THE XPRIORI REPORT

In EDiscovery and Data Classification: Why should you work with Xpriori?

Human beings, we organize things according to their sameness or similarity. In our solutions, we deploy the same concepts digitally in looking at larger aggregations of documents and automate the clustering of documents according to their data similarity – often a million documents sort to a few thousand clusters. For data classification, the result is: **HUGE LABOR SAVINGS, EXTRAORDINARY ACCURACY, AND MORE COMPLETE VIEWS OF INFORMATION.**

What if you had the ability to automatically classify documents by data similarity for both electronic and scanned paper documents and your organization could see all of the document types , (as opposed to file types) that exist in your storage and other repositories?

This means that you can setup rules and retention policies based on the document type and automatically apply them to all documents of that type throughout your enterprise - all done from your normal Line-Of-Business (LOB) application (Documentum, OpenText, Sharepoint, Alfresco, OnBase, etc).

We don't get "confused" by file types. It doesn't matter whether a document originated from Word, PDF or a scanned paper copy. We group documents based upon their data similarity taking into consideration graphical, textual and zonal considerations.

On an automated basis without much by way of human intervention, this has enabled companies to turn unstructured documents into groups of structured document types such as: invoices, sales agreements, service agreements, etc. At the same time, these companies have been able to significantly enhance their storage taxonomies based on cluster content – in one case virtually doubling the number of classes in the taxonomy assuring easier access to enterprise information.

But there is more…

- Deduplication procedures are not restricted to discrete file types. So often people have taken a significant document and converted from a word file, to a pdf, to a jpeg etc. – we identify duplicates and near duplicates across file types;

- Retention/reduction/remediation decisions about millions or billions of documents by simply looking at the first few and last few exemplar documents in each document type cluster. You know that the decision being made is no better and no worse that the examples that are reviewed. Hundreds or thousands of groups can be quickly and accurately reviewed that effect millions or billions of data similar documents universally;

- Comprehensive document retention, reduction and remediation policies can be created - from a position of awareness - rather than interviews with stakeholders who can't possibly remember everything that they have, much less what their predecessors passed on;

- You can always use traditional approaches to "sub-divide" a cluster for finer tuning. Our methodology takes deduplication and parent / child into account as well, making it an order of magnitude more comprehensive than any current method; and …

  There is still more….

- Code automatically generated to create the clusters is preserved and can be subsequently applied to newly introduced information – enabling implementation of a mostly automated self-classification system for the new information;

- For example, use this approach to ease the pain of preservation and continued vigilance over information required to be found and tracked in a litigation hold; and

  Finally, you have an approach where:

- Your data content itself informs your knowledge of your data environment – no guesses, real information and knowledge derived from your own content;

- You can approach resolution of many of the problems of big data – data volume, data variety and data velocity or extraordinary creation rate – with largely automated approaches that assure good decisions of what to retain, that assure continued enhancement of the taxonomy to support the storage and that assure accuracy and confidence in the timely retrieval of the right information for stakeholders;

- You can move to implementing a cost effective, sustainable, objective and automated program of data classification based upon iterative clustering and adjustments to storage taxonomies or requirements of a specialized repository, such as for litigation, based upon the content of the information itself.

  Give us a call to discuss your needs. No doubt our solutions based approach can resolve most of them.