The CENSAM Cyber-infrastructure (CCI) provides a framework for data management of scientific data to extend the longevity of their data, to facilitate data discovery, and to provide data access for authorized users. A survey to collect functional requirements from potential users of the CCI was conducted. From the survey it became clear that most of the CENSAM research workflows are purely file-based and, therefore, “there is need for file-based (rather than database) data management and provenance tracking.” Scientific data is created and derived from other processes according to process workflows. “Unlike a data repository using relational database management system, data generated in purely file-based system is hard to track provenance during the life cycle of scientific data. In order to solve this issue, a content management system is customized to manage custom content type and its metadata which are used in scientific research.”

**Motivation**

Metadata extractors offer server-side extraction of values from added or updated content. One of the default actions that can be triggered in a CMS space is Extract Common Metadata. This action will look at the metimetype of the document that triggered the rule and request an appropriate MetadataExtractor from the default MetadataExtractorRegistry. Each extractor is registered to handle a set of metimetypes. The extractor pulls a set of values from the document. The extractor uses a set of properties to map the extracted values to the document’s metadata. If the property was declared as part of an aspect in the model, then the aspect is also added to the document. External metadata extractor can be integrated with the CMS. Custom metadata extractor finds added or updated content using Alfresco repository API, parses the content and pulls a set of values defined in custom content model. The external metadata extractor can be triggered separately from the CMS access interface.

**Content Management System**

A content management system (CMS) is the collection of procedures used to manage workflows in a collaborative environment. The procedures are designed to do the following:

- Allow for a large number of people to contribute to and share stored data
- Control access to data, based on user roles (defining which information users or user groups can view, edit, delete, etc.)
- Aid in easy storage and retrieval of data
- Reduce repetitive duplicate input
- Improve the ease of report writing
- Improve communication between users

CMS are frequently used for storing, controlling, revising, semantically enriching, and publishing documentation. Serving as a central repository, the CMS increases the version level of new updates to an already existing file. Version control is one of the primary advantages of a CMS. We evaluated the latest implementations of content management systems and conducted a functional analysis of leading candidates, including Alfresco and KnowledgeTree. The Alfresco content management system codebase, its API, and its extensibility that permits the creation of custom content types were investigated to verify the suitability of the Alfresco architecture for linking provenance information to custom content types. The Alfresco architecture takes a layered approach as shown in Figure 2.

**Custom Content Model**

A content model describes the data being stored in the repository under control of content management system. Each CENSAM project in different domain uses different content type. Workflow analysis of a CENSAM project was conducted to figure out characteristics of content type used in research project as shown in Figure 3. Here is a list of key information the content model provides CMS:

- Fundamental data types and how those data types should be persisted to the database. For example, without a content model, CMS wouldn’t know the difference between a String and a Date.
- Hierarchical data types like “content” as well as common content types like “Sensor log” or “Sensor data”.
- Out-of-the-box aspects like “auditlogable” and “classifiable” as well as custom aspects like “complaintable”, “Properties (or metadata) specific to each content type”. Constraints placed on properties (such as property values that must match a certain pattern or property values that must come from a specific list of possible values.
- How to index content for searching.
- Relationships between content types.
- Custom content models are built using a small set of building blocks: Types, Properties, Property types, Constraints, Associations, and Aspects. Figure 4 shows the custom content model designed from the analysis of process workflow from urban meteorology research.

**Metadata Extraction**

Metadata extractors offer server-side extraction of values from added or updated content. One of the default actions that can be triggered in a CMS space is Extract Common Metadata. This action will look at the metimetype of the document that triggered the rule and request an appropriate MetadataExtractor from the default MetadataExtractorRegistry. Each extractor is registered to handle a set of metimetypes. The extractor pulls a set of values from the document. The extractor uses a set of properties to map the extracted values to the document’s metadata. If the property was declared as part of an aspect in the model, then the aspect is also added to the document. External metadata extractor can be integrated with the CMS. Custom metadata extractor finds added or updated content using Alfresco repository API, parses the content and pulls a set of values defined in custom content model. The external metadata extractor can be triggered separately from the CMS access interface.

**Access Control**

Access control system is basically composed of three elements as subject, operation and object. Abstract elements such as role, activity and view are added to extend its flexibility while conditional attributes such as organization and spatial-temporal properties are added to elaborate the policy in a more fine-grained way. The access control functions are provided as services in order to incorporate with other existing organizations, systems, services, or functions. Trusted users can plug the services to their own applications, and in virtue of the central access control model, policies from different applications can be adapted to each other. The diagram on the right depicts the key categories of services.

**Conclusion**

In order to verify the file-based content management concept, a prototype of scientific content management system was implemented. Here is the list of accomplishment and lesson learned:

- Analysis of research process workflow was useful to understand custom content type and design custom content model.
- Metadata extractor for custom content type was implemented to automate registration for custom content.
- Consideration of implementing a project-wide root content type from which all other content types in the repository inherit is useful.
- In addition to the potential performance and overhead savings through the use of aspects, aspects promote re-use across the model, the workflow logic, and the presentation layer.

Continuous progress and expansion will be achieved with following future work.

- Integration with network attached storage
- Extend custom content type
- Develop custom content model for different domain in CENSAM research projects

**Acknowledgements**

The research described in this project was funded in whole or in part by the Singapore National Research Foundation (NRF) through the Singapore-MIT Alliance for Research and Technology (SMART) Center for Environmental Sensing and Monitoring (CENSAM).