Ovation Scientific Data Management System®

Comprehensive data management and collaboration system for biological research

Barry Wark, PhD
Physion Consulting

Modern biologists are faced with extraordinary data challenges. In particular, today's computer-intensive research requires that scientists manage an unprecedented volume of raw and analyzed data in a broad set of formats. Ad-hoc data management systems, prevalent in many academic research labs, inhibit scientists' ability to explore complex data sets and limit collaboration between lab members. Lack of interoperability between data management practices and the difficulty in transforming data for public consumption further prevent broader sharing of data within the scientific community.

The Ovation Scientific Data Management System™ is a state-of-the-art software platform designed specifically to address the needs of modern scientists. Ovation manages data storage, maintains relationships between data elements, reduces the barriers to exploration of complex data, simplifies incorporation of new data into existing analysis pipelines, supports annotation, stores analysis results directly in the database, and provides powerful collaboration and data sharing tools. With its exceptional power and flexibility, Ovation is revolutionizing the way scientists analyze, share, and think about their data.

Evolving data management in basic science

How has the scientific workflow changed as a result of modern data acquisition?

The laboratory notebook was once the epitome of scientific record. It provided a format for researchers to record, annotate, analyze, and share their observations. Today, technological advances in data acquisition, information storage capacity, and computing power allow modern researchers to collect data at a scale that would have been unfathomable by previous generations of investigators. Even computer-based lab-notebook strategies are inadequate for managing data sets that measure in terabytes and incorporate everything from 3-D time lapse imaging streams to

transcriptional profiles. In addition, modern systems-oriented research projects often acquire physiology, imaging, video and genomic information from overlapping research subjects. Maintaining the relationships between these data modalities becomes a significant challenge if the each modality is stored as an individual data file, separated from its metadata, experimenter's notes, and analysis results. As analysis becomes more complex, intermediate analysis results further complicate the data management picture to the extent that tracing the flow of raw data through to final analysis becomes difficult or impossible.

Today's scientists embrace the opportunities presented by advances in data acquisition and analysis. The time has come for scientific data management tools to evolve their capabilities to meet the advanced needs of a cutting edge scientific community.

How does Ovation organize and store diverse scientific data?

The Ovation Scientific Data Management System is a comprehensive database for all of your scientific data and analysis results. Ovation's database engine is designed specifically for academic scientists with a data model that matches the natural hierarchy of elements in scientific data sets (Figure 1). This design enables all of a user's data to be accessible for analysis simultaneously, rather than limiting data mining by project, experimenter, or date, as do many ad-hoc data analysis systems. Unlike a relational database engine that stores data in 2D tables—like Excel spreadsheets—Ovation's object database stores objects directly, reducing the computational and programmer overhead to interact with the data.

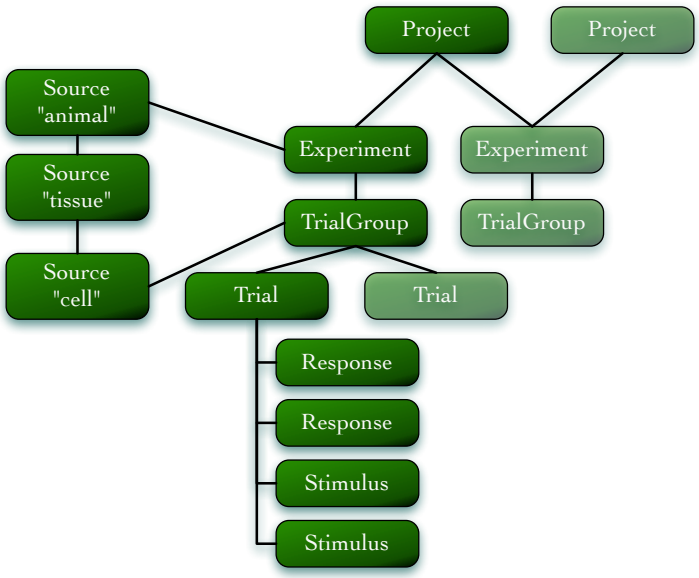

Figure 1 The Ovation database naturally represents the hierarchy of scientific data. The subject of an experiment is represented by one or more Source objects that naturally represent the relationships between organism and the biological source of data such as a single cell. Stimuli and Responses store information about

the stimuli presented and the data acquired during a Trial, which is represented in its broader scientific context, from enclosing TrialGroup to the larger Project.

In the rare cases where Ovation's data model does not natively support information that you would like to store in the database, you can attach files to any object in the database (Figure 2). These files are stored *within* the Ovation database but can be copied to a user's local computer for editing or viewing. File attachments allow you to store propriety binary data files, scanned handwritten notes or files produced by other software within the database, maintaining the relationship between those files and their associated data and context.

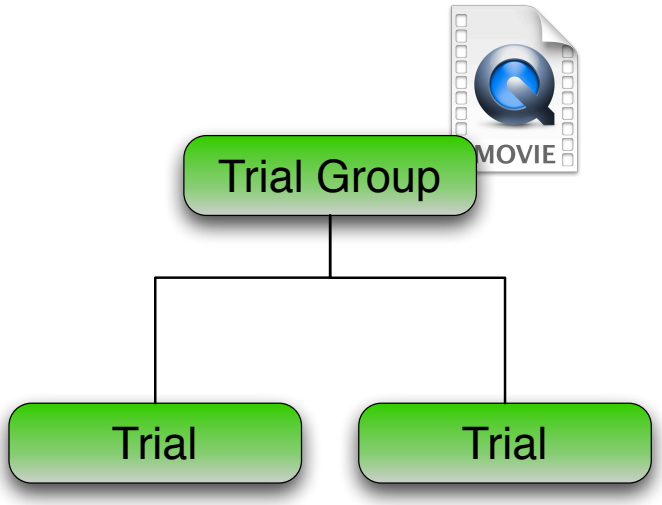

Figure 2 Ovation supports attaching files to any object in the database. This diagram shows a section of the database with a file containing the video of a group of trials attached to the TrialGroup object that represents that group of trials.

Because Ovation stores attached files within the database, these files are backed up with the rest of the Ovation database (see "How can Ovation help backup scientific data?" below).

Acquiring data is only the start of the data lifecycle. As researchers work with data, their observations and annotations enrich the data set. Ovation seamlessly incorporates these annotations, allowing researchers and their collaborators to build richer and more complete interpretations of their data. Annotations supported by Ovation include keyword tags, key-value pairs, notes attached to any object, regions in time or space, and "derived responses" that are calculated from raw data, and complete analysis records. All annotation data in the Ovation database is stored per-user. Thus multiple users may annotate the same data without stepping on each other's toes. Users can view and query annotations added by a particular user or for annotations from any user. The many types of annotations managed by Ovation are described below:

Keyword tags can be used to categorize data along many dimensions. Some uses of keyword tags amongst existing Ovation customers include tagging individual trials or groups of trials that exemplify a particular behavior or observation, making it easier to find those trials when making figures. Users also tag data that they want to share with a colleague (example: "please take a look at everything tagged 'cool").

Key-value metadata can be used to add individualized attributes to objects that make your project's querying or analysis tasks easier. For example, users might add a behavioral category such as "aggressive" or "passive", with key "behavioral-category", to a Source object representing an animal after analyzing a given trial. It is then easy to query for all Sources that have the "aggressive" value for the "behavioral-category" key.

Notes and timeline annotations can be used to mark points or regions on an experiment's timeline that are relevant to a project or analysis. Regions may be used, for example, to mark regions of the timeline where an animal is exhibiting a particular behavior, as observed on video. This timeline annotation can be associated with both the video and simultaneously recorded audio or physiology recording(s). It is then trivial to find all the physiology data that falls within the region(s) of the timeline annotated as exhibiting that behavior. As with all objects in the database, annotations may be tagged with keyword tags, allowing arbitrary groupings of annotations.

Space and space-time annotations can be used to mark one or more regions on an image or movie. Like timeline annotations, these annotations may be grouped by keyword tag labels and may be associated with multiple objects. It is then easy to query for all images that have an annotation marking a particular feature.

Derived responses are associated with a single trial and store values or responses derived from that trial's data. These derived values such as filtered extracted spike clusters or tracked motions paths in video may be time consuming to recreate. Instead of forcing users to store these derived values in separate intermediate files, divorced from their original data, derived response annotations allow users to store this information in the database for fast retrieval. The derived data is stored along with the code and parameters used to derive that value so that you and others can easily recall how it was created. You may have any number of derived responses for each trial. Since derived response objects are stored per-user, you will not interfere with other users' derived values.

Analysis Records associate the raw data, code, parameters and results of an analysis. Analysis Records thus allow current and future researchers to determine exactly how an analysis was performed from raw data through to final figure. The analysis can also automatically be re-run using the original

data and parameters, incorporating new data acquired since the original analysis or using new parameters. See "How does Ovation track the results of analysis" below.

Annotations may be modified at any time, so feel free to explore your data and take notes as you go. Of course, only the user that owns an annotation may modify that annotation. So you are free to modify (or delete) your own annotations without fear of modifying another user's annotations or workflow.

How can Ovation help backup scientific data?

Ovation provides a simple backup tool that lets you make a full database backup or backup just the data that has been added or changed since the last backup. These incremental backups can be stored locally, offsite, or on a cloud storage provider such as Amazon's Simple Storage Service for true disaster recovery protection. Ovation stores all data and attached files within the Ovation database, so this single backup can capture data, annotations, and analysis results for an entire lab or institute.

How do you import to and access data from the Ovation database?

Ovation provides native interfaces (APIs) for common scientific programming languages and environments including Matlab[1], Python, and Mathematica[2]. Interfaces for more Java[3], and C/C++ are also available. Although Ovation provides a number of user interface options, the programming interfaces allow you to use all of the features of the Ovation system directly from your code. If you already have Matlab, Python or Java code to read your existing data, you can easily write that data directly to Ovation from those languages once you've read in the raw data.

You can read data from Ovation directly into in any of several programming environments. Within Matlab, for example, objects in the database (e.g. Experiments, Trials, Responses, etc.) are presented as native Matlab objects, which expose the attributes of the database objects and their relationships to other objects in the database. Ovation handles the mapping between languages, platforms and processor architectures so that data imported with one language can be read from any other. Ovation supports OS X 10.6 and later, Windows 7 and later and RedHat Enterprise Linux (RHEL 5.5 or Fedora Core 12 and later) for both clients and servers.

---

[1] Matlab is a registered trademark of The Mathworks, Inc.
[2] Mathematica is a registered trademark of Wolfram Research, Inc.
[3] Java is a registered trademark of Oracle Corporation.

Lowering the barriers to exploration

As data sets become larger and more complex, finding data of interest becomes more difficult. The complexity of data exploration inhibits researchers from trying a new analysis, adding new data to an old analysis, or finding old data for a new analysis. One of the game-changing features of Ovation is its ability to lower the barriers to exploration. Finding the data you want becomes simple.

How can scientists easily explore and query large, complex data sets?

The Ovation database engine supports easy browsing within the object graph of data. By maintain the relationships between individual data items and their larger context, Ovation supports easy traversal from any data item to its related items. For example, traversing from a single response to related response modalities, or to the previous or next trials is a single operation. Ovation's data engine manages a memory cache on your computer so that you can browse through a dataset larger than your system's memory without any problem (even in Matlab!).

Ovation includes a revolutionary query engine. Queries can include standard data attributes such as the protocol of a trial, as well as key-value metadata and other annotations. The full power of the Ovation query engine is available from a graphical query builder (Figure 3). You do not need to learn a query language like SQL. Figure 3 shows a sample screen capture from Ovation's query editor tool. The shown query predicate specifies all trials acquired during a trial group with label "Control", that used the "my.protocol" protocol, had a protocol parameter called "frequency" whose value is greater than or equal to 10, and that have been tagged with the "example" keyword tag by the current user.

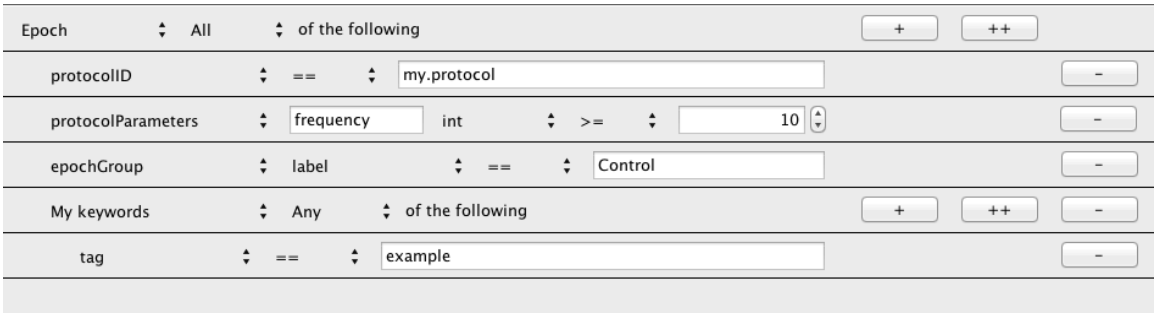

Figure 3 Complex queries can be built using an intuitive query builder.

Despite the power of Ovation's query engine, queries in Ovation are fast. A typical search time to query the data for an entire project (or publication)—regardless of the complexity of the query criteria— is ten seconds. As the database size outgrows

a single data server, Ovation's query engine naturally scales queries to use the power of all available database servers.

Is it possible to query context?

Biology is complex and often context-dependent. Perhaps the most profound and difficult questions in many fields thus involve finding data that was acquired in a given context. For example, consider a particularly interesting finding that you hypothesize is the result of the particular combination of subject age, time of day and the preceding sequence of trials. Finding data with the same context—subject age, time of day and preceding sequence of trials—is a difficult (if not impossible) task in traditional relational database engines such as MySQL. In comparison, Ovation's query engine puts this dataset at your fingertips.

Ovation's powerful query engine means that the barriers to exploration of your dataset disappear. Large meta-analyses and queries are enjoyable, not just possible. Go ahead…think big!

Powerful analysis pipelines

How does organizing data with Ovation improve the efficiency of existing analysis workflows?

With Ovation, you continue to use your existing analysis pipeline and analysis code. Where you currently read data from data files, your analysis pipeline simply pulls data from the Ovation database, retrieving objects as native objects in Matlab, Mathematica, Python, or Microsoft Excel's PivotTable plugin. Furthermore, with one simple change at the beginning of your analysis pipeline, you can easily incorporate new data as it is collected. Ovation-style analysis begins by querying the database for all trials that match a given predicate (Figure 4).

Figure 4 You continue to use your existing analysis pipeline with Ovation. Where you previously searched the folders on your computer for relevant files, then read data from those files, you get data into your existing pipeline via a simple query in Ovation.

Performing a query in Matlab is a single line of code:

```
iterator = context.query(Trial, <…predicate…>);
```

which returns an iterator object for the result set. If the analysis then proceeds by iterating the result set performing the desired analysis on each element of the result set:

```
while(iterator.hasNext())
    currentTrial = iterator.next();
    <…perform analysis on current trial…>
end
```

the analysis *automatically* incorporates new, relevant, data as that data is added to the Ovation database (Figure 5).

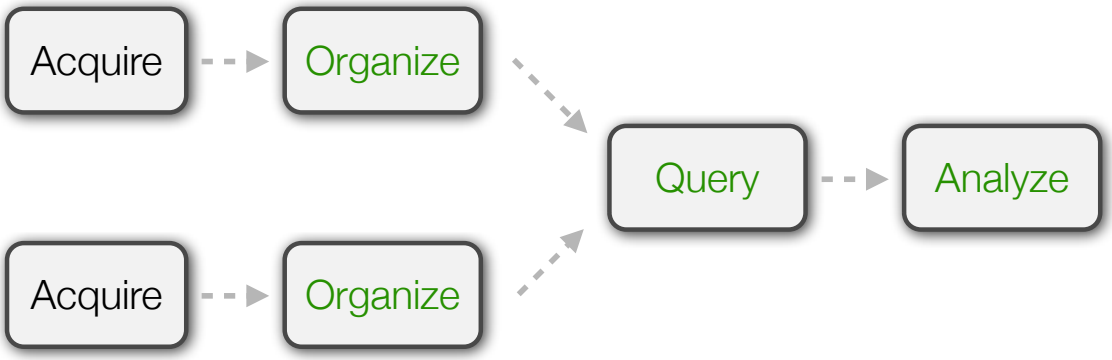

Figure 5 Ovation-style analysis pipelines that incorporate a query stage automatically incorporate new data as it is added to the database.

For demanding computational analyses, Ovation supports synchronizing a portion of your data set to a local or cloud-based computational cluster. Imagine being able to unleash unlimited computational power at the touch of a button!

How does Ovation track the results of analysis?

Isolated analysis products (data files, values, annotations, etc.) block scientists' ability to easily track data from acquisition through analysis. This barrier to reproducing analysis results or trying alternative parameters, data, or analysis algorithms ultimately injures the scientific community. Ovation's Analysis Records break down these barriers by associating the original data, analysis code, parameters, and final results (as numeric data, binary files, images or PDF files) of an analysis (Figure 6). Analysis records are annotatable, searchable, and shareable just like any other object in the database.

From an Analysis Record, it is trivial to recreate the *exact* analysis described the analysis record. In Matlab, Ovation can also automatically re-run the analysis using alternate parameters, a new data set, or a new algorithm. Analysis Records represent the end of difficulty reproducing analysis results within a lab or between collaborators.

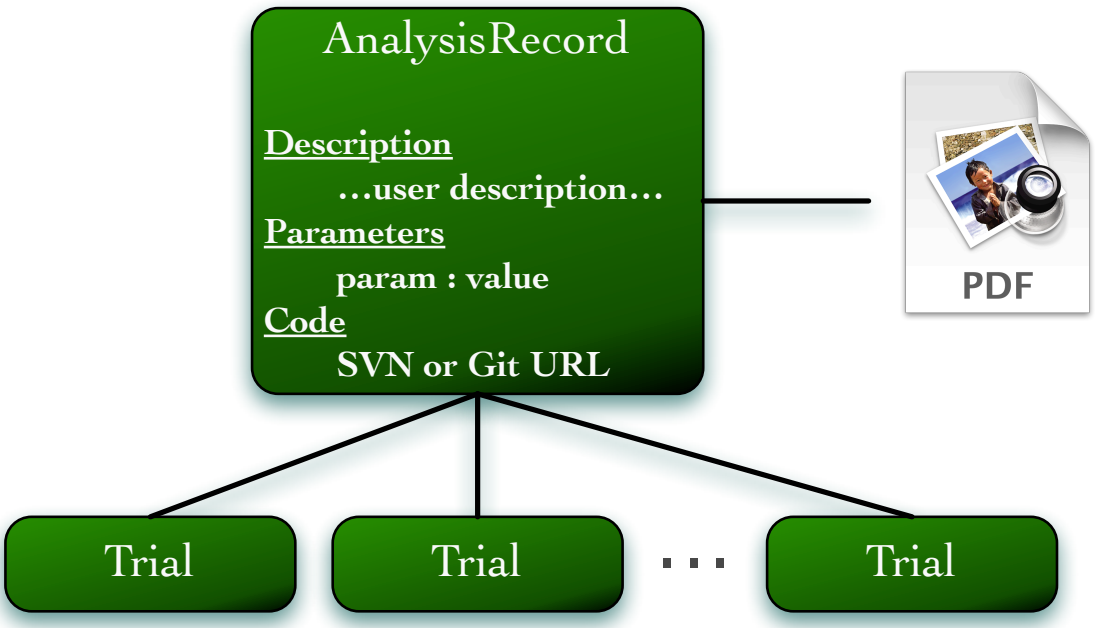

Figure 6 Analysis Records associate analysis code (stored in Git or Subversion), parameters of the analysis, results such as PDF figures and the Trials that were the input for the analysis.

Lowering the barriers to collaboration

Multi-user collaboration

Every object in the Ovation database has a globally unique identifier. This simple innovation allows users of a common database to share data by simply sharing a link to any object in the database. For example users may send links to a single response, a trial, an entire experiment or a full analysis. Upon receiving a sent link, users can connect to the Ovation database and retrieve the object identified by the link. Of course, the retrieved object maintains its relationships to other objects in the database so a link to a single response does not divorce that data from its trial or the broader experimental context.

The Ovation database engine supports multi-user access to a centralized data store so an entire lab or institute can work with the same database. Each user creates one or more connections, called data contexts, to the database, which they use to read or modify objects in the central database. All users' data contexts are automatically updated with the most recent database activity. This concurrent connection model supports live multi-user collaboration—you can view the same data and analysis, working with a colleague across the room or across the Internet.

Ovation's multi-user annotation capabilities build upon this simple, effective, sharing platform. Shared annotation and analysis is easy with each user working

within their own data context. Workflows that require forming a consensus annotation, such as anatomical studies with multiple annotators, are simple and efficient using Ovation's collaboration features.

Is it possible to provide tools to scientists that both enhance the efficiency and enjoyment of their work and facilitate collaboration with the broader scientific community?

Ovation was developed from the ground up by scientists and programmers who understand the day-to-day needs of academic researchers. The system provides dual benefits to researchers: relief from the individual challenges of data management and a common data platform to meet the needs of collaborative researchers, labs or institutions. Ovation provides a platform for developing efficient acquisition, analysis, and collaboration tools that make users' daily lives in the lab easier and more enjoyable. With Ovation, data in the lab is already in a standardized format and is necessarily properly annotated (Ovation validates data and will not accept incompletely specified data). Thus with no extra effort, Ovation users can share data with the broader community in a standard format, with full annotations. Ovation currently supports HDF5, Microsoft Excel and XML data export.

Deployment and support

Physion supports Ovation customers in adapting their existing analysis pipelines and data sets to function seamlessly with Ovation. With purchase of a lab Ovation license, Physion engineers and consultants will provide onsite customer training and system deployment. While working with your lab, Physion consultants will devise a plan to import existing data sets into your Ovation database, develop code to perform the import, and validate it's performance. Our deployment support does not end until you are satisfied with Ovation.

After deployment, ongoing support and maintenance is available on a yearly contract basis. Physion is proud to provide world-class customer support via phone or email during standard business hours for customers of the Ovation maintenance program. Customers of the program also receive all upgrades and new versions of Ovation while participating in the maintenance program.

To learn more about Ovation or to schedule a demonstration, please call (617) 299-9520.