Solidatus – Ingestion Methodology

Summary

Solidatus is an award-winning web-based data lineage application which enables teams to collaboratively map and analyse their data landscape at scale.

The rich suite of connectors and the Solidatus API enable organisations to finally fully automate data lineage documentation from various sources, including data governance tools, spreadsheets, data dictionaries, databases and ETL tools. All changes are versioned, audited and can be augmented with manually maintained metadata and lineage.

The powerful governance and version control features of Solidatus enable and facilitate enterprise-wide collaboration, giving subject-matter experts (SMEs) across an organisation the power to take ownership of the definition and maintenance of metadata within their respective areas of expertise. This may then be shared to ensure consistency, comprehensive understanding and adoption across the entire firm, providing an essential foundation to best-practice data governance.

The Methodology

Solidatus has been uniquely designed to model the most complex of systems with truly and engineering focus to ensure rapid ROI for clients as well as simplicity and longevity of use. Many systems claim fully automatic lineage of code and current systems landscapes, this is a nirvana that cannot and will not ever be achieved. The reason it can never be achieved is the same reason that it was not automatically coded or created. The sheer complexity that exists across programming languages, the limitless coding styles and levels of proficiency of programmers coupled with the infancy of the software development discipline and the constant change has led to the complex lineage issues we now face.

Instead, Solidatus approaches the issue of modelling highly complex systems with the same methodology that Google use for populating Google Maps, arguably one of the most challenging mapping tasks of all – documenting the Earth. Rather than claim automatic mapping, Google employed the following technique which we mirror.

Firstly, *technical ingestion* involves connecting via technical means to as many disparate sources as required and harvesting metadata and lineage data. In Google's case, this is ingesting and digitising traditional street maps and street view data. A list of technical ingestion methods that Solidatus has natively or offers through expert technical partnerships is listed at the end of this document. This extensive list of data connectors, as well as Solidatus' uniquely simple interface, allows the technical ingestion phase to be accelerated by up to 60%. Solidatus provides real modelled data in hours and days, not weeks and months.

The second phase of modelling is the *collaborative ingestion* by subject matter experts. Following the Google Maps analogy, this is utilising the hundreds of millions of users (which they call *local guides* not SMEs) to add and keep up to date the non-technically available data for their known local neighbourhoods. For example, Google Maps asks its local guides questions such as, "does this restaurant have parking?" Solidatus uses the same approach, utilising the organisation's SME's to capture and keep up to date data they are familiar with and puts the responsibility of managing change in the hands of the people making change. This divide and conquer methodology gives organisations centralised control with de-centralised execution.

Finally, the Solidatus metadata repository, with its audited and versioned workflow, gives analyst, developers and business users the ability to support waterfall, agile and DevOps change methodologies. In addition, the Solidatus API provides developers with programmatic access to all of the functionality

exposed through the GUI. This allows microservice architectures to connect, push and pull data from Solidatus automatically without the need for manual intervention.

List of connectors

Direct connectors

Excel/CSV	Structure (e.g. schemas) Lineage (e.g. mappings) Metadata (e.g. properties)
Solidatus JSON schemas	Structure, lineage and metadata
XML	XML file structure and XML attributes
JSON	JSON file structure and values
Databases	JDBC/ADO.NET/ODBC – e.g. most database vendors Scans database and extracts database tables, columns and metadata. Can detect lineage from tables to views for some database vendors (e.g. SQL Server) Examples: ▪ Teradata Database ▪ Oracle Database ▪ Microsoft SQL Server ▪ SAP ASE (Sybase) ▪ Hive ▪ IBM DB2 ▪ Impala ▪ PostgreSQL ▪ Cassandra ▪ InterSystems Caché ▪ Cloudera ▪ MySQL ▪ Neo4j ▪ IBM Netezza
Collibra DGC	Full end-to-end one-click integration with Collibra to pull from, visualise, manipulate and push data to Collibra DGC in bulk. Import communities, domains, assets and attributes into Solidatus to automatically layout and visualise a community's constituent domains according to the relationships between assets. Mass manipulation, socialisation and validation of assets in a continually version controlled environment before pushing back to Collibra DGC.
Google Cloud	DataFlow BigQuery PubSub Cloud Storage Data Catalog
Informatica ETL mappings	Scans ETL mappings Imports source and target structure Imports transformational lineage

Microsoft SSIS	Scans ETL packages Imports source and target structure Imports transformational lineage
SPARQL	Connects to graph databases over SPARQL endpoints Executes user defined SPARQL queries Imports structure, lineage and metadata from RDF data Example: ▪ Stardog ▪ OpenLink Virtuoso ▪ Blazegraph ▪ MarkLogic ▪ Amazon Neptune
Autosys Process Automation	Parses Autosys JIL files Produces batch job flow diagrams Imports job execution metadata Shows job lineage
Control-M Process Automation	Parses Autosys JIL files Produces batch job flow diagrams Imports job execution metadata Shows job lineage
Solidatus API	Allows organisations to fully automate the data lineage documentation from various sources, including data governance tools, spreadsheets, data dictionaries, databases and ETL tools. This is versioned and audited and can be augmented with manual annotation and collaboration.
Apache Pig	Import an Apache Hadoop Pig script to automatically load full attribute-level end-to-end lineage of the job with all associated metadata.

Partner integrations

Partner 1: **Code scanning and integrations**	Informatica PowerCenter Talend Java SAP IS Sqoop COBOL (Coming soon) Informatica EDC Informatica IMM Collibra DGC IBM IGC TopQuadrant EDG
Partner 2: **SQL code scanning**	Teradata PostgreSQL Microsoft SQL Server Oracle Amazon Redshift Snowflake Greenplum Apache Spark Netezza Hive Cloudera IBM DB2 MySQL SAP HANA
Partner 3: **Mainframe scanning**	Adabas ADS/OL C++ CICS CSD CICS Tables COBOL Delta DL/I Easytrieve Focus Fortran (!) Ideal IDMS IMS JCL Link Decks Load Modules Mantis Model 204 Natural PL/1 SQL Syncsort Telo

Partner 4: **Complex Systems Scanning**	SAP Salesforce Siebel PeopleSoft JD Edwards Oracle E-Business Suite MS Dynamics AX 2012