XPRIORI

THE XPRIORI REPORT

An Actual Case on Unlocking Value in Information – A Large Petroleum Company Uses of Data Similarity Technology in Large Scale Classification Projects
Part 3

Introduction

As we indicated in Parts 1 and 2, in recent months, a large Petroleum Company has launched a series of projects using new automated document clustering technologies to evaluate and organize these largely digital information assets with a view to creation of value and competitive edge. The Company is undergoing a revamp of its enterprise content management strategy and philosophy at the functional and business asset level. As is the case with many organizations, the primary objectives of this exercise include enterprise-wide cost containment, risk reduction and the extraction of value from content while enabling more effective use and management of the asset.

In Part 1, we discussed the goals and values for the project; the need for new technologies to supplant the slow pace of manual review and the small but expert team that was capable of using the new technologies to meet the goals, values and objectives. In Part 2, we addressed the team's ability to meet the overarching goals for the system of *Accuracy, Findability, Consistency and Governance* at the various stages of creation, use and storage of information at the Company. We also offered a number of keys to understanding the process and deployment of the technology.

In this Part 3, we address the two phase work flow deployed at the Company and provide some details of the step by step process that was followed. In Part 1, we addressed the structure of the team and repeat some of that discussion immediately below.

The Team

The Team. The typical team assembled for a large scale classification project must include business subject matter experts who are familiar with the domain, vertical and content types related to the documents as well as how they are used. At the Petroleum Company, the team included the following:

- Information Technology ("IT") project managers;
- Compliance specialists;
- Electronic Content Management ("ECM") application developer/engineers; and
- Document controllers ("DCs") who have significant familiarity with the type of documents being collected.

The Document Controllers ("DCs") were critical and were all subject matter experts ("SME's"). They include persons familiar with the documents in the dataset In the Petroleum industry, as in the use case, DCs can include subject matter and domain experts in a particular discipline, i.e. energy exploration and production (E&P). In the use case, the document controllers had vast amounts of experience in the documentation and processes related to wells, fields, drilling operations and supporting functions. Their qualifications varied from individuals with strong clerical skills to PhD level geophysicists.

What was actually done?

The work was broken down to *two phases*: **(a)** **Phase One:** an agnostic collecting, culling, denisting and de-duplicating files from targeted file shares within the enterprise with limited regard to their discrete content – and as a result, clustering not yet applied ; and (b) **Phase Two:** content sensitive sorting including application of automated text and visual similarity processing of documents to clusters for analysis and classification to: (i) apply, modify and re-apply a broad and hierarchical classification taxonomy; (ii) identify documents that will not fit the existing taxonomy and modify the taxonomy to include them at an appropriate place; and (iii) create an operational system that will automatically do the same with any newly introduced documents as they are ingested into staged storage.

A. Phase One Detail: Data Collection and Initial Analysis

The Phase One effort was typical, e.g. (a) eliminate system files, (b) exact duplicates, (c) organize files by file type, (d) index them for common search tools, and (e) develop and apply some business rules for higher level classification. At the Petroleum Company, some rules already existed and were the basis for aligning a certain number of documents with the existing taxonomy – low hanging fruit if you will. To the extent possible, those documents could be stored and used depending upon stage of use or creation. In the case of legacy information, the storage is more likely to the “published” documents or, at the Petroleum Company, a Documentum store. There were significant numbers of documents that required further attention after completion of Phase One.

A few words about the Collection Process: at the Petroleum Company, the DCs and their project managers worked to identify potential authoritative data sources. Thus, they interview various people who might be involved as leaders or custodians of data. In this regard, among other efforts, they sought advice from their “Big Data” people, a group with data mining and other similar functions. They helped identify databases, etc. that contain authoritative data. Once the DCs completed clustering/classification/attribution activities described below, they created a register that they used to validate against the data source and to enhance the attributes to the documents that were collected.

Phase 1 activities are illustrated by the following workflow diagram and in the notes to each of the activities below:

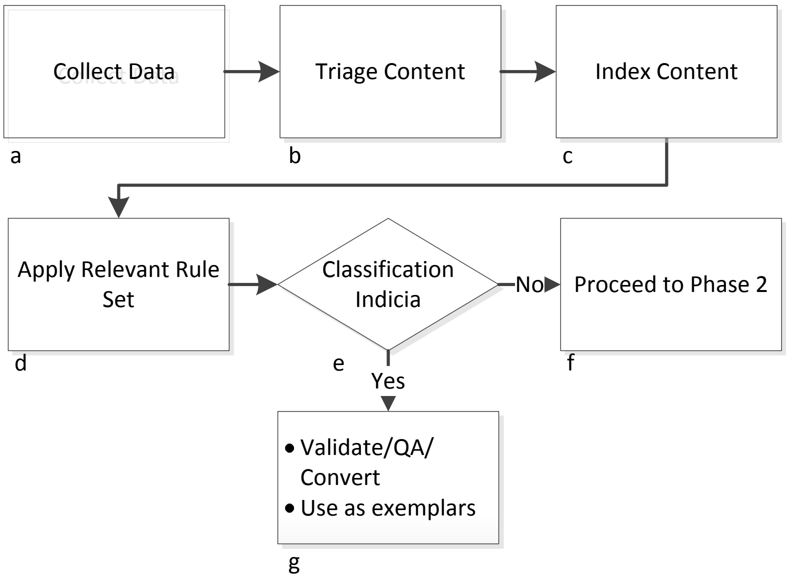


Figure 1

a. **Collect Data** – data residing on business unit file shares is collected in ways that do not alter the file dates or content. Data is collected in a forensically sound fashion, preserving original metadata such as the time and date stamps of the files. Collection included file de-duplication and removal of known superfluous system files.
b. **Triage Content**— the data is then organized and ordered by data type, i.e. Word documents, CAD files and end user software created file types. Xpriori works with the client to prioritize the analysis of the collected data depending on client objectives.
c. **Index Content** – the indexing process makes the files and file properties searchable through common search tools and specialized tools for particular types of data; it also helps identify files that need to be subjected to an OCR process. This process enables text based and pixel based visual clustering. It also facilitates the extraction of attributes, i.e. P.O. numbers, company names and other important content.
d. **Apply Rules to Documents and Classification Indicia** – Business rules are created and then configured for use as basic Boolean and other search rules, categorized consistent with the storage taxonomy, and then applied to the data.
e. **Classification Indicia** - At this stage, there will be some percentage of documents that fall into classification categories and, as such, may not require further classification related activities. In some instances these documents will be ready for ingestion in an ECM system. There will also be some percentage of documents that need further processing in order to be indexed and classified. For those documents that have been classified they can then go through the validation process (block g in Fig 1.) and be used exemplars going forward.

Subsequent to the application of rules in item "d" above, we typically have a sense of how much content can be immediately classified without further processing – the identification of low hanging fruit. The most common form of a rule is expressed as a Boolean keyword search. For example if an analyst wanted to find all well logs and related closeout reports, they would simply construct a query similar to the following: (well W/2 log) AND closeout W/2 report. This search example would find all documents that contain the word "well" within 2 words of "log" and "closeout" within 2 words of "report".

f-g. Validation and QA. Once the results are returned, they are validated then subjected to a comprehensive QA process which includes:

- **Visual confirmation by subject matter experts**.
- **Keyword testing.**

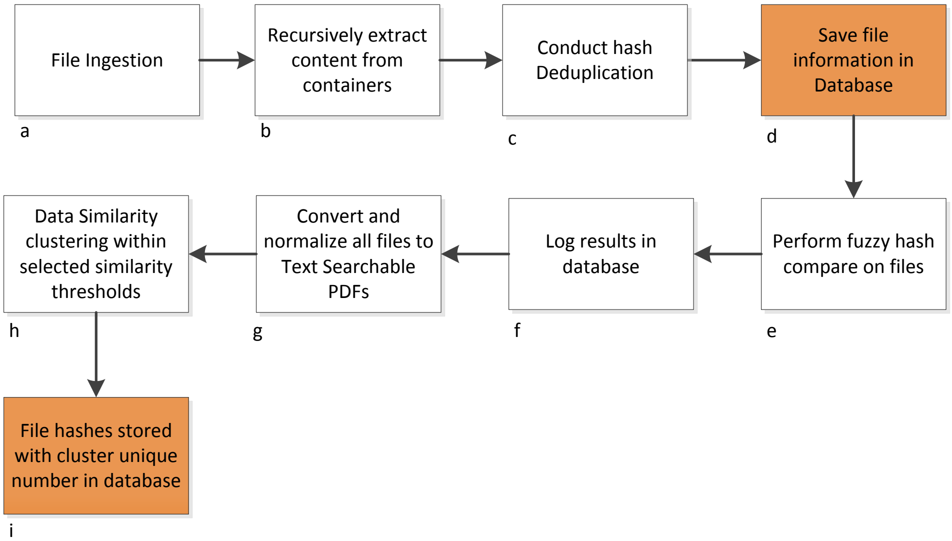
Conversion of unclassified files to normalized PDF format. To further analyze the content, Phase 2 processing requires that the files be converted to a normalized PDF format. Once this is done, the files are then re-indexed and subjected to the processes described in ”d” and “e” above, and then passed to Phase 2 processing.

B. Phase Two Detail: Recursive Extraction of Documents from Container Files and Deployment of Data Similarity Analysis For Clustering

The remaining documents typically require at least two types of transformation to be susceptible to the automated processing for clustering. Many of the documents not processed in Phase I will be in container files – ZIP, PST, TAR files etc. – and have to be removed from the containers. Once the removal is completed: (a) hash based deduplication procedures are applied; and (b) all files that remain are converted to a normalized PDF format.

Application of hash based deduplication further reduces duplicative content and identifies unique files that need to be extracted from a parent object in order to determine whether they fall into a particular classification category. This conversion enables the clustering algorithms to operate across all of the documents consistently. Please note that over time, it is common for text base files to have been converted to other formats really creating duplicates that will not be identified as such prior to application of this process. Visual similarity analysis also enables the identification of non-textual symbols as part of the process. Original files are maintained and the converted files are always available.

The following diagram, Figure 2, and explanatory notes illustrate both processes.


File Ingestion
a
Recursively extract content from containers
b
Conduct hash Deduplication
c
Save file information in Database
d
Perform fuzzy hash compare on files
e
Log results in database
f
Convert and normalize all files to Text Searchable PDFs
g
Data Similarity clustering within selected similarity thresholds
h
File hashes stored with cluster unique number in database
i

a. **Files Ingestion** – files that come out of Phase 1 block “f” are pulled into a tool that recursively extracts files from container objects.

b. **Recursive Extraction** – this process targets compressed files such as ZIP, RAR, TAR, PST’s and embedded objects in files;

   - ZIP, RAR and TAR – these are known compressed file types that may contain other compressed files types or objects that need to be opened before they can be fully de-duplicated against the rest of the data collection.

- PST's, LOTUS NOTES, etc. – these file types are containers that have other objects that should be extracted for comparison against the overall data set.

c. **Hash de-duplication** – similar to the process used in Phase 1.
d. **Save information in the project database – all hash values of files included the absolute and relative location of the files is stored** in a database for audit and chain of custody purposes.
e. **Fuzzy hash compares** – this process compares certain metadata attributes of files to assist in the clustering process.
f. Save the information in the project database
g. **All files are converted to a normalized format, PDF**, such that the text clustering algorithms can be run across the entire corpus of collected, recursively extracted and text enriched content. Automated OCR ("Optical Character Recognition") procedures are applied where the state of the information requires.
h. **Data Similarity Clustering** is applied; visual similarity will disclose same documents appearing as multiple file types – a word file that has been converted to pdf with both files still in the collection; this enables **further elimination of duplicates; also visual**
i. Hash, metadata values, and other pertinent information are **stored to a project database**.
j. Attributes can be extracted and associated with particular documents as tags or metadata; more on this in Part 4 of this series.

Both text and visual similarity clustering operate based upon a percentage of similarity that can be adjusted by the DCs. The DCs manually do a compare of results at varying percentages and obtain a consistent level of comfort in the result. The process enables the deployment to human judgment at the right point – the point at which similar documents suggesting common attributes have been identified, with duplicates and system files culled.

The system presents clusters for the DC to associate with the existing taxonomy and/or to render new orders of classification; and, to discover duplicates occurring where the same document appears in multiple file types.

The speed of process is further enhanced by the preservation and continued application of the coding that supported the creation of any cluster. At the discretion of the user, the coding can be applied to all new information introduced to a project or beyond. New documents are automatically aggregated to existing clusters. This speeds by a large factor the management of large projects or continuing introduction of new documents to a dataset.

At the Petroleum Company, as mentioned above, sometimes the clusters presented required further analysis and subdivision to achieve collections with sufficient homogeneity to be associated as a group with the taxonomy. There is really a "crossruff" between deployment of the visual/textual similarity tools and other tools such as Boolean search, keyword search, and file metadata to identify content. In these circumstances, the later referenced tools are applied to clusters that are typically much smaller sets of documents. The tools operate far more effectively on the smaller sets.

Finally, in Phase Two, the DCs apply descriptive tags to the documents within a cluster or set of clusters. These tags function to associate the documents to the taxonomy and to identify any other common attributes associated with them. For example, common attributes such as dates, authors etc., and are pulled forward from the document metadata. Metadata is expressed as tags as well. This process is called "attributing" converting the noun to a verb to identify the process. Corporate policy is reflected in the substance of the tags used. Part 4 of the Case Study will provide more information on attributing as part of this case study. At the Petroleum Company, their result was significant improvements to the taxonomies as well as new ways of looking at existing information.

With the foregoing processes completed, the documents are stored consistent with their phase of use and creation.

Conclusion: Three Major Value Propositions Are Presented

While this Part 3 discussion deals primarily with handling legacy information at business units of the petroleum company, there are a number of fundamental value propositions that arise. We mention them briefly here but will develop use cases around them in future editions of the Xpriori Report.

Value Proposition 1 - Data content informs our knowledge of our environment

As illustrated in the foregoing, clustering algorithms now help us understand our data categories in ways heretofore unachievable when working with big data. Clusters can be created without human definition for review and culling or accepting. Much manual effort is avoided and all content is considered. Prior to having the ability to use algorithms to help vast quantities of data self-contextualize itself to a user, people would come at their data from pre-defined point. This pre-defined view of data is grounded in presumptions about the data which by its very nature creates a “data horizon” (data risk or value that is not “visible”, addressable or otherwise readily usable by an organizational stakeholder lives below their data horizon). Having data describe itself to the user allows the user to see data in its complete context. Data blind spots for which there is no or insufficient classification are revealed in terms of relevance to other known data objects or documents within the corpus of content examined.

Value Proposition 2 – Taming big data

What is big data? Big data is characterized by an acronym describing 3 key variables which has been coined as V^3

- Data Volume – large data volumes. This is a relative term that changes based on innovations in storage device areal density (the number of bits that can be stored in a given area on a device).
- Data Variety – the kinds of data, i.e. unstructured, structured, semi-structured and newer polymorphic content formats.
- Data Velocity – the rate at which content is created.

The ability to identify substantively similar and duplicative documents gives organizations the ability to select “the” business records that should be kept in an ECM archive. The impact on storage budgets can be significant. The amount of storage needed for an organization can now be projected with pinpoint accuracy. The key metrics that allow us to do this can be generated from metadata; storage growth year over year and document duplicates based on clustering.

Value Proposition 3 – Sustainable, objective and automated data classification based on iterative clustering and informed and automated modification to the storage taxonomy

The greater the corpus of information that is clustered, the more we know about an organization’s document and content types. The process is automated and provides objective review of all content. Each ratified (classified) grouping of document types within an organization becomes the classification exemplar for new documents that enter that particular managed storage environment. New information added to a collection will be subjected to the same code for classification and organization. The ROI comes from dealing effectively with the large increases in unstructured information experienced by all organizations; from development and validation of the storage taxonomies based upon similarity of content; from automating classification of information from external sources such as new entities acquired through M & A; and more.