

The Value of Automated Data Preparation & Mapping for the Data-Driven Enterprise

Creating a sustainable bond between data management
and data governance to accelerate actionable insights
and mitigate data-related risks

Data Preparation & Mapping: Why It Matters

Business and IT leaders know that their organisations can't be truly and holistically data-driven without a strong data management and governance backbone.

But they've also learned getting to that point is frustrating. They've probably spent a lot of time and money trying to harmonise data across diverse platforms, including cleansing, uploading metadata, code conversions, defining business glossaries, tracking data transformations and so on. But the attempts to standardise data across the entire enterprise haven't produced the hoped-for results. How, then, can a company effectively implement data governance—documenting and applying business rules and processes, analysing the impact of changes and conducting audits—when it fails at data management?

In most cases, the problem starts by relying on manual integration methods for data preparation and mapping. It's only when companies take their first stab at manually cataloging and documenting operational systems, processes and the associated data, both at rest and in motion, that they realise how time-consuming the entire data prepping and mapping effort is, and why that work is sure to be compounded by human error and data quality issues. They can't afford any mishaps if they're to effectively promote business transformation, as well as to fulfill regulatory and compliance mandates.

It's obvious that the manual road is very challenging to discover and synthesise data that resides in different formats in thousands

of unharvested, undocumented databases, applications, ETL processes and procedural code. Consider the problematic issue of manually mapping source system fields (typically source files or database tables) to target system fields (such as different tables in target data warehouses or data marts). These source mappings generally are documented across a slew of unwieldy spreadsheets in their "pre-ETL" stage as the input for ETL development and testing. However, the ETL design process often suffers as it evolves because spreadsheet mapping data isn't updated or may be incorrectly updated thanks to human error. So questions linger about whether transformed data can be trusted.

The sad truth is that high-paid knowledge workers like data scientists spend up to 80 percent of their time finding and understanding source data and resolving errors or inconsistencies, rather than analysing it for real value. Statistics are similar when looking at major data integration projects such as data warehousing and master data management, with data stewards challenged to identify and document data lineage and sensitive data elements. So how can businesses produce value from their data when errors are introduced through manual integration processes? How can enterprise stakeholders gain accurate and actionable insights when data can't be easily and correctly translated into business-friendly terms? How can organisations master seamless data discovery, movement, transformation and IT and business collaboration to reverse the ratio of preparation to value delivered?

Data Preparation & Mapping: Why It Matters *(continued)*

What's needed to overcome these obstacles is establishing an automated, real-time, high-quality and metadata-driven pipeline that will be useful for everyone, from data scientists to enterprise architects to business analysts to C-level execs. Doing so will require a hearty data management strategy and technology for automating the timely delivery of quality data that measures up to business demands. From there, they need a sturdy data governance strategy and technology to automatically link and sync well-managed data with core capabilities for auditing, statutory reporting and compliance requirements as well as to drive business insights.

Data Preparation & Mapping Defined

Data preparation involves exploring, combining, cleaning and transforming raw data into curated datasets for self-service data integration, data science, data discovery and BI/analytics. It provides faster time to insight and improves trust.

Source:
Gartner

Data preparation includes data mapping, the process by which different data models are linked to each other using a defined set of methods to characterise the data in a specific definition. This definition can be any atomic unit, such as a unit of metadata or any other semantic. Data mapping is the first step in data integration.

Source:
Techopedia

Data Preparation & Mapping's Role in Data Management & Governance

- Discovers and synthesises data in different formats in thousands of unharvested, undocumented databases, applications, ETL processes and procedural code.
- Speeds the development of a real-time, accurate picture of the data landscape, including “data at rest” in data warehouses and data lakes and “data in motion” as it is integrated with and used by key applications.
- Improves data quality by depicting lineage and transformation to help facilitate strategic decision-making and meeting compliance mandates.

Data Preparation & Mapping: How to Do It

erwin Data Catalog (DC), a 100-percent metadata-driven and web-based solution, combines data management and data governance processes in an automated flow through the integration lifecycle from data mapping for harmonisation and aggregation to generating the physical embodiment of data lineage—that is the creation, movement and transformation of transactional and operational data. Its hallmark is standardising on a consistent approach to data delivery (business glossaries connect physical metadata to specific business terms and definitions) and metadata management (via data mappings). This is the path to getting more control over and more value out of all of a company's data.

Its critical capabilities are:

→ **erwin DC takes an automated metadata-driven approach to data mapping.**

Metadata can be automatically harvested from more than 80 data sources. The solution's automated approach to data mapping and management means that the metadata of source and target systems can be used to accelerate the creation of source-to-target metadata-driven mapping. There's no manual and time-consuming work to move things along from the pre-ETL stage, as analysts don't have to type in column names and business rules, nor are there worries about costly errors that can happen using traditional Excel-based approaches that require employee input.

Instead, a drag-and-drop, web-based environment makes it a snap to pick and choose from among sources and columns for mapping. As part of the picture, a central metadata repository offers a way to build and manage version mappings.

These features promote better data quality and standards surrounding the entire data mapping process, showing lineage and transformation of data across the enterprise, for easier manageability and more collaborative work efforts on mappings without overwriting or confusing multiple versions. When there's support for most ETL and data asset technologies with centralised metadata-driven mapping, data transformation and movement will be done in tight alignment with business-driven requirements from a single, unified platform.

→ **It eases the process of ETL code generation to accelerate ETL delivery.**

With **erwin DC**, staff members don't need to have specialized knowledge of database procedural code as well as specific ETL tools to handle code generation. That comes courtesy of the product's plug-in code automation framework that automates the generation of data integration components for a number of ETL solutions that an enterprise may already be using—such as those from Informatica, DataStage, Talend and erwin—based on job design best practices. At the same time, it automatically generates data movement and other data integration code and components, reducing development cycles as well as coding and logic errors.

Companies can easily reverse-engineer ETL/ELT components into mapping documents and lineage documentation; automate data vault hub, link and satellite code generation; and generate SQL code, stored procedures and DDLs from metadata to build warehouses and marts.

→ **It aligns with erwin's vision of role-based and role-aware access for users.**

erwin DC provides a meaningful user interface by publishing data to people and roles in ways that are germane to them—that is, which matter to their analytics processes. It keeps an eye on security, of course, controlling access by specifying what authorisation applies to which individuals in what roles.

Collaboration across roles is a staple, too, making it possible to bring together diverse project teams, including business analysts, data architects, ETL developers, testers and project managers, in a highly collaborative environment to facilitate faster decision-making and approvals.

→ **It makes it easier for business users to know what data the enterprise has, where it is, how it has changed through integration, and whether that affects any business or analytics processes in relation to data governance.**

It can be difficult for enterprises to have a complete view into and understanding of their data, including being able to reach down to its source origins and see changes to it over time—changes that could affect everything from gaining reliable business insights to conducting true audits. Concerns are minimised when data management capabilities such as a centralised design and repository of data elements mapped to their sources, automated tagging of sensitive data at rest or in motion, change activity logging and version mapping are available to support data discovery, visibility and upstream and downstream lineage. Impact analysis to show links from physical assets to business representations ranks high among the insight capabilities **erwin DC** provides, offering a way to instantly identify the impact of change to a warehouse attribute.

When stakeholders can take advantage of having access to end-to-end lineage between repositories and can view data flows from source systems to the reporting layers, including all intermediate transformations and business logic, they're poised to make better decisions more quickly.

→ **It integrates with erwin Data Modeler, Business Process and Enterprise Architecture to cover the chain of data management and governance.**

Data modeling benefits from **erwin DC's** ability to automatically map integrated and cataloged data to and from models, where they can be stored in a central repository for re-use across the organisation. It supports the goals of business process modeling to reveal the workflows, business capabilities and applications that are dependent upon particular data elements by ensuring that the data assets used by these functions are appropriately governed components of an integrated data pipeline—one that rests on automated data lineage and business glossary creation. It supports data flows and architectural diagrams within enterprise architecture by automatically assessing and documenting the current data architecture and providing and continuously maintaining business glossary ontologies and integrated data catalogs.

→ **It emphasises speed.**

Data management pros can expect to spend most of their time on data preparation—up to 80 percent. That can be significantly reduced when data mapping is automated throughout the enterprise data integration lifecycle. That is facilitated by **erwin DC's** web-based drag-and-drop approach that accelerates pre-ETL mapping. With **erwin DC**, mappings can be accelerated by 70 percent due to increased manageability and control over the process. With more accurate data preparation, a business has the ability to double the analysis it can do in shorter timeframes with fewer people. Overall, these functions drive faster time-to-value as well as greater accuracy for data movement projects.

→ **It seamlessly handles the relationship between data management and data governance.**

erwin DC enables the consolidation of all metadata from supported systems in one repository, synchronising and cataloging metadata assets and their flows across the enterprise with core data management capabilities in real-time in order to lay the groundwork for data governance.

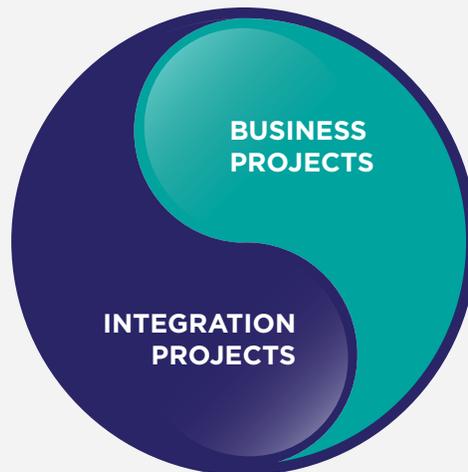
Its central metadata repository provides the foundation for the automated discovery of lineage and sensitive data elements. It also provides the means to managing data dictionaries. Its automated approach to gain better control and management of the data

mapping process determines master data sets, discovers the impact of potential business glossary changes across the enterprise; audits and scores adherence to governance rules, discovers risks, and appropriately and cost-effectively applies security to data flows. Automated ETL patterns that go beyond the spreadsheet and schema population ensure that the business is always ready to show an audit trail of the change process and clearly articulate who did what and when through the system development lifecycle, both to internal auditors or external parties conducting compliance reviews or performing statutory auditing. This is inclusive of Basel, HIPAA, SOX, Solvency 11 and other government regulations.

Use Cases for Data Mapping

INTEGRATION PROJECTS

- Data Warehouses/Data Marts
- Operational Data Stores
- ERP/Data Conversions (SAP, JDE, Oracle Applications, etc.)
- Data Migration/DBMS
- Re-Platform (SQL Server to Oracle, etc.)



BUSINESS PROJECTS

- Reporting Specifications
- HIPAA Code Conversion (Healthcare)
- ICD Code Conversion (Healthcare)
- Swift/ISO Conversion (Financial)
- Regulatory Compliance (GDPR, SOX, Solvency II ... etc)
- Operations Consolidation
- Business Process Management
- Services Oriented Architecture (SOA integration)
- Data Federation

INITIATIVES

- Master Data Management
- Metadata Management
- Data Governance & Data Quality

Data Management & Data Governance: Together Forever

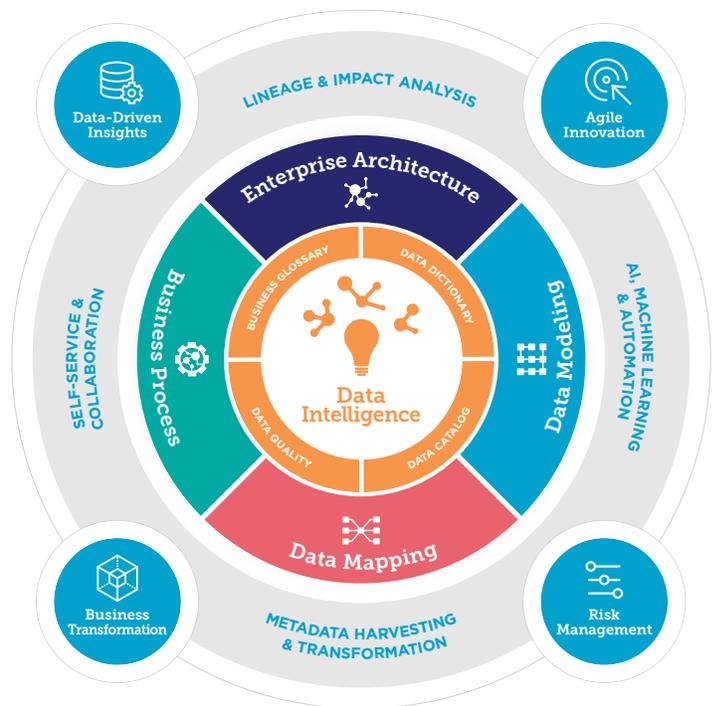
With **erwin DC** integrated into the **erwin EDGE Platform**, erwin now has the distinction of owning every critical piece of the data management and data governance lifecycle, streamlining customers' ability to gain value from these assets. The **erwin EDGE** platform has become the most connected and comprehensive data management and data governance solution available. It boasts the broadest set of metadata connectors and automated code generation, data mapping and cataloging tools, as well as governance enablers that take into account the need for consistency around business terms and enterprise policies, processes and tools to define and control access to data by roles and across workflows, which is critical for security purposes.

Working hand in hand, the data management and data governance capabilities of **erwin DC** fulfill the need for companies to have a real-time, accurate picture of their data landscape, including "data at rest" in databases, data lakes and data warehouses and "data in motion" as it is integrated with and used by key applications. And its features provide a way to control that landscape to facilitate insight and collaboration and limit risk.

With a metadata-driven, automated, real-time, high-quality data pipeline, all stakeholders can access data that they now are able to understand and trust and which they are authorised to use. At last they can base strategic decisions on what is a full inventory of reliable information. The integration of data management and governance capabilities also supports industry needs to fulfill regulatory and compliance mandates, ensuring

that audits are not compromised by the inability to discover key data or by failing to tag sensitive data as part of integration processes.

Data-driven insights, agile innovation, business transformation and risk management are the fruits of enterprise modeling, data preparation/mapping and data literacy capabilities around a data governance hub. Only erwin connects all the critical enterprise data and business architectures for the intelligence required to reduce risks and realise results.



The erwin EDGE's integrated capabilities produce the data intelligence modern enterprises need to fuel their initiatives.

An Overview of erwin Data Catalog

Automates the Data Mapping Process

- Metadata + Data Mapping Repository
- Drag-and-Drop, Build-and-Version Data Mappings

Enables Control, Visibility & Traceability

- Groups and consolidated STMs
- Mapping Comparison and Impact Analysis

Integrates with Leading ETA Tools

- Informatics | Datastage | SQL Server SSIS



As the data governance company, erwin provides enterprise modeling, data cataloging and data literacy software to help customers discover, understand, govern and socialize their data to mitigate risks and realize results. The erwin EDGE platform facilitates IT and business collaboration in driving actionable insights, agile innovation, risk management and business transformation. We help government agencies, financial institutions, healthcare companies and other enterprises around the world unlock their potential by maximizing the security, quality and value of their data assets.

Connect with us at
sandhillconsultants.com



© 2019 erwin, Inc. All rights reserved. All trademarks, trade names, service marks, and logos referenced herein belong to their respective companies.