The Ultimate Guide to Self-Service Data Preparation

Whether you call it data mining, extracting, cleaning, joining, blending or masking - it's all data preparation. Professionals can spend hours a week working with their data, trying to join data from disparate sources, re-keying info from static reports or PDFs, or formatting data for accurate reporting. A self-service data preparation tool, such as Datawatch Monarch, empowers users to import, transform and export data in less time, often with the ability to automate processes for even more time savings. Learn more about how easy it is to do more with your data in this Guide to Self-Service Data Preparation.
Chapter 1: The 5 Overlooked Requirements for Self-Service Data Preparation

1. **Access to Multi-Structured and Streaming Data**
   Often times the most important information you need is locked away in multi-structured documents and sources that seem impossible to use without rekeying the data. You may have production reports or third party documents that would add tremendous analytic value if only there was an easier way.

2. **Data Masking**
   As data volume grows exponentially, businesses are faced with the challenge of scaling their protection of sensitive information to meet government standards and to defend against theft. Data masking has emerged as the leading solution to accomplish this across all industries and verticals. Data masking is the process of creating a new version of an original data set that maintains the data formatting, but generates new data values. Hiding the original data in this way allows it to still be usable for analytics, training, or testing, while ensuring that the underlying data is only visible to approved users.

3. **Automated Processes**
   Every time you prepare data, you should look to automate the data preparation routines whenever possible. And you should share your work with others in your organization so you can have more consistent data and be more productive.

4. **Reduced Risk with Governance**
   The move to self-service is all about speed and agility for the business user. But IT has given up some control and with that comes increased risk. You need to reintroduce some governance but do so in a frictionless way, accounting for the fact that much of the data preparation is against non-managed data sources—like CSV extracts, PDF reports or third party data.

   Look for an enterprise solution to securely store, manage and control access to prepared data sets, reusable models, visualizations, and dashboards.

5. **Ease of use**
   Work directly with your data and use 100 prebuilt manipulation functions, add calculated fields and mask data without learning a new script language or complex flow diagrams. Eliminate the need for V-look ups, macros, or pivot tables.
Chapter 2  

Bridge the Gap Between Business Agility and Governance

Self-service data preparation tools are rapidly being recognized as a necessary element to any data discovery or advanced analytics implementation.

Bridge the Gap Between Self-Service Data Preparation and Enterprise Governance

Most organizations have well formed strategies to govern data that reside in managed systems like enterprise applications or data warehouses. One of the biggest benefits of self-service analytics lies in the ability to rapidly combine and analyze data from a variety of sources. However this approach also provides a serious governance challenge in that it is estimated that half of this data comes from sources not typically managed by IT. For example, analysts will pull from sources including CSV or text extracts from transactional systems, personal spreadsheets, third party reports, semi-structured content and more. Issues then arise around version control, data breaches, reconciliation, auditing and more. An enterprise data preparation platform is required to address these governance risks in a frictionless manner so the business users has the speed and agility they require.
Governing “Non-Managed” Data

Many organizations understand the need to create a content repository, which lays the foundation for governing "non-managed" data. The repository should include key governance such as:

- **Data Retention**  At a minimum you should have document version control for consistency, but you may also need to persist and archive the source data/documents to meet regulatory or business requirements.

- **Data Masking**  The most common cause of data breaches comes from internal employees. Data discovery tools are a great way for users to build and share information, but many times the underlying data is not protected and includes personally identifiable data (like Social Security numbers), personal sensitive data (like medical procedures) or commercially sensitive data. There are a number of industry and government regulations where noncompliance has an estimated cost of $5.5 million per breach and can even result in legal action against you and your organization. And the data can be un-masked by authorized users.

- **Data Lineage**  Because you have a repository for the underlying source content, you should then provide complete data lineage and drill right down to the cell level of the source document. This capability is critical when you need to audit or reconcile data.

- **Data Curation**  Frequently used data sources or automated data preparation routines should be shared with others based on user roles. This way you instill confidence and consistency in the data used to make important business decisions.

- **Role-based Access**  Prepared data sets should be segmented based on user roles to ensure that the right subset of data is delivered to authorized users.

- **Auditing**  A reputable system will provide complete audit logging and reporting.
Chapter 3  The Value of Extending Self-Service Data Preparation Through Automation

You know how difficult it can be to access, prepare and combine data from a variety of sources. But the real value from data discovery and advanced analytic tools comes only when you bring together the right information, in a timely and trusted manner. You know you need a self-service tool that will empower you to do this quickly, so you can spend less time preparing data and more time analyzing it. The value you bring to your team is in the insights you provide, not the mundane time spent in preparation. And you’re not the only person in your organization struggling with this task.

Adding automation to your data preparation drives significant time and cost savings while still preserving the agility that you need.
Prepared data sets can also be delivered to other systems like data warehouses or departmental data marts. The new data can be added to existing systems easily using industry standard database drivers to multiple different systems.

Frequently used data sources or automated data preparation routines can easily be shared and discovered with other users. This way you won’t be recreating the wheel - it already exists. Or you can take another trusted process or prepared data set and simply augment this data set for your own use.

When new source data becomes available, your data preparation process can automatically be invoked. For example, if you use data from customer invoices, you can direct the system to “listen” for any new invoices in a number of places, such as a content management system, a shared directory or even your email inbox. This way you’ll always have the most up to date information.