actually a known PlaceName. But, it was not classified under any existing sub-concept of the concept **Gulf of Mexico**. To add this new term, the user chooses the PlaceName option that results in a selectable list of all place names that are already defined in FNC Ontology, out of which, the user selects **Gulf of Mexico** and clicks on the Confirm button. The User Assisted Annotation component updates the ontology with the new knowledge and it establishes an association between the two file records containing the attribute "*gulfofmaxico*" and concept **Gulf of Mexico** using an OWL property *hasPlaceNames*. The user can see the updated information by filtering on the concept *PlaceName*s. As shown in Figure 9, there is a new concept **Gulf of Mexico**, which has now five records associated with it, whereas the list of **Unknown** concepts does no longer include the term *gulfofmaxico*.

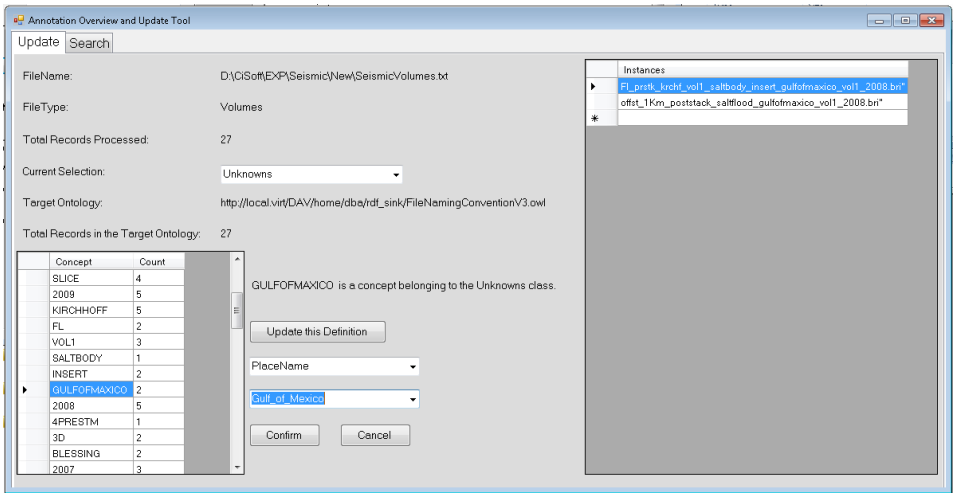


Figure 8. Updating FNC Ontology with "gulfofmaxico" Term

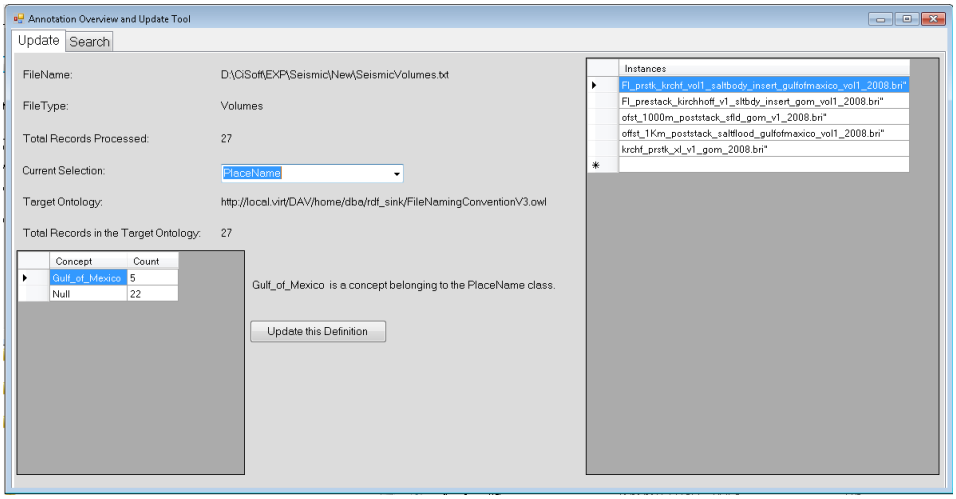


Figure 9. Updated Search Results for "Gulf_of_Mexico" Concept

C. Advanced search capability

Figure 10 depicts advance search capability based on "Search by Ontology" and "Search by File Name" options, demonstrating the semantic web capability that is achieved for a user who is unfamiliar with semantic web technologies. "Search by File Name" provides capability to perform keyword based search. Search by Ontology" allows users to interactively explore the ontology in the form of selectable tree view that allows user to expand, collapse and select at different levels. Based on one or multiple concepts selected from hierarchy, the matching files are listed in the middle-right side of the UI. By clicking any of the listed file, user can further retrieve the complete annotated record -as seen in the lower-right part of the UI screen. Users can review and update, or delete metadata as needed.

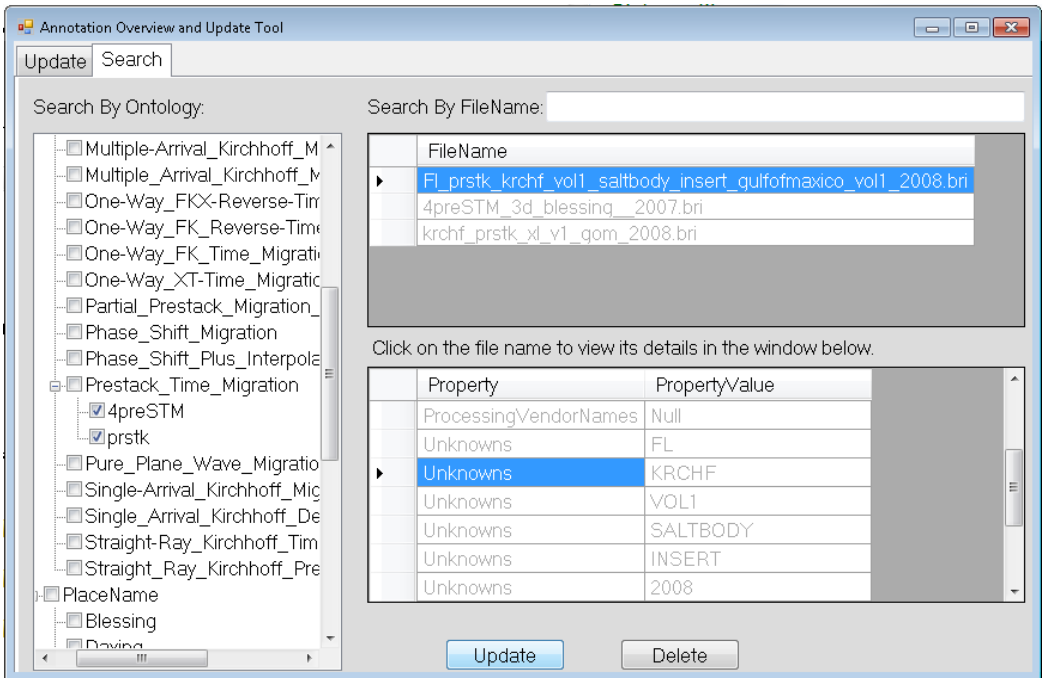


Figure 10. Ontology Search and Update Capability

Conclusion

In this paper, we discussed data management problem faced by exploration and production (E&P) organizations and emphasized on the problem related to large volumes of unstructured data generated by users and stored with no proper

metadata. We investigated seismic image processing, interpretation and analysis workflows as they provide unique opportunity of recovering metadata from various user-selected terms included in their file names.

We argued that Semantic Web techniques can provide useful support for solving this problem, however its adoption for given problem introduced several interesting research challenges. We observed that development of a new ontology or selection of existing domain ontology could not serve this problem due to constantly evolving nature of file naming practices. Additionally, development of ontology and use of semantic search both assumes familiarity with the underlying techniques that may prevent large-scale adoption.

To address these issues, we proposed a semi-automatic approach that provides assistance to users in creating and maintaining a File Naming Convention Ontology that subsequently facilitates the data curation process. This ontology is employed to carry out automated and interactive annotation of filenames thereby generating semantically enriched metadata. Our proposal leverages recent advancements in Semantic Web, Linguistic Processing, Named Entity Recognition and Classification techniques in developing an intuitive and user-friendly system.

They key contribution of our approach is the recognition and realization of unique opportunity to simultaneously generate and update ontology and metadata from the filenames where all user-supplied terms are processed and annotated. We proposed to create a unique instance in ontology repository for each file being annotated that subsequently helps in detecting duplicate copies and multiple versions of same files. In this paper we demonstrated, advanced capability to search seismic images based on various domain specific criteria - like project name, processing algorithms, processing techniques, processing organization, and similar other keywords. . Due to runtime update in Ontology, new filters are instantaneously made available in search capability. Such capability significantly reduces the effort required in searching for right datasets in large repositories. While we achieved data curation for Seismic datasets, this capability is not specific to the seismic domain. The same results can be reproduced for other types of datasets by including additional domain specific metadata attributes unique to the target datasets.

With the promising initial results, we are now planning to extend our work in several potential directions. First, an obvious extension is to include support for other data intensive workflows within E&P life cycle. Considering the metadata standard compliance and mandatory reporting regulations, another important future direction is to incorporate support of such regulatory requirements by automated generation and maintenance of standard compliant metadata content. In addition to filenames, there are several other artifacts that need to be supported for annotation - limited not only to the name, but also including the content. This may include web pages, email communications, short messages, micro-blog entries, presentation files, reports etc.

Acknowledgement

This work is supported by Chevron Corp. under the joint project, Center for Interactive Smart Oilfield Technologies (CiSoft), at the University of Southern California.

References

[1] S. Gibbons, "Harnessing the information mammoth: New advances in data management," in Abu Dhabi International Petroleum Exhibition and Conference, 3-6 November 2008, Abu Dhabi, UAE. Society of Petroleum Engineers, 2008.

[2] K. Edehe, "Optimization of information/data quality amongst dispersed project teams and management groups in the nigerian oil and gas industry," in Nigeria Annual International Conference and Exhibition, 6-8 August 2007, Abuja, Nigeria, SPE, Schlumberger Information Solutions. Society of Petroleum Engineers, 2007.

[3] B. Burda, J. Crompton, H. Sardoff, and J. Falconer, "Information architecture strategy for the digital oil field," in Digital Energy Conference and Exhibition, 11-12 April 2007, Houston, Texas, U.S.A. Society of Petroleum Engineers, 2007.

[4] F. Verhelst, F. Myren, P. Rylandsholm, I. Svensson, A.Waaler, T. Skramstad, J. Orns, B. Tvedt, and J. Hydal, "Digital platform for the next generation io: A prerequisite for the high north," in SPE Intelligent Energy Conference and Exhibition, 23-25 March 2010, Utrecht, The Netherlands. Society of Petroleum Engineers, 2010.

[5] C. Bravo, L. Saputelli, J. A. Castro, A. Ros, F. Rivas, and J. Aguilar-Martin, "Automation of the oilfield asset via an artificial intelligence (ai)-based integrated production management architecture (ipma)," in SPE Digital Energy Conference and Exhibition, 19-21 April 2011, The Woodlands, Texas, USA. Society of Petroleum Engineers, 2011.

[6] M. Garbarini and B. Catron, Robert E.and Pugh, "Improvements in the management of structured and unstructured data," in International Petroleum Technology Conference, 3-5 December 2008, Kuala Lumpur, Malaysia. International Petroleum Technology Conference, 2008.

[7] D. M. O'Neill, L. R. Walls, E. VanSickle, and B. Baker, "New approach to knowledge capture: Center of excellence for sand control completions as a model," in SPE Intelligent Energy Conference and Exhibition, 23-25 March 2010, Utrecht, The Netherlands. Society of Petroleum Engineers, 2010.

[8] T. Alsos, A. Eide, D. Astratti, M. Pickering, Stephen and- Benabentos, S. Dutta, Nader andMallick, G. Schultz,

L. denBoer, M. Livingstone, M. Nickel, L. Sonneland, J. Schlaf, P. Schoepfer, M. Sigismondi, J. C. Soldo, and
L. K. Stronen, "Seismic applications throughout the life of the reservoir," Oilfield Review, vol. 14, no. 2, pp. 48–65, 2002. [Online]. Available: http://www.slb.com//media/Files/resources/oilfield review/ors02/sum02/p48 65.ashx
[9] J. B. Bednar, Modeling, Migration and Velocity Analysis in Simple and Complex Structure. Panorama Technologies, Inc., 2009. [Online]. Available: http://www.panoramatech. com/papers/book/index.php
[10] M. H. Badar, "Application of advanced volume interpretation (avi) workflows to improve data quality for rapid interpretation," in SPE/DGS Saudi Arabia Section Technical Symposium and Exhibition, 15-18 May 2011, Al-Khobar, Saudi Arabia. Society of Petroleum Engineers, 2011.
[11] F. Bosquet and J.-C. Dulac, "Advanced volume visualizationnew ways to explore, analyze, and interpret seismic data," The Leading Edge, vol. 19, no. 5, pp. 535–537, May 2000. [Online]. Available: http://tle.geoscienceworld.org/cgi/reprint/19/5/535
[12] P. Neri, "Data management challenges in the pre-stack era," First Break, vol. 29, pp. 97–100, January 2011. [Online]. Available: http://www.pdgm.com/Resources-(2)/Articles---Papers/
Data-Management-Challenges-in-the-Pre-Stack-Era.aspx
[13] T. A. Al-Ghamdi, G. O. Zahdan, H. A. Ali, and I. M.Nahwi, "Automate the process of managing well logs data," in SPE/DGS Saudi Arabia Section Technical Symposium and Exhibition, 4-7 April 2010, Al-Khobar, Saudi Arabia. Society of Petroleum Engineers, 2010.
[14] J. Leidig, E. A. Fox, K. Hall, M. Marathe, and H. Mortveit, "Simdl: a model ontology driven digital library for simulation systems," in Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries, ser. JCDL '11. New York, NY, USA: ACM, 2011, pp. 81–84. [Online].Available: http://doi.acm.org/10.1145/1998076.1998091
[15] NASA, "Advanced microwave scanning radiometer - eos (amsr-e) data management plan," National Aeronautics and Space Administration, Tech. Rep., 2001. [Online]. Available: http://weather.msfc.nasa.gov/AMSR/data management plan.html
[16] SURA, "Sura coastal ocean observing and prediction (scoop)-filename conventions," Southeastern Universities Research Association, Tech. Rep., 2006. [Online]. Available:http://scoop.sura.org/documents/naming convention final 5-3-06.pdf
[17] D. Nadeau and S. Sekine, "A survey of named entity recognition and classification." Lingvisticae Investigationes, vol. 30, no. 1, pp. 3 – 26, 2007.
[18] V. Levenshtein, "Binary codes capable of correcting deletions, insertions, and reversals," in Soviet Physics Doklady, vol. 10, no. 8, 1966, pp. 707–710.
[19] A. Popa, K. Horner, S. Cassidy, and S. Opsal, "Implementing integrated solutions for reservoir management: San joaquin valley case study," in SPE Western North American Region Meeting, 7-11 May 2011, Anchorage, Alaska, USA. Society of Petroleum Engineers, 2011.
[20] C. Lagoze and K. Patzke, "A research agenda for data curation cyberinfrastructure," in Proceeding of the 11th annual international ACM/IEEE joint conference on Digital libraries, ser. JCDL '11. New York, NY, USA: ACM, 2011, pp. 373–382. [Online]. Available: http://doi.acm.org/10.1145/1998076.1998145
[21] B. Shevade and H. Sundaram, "A visual annotation framework using common-sensical and linguistic relationships for semantic media retrieval," Adaptive Multimedia Retrieval User Context and Feedback, vol. 3877, no. 20, pp. 251– 265, 2006. [Online]. Available: http://www.springerlink.com/index/M22164PQ037422V1.pdf
[22] J. Wielemaker, A. T. Schreiber, and B. Wielinga, "Supporting semantic image annotation and search," Annotation for the semantic web, vol. 96, p. 147155, 2003. [Online]. Available: http://www.cs.vu.nl/guus/papers/Wielemaker02a.pdf
[23] S. Rinc, "Towards standard semantic image annotation and search," 2008 International Workshop on ContentBased Multimedia Indexing, pp. 484–488, 2008. [Online]. Available: http://ieeexplore.ieee.org/lpdocs/epic03/wrapper.
htm?arnumber=4564986
[24] R. Soma, A. Bakshi, V. Prasanna, , W. DaSie, and B. Bourgeois, "Semantic web technologies for smart oil field applications," in Intelligent Energy Conference and Exhibition, 25-27 February 2008, Amsterdam, The Netherlands. Society of Petroleum Engineers, 2008.
[25] P. Perry, W. Wise, D. O'Neill, and M. Jensen, "Leveraging a technical domain taxonomy to enhance collaboration, knowledge sharing and operational support," in SPE Digital Energy Conference and Exhibition, 19-21 April 2011, The Woodlands, Texas, USA. Society of Petroleum Engineers, 2011.
[26] J. Zhao, C. Chelmis, V. Sorathia, V. Prasanna, A. Goel, Recovering Linkage Between Seismic Images and Velocity Models, SPE Western North American Regional Meeting, 2012.