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 2

1. INTRODUCTION

As we indicated in our first installment, 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 address 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. The Company outlined objectives for each stage, and the team deployed the new technologies to meet those objectives at each stage.

2. STRUCTURAL CONSIDERATIONS: Location and Various Requirements of Documents during Content Life Cycle

The Company maintains its documents/content at different places in the corporate network environment depending on kinds of information and their association with stage of creation, use and retention – the :

Stage	Location
(1) Early Stage Work in Progress Documents	(1) Unmanaged Global File Shares
(2) Department Documents	(2) Managed File Shares
(3) Project Documents	(3) Managed Share Point
(4) "Published" Documents.	(4) Documentum or Stored Documents

The Company identified four areas of storage, use and retrieval, coupled with standard attribution and functionality, to be used at progressive stages of the data classification process. The

overarching goals of managing collaborative SharePoint site, classifying and managed and unmanaged file share unstructured content and ultimately publishing managed documents and metadata were achieved by the Company's ascribing and meeting the following functional and qualitative protocols and objectives for file shares associated with each stage: (1) Future Policy, the applicable time during which the policies are effective; (2) Permission Management, the applicable enterprise rules for storage and use; (3) Structure and persons who might have access and use; and (4) Types of files covered together with various characteristics either required or available. The application of these criteria to the four stages and locations outlined in the document/content life cycle is outlined immediately below.

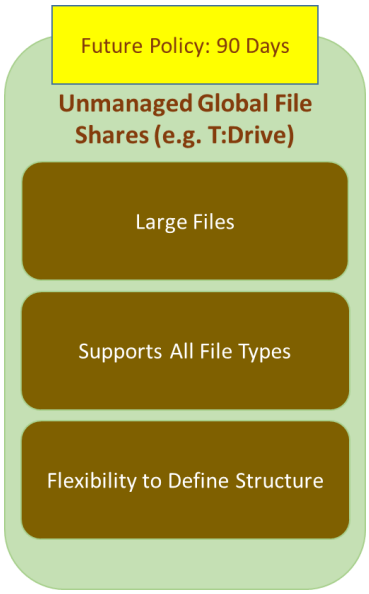

Figure 1

- Permissions Flexibility
- User Based Storage Provisioning

Stage 1 – Early Stage Work-in Process -- Unmanaged Global Files Shares (e.g. T:\ drive)

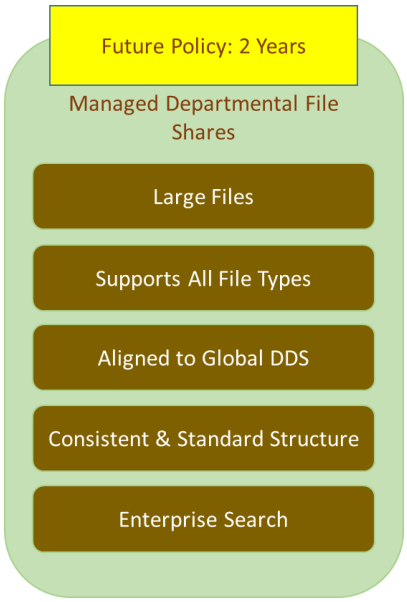

Figure 2

- Permissions Flexibility
- Nominal Required Training

Stage 2 –Department Documents –Managed Department File Shares

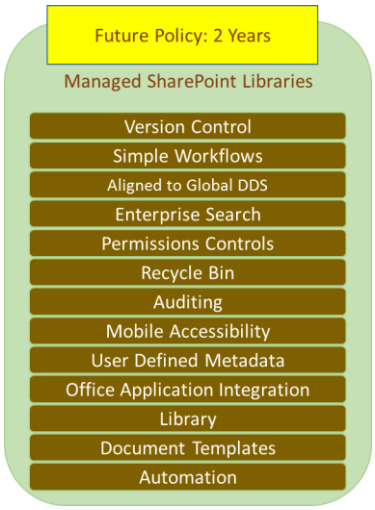

Figure 3

- Minimal Required Attribution
- Specialized Training
- Built in Governance

Stage 3 – Project Documents – Managed SharePoint Libraries

Figure 4

- Up to "n" Required Attributes
- Specialized Training Required
- Named Document Controller(s) per Asset or Function

Stage 4 – Published Documents – Documents Stored to Documentum According to Taxonomy

3. GENERAL CONSIDERATIONS – MIGRATION OF DOCUMENTS

To meet the Company's goal, significant amounts of legacy information has had to be moved through various processes to have it reside ultimately in "published" storage and available for enterprise wide use and collaboration. The Company also has deployed a centralized stack of Electronic Content Management (ECM") tools or applications to help users find what they need. In the use case in question, the underlying data set was comprised of documents related to the field operations and management associated with oil and gas exploration and production. The information, by and large unstructured in nature, contained mostly large quantities of oil logs, maps, cad drawings, and other documentation largely non-textual nature. The initiative involved the collection, review, remediation and storage to Documentum of large amounts of this unclassified information – approximately 1.9 million documents -- that existed in various file shares on the Company servers. Initially, the Company worked through a pilot on approximately 100,000 taken from a couple of file shares that contained part of the 1.9 million.

4. THE RESULTS.

The documents processed have yielded real value. In fact, initial assessment on documents classified to particular aggregations of acreage has demonstrated positive results on *finding new drilling opportunities from information that is, in part, decades old.* Also, 90% of information subjected to the process has been culled and culled defensibly in compliance with legal process.

There were many superfluous documents. The Company has found that clustering leads to aggregations of information more or less homogeneous in nature. As a result, search technologies available to users are working better and yielding better results. Even sophisticated tools such as "concept search" and "predictive coding", often deployed in eDiscovery projects, could now be applied against smaller and more homogenous datasets with a common content based organizational presence. This has meant quicker response time, elimination of false positives in the search returns and quicker understanding of the context in which the information was created and used. To repeat, initial assessment on documents classified to particular aggregations of acreage has demonstrated positive results *on finding new drilling opportunities from information that is in part decades old and that's Real Value*.

The program continues by developing new projects covering information being used or retained by other aspects of the business or other business units or functions.

5. KEYS TO UNDERSTANDING

There are several keys to understanding this process:

- The clustering process is automated and suggests clusters of documents that are either visually similar or textually similar based upon various algorithms;

- The results of the processed are tuned to a percentage of similarity that produces results that you can review and accept or decline.
- The process is to start small – that is to start with a smaller batch from the larger collection that you want to assess and classify, and the process will apply what it has learned to subsequent batches as processed;

- The clustering will suggest changes to your taxonomy;

- Particularly where you have documents not stored according to any rules, you can expect a significant cull of superfluous information;

- The process should be engaged with a smaller team –in the current case about 10 – but the team should be made up with subject matter experts, some technology support and people somewhat familiar with the business source of the documents;

- The process groups the documents to the particular contexts identified the team, making the information more useful and findable;

- The process should be governed by an overarching set of principles and goals; in this case, *Accuracy, Findability, Consistency and Governance;* and should be applied to the various stages of the content lifecycle of the organization.

- The process is best applied to discrete business units and/or its processes and workflows;

- What is learned at one stage is always carried forward to the next without the need to redo or recreate classification code; what has been created is carried forward.