

Data Centric Knowledge Management System

Product Literature

prepared by 3rd Millennium, Inc.

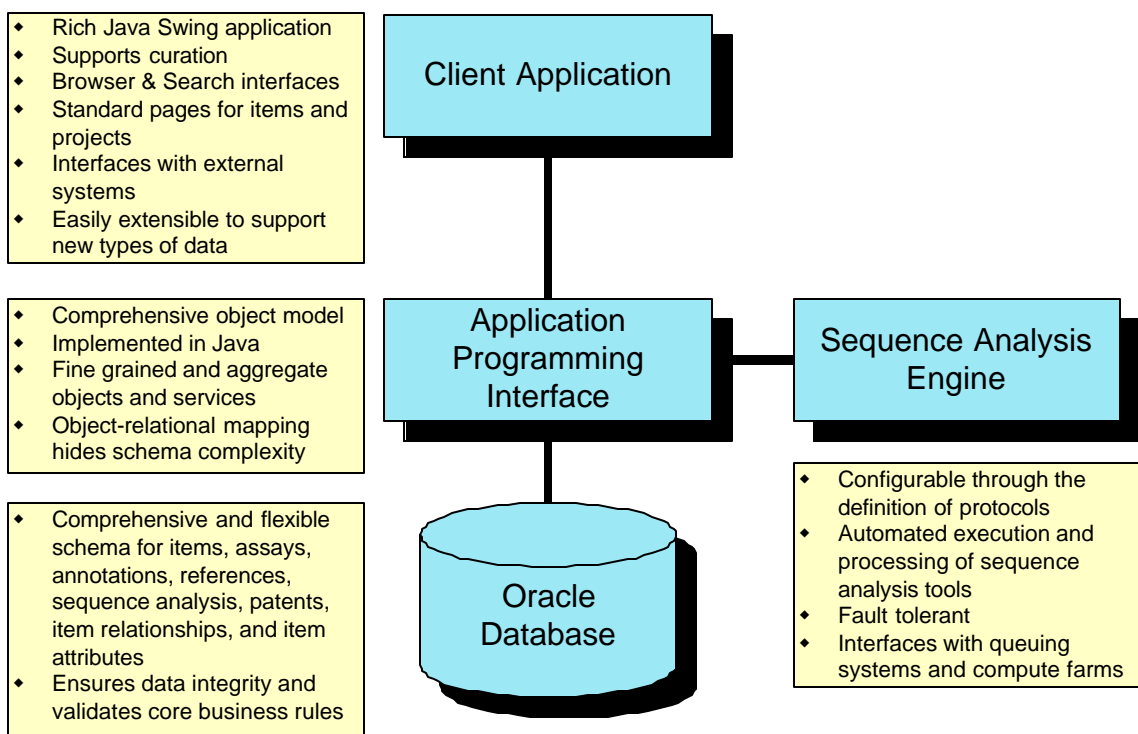
December 4, 2003



1 Overview

The purpose of the Data Centric Knowledge Management System (KMS) is to centralize knowledge generated by scientists working within and across functional areas, and to organize that knowledge such that it can be easily accessed, searched, browsed, navigated, and curated.

Rather than a complete solution, KMS is a foundation system and a framework for creating knowledge management solutions that support biopharmaceutical R&D. It is specifically designed to organize knowledge about drug targets, lead compounds, and about R&D programs. It comprises an Oracle database, and Application Programming Interface (API), a client application, and a component for automated analysis of biological sequences. Through extensions, customizations, and integration with other systems, KMS can provide useful functionality to a wide range of biopharmaceutical organizations pursuing different discovery paradigms.



KMS can be used across R&D programs to:

- ◆ Promote controlled sharing of a company's knowledge related to biological and chemical entities.
- ◆ Make knowledge persistent and durable in light of organizational and personnel changes.
- ◆ Support the tracking of intellectual property, especially for organizations engaged in multiple collaborations.
- ◆ Promote better management and decision-making by providing information across projects in a consistent way in order to measure progress.

2 KMS concepts

The system is designed primarily around a few key concepts:

- ◆ **Projects** – Data are organized in Projects. A project is a grouping of investigations all aimed at addressing a specific scientific question. Projects are organized in project families.
- ◆ **Items** – Supported items are: Species, Strain, Chromosomal Region, Marker, Gene, Intron, Exon, 5'-UTR, 3'-UTR, ORF, SNP, Allele, Gene Product, Biosequence, Compound, Tissue, Population, and Phenotype. Items are also referred to as Knowledge Items.
- ◆ **Assay Results** – Laboratory or computational investigations with associated results are described as assays within KMS and have one or more items associated with them. Examples of assays include Northern blots, SAGE analysis or yeast-2-hybrid analyses.

3 Key Features

The system includes the following features:

- ◆ **Browser** that organizes all data available to a user in a tree structure, allowing the user to navigate through data without specifying search criteria
- ◆ **Search interfaces** for:
 - Assays – searchable by project, assay type, creation and modification date, and user
 - Items – searchable by project, type, assay condition, assay data, attribute, annotation, comment, external database identifier, priority, patent, sequence analysis result
- ◆ **Item Pages** that include: Summary information, Annotations, Assays, References, Patents, Related Items, and Attributes
 - Specialized gene page that includes Sequences and SNPs
- ◆ **Assay Pages** that include: Status, Rationale, Goal, Results, Comments, Conclusions, relationships to Knowledge Items, Attachments, Related Assays, and Assay Conditions
- ◆ **Assay Groups** that track the status of assays performed on knowledge items.
- ◆ **Project Pages** that include: Summary Information, Annotations & Attachments, and Knowledge Items.
- ◆ **Curation interfaces** for: Names, External Links, Annotations, References, Comments, Intellectual Property Analysis, and Attributes

- ◆ **Administration interfaces** for: controlled vocabulary, templates, users, and privileges
- ◆ **Attachments**: files can be attached to annotations, assays, items, projects, and patents. Attachments can be viewed in a suitable application, if available, on the user's desktop.
- ◆ **Links** to external resources: the system utilizes a flexible model for storing external database identifiers, associates these identifiers with items, and defines URLs for retrieving the content associated with the identifiers. Since the client application is written with Java/Swing, HTML pages can be rendered within the application or in an external web browser.
- ◆ **Security** model for assigning privileges to users based on types of data, projects, and project families. The access level supported are: None, Read, Write, Edit, Knowledge Reader, Knowledge Administrator, and System Administrator.
- ◆ **Sequence Analysis Engine** component: automatically processes biological sequences through bioinformatics tools, extracts relevant content from the results, and associates the results to items. The engine includes wrappers for more than 10 commonly used tools such as Blast, motif, pfam or psort and a patent search pipeline for genes.
- ◆ **Support for assays** that can be modeled as a multi-dimensional matrix relating Knowledge Items to measurements performed on the items.
- ◆ **Genomic Map** coordinates for genomic items.
- ◆ **Application Programming Interface** implemented in Java. Includes fine-grained and aggregate services that provide complete coverage of the database schema.
- ◆ **Cross-platform Client Application**: client application can be run on platforms that support Java 1.4 or greater.
- ◆ **Includes complete source code**

4 Uses of KMS

There is no single typical usage of the KMS. Rather, people with different roles are expected to leverage different features in the system. Senior scientists will make extensive use of the pages that display comprehensive summaries of information available for items, e.g., genes. The pages will present data generated by several workgroups and from the public domain, e.g., sequence data, computational biology results, SNP data, links, attachments, annotations, assay results, and references. In addition, the pages contain inferences made by researchers from the data. The pages offer a convenient point of entry for gathering the information required to investigate research questions and to formulate new research questions.

Computational biologists can leverage the data and the APIs for KMS to create agents and data mining tools that derive new explicit knowledge from the complement of data and the relationships between data available in the system. For example, automated

mining of literature data may be implemented to extract references relevant to specific research interests. From these references, new knowledge may be inferred that can be loaded into KMS. Examples of such knowledge may include co-occurrences of genes with targets of interest in full-text articles or abstracts, or relationships between disease states and genes. In addition, computational biologists will utilize the assay component to record the result of the analyses they perform to identify putative targets.

Program managers and senior managers can utilize KMS to track the status of projects and the advancement of work on high priority items, e.g., targets. To facilitate such work, KMS provides summary information about projects and allows recording of status information. For example, the status of assays for a subset of genes within a project is available in a project page. Furthermore, meeting minutes or summary information such as the association signal across a region may be shown in a project page.

5 Implementing and extending KMS

Implementing KMS in your organization requires careful planning. Before the system can be used productively, procedures must be defined and implemented for loading data into the system. There are two main aspects of loading data into KMS: 1) identifying data sources and relevant information within the sources; 2) defining rules that are validated when data are loaded. It is important to consider both the initial load of data to seed KMS and ongoing use. In some organizations, a curator may be responsible for processing ongoing item creation requests. For some uses, regular automated creation of items from well-defined and high-quality data sources may be appropriate.

There is considerable variation in the industry of both data sources relevant to organizations, and of the rules that should be applied to the data. Therefore, KMS does not provide a standard set of loaders. Rather, the API for the system includes specific services that facilitate the creation of loaders. These services facilitate loading of assay data, items, and of data associated to items (e.g., annotations or relationships). For example, the KMS API handles and validates “preferred” item names, which are guaranteed unique either globally in the system or for a given project. The KMS code base includes a simple loader for LocusLink link data that can be used as an example.

KMS was designed using a modular component approach and can be extended easily beyond loaders. Examples of possible extensions include:

- ◆ Integration of a genome browser for visualizing the coordinates of genomic items stored in the system
- ◆ Integration of a small molecule viewer for visualizing and interacting with compound structures
- ◆ Extending the patent search pipeline to items other than genes and sequences (e.g., compounds).
- ◆ Creation of an alert mechanism where users are automatically alerted of new content based on interests they register with the system
- ◆ Adding support for new type of items (e.g., haplotype items)

- ◆ Integration of a text mining tool that operates on the body of references stored in the system and creates new relationships between items
- ◆ Creation of a component that supports the filing of patents by extracting and formatting relevant information from the system

6 Documentation

The following documentation is provided with KMS:

- ◆ Product Literature (this document)
- ◆ System help: includes information for end users and for system administrator. The system help is provided as a stand alone document and on-line within the client application
- ◆ Data model and data dictionary
- ◆ API object models and sequence diagrams
- ◆ Design documents for specific components in the system
- ◆ Installation manual
- ◆ Acceptance Test Plan scenarios
- ◆ Licensing terms

7 Platform and system requirements

To install KMS, the following must be available:

- ◆ An Oracle 8i or 9i instance (see otn.oracle.com)
- ◆ The Ant build tool. Ant can be downloaded from ant.apache.org
- ◆ Java J2SE 1.4 SDK. The SDK can be downloaded from java.sun.com/j2se/downloads.html
- ◆ To run the KMS client, a Java Runtime Environment (JRE) version 1.4 or greater must be available. The JRE can be downloaded from java.sun.com/j2se/downloads.html
- ◆ 128 MB RAM available on the client
- ◆ The sequence analysis engine requires a Unix server

8 Contact information

For questions regarding the Data Centric Knowledge Management System, please contact: consulting@3rdmill.com