



**PointCross Solutions**

[www.pointcross.com](http://www.pointcross.com)

[pointcross.wordpress.com](http://pointcross.wordpress.com)

[get\\_started@pointcross.com](mailto:get_started@pointcross.com)

# CONTEXTUALIZING LEGACY INFORMATION IN AN PHARMACEUTICAL ENTERPRISE

## THE ISSUE OF LEGACY INFORMATION

One of the topics of discussion during our recent meetings with customers is the question of how metadata of very large quantities (in the billions of files) of legacy documents, regulatory data stores, and files in the company's shared file stores as well as managed content systems can be properly categorized, their metadata extracted, and all content related to specific business topic related so that they can support scientific search, cross-study meta-analysis, and meet other important needs of pharmaceutical companies such as legal discovery, knowledge management, key business processes and regulatory compliance. Big Pharma is especially concerned not about only legacy data that resides internally, but sourced from the companies with which they merge or may acquire. This is not a trivial problem. Many companies have been known to expend enormous amounts of time and money hiring service companies to "clean-up" the data so that everyone can start with a fresh slate. This is not practical answer – data is constantly changing, external people do not have the institutional knowledge, and the clean up must be a self-sustainable process that will continue into the future.

This write-up is intended to share some of our, PointCross's, thoughts on the subject and our suggestions on how such contextualization of information can be done with consistency, scalability and sustainable results. We contemplate the use of a set of tools including technical search tools and data handling tools, as well as a process that pushes legacy information into contextually relevant structures while business people pull information into their business contexts and SharePoint Sites.

## IF METADATA IS IMPORTANT – CONTEXTUALIZATION IS ESSENTIAL

This section makes the case for relating content placed in passive, or managed, storage systems to business topics around which people work, communicate, and make decisions. In later sections we will discuss how this process can be set up and executed.

When we talk about contextualization, we mean the process of uniting disparate information types that share a common business purpose for their existence, or that may have served a common business topic at one time, under a "context" that represents that purpose or topic. The "context" is one of the foundational concepts underlying our Orchestra™ ontology engine due to its simplicity, flexibility, and universality. We use "context" in the same sense as the familiar English definition: "the conditions and circumstances that are relevant to an event, fact, etc.," where 'conditions' include the business rules, roles, and people associated with the event or fact; and where circumstances



includes content, data, decisions and any intellectual effort that relates to the purpose of the event, fact, or entity. In business we see “contexts” as giving people a shared purpose or goal; a shared space where they also share data and information and engage in intellectual discussion, tacit communications, and contemplation in aid of making critical or strategic business decisions; where they make decisions that are made aware to all who share this context; and where all people who share a context have common situational awareness. People who work in a context may have disparate roles, responsibilities, and disciplines. They only come together when they have a shared purpose.

In Orchestra, we treat contexts as virtual objects that represent real world business topics, events, entities, studies, assets, projects, deals, complex co-authored documents, laboratory/manufacturing facilities and their components, and constructed scenarios – for that matter anything where business decisions will be made based on a combination of data, information, tacit communications of insights and judgment. Therefore these contexts are fertile virtual spots around which rich metadata can be clustered; new metadata is constantly added, new relationships are formed between the context and other contexts; between the content within a context and other contexts; and between people who occupy roles that relate to a context and other people, contexts, and content. Contexts are intimately related to the core business and everything including data and content (including intra-web, documents and files); communications, decisions and recorded contemplation by people are captured within these contexts. Contexts are structured into hierarchical structures called “Trees”; and trees can be grouped to form “Projects” in the lingo of Orchestra. Groups of trees can also be collected within “Folders” that can then be grouped under a “Project.” Projects are grouped into a single “Business Base” that represents a self-governing business entity within whose boundary their information is kept secure, separate and sacrosanct from an IT perspective (no data tables are shared across “Business Bases”).

Contexts can be related to elements in domain trees, or taxonomies. An unlimited number of these taxonomies may be maintained – both local folksonomies as well as controlled taxonomies. These are used for attribution, matching extracted words and metatags to create keywords, and to create tacit relationships based on people communicating (emails, blogs and meetings). Links between taxonomy elements and contexts, their content, or data sources form virtual scaffoldings around the work products of people (content, communications) for users, search engines, and process analytics to navigate and locate people, information, or the next steps of their workflows. Contexts have attributes and properties – metadata that describe them – partly defined by the taxonomy elements they themselves are linked to, and partly by inferring the attributes of the content within them; all of them continually grow with continued activity within these contexts. Multiple layers of metadata can be built around the content within these contexts. A conceptual list of the layers of metadata currently contemplated is shown in the schematic below.

Source Types	Header	Extracted/Added	Usage	Versioning	Relationships
Documents	Name, Created by, Data/Time Stamps, Size, Type, Last Changed By, Source, Location, ...	Metatags (Proposed/Approved/Matched), Attributes (Security, Records, eDiscovery), Matches with Taxonomy elements, Public/Private	Updated by/for, Fetched, Sent to, Made Available to (Blind Search Workflow)	Revision History & Header/Extracted & Added/Usage	Context, Taxonomy (-ies), Element(s), Linked/Move/Copy log data, Location of Duplicates, Documents linked to the same metatags, People, Roles, ...
Emails, Meetings, Blogs	Who (from, to, cc, bcc, reply to), Subject	Extracted metadata from email body	Viewed by, Read By, Added to Thread and by, ...	Thread location, Original/Reply/Forward	Context, Taxonomy (-ies), Element(s), Linked/Move/Copy log data, Location of Duplicates, Documents linked to the same metatags, People, Roles, ...
Structured content and data stores	File Name/Data Source and System, Type, Size, Who, Source Location	Attributes, Type, Data File Name (if applicable), Tags (if available), Parametric Data Analyzed & Indexed, Links to related content	Last Updated By, Time/Stamp, Context accessed by, ...	Change Tracker	Context, Taxonomies, Tags (if available) OR data points to linked attributes, People, Roles, Content



In a contextualized ontology these multiple layers of metadata can be built and self-sustainably maintained as people go about their work. Each layer of metadata reflects higher levels of understanding about the content they describe. The ontology lets users and software applications that use these disparate layers of metadata to navigate through information networks, networks of business topics and processes, geo-spatial relationships, or social networks. Absent such a context based ontology there is no mechanism for building beyond the basic explicit header information that a file might yield, or that a search engine's indexer might extract. The heavy lifting beyond that will need to be done in a contextualized environment where many layers of relationships can be expressed.

Traditional file storage systems where legacy information is stored carry the basic header type of metadata. Using text extractors, indexers and search engines it is possible to get to the next level of "extracted" metadata including metatags. Managed storage devices such as DMS and CMS could possibly add new metadata that reflect changes and versioning of the files, but not any more levels of metatags. This is the reason that DMS, CMS, and shared drive facilities have not contributed to the knowledge value of the work products in companies. There is little or no direct tie-in between the workflows or collaboration that scientists, business people, and regulatory personnel go through and the way their touch points on files and content are registered as metadata. Therefore in traditional file storage systems, the metadata remains relatively static and it does not grow with the work done in business.

#### **CONTEXTUALIZATION - ISSUES TO KEEP IN MIND**

The goal is to be able to put legacy data and information into sufficiently granular business contexts that are useful at the enterprise level, divisional level, project/study level, or even for a small collaborating team.

We look at the existing legacy content as "supply" (or push) and the context that they will ultimately be related to as "demand" (or pull). Contextualization of Legacy Content is all about flowing content from supply to meet demand. Processes or people who define contexts cause each new entity (projects, deals, studies, etc.) they mature to create one or many contexts. Each of these contexts seeks legacy content related to them, or important to them. This creates pull, or demand, for content so that they may be available to the content.

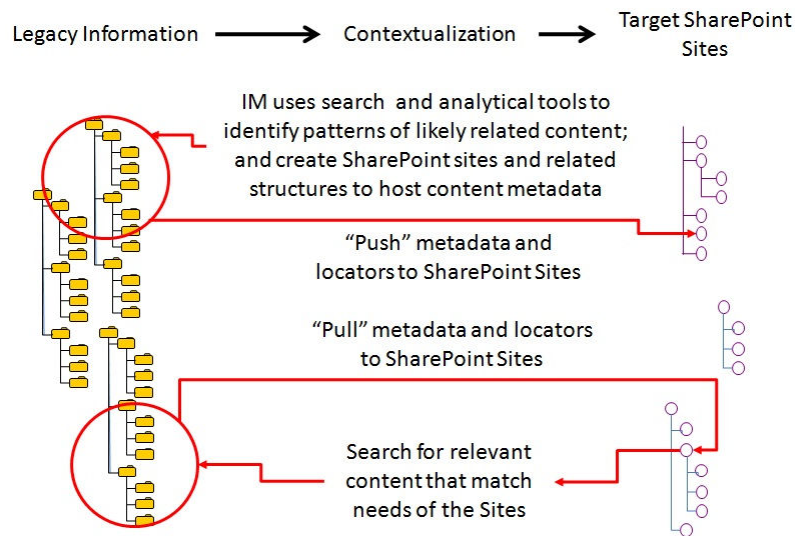
Information Management (IM) at the enterprise or functional level will create a "push" for legacy content found in the current stores and then relate them to destination contexts decided by IM as being of importance to their constituencies.

Both of these contextualization forces can, and should, act concurrently. They should also be able to work autonomously to meet their specific business drivers. Think of it as two sets of processes that are continually contextualizing the legacy data and information. This is also entirely consistent with the real world situation because new content may be added or modified as work progresses; and the crawlers, combined with the push and pull processes, ensure that the cleansing and contextualization activities are continual.

The idea of doing a single mega-effort of clean-sweeping all legacy information is a surefire formula for failure. We have seen many initiatives where Pharma companies attempt to rationalize and catalog their vast quantities of study data stores. It provides excellent job opportunities to third party service providers who charge by the hour. However, neither do these service providers bring institutional knowledge about the company's data, nor do they know the organization

well enough to use any social networks to resolve conflicts or better understand the genesis of some data, but most importantly they do not come armed with any automation tools that can help by crawling, applying analytics, or providing cues to the classifiers about what ought to relate to what. And, even if they did have some tools, they wouldn't have the ontology facilities to establish, build, and maintain those relationships in a self-sustaining way.

## Pushing & Pulling Content Metadata into SharePoint Sites

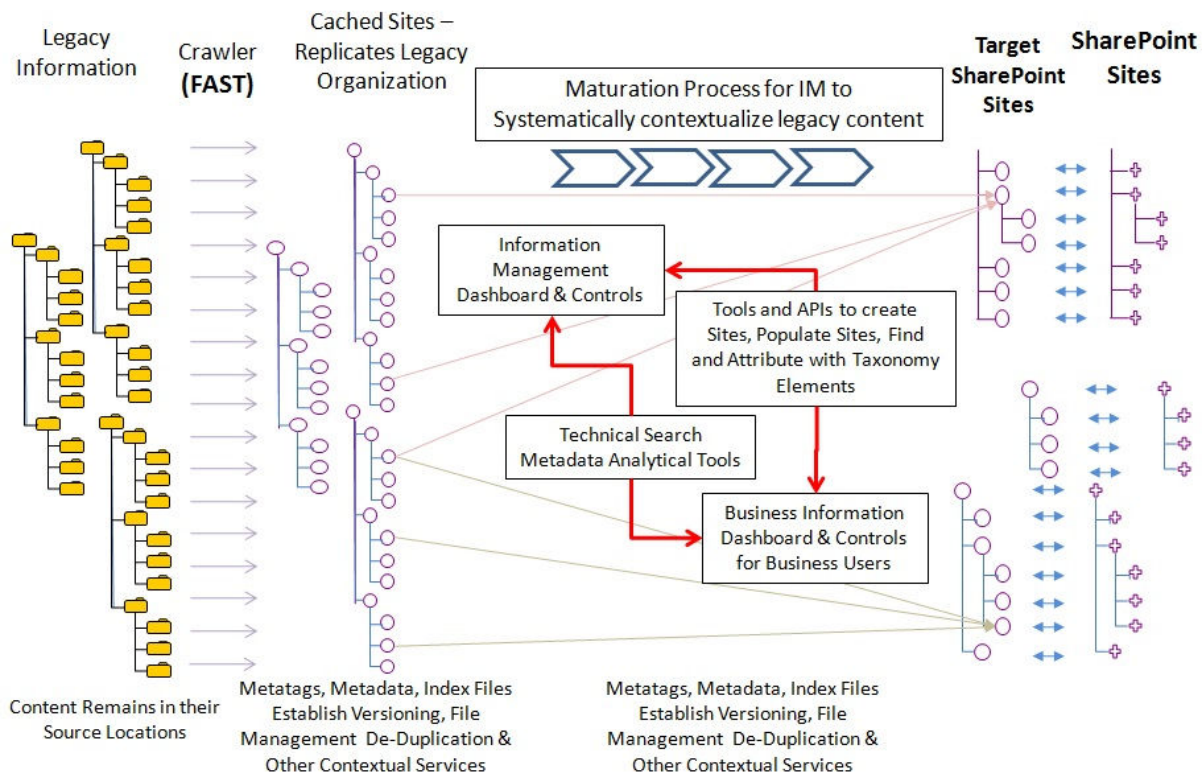


### PROCESS INFRASTRUCTURE FOR CONTEXTUALIZING LEGACY CONTENT

Two processes should be enabled to create a sustainable way of contextualizing legacy and contemporaneous content that gets added to shared-drives and managed document and content management systems. The legacy information is shown as a collection of folders on the left side. The target contexts which Information Managers create for the various business units; and the contextual structures that individual teams and departments might create are shown on the right (titled “Target Contexts”) of this diagram. Note that a set of mirrored SharePoint Sites are shown on the extreme right side with a bi-directional arrow between the contexts and the “Sites” to represent the auto-synchronization (the Gluon<sup>+1</sup> adapter) between Orchestra and SharePoint. This is for a later discussion about SharePoint and how it gets incorporated into this contextualized metadata enriched environment.

The legacy information on the left may contain billions of files in millions of folders all over the corporate network. Shown right next to it is the “Crawler”. The crawler consists of automated software run process of scanning folder paths assigned to it. The software will need to run in a set of parallel threads in multiple servers dedicated to this task temporarily.

# Process for Contextualizing Legacy Content into SharePoint Sites



The crawler performs a number of tasks:

1. For every folder in the legacy area, create a corresponding Context in an Orchestra “tree” which represents the path of the folder;
2. Extract basic header information that is associated with files – see the first metadata level in the table above;
3. Extract text from the files using the appropriate filter for the files. The crawler will place the extracted text into a file within the Orchestra’ File Management System (FMS). Text is usually less than 5% to 10% of the total file size so this extra storage should not be a large penalty. It is possible to avoid saving this raw text; but at loss of some convenience;
4. Create a linked image of the document or file in the Orchestra context under its Document Management Module. Set the Version of this document as 1.0. From now on the crawler will keep track of the legacy folder and this specific file and it will reflect changes as versions. Of course, in unmanaged DMS such as Explorer files, there is no trace left of the old files when over-written and so while Orchestra can keep track of changes it cannot provide access to the older, over written, documents. However, the FMS will contain the raw text of the older version and it can act as a reliable cache;
5. Absorb all the available metadata from the header and the hash fingerprint of the file into the Orchestra document header metadata set;
6. Generate a hash (MD5) for each file and register these with Orchestra. The hash is a unique code generated for a file. It allows Orchestra to recognize duplicate files found in other paths or folders. It is also used to confirm if a specific file in a folder has actually changed since the last crawl;
7. Use the technical search indexer to index the extracted text files and generate proposed metatags. These are words found in the text such as unique formations of alpha-numeric text, nouns, proper names, abbreviations

and others that are not part of an explicitly set stop-word list (a list of standard words like prepositions and verbs that are not to be indexed – something that will be done in consultation with customer). These proposed metatags will be associated with the document in Orchestra (actually the link to the document);

8. Automatically compare the extracted metatags from the document files and compare them with the elements of those taxonomies registered for the purpose of validated metatags. The Orchestra environment can have an unlimited number of taxonomies both local (folksonomies) as well as controlled taxonomies (such as the NCI's Enterprise Vocabulary Services). Each of these elements can also have synonyms, abbreviations, and AKAs;
9. Periodically scan paths that have been crawled and look for changes in the files.

The Orchestra contexts where the legacy content is held can have folders exactly mirrored as shown in the diagram; or it is possible to decide that all folders up to, for example, level 3 from the root of the shared drives be replicated as contexts while the remaining folders below these folders will be replicated as folders within the contexts. See the diagram below for these two approaches. Both are supported. If there are likely to be a lot of frivolous folders created by users in the legacy system it might make more sense to treat them as folders. This is not a critical decision but the flexibility does exist.

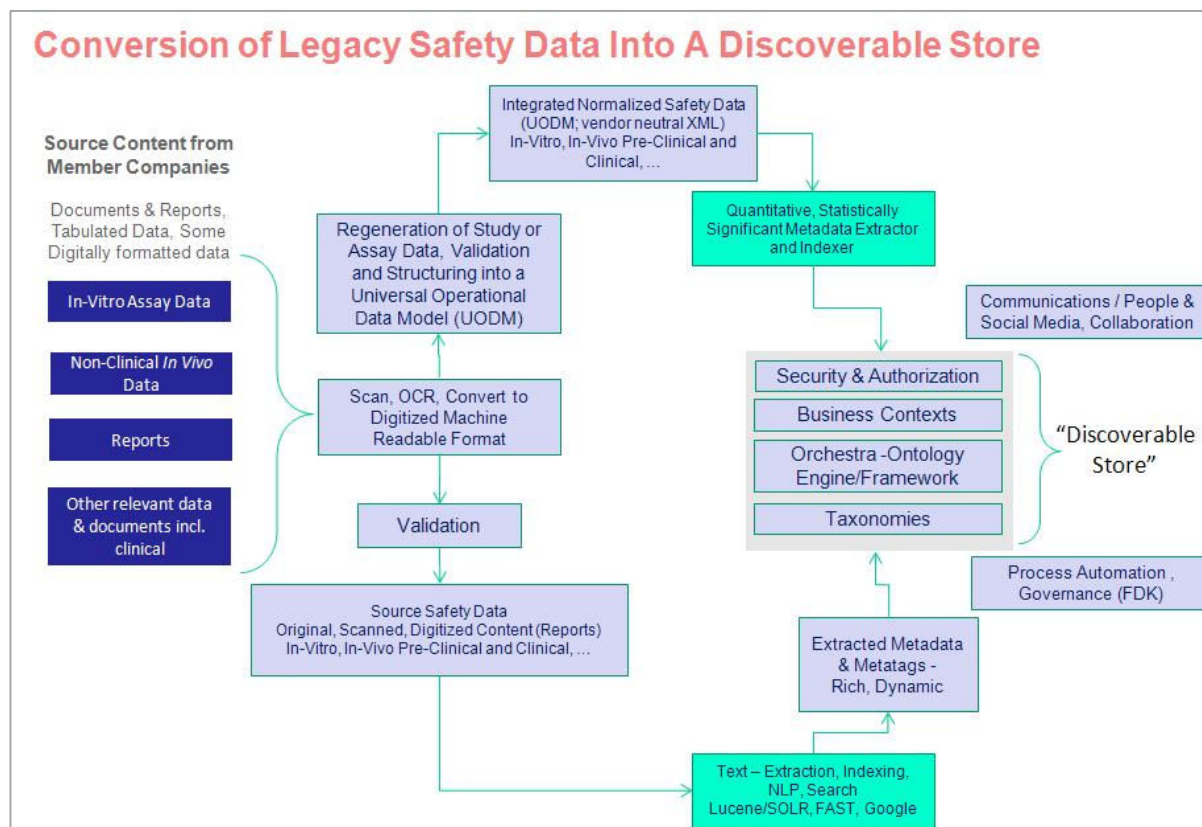
It is also possible to have users' personal C: drives or other folder paths included as legacy data to be contextualized. Usually people prefer to use their "personal space" for such contextualization.

Initial crawling will take significant time. Depending on the number of servers available for this initial crawling this could take many weeks and with the validation and verification that goes with it, the elapsed time could exceed a few months.

## **CONTEXTUALIZING LEGACY PAPER CONTENT**

Pharma companies have considerable volumes of tabulated data and content in legacy paper format. There is increasing interest in adding these resources to internal searchable data stores; in some cases, the need to share such data with partners or consortia is another business driver. The challenge is not to simply scan/OCR and digitize paper content, but to both normalize any tabulated data within paper reports into a Universal Operational Data Model (UODM) for parametric search, and extract text and relationships from the digitized content to support free text search. We discuss some of the related challenges with normalizing existing data from electronic formats in a related paper on Semantic Data Exchange; the same infrastructure along with controls for verification and validation can also be leveraged for transforming paper content into valuable assets. The schematic below summarizes this process:

## Conversion of Legacy Safety Data Into A Discoverable Store



## INFORMATION MANAGEMENT DASHBOARDS

IM dashboards are intended for IM staff to:

1. Monitor the crawling status;
2. Change the priority of crawled paths and URLs;
3. Use technical search tools to find clusters of content that are likely to be related to a single topic, or context:
  - a. Analyzing the basic header metadata and source location data to look for clustering around time periods;
  - b. Analyzing content for correlations such as references to compounds, studies, submissions, organizations/sites or other attributes;
  - c. Looking for correlation of people names (as authors, or modified by);
  - d. Looking for correlation and common metatags (extracted) ;
  - e. Identify likely contexts and clusters of contexts, and be able to create target context names;
  - f. Reconcile potential gaps, conflicts, or discrepancies in source information
4. Create contexts and establish mapping between the source content (legacy) and the target contexts;
5. Designate roles and levels of authorization to those roles, and assign them to the contexts;
6. Assign people to roles within the contexts; to assist with this, people can be found using the mapping of the contact data base within Orchestra which is synchronized with Active Directory or other LDAP systems;
7. Perform other activities that can be configured using the web services and APIs of Orchestra

Keep in mind that many documents from many cache contexts (i.e. legacy folders) can be mapped to a single target context. It is also possible for a single content file to be mapped in this manner to many target contexts. It is also

possible that new contexts might map already-mapped documents into them. Therefore, this is a many-to-many mapping between the legacy structures to the target contexts.

The IM dashboard is driven by a cache of data from the analysis and decisions being made. It holds all the mapping that has been asserted including the "pulled" mapping that business users might assert from their contexts. There is no need for a special database to hold this mapping; the mechanism for holding the links will manage this data store.

Statistics of the links and their currency can be used to identify reference documents of value. Over time these dashboards can also provide KPIs and value content based on their usage and reference.

It's a good idea to establish a process for governing the continued migration and contextualization because it creates a base from which compliance and information security can be managed. The process and dashboards capture requests from various quarters - people, departments, librarians and the processes make it possible to handle them consistently and in a timely manner.

How IM should be organized and governed is entirely up to each company. These concepts have just been presented as a way to push legacy content so that they can yield better insights and relevant content may be found using tacit search attempts.

## **PULLING LEGACY CONTENT TOWARDS CONTEXTS**

---

People or teams of people decide to organize their way of working, communicating, and organizing their work products and content. They begin with knowledge about their contexts. But they may not have all the content; in contrast the IM staff has content but not the contexts. These business users can use their dashboards to create their contexts, staff them, and establish links to taxonomies and provide attributes that result in their definition. They can also create these contexts manually however but that is tedious.

Another common tendency is that people or teams will define their contexts – usually manually – and then simply start using them by uploading documents and writing emails, etc. This means that in a short while these contexts yield considerable metadata and attributes in the form of extracted metatags from the content within. When applied against, and matched with, taxonomies these metatags can be formalized as matches are found. Now – once these matched metatags are identified (words extracted from the documents that match with elements of registered taxonomies), two things happen:

1. It is possible to find other content that is linked to those matched taxonomy elements. This means that if a significant number of metatags in a document show they have a number of linked documents (common metatags) and they are in another target context – as opposed to a cached context – these contexts are worth looking at by the users as they may already be relevant to their work; and
2. If the content within the context point to documents in the cached contexts (the contexts that mirror the legacy explorer or managed systems), then the content is likely worth consideration for inclusion in these recently formed contexts. In other words, people are creating a pull for content within the legacy stores into their contexts.

Changes caused by such “pull” will be visible to IM from their dashboard. In summary then, both pull and push activities of contextualizing will be assisted with tools from the technical search facilities as well as the tools and interfaces from the Orchestra ontology engine.

## SHAREPOINT ORCHESTRA SYNCHRONIZATION

On the extreme right side of the diagram above we see SharePoint “Sites” being synchronized with Orchestra Contexts. Keeping with the theme of leaving the legacy content in their original stores, it is also possible to keep the content in SharePoint Sites. However there are a number of issues that are worth considering in the spirit of self-sustainability, so that users do not end up placing the source content in an odd format/location. We can help with these questions:

1. Will SharePoint be the final repository for ALL legacy content?
2. Will SharePoint be one of the Content management systems in the enterprise?
3. Will SharePoint be treated as a “System of Record”?
4. Will SharePoint be the system to manage the “Most confidential” content within the company?

A number of additional considerations must be put in place for ensuring data integrity, controlling site and content proliferation – such as locking out ALL access to the SQL data base from any point other than through the SharePoint Application; and putting in place a rigorous, granular, authorization model that can robustly manage the delegation of authority and security as well as perform site management functions.

Based on where SharePoint is at today and taking into account the SharePoint 2010 and Office 2010 capabilities, if we were to deploy today, we would configure new content that comes into SharePoint to remain in it; then we would cache this content within the virtual mirrored contexts within the target contexts. (This is in the same way that the cached contexts carried the metadata and links to the source content.). All other content in the target contexts can be mirrored in SharePoint. There are a number of very specific details that need to be considered in the SharePoint synchronization that is not central to the issue of contextualization of legacy information and those are addressed separately in our other white papers.

Orchestra includes FDK – Funnel Development ToolKit – for building strategic stage-gated and choreographed workflows using simple configurators. FDK allows processes and solutions to be built in a matter of 5 to 10 days. Training is imparted to the customer IT department as well as partners. We have Pharma companies who will build and maintain their own solutions built on Orchestra and our Gluon infrastructure for SharePoint using FDK.

## ABOUT POINTCROSS

PointCross is a global provider of advanced strategic business solutions to knowledge-rich markets, including the pharmaceutical industry. Our Integrated Drug Development Suite (IDDS™) specifically addresses the pharmaceutical industry’s key R&D business needs. At the heart of IDDS is Orchestra+Solo™, an adaptive, contextual knowledge

environment and personalized client that orchestrates core business processes. This robust solution set delivers the following capabilities:

- ☑ Single point of access to contextualized tacit and structured knowledge across the enterprise, with search and guided navigation within and across contexts;
- ☑ Robust search and orienteering capabilities across studies, emails, documents, meta-data and more across the entire organization, CROs and partners
- ☑ Flexible, fool-proof IP security based on contexts and roles, determined by business rules;
- ☑ Predictive analytics for clinical and non-clinical data;
- ☑ Secure multi-party workflows for knowledge sharing and business-social networks within and across companies;
- ☑ Semantic Data Exchange (SDE) for vendor-neutral data exchange, normalization and unit harmonization;
- ☑ Business development, in/out-licensing, e-discovery, audit, regulatory submissions, compliance, and more;
- ☑ Scalable architecture and development toolkits for additional capabilities.

PointCross represents a new way of doing business. We deliver business ready solutions in 1/10th the time and a fraction of the costs compared to standard technologies, while offering strategic advice from people who know the pharmaceutical industry.

We are headquartered in the California Bay Area. We are a CDISC solution provider to the Pharmaceutical Industry, and a Microsoft Gold Certified partner. We also have a global network of service, consulting and infrastructure partners.



For more information, visit us at [www.pointcross.com](http://www.pointcross.com) and call us at (650) 350-1900. Also, check out our blog at <http://pointcross.wordpress.com>.