Data Management Face2Face

1:00 to 2:30 PM	**Hot Topics - Data Management**	**Klara Jelinkova**
	There is a rich set of metadata and middleware needed to support the data classification we are starting to put in place for at least business applications. And the need will become even greater once we develop the same classifications for all of our institutional data including research data. Institutions seem to have a pretty good handle at least theoretically on business data. But once we start to cross over to other assets such as data associated with research, teaching and learning most IT organizations seem to give up. That type of data is viewed by most IT organizations in our institutions as someone else's problem. However it is an important IT and security problem. With collaborative research taking place everywhere how do we classify the research data, protect it while the research is going on and then enable everyone to see after the research is published? How do we collaborate with the libraries on this one? How do we arrive at data management policies that cut across the whole institution? What are some of the examples of institutions doing it successfully today?	

Session Details, Presentations and Notes

Presentations and Links

Data Management Discussion (Klara.ppt)

This session was a discussion session around the issues of data management.

> *Data Management Discussion:*
>
> ***Key Issues:***
> *• Data Architecture, Analysis and Design*
> *• Data Security Management - data access and security*
> *• Reference and Master Data Management - making data available rather than copying data*
> *• Data Warehousing and Business Intelligence Management - normalizing the data across the data warehouse*
> *• Document, Record and Content Management -*
> *• Meta Data Management -*
> *The difference between Structured Data (data in authoritative systems, usually in a database) and Unstructured Data ( ). The Structured Data was designed by DBA. These can proliferate silos. Complex queries are difficult to build and brittle. The metadata and taxonomy as delivered is often "accepted" without thought as the enterprise definition and taxonomy. They also include open fields to store what ever you want.*
>
> *Unstructured data is individually generated, often in file systems, often without much metadata that is meaningful to enterprise. The rich media formats cannot be easily mined to discover content. Management is a nightmare with a proliferation of stores and types of content.*
>
> ***Structured Data Gaps:***
>
> *Data Warehouses: it was sold as a way to build a bridge across the silos. The queries are difficult to construct and often take a lot of effort to get written. It is hard to deliver the complex queries. All the business logic is missing that is used to develop the data and queries. There is a gap in the definitions and the data in the warehouse. You can define student 12 ways so any query could have 12 answers.*
>
> *There is no business rules repository that lets you figure out how things are defined. You can build business rules into the database and into the application code. The farther you get from source, the farther you get from the business rules and the definition and intent for the data.*
>
> *Data Warehouse is used to buffer the source system from queries.*
>
> *When we give out reporting tools to individuals in offices, then it locks you into schemas in the data warehouse. As people develop their queries, it locks down the database table structure. If you change the schema to make more enterprise sense, then many distributed queries suddenly break. There are also "experts" who are vested in their interests in the complexity of the data warehouse. When you streamline and change the process and the queries, you actually threaten the experts.*
>
> *LDAP as an example: We bring data from a bunch of sources, we then normalize the data and present it in standard queries for consumption at large.*
>
> *A place to start: things that go into an executive dashboard.*
>
> *Access To Data project that turned into a drive to get large data sets into Excel on the desktop so they could drill around on their own.*
>
> *Privilege Management: Authorization in application based on name NOT on an institution role.*
>
> *At UW-Madison, we manage privileges by sneaker-net. We don't have access to metadata so that we can generate privileges based on roles. We don't have a way to delete someone from all of the systems when they leave or change roles. The roles of people have states that we have to move them through.*

There are multiple organization charts that come into play when you try to define the role(s) the person which can actually be different at the application roles. Every application also has roles defined and applications do RBAC. But there needs to be an external system where you manage these people and roles. There are two views: one is that there has to be application centric views of roles and privileges, the second is that there could be a set of pre-defined roles that come with a suite of privileges.

There are a set of RULES which are different than the roles. The rules must be stored in a repository as well.

Unstructured Data Gaps:

Electronically recorded lectures, talks etc: We gather some metadata when we create the file like it is the third lecture, created on this date, etc. We cannot scan these files to get rich metadata.

Unstructured Data Management Architecture from IBM. It is cycle-intensive. It looks at 10 second clips of music and adds metadata (like it is "happy music"). The idea that you can just grind at the problem with power might work for a while. There are vendor(s) who are working in this spaces.

Just knowing what data exists is an important step. Storage is just as important. How long do you archive, repose the data? At what level of storage should you storage? The librarians are building dark archives. They are storing data in hopes that some day we will be able to "do something with it". The metadata harvesting and management tools are immature.

Digitally Signatures: When we throw stuff out onto the web or into distributed storage, how do we mark the content so we can mine the archives. "If there was a point to doing it, people might do it." Not many people see the value in deploying the systems.

Wikipedia claims that authors are professors who aren't so their stuff will be taken more seriously. The ability to express our university membership out in the world at large becomes more important.

Students will be coming to us with digital identities. They will want to use those identities and we will become another fob on their keychain that they use in the world at large. We may not be the source of their identities in the future.

All of the data is going to live someplace. We will not be holding it all but we will need to be able to assert our IP over the data wherever it lives. Look at the RIAA and their ability to enforce their IP across multiple platforms.

Standardized media formats:

E-discovery: When you have an E-Discovery request, it is no longer personal data or institutional data. What is the impact of distributed storage and the Web2.0 applications on e-discovery requests. Where is the liability? Who will be sued? Don't change data management practices to because of e-discovery.