PNNL-14567

Proteomics Research Information Storage and Management (PRISM) System

G. R. Kiebel
G. A. Anderson
K. J. Auberry
D. A. Clark
M. E. Monroe
N. Tolic
P. J. Cowley
R. D. Smith

February 2004

Prepared for the U.S. Department of Energy
Under Contract DE-AC06-76RL01830

Pacific Northwest National Laboratory
Richland, Washington 99352

Summary

The Proteomics Research Information Storage and Management (PRISM) System is an integral part of a new proteomics research facility at the Pacific Northwest National Laboratory (PNNL). PRISM was developed at PNNL to manage and automate the processing of the large volume of data generated by the new proteomics facility and the mass spectrometry laboratory at the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL).

Proteomics is the study of proteins expressed by cells, tissues, or organisms under a specific set of conditions. A well-known technique for observing proteins in biological samples uses enzymatic digestion to cleave proteins into characteristic peptide fragments and then detect these peptides through mass spectrometry. Computational analytical methods are then used to infer the presence of parent proteins from the observed peptides. PNNL uses a promising variation of this technique based upon the use of very high accuracy mass measurements in combination with liquid chromatography (mass tag process).

PRISM compiles lists of proteins and peptides observed in the samples processed by the proteomics facility. These lists are kept in multiple databases that collect the results for given organisms and research campaigns and allow researchers to query them based on a given specific biological sample, or upon a group of samples representing specified sets of conditions chosen by the researchers in their experiments. These queries help researchers to determine what proteins an organism creates under different conditions of growth and stress to better understand how cellular biological mechanisms work. PRISM also maintains a complete pedigree of all the data that it processes, which can be followed all the way back to the original mass spectra data.

PRISM is composed of distributed software components that operate cooperatively on a network of commercially available PC computer systems. It collects data files directly from multiple mass spectrometers in the laboratory, stores and tracks them, and automatically extracts data from them and processes it into information about proteins. PRISM also collects and maintains information about the biological samples used in research experiments and the laboratory protocols and procedures used to prepare them. Finally, PRISM allows users to readily locate and examine the information that it contains through a web-based interface, and allows some direct access to external information systems and visualization tools.

PRISM has been in continuous operation since March 2000 and has operated reliably with very little downtime, planned or unplanned. PRISM has enabled high production rates while reducing the demand on staff time for data management. Because PRISM operates around the clock, it makes optimum use of the available software tools and computers, which reduces the overall system cost, especially for software that must be purchased.

Many large and small modifications to PRISM have been made during its operation, both as a natural consequence of changes in requirements due to the dynamic nature of the research that it supports and as part of a planned evolution in capability. The architecture of PRISM has proven to be robust in the face of these modifications due to its modularity and loose coupling between the system components. Given changes tend to affect only a single component. Planned changes have largely been grouped into major upgrade phases:

Phase 1: Data capture, file storage and tracking, and archiving (May 2000)

Phase 2: Automation of analysis jobs (March 2001)

Phase 3: Mass Tag System (March 2002)

Phase 4: Distributed Storage Servers (April 2003)

PRISM has successfully supported multiple proteomic mass spectrometry research campaigns at EMSL, operating so effectively that it is relied upon and generally taken for granted by most of the research staff. The development team expects that the basic architecture of the Data Management System (DMS) will continue to meet the demands of the research needs for the foreseeable future. These include increased throughput due to mass spectrometer automation, accommodation of new and improved analysis tools, and the capability to handle the ever-increasing volume of data.

Abbreviations and Acronyms

AMT	accurate mass tag
CMT	confirmed mass tag
EMSL	William R. Wiley Environmental Molecular Sciences Laboratory
DMS	Data Management System
FTE	full-time equivalent
FTICR	Fourier Transform Ion Cyclotron Resonance
FTP	file transfer protocol
HMMA	High Mass Measurement Accuracy
HTTP	Hyper Text Transfer Protocol
JDBC	Java Data Base Connectivity
LC	liquid chromatography
MS/MS	tandem mass spectrometry
MT-subset	mass tag subset
MTS	mass tag system
NET	normalized elution time
ORF	open reading frame
PMT	putative mass tag
PNNL	Pacific Northwest National Laboratory
PRISM	Proteomics Research Information Storage and Management
RAID	redundant arrays of inexpensive disks
SAM	Scientific Archive Management
SQL	Structured Query Language
TIC	total ion chromatogram
TOF	time-of-flight
UMC	uniform mass class

Contents

Summary iii
Abbreviations and Acronyms v
Introduction 1
Mass Tag Process 2
Description 4
  Overall Structure and Function 4
  PRISM Operation 5
  DMS Details 8
    Tracking Entities 8
    Manager Programs 13
  MTS Details 15
  Information Access 18
    Main Reference Chain 18
    File Sharing 19
    DMS Web Interface 20
    MTS Web Interface 20
Performance 21
Conclusion 24
Acknowledgements 24
Reference 24

Figures

Figure 1. Mass tag process 2
Figure 2. Basic structure of PRISM system 5
Figure 3. Mass tag process and DMS data flow 6
Figure 4. MTS data flow 6
Figure 5. Tracking entities used by DMS 9
Figure 6. Auxiliary information (showing association with tracking entities) 10
Figure 7. DMS structure 11
Figure 8. Organization of data file storage 12
Figure 9. Basic manager program operation 13
Figure 10. MTS structure 17
Figure 11. Main reference chain linking PRISM databases 19
Figure 12. Cumulative counts for experiments, datasets, and analysis jobs 22

Tables

Table 1. MTS selection criteria applied to mass tag databases 18
Table 2. Statistics for selected mass tag system databases 22

Introduction

The Proteomics Research Information Storage and Management (PRISM) System is an integral part of a new proteomics research facility at the Pacific Northwest National Laboratory (PNNL) for research on a number of different organisms. PRISM was developed at PNNL to manage and automate processing of the large volume of data generated by the new proteomics facility and the mass spectrometry laboratory at the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL). This report describes how PRISM is designed, how it evolved, how it functions, and how it has performed from inception to the current date.

Proteomics is the study of proteins expressed by cells, tissues, or organisms under a specific set of conditions. A well-known technique for observing proteins in biological samples[i] uses enzymatic digestion to cleave proteins into characteristic peptide fragments and then detect these peptides through mass spectrometry. Computational analytical methods are then used to infer the presence of parent proteins from the observed peptides. The Macromolecular Structure and Dynamics (MSD) organization of the Pacific Northwest National Laboratory (PNNL) has developed a promising variation of this technique[ii] based on the use of diverse types of mass spectrometers in combination with liquid chromatography (referred to as the "mass tag process" in this report). A production-scale proteomics research facility using the mass tag process has been established in the William R. Wiley Environmental Molecular Sciences Laboratory (EMSL)[1] to study multiple organisms.

PRISM collects data files directly from multiple mass spectrometers in the laboratory, stores and tracks them, and automatically extracts data from them and processes it into information about proteins. It also collects and maintains information about the biological samples used in research experiments and the laboratory protocols and procedures used to prepare them. Finally, PRISM allows users to readily locate and examine the information that it contains, and allows other information systems to access appropriate portions of it.

PRISM compiles lists of proteins and peptides observed in the biological samples processed by the proteomics facility. These lists are kept in multiple databases that collect the results for given organisms and research campaigns. Researchers may use a web based interface to query them based on a given specific biological sample, or upon a group of samples representing specified sets of conditions chosen by the researchers in their experiments. These queries help researchers to determine what proteins an organism creates under different conditions of growth and stress to better understand how cellular biological mechanisms work. PRISM also maintains a complete pedigree of all the data that it processes, which can be tracked from the end results to the original sample.

PRISM is composed of distributed software components that operate cooperatively on a network of commercially available PC computer systems. It uses several relational databases to hold information and a set of autonomous programs that interact with these databases to perform much of the automated file handling and information processing.

---

[1] The William R. Wiley Environmental Molecular Sciences Laboratory is part of the U.S. Department of Energy's Pacific Northwest National Laboratory, which is operated by Battelle, Pacific Northwest Division.

Mass Tag Process

Figure 1 shows a simplified view of the mass tag process.

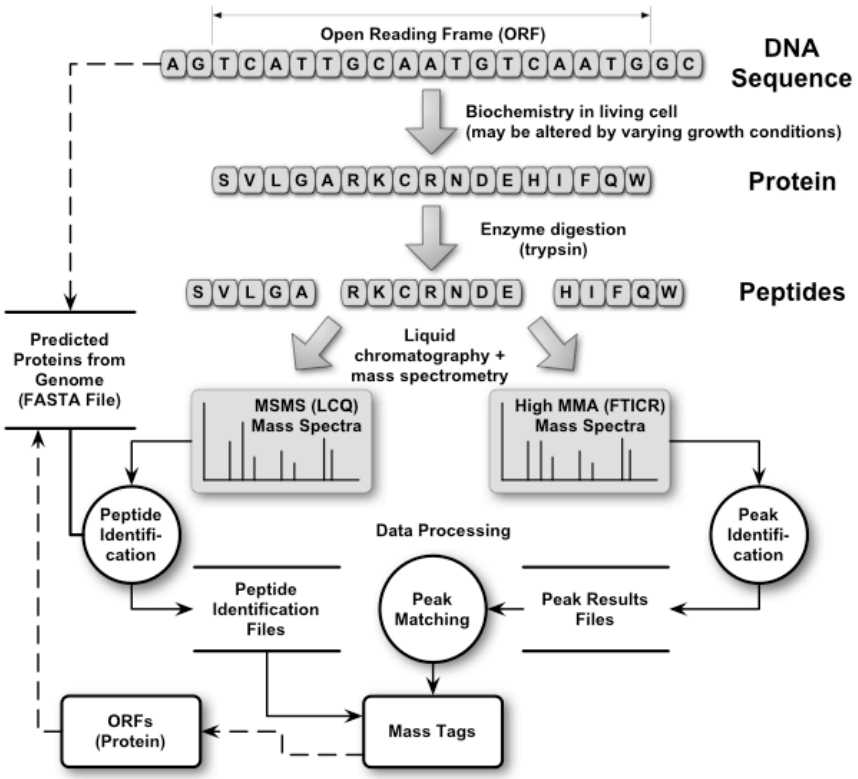

Figure 1. Mass tag process

Mass tags are idealized representations of peptides. Their key properties are their amino acid sequence and their mass (normalized to the monoisotopic condition). Mass tags also reference the open reading frame (ORF) in the organism's genome that is believed to code for their parent protein. There may be more than one ORF associated with a given peptide, so the ORF reference is a fundamental property of the mass tag that is not necessarily unique.

The following are key features of the mass tag process:

1. Preparation of samples using enzymatic digestion of biomass to convert proteins into peptide fragments (trypsin is the predominant choice of digestive enzyme)

2. Injection of samples into mass spectrometers through identical liquid chromatography (LC) columns to provide a discrimination parameter based on elution time[2]

[2] Different peptides pass through the LC column at different rates and emerge (elute) into the spectrometer at different times. Spectra are taken by the spectrometer at periodic intervals so each spectrum is associated with a particular elution time.

3. Use of diverse types of mass spectrometer:

- **Tandem mass spectrometry (MS/MS).** These instruments are commercially available ion-trap devices and are referred to in this report collectively as MS/MS instruments. MS/MS generated spectra are used to produce peptide identifications through an analysis process that matches patterns of mass/charge peaks against a list of peptides predicated from the organism's genome.

- **High Mass Measurement Accuracy (HMMA) spectrometry.** These instruments are typically Fourier Transform Ion Cyclotron Resonance (FTICR) spectrometers and are highly specialized and usually customized. Some time-of-flight (TOF) instruments are being evaluated for this function also. HMMA-generated spectra are used to produce monoisotopic peak results having very high mass measurement accuracy.

4. Creation of mass tags by selection of peptide identification from MS/MS results

5. Confirmation of mass tags by peak matching with HMMA peak results using mass and elution time.

Levels of confirmation are assigned to mass tags as part of the process based on their matching with peak results and the uniqueness of their DNA coding sequence in the organism's genome:

- **Putative Mass Tag (PMT)** – mass tag has only MS/MS peptide identifications and no associated peak results (this is the initial level of all mass tags).

- **Confirmed Mass Tag (CMT)** – mass tag has one or more associated HMMA peak results.

- **Validated Accurate Mass Tag (AMT)** – confirmed mass tag whose DNA coding sequence appears in exactly one ORF in the organism's genome.

Once a sufficiently complete set of AMTs has been established for an organism, the need for tandem mass spectrometry is eliminated and new samples can be processed only with HMMA mass spectrometry. HMMA spectrometers can extract more information from a given sample than ion-trap instruments due to their greater sensitivity, dynamic range, and accuracy. This is an important consideration for high-throughput operation.

Description

Overall Structure and Function

As shown in Figure 2, PRISM is composed of two major subsystems: the Data Management System (DMS) and the Mass Tag System (MTS). DMS is responsible for supporting the "front end" of the mass tag process—tracking the laboratory work and mass spectrometer runs, and processing the raw spectra into analysis results. The MTS is responsible for supporting the "back end" of the mass tag process—compiling lists of proteins and peptides observed in biological samples and allowing researchers to query those lists. MTS performs an intricate set of data processing functions to accomplish this. DMS is largely transaction based and it is designed to respond to different events as they happen. MTS is largely analytical and it is designed to process information in batches as new entries are made in DMS.

DMS functions:

- Capture raw spectra files from mass spectrometers
- Process raw spectra files into analysis results files
- Store and track data files
- Collect and manage lab information (samples, experiments, protocols, etc.)
- Allow users to track processing of samples
- Allow users to locate and access data files.

MTS functions:

- Create lists of proteins and peptides observed for an organism
- Allow users to examine lists.

DMS and MTS are composed of databases, web-based user interfaces, and autonomous programs (called "managers") that perform various automated processing tasks.

DMS currently runs on a Dell PowerEdge[iii] 6450, Dual 733 MHz Pentium[iv] III Xeon[iv] running Windows[v] 2000 Server. The MTS currently runs on a Dell PowerEdge 2550 with dual 933 MHz Pentium III Xeon processors. Several auxiliary processors are used for certain data processing tasks. The databases are all implemented with Microsoft SQL Server 2000.

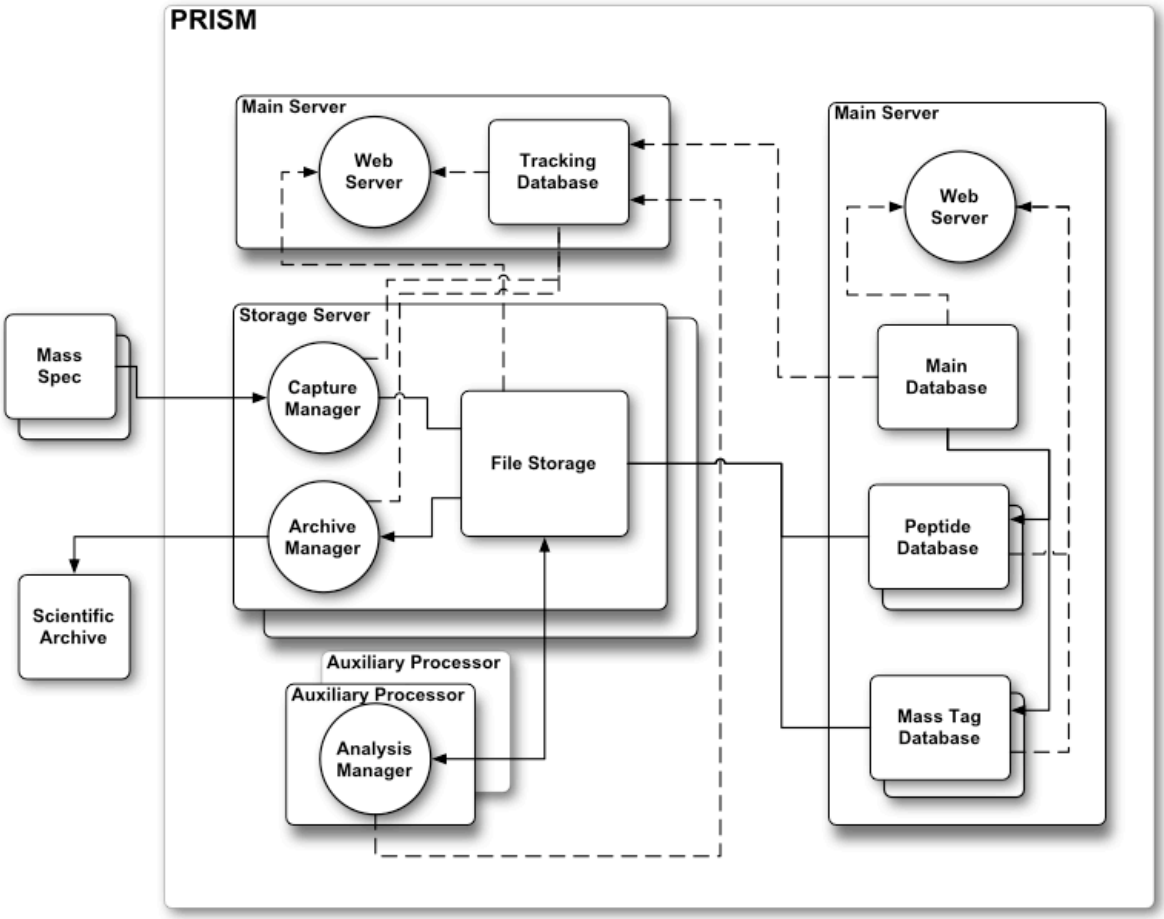

Figure 2. Basic structure of PRISM system

PRISM Operation

Figure 3 and Figure 4 illustrate the operation of PRISM. Figure 3 shows how the external process flow and internal information flow are linked for DMS. Figure 4 shows the internal data flow for MTS.

When samples from a new organism are to be processed by PRISM, one or more campaigns must first be established by the principal investigator. A campaign is a "tracking entity" in PRISM that represents a major line of investigation. A given organism may be represented by one campaign or several depending on how the research is to be performed.

Biological material is prepared by lab technicians. This process typically includes factors such as growth conditions and may include stressors. PRISM offers the opportunity to record this information in a "cell culture" tracking entity; however, this is at the discretion of the researcher. A cell culture belongs to a specific campaign.

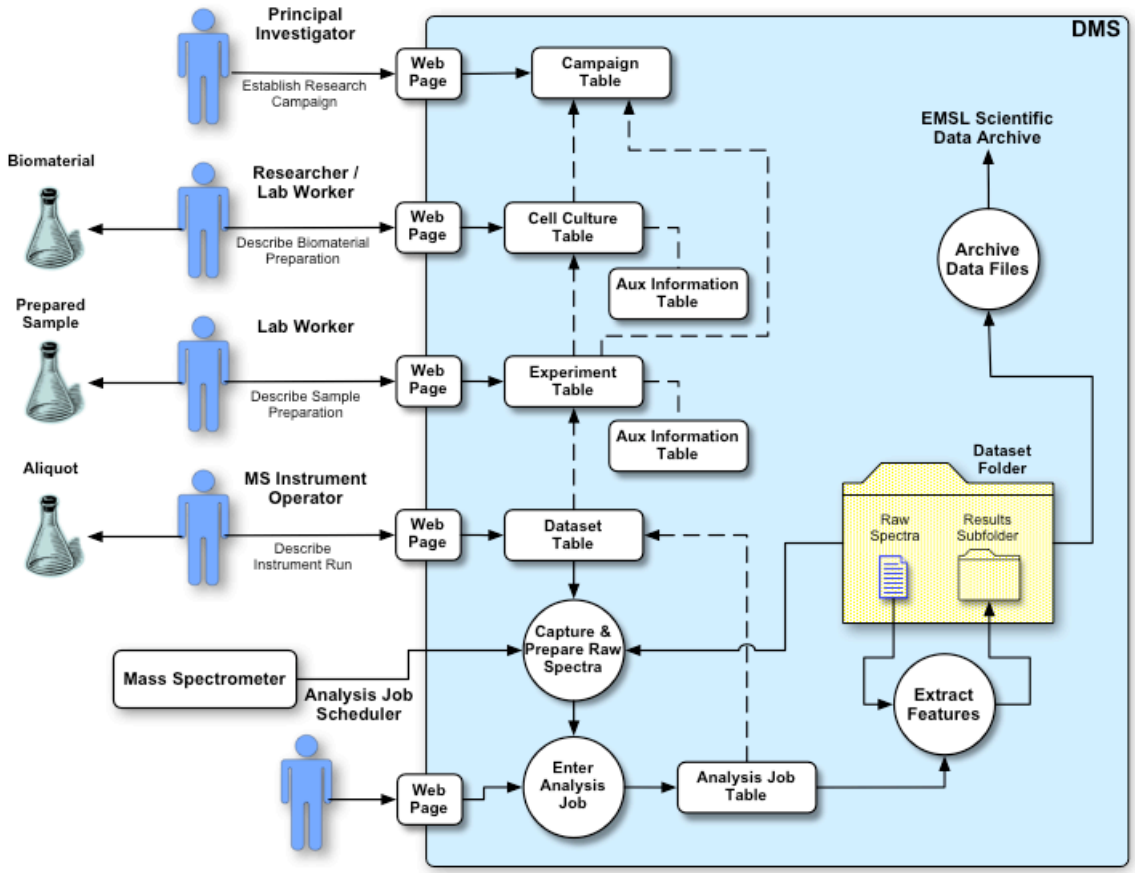

Figure 3. Mass tag process and DMS data flow

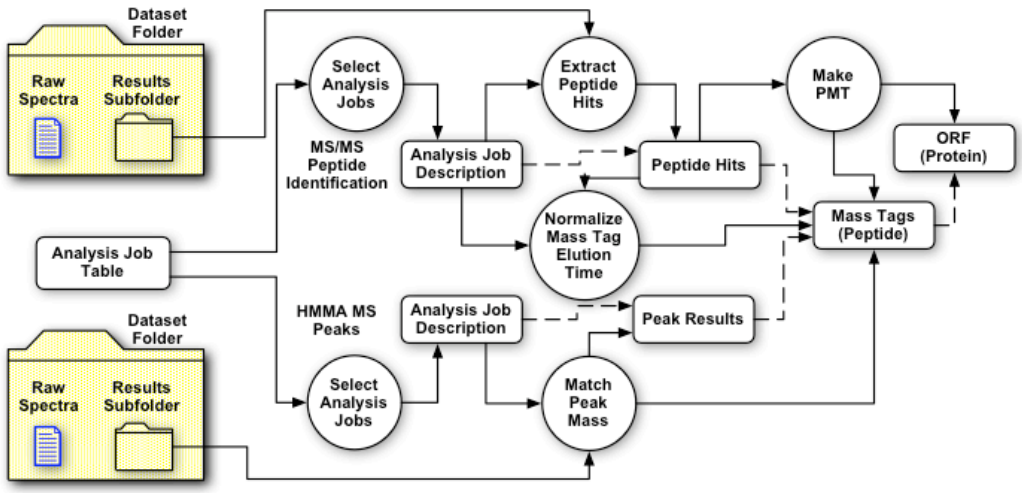

Figure 4. MTS data flow

When cell or tissue growth is terminated, the biomaterial must be processed into a prepared sample before it can be submitted for mass spectroscopy. This sample preparation typically includes factors such as cell membrane rupturing the cell wall (lysis) and enzymatic digestion of proteins into peptide fragments. PRISM requires that basic information about the sample preparation be entered into an "experiment" tracking entity. It also offers the researcher the opportunity to record much more detail about the preparation process if desired. An experiment belongs to exactly one campaign, and may be associated with one or more cell cultures that belong to the same campaign.

A mass spectrometer run is performed on an aliquot drawn from the prepared sample. Each spectrometer instrument uses a LC apparatus to separate the sample into components that elute in different order from the exit of the separation system into the injection port of the instrument. The instrument is configured to make multiple scans producing individual spectra for each elution fraction. When the run is complete, the operator creates a "dataset" tracking entity that initiates a subsequent automatic chain of events. The entry of a new dataset causes a manager program to copy the raw spectra files for the instrument run into a dataset folder that it creates in PRISM file storage space. Raw spectra from some spectrometers may require further processing to prepare them for subsequent analysis. If so, another manager program will perform such preparation immediately after the raw spectra files have been captured into the dataset folder. Following capture and preparation, PRSIM evaluates a dataset tracking entity against a set of predefined criteria. An analysis job is automatically scheduled for each criterion that is satisfied by the dataset.

An analysis job extracts features of interest from raw spectra. Each job applies a single software tool against the raw spectra of a single dataset and produces a single set of results. These results are in the form of files and they are stored into the dataset folder in their own subfolder. Each one is given a unique name derived from the job number. It is typical to run more than one analysis job against a given dataset and so there are often multiple analysis results subfolders in a dataset folder. PRISM provides several different analysis tools to be used in analysis jobs, including those that identify peptides from MS/MS spectra by genomic database search, those that extract monoisotopic peaks from HMMA MS spectra, and those that build a total ion chromatogram (TIC) of the spectra. In addition to being scheduled automatically by the capture process, analysis jobs may be manually created either singly or in batches. DMS provides a pool of auxiliary processors that are available to work on analysis jobs. A simple priority value is used to resolve the order in which jobs are assigned to available processors when there is a backlog.

The MTS collects results from multiple analysis jobs to build lists of proteins and peptides observed in biological samples. The process first establishes a set of mass tags (idealized peptides) from analysis results that identify peptides based on genomic database searches. It also establishes the list of proteins (ORFs) to which the mass tag peptides belong and maintains a cross-reference between the two. Suitable analysis jobs are selected from DMS according to a set of criteria and then suitable peptide hits are extracted from the analysis results of each selected job. The extraction process checks each peptide hit against a set of criteria before accepting it into an internal list. Each acceptable peptide hit is associated with the mass tag that matches its amino acid sequence. New mass tags are created for peptide hits that don't match any existing ones.

A normalized elution time (NET) is calculated for each mass tag based on the peptide hits associated with it. This is done by an algorithm that tries to eliminate the differences due to variances in the LC apparatus from one run to the next by first calculating a set of linear conversion parameters for each analysis job and then applying the conversion parameters to the

scan number for a given peptide hit to generate its NET. The NET for a mass tag is calculated by averaging the NETs for all of the peptides associated with it. Once a sufficient initial set of mass tags has been established using MS/MS analysis results, the process of matching monoisotopic peaks from high-accuracy MS analysis results to mass tags may begin. Suitable analysis jobs are selected from DMS according to a set of criteria. Peaks from results files of the selected analysis jobs are trial-matched against the list of mass tags based on mass and elution time. The actual matching process uses clusters of peaks called uniform mass classes (UMC). Peaks in a UMC are considered to represent the same peptide eluting over several adjacent scans. The trial-matching process includes normalization of elution time for each peak in the UMC and an overall elution time for the UMC. It is iterative and will eventually converge on a mass and elution time value. A UMC that matches a mass tag within the allowed tolerances is associated with that mass tag.

DMS Details

Tracking Entities

Tracking entities contain information about the mass tag process and data files. There are several types of tracking entities and each type tracks a particular object or step in the process. This information can include such things as background information on the experiment, growth and stress conditions for the initial biomass source from which the protein sample resulted, sample digestion and processing conditions, as well as experimental factors such as chromatographic separation conditions or peptide identification parameters. Tracking entities associate with each other in a natural hierarchy as shown in Figure 5.

Tracking entities are implemented as tables in the DMS tracking database with a different table for each type. The content of the tables is referred to as tracking information and it is considered to be essential and is required for PRISM to support the mass tag process. Tracking entities are created and updated by means of web page entry forms.

Campaign: An association of related experiments that represent a major line of investigation.

Cell Culture: Represents the initial source of biomass for any given a protein sample, whether prokaryotic or eukaryotic, as well as the environmental conditions, growth media, and any stressors placed upon it during its growth. A cell culture belongs to a particular campaign, but may be associated with several experiments.

Experiment: Describes the protocols and handling used to prepare an enzymatically digested peptide sample for analysis by mass spectrometry. An experiment belongs to a single campaign, but may be associated with several biomass sources to allow for metabolic labeling experiments.

Dataset: Describes everything related to a single mass spectrometer run, including details of the separation method and the names associated with the raw data files. A dataset belongs to a single experiment, but an experiment may give rise to multiple datasets.

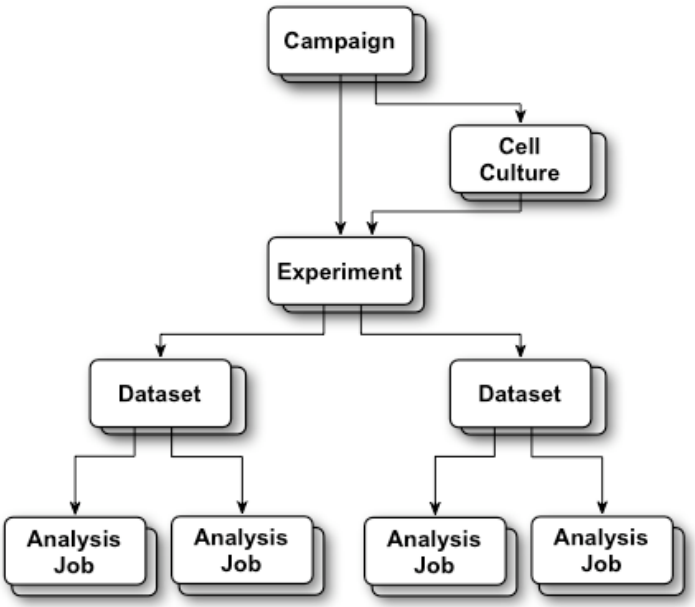

Figure 5. Tracking entities used by DMS

Analysis Job: Represents a single post-acquisition analysis run performed on the raw data by various analysis programs. This can represent TIC generation, a peptide identification run for ion trap MS/MS data, or isotopic peak deconvolution for HMMA-MS data. An analysis job is always performed on exactly one dataset; however, a dataset may have multiple analyses performed on it. Analysis jobs may be created automatically or manually, both singly and in batches. A simple priority number is assigned to each job to control the order in which they are assigned to available processors.

DMS also allows auxiliary information to be recorded and associated with a particular tracking entity (this is only available for cell culture, experiment, or dataset tracking entities at present). This auxiliary information is optional and is not required by PRISM to support the mass tag process, but it exists to provide the researchers with an additional place to record supplemental data that would typically be recorded in their lab notebook. Figure 6 shows the arrangement of auxiliary information and tracking entities.

Because tracking information is used constantly in the automated tasks that DMS performs, its database implementation needed to be as efficient as possible. However, the structure of that information was expected to change only when required by changes in the underlying mass tag process. Accordingly, the detailed structure of the tracking information was "hard coded" into the schema of the tracking database. Each tracking entity is represented by its own table in the database with columns for each item of information that it records. Associations between entities are expressed as foreign key relationships that the database enforces. This allows access to be handled through straightforward structured query language (SQL) statements in the DMS code.

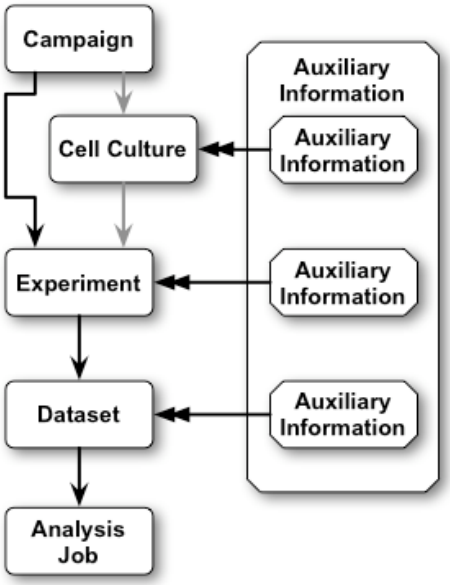

Figure 6. Auxiliary information (showing association with tracking entities)

Auxiliary information needed to be more flexible than tracking information. Researchers should be able to specify unique items of information to be recorded for their particular campaign, or perhaps even for a particular experiment, and should also be able to add new items as the research progresses. However, it is not necessary that access of auxiliary information be highly efficient since it is not used directly by DMS in supporting the mass tag process. Therefore, an implementation for auxiliary information was chosen that emphasizes flexibility over efficiency. The tracking database permits the definitions of auxiliary information items to be configured as data in the database rather than by the schema of the database itself. Changing auxiliary information definitions does not affect the software structure and does not require a software change control process; however, the definition of auxiliary information must conform to a fixed pattern required by DMS. The basic unit of auxiliary information is the *item*. An item has a name and it represents an instance where a single scalar data value can be recorded against a specific tracking entity. Auxiliary information items are organized into *subcategories* that are in turn organized into *categories*. Categories and subcategories are used to express the natural structure of the auxiliary information (albeit in a somewhat constrained fashion). All actual instances of auxiliary information items are kept as rows in a single table in the tracking database. Each row contains the identity of the auxiliary information item, a reference to the tracking entity to which it is associated, and the recorded value of the data.

File Storage

Figure 7 shows the structure of DMS in more detail. In particular, it shows all of the manager programs that perform the automated file handling and data analysis functions for DMS.

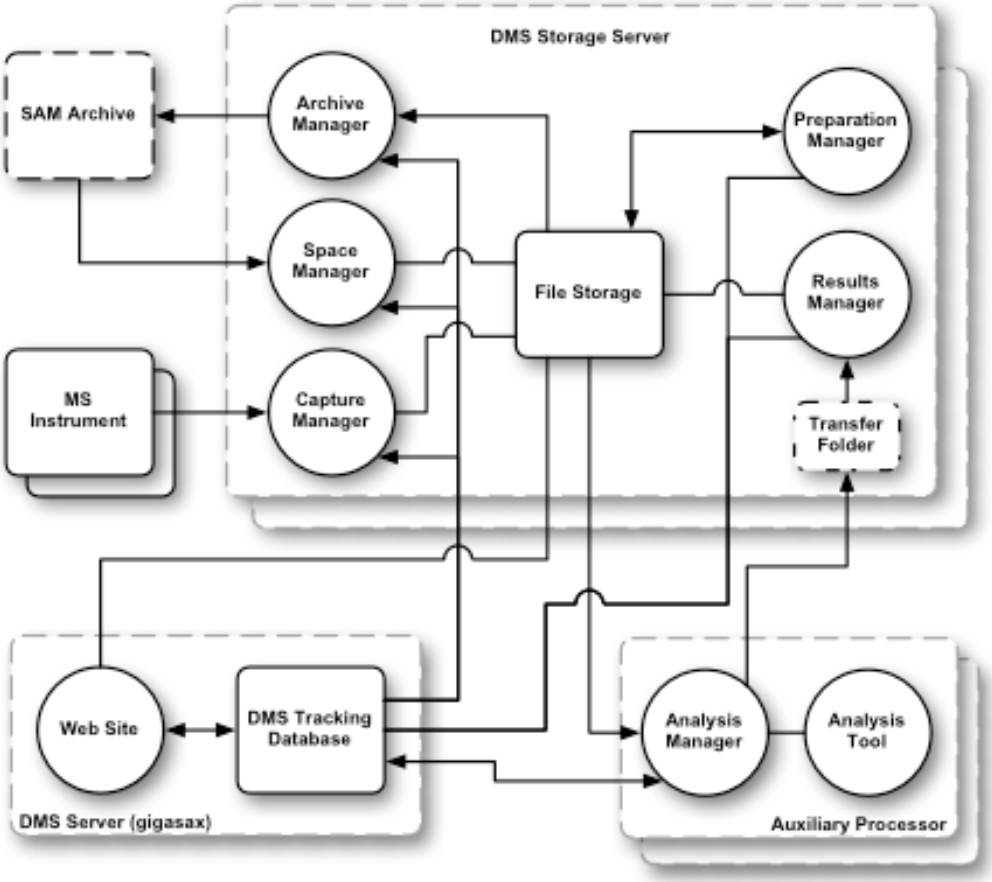

Figure 7. DMS structure

DMS provides a set of storage servers for storing all the data files that PRISM manages and for running all the manager programs that perform file handling. Because of the large size of the raw spectra files, each HMMA spectrometer is serviced by its own separate storage server. All the ion trap spectrometers are serviced by a common storage server. DMS file storage space is organized into a hierarchy of volumes and folders as shown in Figure 8.

Volumes (virtual drives) make up the top layer of the hierarchy. Each volume is composed of multiple disks that are configured into RAID 5 in hardware. RAID 5 protects data files from damage by single point hardware failures and this protection was considered more than worth the attendant loss in storage capacity. New storage capacity is always added to the system in the form of new volumes.

Instrument folders make up the next layer of the storage hierarchy. Each instrument folder contains only data files related to one specific instrument. This design decision was made because the size of the data files varies widely between the different types of mass spectrometers. HMMA instruments produce hundreds of files totaling several gigabytes per run while ion-trap instruments produce a few files totaling only a few megabytes in total size. The larger drive

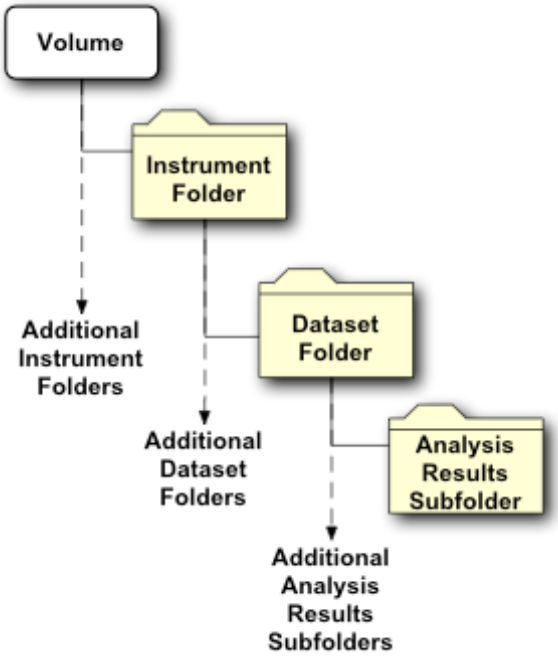

Figure 8. Organization of data file storage

volumes are used to contain the instrument folders for the HMMA instruments. These are more closely monitored because even with volume sizes maximized, they fill up more quickly than other types.

Dataset folders are created inside an instrument folder. These contain all the files from a single run on that instrument. At first, this only includes the raw spectra files from the instrument run itself. Subsequently, analysis results files are generated from the raw spectra by analysis jobs.

Analysis results folders are created inside a dataset folder. These contain the set of analysis results files from each analysis job performed on the raw spectra files of the dataset.

Permission settings on storage folders are used to protect the integrity of the files on DMS. Network sharing and the web interface are used to give users access to the files, but this is restricted to read-only permission. Only the manager programs are allowed to add, delete, or modify the contents of storage folders, and they always coordinate their activities with the tracking database.

Only the manager programs have write permission into the file storage folders. This protects the data files from being improperly altered and assures that the tracking database is always synchronized with the data files. Network sharing is done at the instrument folder level - volumes are not shared and are only visible internally within DMS.

Manager Programs

Data file handling and automated processing is performed by the set of manager programs.

- Capture Manager
- Preparation Manager
- Archive Manager
- Space Manager
- Analysis Manager
- Results Manager

Multiple instances of a given type of manager programs can be running at one time. Manager programs operate independently of one another and interact only with tracking entities in the DMS tracking database. In particular, their actions are governed by state variables in the tracking entity tables. Figure 9 shows the basic operation of a manager program. When a manager program has been idle for some period of time, it calls a stored procedure in the DMS tracking database requesting a task that it can perform. If an appropriate tracking entity exists in the correct state, a task is assigned to the calling manager and the information necessary for it to perform the task are returned in the request call. The manager will probably interact with file storage as it performs the task. Some managers also make use of an external software tool. When the manager has completed its assigned task, it calls another stored procedure to report success or failure.

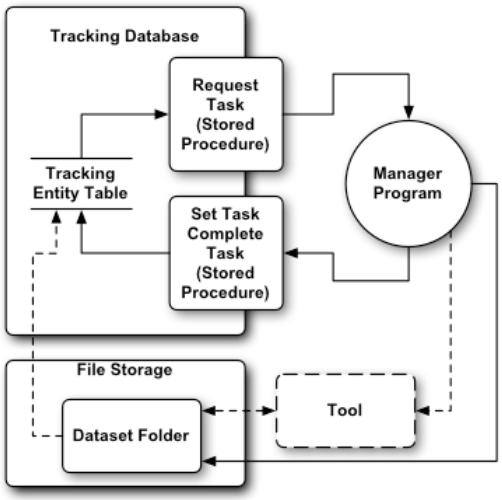

Figure 9. Basic manager program operation

Presently, all of the manager programs are written in Visual Basic.Net[v] and are based on an object-oriented design pattern that is shared among all manager programs. Each manager program implements its own unique code in subclasses derived from common abstract base classes. Most utility and housekeeping code (such as logging) is directly shared by all the manager programs. The same basic design pattern would be followed if it became necessary to port one or more manager programs to a non-Windows operating system. Java would be a very likely candidate for the implementation environment in that case.

A brief discussion of each manager program follows.

Capture Manager Program
The capture manager program is responsible for creating a new dataset storage folder in data file storage and copying raw spectra files from the spectrometer to it. Some instruments (whose workstations run the Windows operating system) provide access to their raw data files via shared folders over the network while other instruments (whose workstations run some form of Unix) provide access via File Transfer Protocol (FTP). The Capture Manager can use either method. It is simple to add a new instrument to DMS and only requires making a single entry in the tracking database that identifies the instrument and describes its type, method of file transfer, and the directory from which the raw data files are to be copied.

The dataset tracking entity has an attribute that can be used to assign an overall level of further interest. If it is determined that its raw spectra are of no further interest, the dataset can be designated as "No Interest" and no further automated processing will be performed on it. This is usually done after the dataset has been captured and perhaps undergone some initial analysis. It is also done when the instrument run was bad, but a record must be kept that it was done (for accounting purposes).

Preparation Manager Program
Raw spectra files from some kinds of spectrometers require some initial reformatting before analysis jobs can be performed on them. The preparation manager program handles any reformatting tasks that may be required. For example, spectra files from the Bruker FTICR instruments are in a different format than those from the Finnigan FTICR instruments. Finnigan instruments predate Bruker instruments at the proteomics facility so the decision was made to reformat Bruker data into the Finnigan format (the Finnigan format also separated the spectrometer scans into individual files that were easier to examine than the extremely large Bruker file). File compression is used on all HMMA raw data files to save space in the DMS data file storage. The preparation manager program performs that function as well.

Archive Manager Program
The contents of the dataset folders are periodically backed up into the Scientific Archive Management (SAM) system[vi] by the archive manager program. The SAM system is an infrastructure system within EMSL that is responsible for the long-term archival storage of scientific data generated at the laboratory. After the initial archiving of a dataset folder, updates are scheduled automatically whenever an analysis job adds a new results subfolder to the dataset folder.

Space Manager Program
The space manager program acts to keep the file storage on its storage server from overfilling by selectively deleting ("purging") raw spectra files when it detects that the amount of free space on the volume falls below a preset threshold. This purging operation only applies to HMMA raw spectra files because they are so much larger (at least one order of magnitude) than all other types of files that are kept in DMS file storage. The space manager selects datasets for purging using a "least recently analyzed" algorithm. The storage server also restores ("unpurges") raw spectra files that have been purged should any new analysis jobs be scheduled against them. This active management of file storage greatly reduces the rate at which more disk space needs to be added to the storage managers (at some modest increased cost in processing time).

Analysis Manager
The analysis manager programs process raw mass spectra captured from the mass spectrometers with one or more analysis software tools according to the type of instrument used and the nature of the experiment being performed. Each analysis manager program operates a single instance of one of the tools. Analysis results are collected into a folder on the auxiliary processor that is given a unique name generated from the unique job identifier. If the analysis job is completed successfully, the analysis manager program will copy the results folder to a transfer directory on the storage server that contains the dataset to which the analysis job is assigned. The results manager program on that storage server will eventually move the copy to the dataset folder.

Operating experience with PRISM has shown that there will typically be some backlog of analysis jobs waiting to be picked up by idle analysis manager programs in the auxiliary processor pool. Analysis jobs have a priority level that is used to determine the order in which they are assigned to requests from analysis manager programs, but this simple mechanism is not adequate to ensure that the highest priority jobs are assigned in a timely manner. The problem is caused by the wide variation in the amount of time that different analysis jobs can take to complete. Some can run in under a minute while others can take several days (or occasionally weeks). Eventually, the auxiliary processor pool can become fully occupied with very long running jobs. If new analysis jobs are submitted at a higher priority, they may have to wait for an unacceptably long time for a processor to become available, or may require that a lower priority job be manually terminated—a wasteful and time-consuming practice. PRISM provides a mechanism for avoiding this problem by using the time that an analysis job is predicted to run as an additional factor in the assignment logic. Particular analysis manager programs can specify the range of predicted run times that they will accept and if at least some percentage of the auxiliary processor pool is configured to not accept jobs predicted to be longer than an hour or so, they will become available to pick up high priority jobs in a timely fashion.

It is relatively easy to scale up the processing throughput capability of DMS by adding more auxiliary processors to the pool. It is also relatively easy to re-task the auxiliary processors in the pool with a different proportion of analysis manager program types to more efficiently address the mix of jobs in the queue.

Results Manager Program
The Results Manager program moves the analysis results subfolder from the transfer folder to the dataset folder.

MTS Details

PRISM compiles lists of proteins and peptides observed in the samples processed by the proteomics facility. These lists are kept in multiple databases that collect the results for given organisms and research campaigns and allow researchers to query them based on a given specific biological sample, or upon a group of samples representing specified sets of conditions chosen by the researchers in their experiments.

MTS uses several different types of databases to modularize the actual implementation. It uses multiple instances of these types to segregate data for different organisms, campaigns, and specific investigations. The following is a list of these database types and their functions:

- Peptide database
    - Selectively imports peptide identifications from MS/MS analysis job results for a given organism from DMS
    - Builds collection of unique sequences from the identifications and maintains cross-reference
    - Evaluates each peptide sequence against a set of enzyme cleavage rules and sets flags for each rule indicating whether or not the peptide satisfies the rule.
- Mass tag database
    - Selectively imports peptide identifications from its associated peptide database
    - Builds mass tag collection from imported peptide identifications
    - Selectively imports peak results from HMMA analysis job results from DMS
    - Normalizes elution time
    - Matches peaks to mass tags based on mass and elution time
- ORF database
    - Provides reference genomic information for proteins.
- MT_Main database
    - Interfaces all other MTS databases and DMS information
    - Performs automated master update cycles.

Instances of these types of databases connect to each other to share information and processing tasks. These connections follow a simple set of rules:

- A mass tag database connects to exactly one peptide database.
- A mass tag database connects to exactly zero or one ORF database.
- A peptide database can be connected to several mass tag databases.
- An ORF database can be connected to several mass tag databases.
- There is only one instance of the MT_Main database.

Figure 10 illustrates typical arrangements in the MTS. There is generally only a single instance of a peptide database for a given organism, but multiple mass tag databases connect to it. If there is only one version of genome for an organism, there will usually be only one ORF database used by the mass tag databases for that organism. It is not uncommon for multiple versions of the genome to exist, especially over time, and in that case multiple ORF databases may be required.

The use of multiple databases instead of a single database was a design decision intended to keep individual database sizes moderate to provide reasonably short times for complex queries. The partitioning of MTS into different types of databases was a design decision intended to reduce the amount of redundancy that might have resulted from the decision to have multiple instances. In most cases, a new instance of a mass tag database is all that is required for a new line of investigation for an organism that is already represented in the MTS. MTS database modularity also helps to contain the effects of modifications to the software necessitated by changes in the mass tag process.

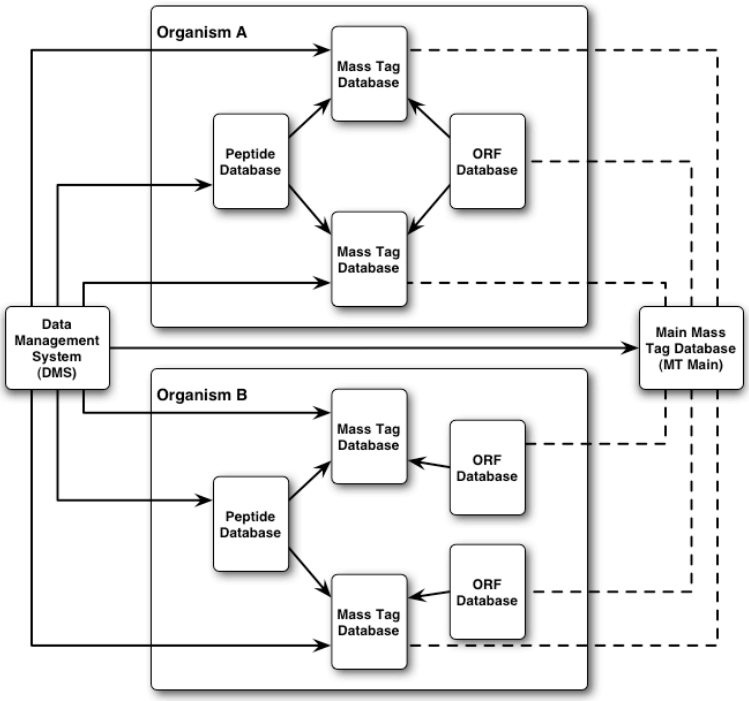

Figure 10. MTS structure

Processing within the MTS databases is performed in batches on a periodic schedule. The MT_Main database manages the overall processing schedule for the MTS. Each peptide database and mass tag database contains a master update procedure that performs all the processing that is done within that database. The MT_Main database calls the master update procedure in all of the currently active databases according to a time schedule, which currently runs each night.

Mass tag databases are designed to offer a high degree of selectivity in which analysis job results they import and process. They also offer the ability for the user to selectively query the information that they contain. This double layer of selectivity allows the researcher considerable flexibility in examining which proteins were observed in biological samples under different conditions. PRISM administrators can create instances of MTS database types as required to support the research. This can typically be accomplished in 30 minutes or less. Table 1 lists the selection criteria that control which analysis job results are imported into a mass tag database. The criteria that control what is imported into a peptide database are much simpler (single organism and list of analysis job result types).

Table 1. MTS selection criteria applied to mass tag databases

Criterion	Type	MS/MS Analysis Job Results	MS HMMA Analysis Job Results
Peptide Database (Note 1)	Source	Source of analysis jobs and peptide hits for import	N/A
Campaign	Single Value	Only analysis jobs that belong to the specified campaign are imported.	(Same as for MS/MS)
Separation Type (Note 2)	List	Only analysis jobs from datasets whose separation system appears in list are imported.	(Same as for MS/MS)
Organism DB File (FASTA) (Note 3)	List	Only analysis jobs that used organism database files that appear in list are imported.	N/A
Analysis Parameter File	List	Only analysis jobs that used parameter files that appear in list are imported.	N/A
Cleavage Filter	List	Only peptide identifications whose peptide sequence satisfies a cleavage filter that appears in list are imported.	N/A
Minimum DeltaCn	Single Value	Peptide identification	N/A
Analysis Result Type	List	N/A (Note 1)	Only analysis jobs producing a result type that appears in list are imported.
Instrument Class (Note 4)	List	N/A	Only analysis jobs from datasets for MS instrument whose instrument class appears in list are imported.

Notes:

1. Restricts jobs by peptide DB import criteria (organism and analysis result type)
2. Normally chosen so that all elution times can be normalized together
3. Must be chosen so that only one genome is represented
4. Can be used to further restrict instruments from which peak results are accepted.

Information Access

PRISM allows users to access information through its web user interfaces (intranet only) or by direct (read only) access to its data file storage. It allows (read only) access to its data file storage and to the DMS tracking database and to the MTS databases by other information systems.

Main Reference Chain

All the databases in PRISM are linked by a chain of reference that ties both peptides and mass tags in the MTS back to their tracking entities in DMS (Figure 11). This report has already described how the automated processing within PRISM both constructs and depends upon this reference chain. The reference chain is in the key data model for accessing information. Going backward through the chain provides the pedigree for a mass tag, allowing all the details to be gathered about how it came to be. Going forward through the reference chain can focus the result on a specific subset of mass tags by applying selection criteria at various points in the chain.

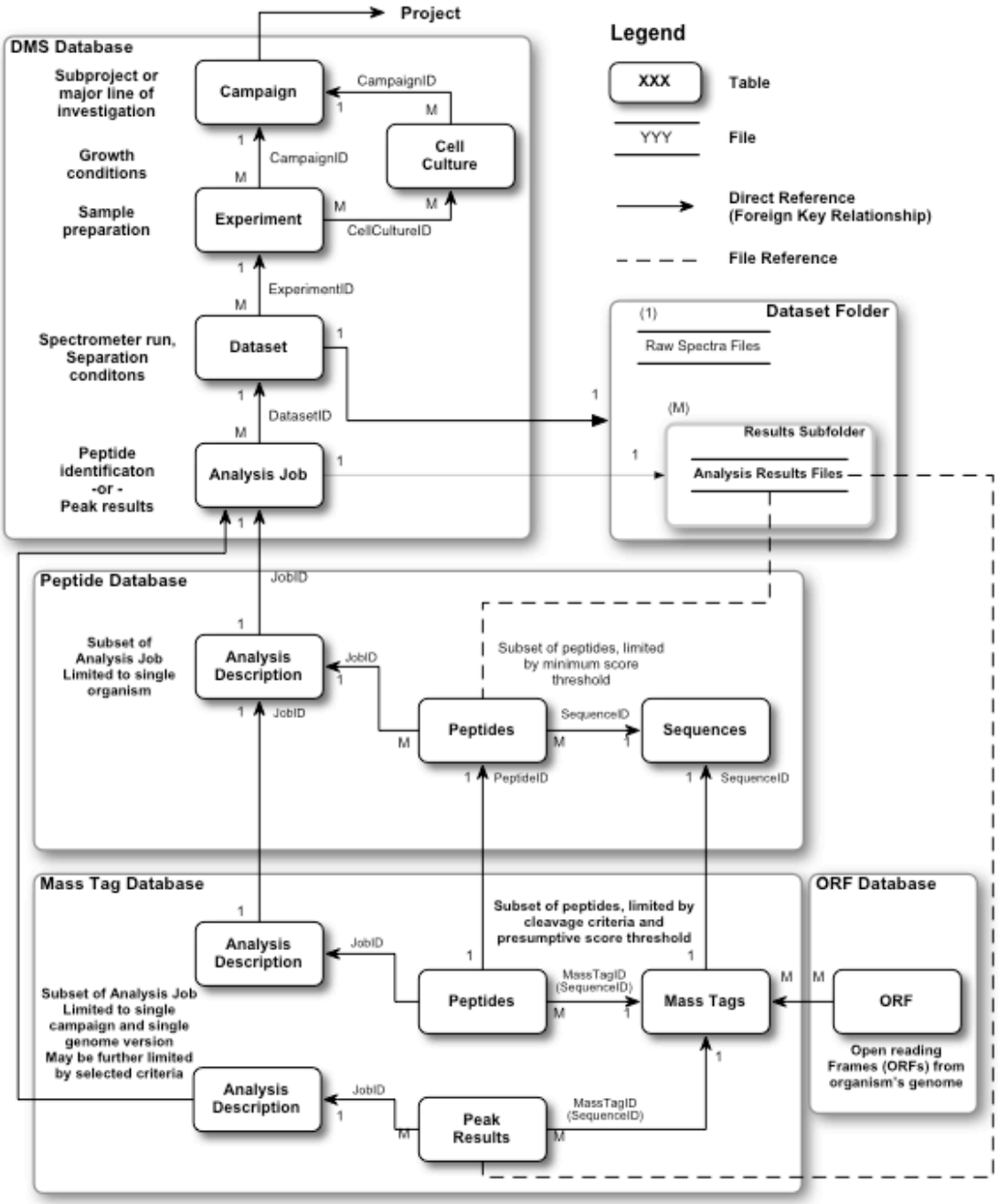

Figure 11. Main reference chain linking PRISM databases

File Sharing

File permissions and folder sharing was discussed under the Data File Storage section of this report. Data files can also be accessed by the user through the web interface. Other data systems are allowed direct, but read-only access, to the tracking database and to the data file storage.

DMS Web Interface

The DMS web user interface provides access to the tracking information and auxiliary information contained in its tracking database. It allows the user to browse and search for specific items, as well as display all known details about an individual item. It also provides several data entry pages used for creating and editing tracking entities and their auxiliary information. Because DMS does not contain any final product, most user tasks are associated with monitoring or controlling the front end of the mass tag process.

MTS Web Interface

The MTS web user interface is primarily concerned with displaying mass tag database information in a variety of ways useful to the researcher. Most of the data display pages are highly cross-linked to provide the user with the capability to drill into the data. Mass tag subsets (MTSubset) are an important and integral part of many of the display pages. The user is given the ability to define and edit mass tag subsets. However, there are no other data entry pages in the MTS web interface since all of the rest of the information in the MTS is derived.

Performance

PRISM has been in continuous operation since March 2000 and has operated reliably with very little downtime, planned or unplanned. Figure 12 is a graph that shows the cumulative count of experiments, datasets, and analysis jobs handled by DMS to date. It has been implemented in major phases:

Phase 1: Data capture, file storage and tracking, and archiving (May 2000)

Prior to Phase 1, a completely manual process was used where data was managed by each researcher. This approach was time-consuming and often made the data difficult to locate. Implementation of Phase 1 of PRISM reduced the time burden imposed on the staff for data management to virtually nothing, and made all the data easy to locate and access.

Phase 2: Automation of analysis jobs (March 2001)

Implementation of Phase 2 had a dramatic impact on the overall throughput rate of the facility, an effect clearly seen in the graph. The equivalent amount of manual labor that would have been required to achieve this rate of production would be enormous. Using a very conservative figure of one hour of staff time per analysis job, more than 12 man-years of manual labor would have been needed in just the first year of operation. Because the analysis manager programs can run around the clock, DMS makes very efficient use of the available software tools and auxiliary processors. This is especially important in cases where the tools are expensive.

Phase 3: Mass Tag System (March 2002)

Prior to Phase 3, all mass tag results were produced by custom programs that generated one-of-a-kind databases. This was extremely slow and expensive. The MTS reduced the time to create a new mass tag database to approximately 30 minutes or less, and all mass tag data now shares a common schema that simplifies access.

Phase 4: Distributed Storage Servers (April 2003)

The original implementation of DMS file storage had all storage attached to a single server that implemented the tracking database, web interface, and file-handling managers. This became a bottleneck as more FTICR spectrometers were added, and the sample run rate was increased for the whole facility. Splitting out the storage servers relieved the bottleneck and established an expandable architecture that can readily accommodate new FTICR instruments.

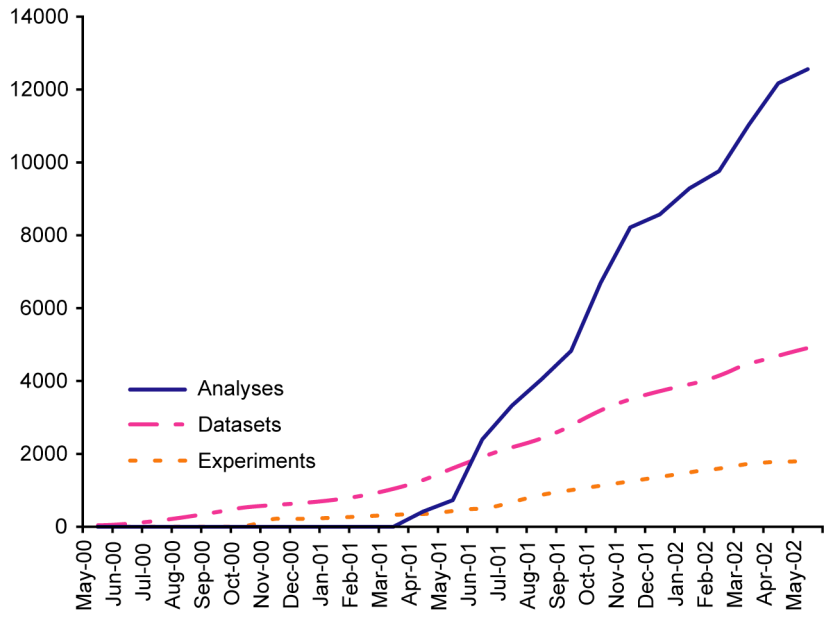

Figure 12. Cumulative counts for experiments, datasets, and analysis jobs

Table 2 shows the statistics for the databases in the MTS for the two organisms that have been the main focus of research to date.

Table 2. Statistics for selected mass tag system databases

Description	**PT_Deinococcus_A8**		**PT_Shewanella_A9**
Analysis jobs imported for TIC_LCQ	1,515		835
Analysis jobs imported for Sequest	3,687		859
Peptides extracted	6,825,120		2,369,514
Unique sequences	4,087,610		1,578,817
Peptides satisfying cleavage rules	1,060,212		366,973
Unique sequence modifications	483		2
Peptides with sequence modifications	1,007		2
Description	**MT_Deinococcus_P12**		**MT_Shewanella_P14**
Imported analysis jobs for Sequest	939		595
Imported analysis jobs for ICR2LS	638		142
Imported analysis jobs for TIC_ICR	394		114
Imported peptides	241,955		203,550
PMTs	81,214		60,354
AMTs	27,315		7,954
Lockers	1,318		655
Imported peak results	1,283,116		100,021
Matched peak results	275,173		27,028

Thus far, no data under the control of PRISM has been lost, corrupted, or misplaced. The tracking information makes locating and accessing both files and metadata straightforward and efficient, and the web user interface has proven to be convenient and comfortable for the users.

Many large and small modifications to PRISM have been made during its operation as a natural consequence of changes in requirements due to the dynamic nature of the research that it supports and a planned evolution in capability. The architecture of PRISM has proven to be robust in the face of these modifications due to its modularity and loose coupling between the major system components. Given changes tend to affect only a single component.

The fundamental operating mode of the tracking database is transaction processing and its table structure and indexing has been designed accordingly. One of the fundamental design decisions was not to incorporate the contents of the files into the database directly, but instead to store them as files and track them with the database. This minimized the size of the tracking database and improves efficiency.

Conclusion

DMS has successfully supported multiple proteomic mass spectrometry research campaigns at EMSL, operating so effectively that it is relied upon and generally taken for granted by most of the research staff. It has proven flexible in the face of research-driven requirements changes while operating reliably as a production system. The development team expects that the basic architecture of PRISM will continue to meet the demands of the research needs for the foreseeable future. These include increased throughput due to mass spectrometer automation, accommodation of new and improved analysis tools, and the capability to handle the ever-increasing volume of data.

Acknowledgements

The Environmental Molecular Sciences Laboratory is located at Pacific Northwest National Laboratory, which is operated by the Battelle Memorial Institute for the U.S. Department of Energy under Contract Number DE-AC06-76RLO1830.

PRISM was developed by the authors of this report. Important contributions were made by Lars Kangas, Patricia Medvick, Mary E. Powers, Kerry Steele, and Eric Strittmatter. Many scientists on the research staff have provided valuable feedback and encouragement to the development team, especially Mary S. Lipton, Ljiljana Pasa-Tolic, and Deanna L. Auberry.

References

[i] Yates, J.R., McCormack, A.L. and Eng, J.K. (1996) "Mining genomes with MS," *Anal. Chem*. 68, A534-A540.

[ii] Conrads, T.P., Anderson, G.A., Veenstra, T.D., Pasa-Tolic, L. and Smith, R.D. (2000) "Utility of Accurate Mass Tags for Proteome Wide Protein Identification," *Anal. Chem*. 72, 3349-3354.

[iii] Dell, PowerEdge are trademarks of the Dell Computer Corporation, Austin, Texas.

[iv] Pentium, Xeon are trademarks of the Intel Corporation, Santa Clara, California.

[v] Windows, Visual Basic, SQLServer are trademarks of Microsoft Corporation, Redmond, Washington.

[vi] Scientific Archive Management, http://www.emsl.pnl.gov:2080/proj/sam/