ERP and CRM metadata for data catalogs: the three challenges

Why they can delay or even derail your Data Catalog project and what you can do about it

Roland Bullivant
Silwood Technology Limited

Contents

Introduction

An increasing number of enterprises are turning to data catalogs to enable them to document, understand and share information about data and how it is used across their information and application landscape. A data catalog has become a critical component of an effective enterprise information management strategy because of its usefulness in supporting data governance and compliance, reporting, analytics and master data initiatives.

> *"Data catalogs enable data and analytics leaders to introduce agile information governance, and to manage data sprawl and information supply chains, in support of digital business initiatives."*
>
> Gartner's 2017 Report, *'Data Catalogs Are the New Black in Data Management and Analytics'*

In the past, some more forward thinking businesses have seen the value in being able to maximise their understanding and use of data. It is only lately however that initiatives which attempt to record and categorise this data have become more mainstream. Much of this has been motivated by increased requirements to be compliant with data and security legislation (e.g. the EU's GDPR or CCPA in California, USA) as well as a result of the growing acceptance that data is an important asset whose value can be better managed and exploited.

Data catalogs should provide a central location for storing information about an organisation's data assets. In addition they should also deliver tools and processes for utilising, enriching, managing and valuing that information. Rather than contain actual data such as sales numbers or production and distribution statistics, a data catalog delivers value by providing a mechanism for technical and business users alike, to make use of the metadata, or data structures which underpin their source systems.

In this respect, the ability to identify and import metadata ("data describing data") to a data catalog is vital.

Every organisation has applications which store data important to it and which could be considered for inclusion in a data catalog. It is common for larger organisations to have hundreds or perhaps thousands of applications and systems that they have acquired over the years. Sources can include unstructured data in the form of text documents and others, RDBMS, files, graph databases, geospatial data and packaged business applications. These contain data vital to the business and their metadata will ultimately need to be brought into the data catalog.

Therefore, during the early stages of a data catalog project, it is important to prioritise which data sources should be those whose metadata are sourced first. Eventually metadata from all the relevant sources will need to be incorporated into the data catalog. All good catalog solutions have a variety of scanners and connectors for identifying and mapping metadata from many different sources.

"Metadata is the core of a data catalog. Every catalog collects data about the data inventory and also about processes, people, and platforms related to data."

Eckerson, '*Ultimate Guide to Data Catalogs*'

In the Eckerson Group's, '*Ultimate Guide to Data Catalogs*':

"The initial build of a data catalog typically scans massive volumes of data to collect large amounts of metadata. The scope of data for the catalog may include any or all of data lakes, data warehouses, data marts, operational databases, and other data assets determined to be valuable and shareable. Collecting the metadata manually is an imposing and potentially impossible task. The data catalog automates much of the effort using algorithms and machine learning to accomplish the following:

- *Find and scan data sets.*
- *Extract metadata to support data set discovery.*
- *Expose data conflicts.*
- *Infer semantic meaning and business terms.*
- *Tag data to support searching.*
- *Tag privacy, security, and compliance of sensitive data."*

There are a variety of other methods of capturing metadata, for example via crowd sourcing or catalog curator-based activities. You may decide to review system documentation, search the internet for information about source application metadata or even hire consultants to help.

However, there are also various classes of system whose valuable metadata is not accessible or usable by those normal methods. In Eckerson's '*Ultimate Guide to Data Catalogs*' they refer to these as "challenging data sources".

Amongst the most difficult of those are the large, complex and often highly customised ERP and CRM packages from SAP, Oracle, Microsoft and others.

They hold large amounts of transaction data critical to the success of many thousands of businesses. It is important therefore, that their metadata should be included in the data catalog.

You might have noticed that the topic of how to provision ERP and CRM metadata into a data catalog during presentations or demonstrations is often ignored, especially by those vendors for whom it is difficult.

Here are the reasons that, without specialist software tools, this task is so difficult to accomplish quickly, accurately and cost-effectively.

> *"The data in these (ERP) systems makes sense and are useful, but only in the context of the hard-coded processes.*
>
> *In short, the data is trapped inside a complex web of thousands of database tables whose integrity is solely controlled by a rigid fossilized collection of software algorithms.*
>
> *If you don't believe me, just ask your SAP support staff for access to directly update (or even read) a data table."*
>
> ***John Schmidt (vice president of Global Integration Services at Informatica Corporation)***

ERP and CRM metadata: the three challenges

1, Discovering the metadata (where is it?)

Given that most ERP and CRM packages are based on a relational database platform you might assume that it would be fairly straightforward to use a database scanner to pull out all the table and attribute names.

In principle this should work, however, there are two key characteristics of ERP and CRM packages which make this an unworkable solution.

Lack of meaningful metadata in the database system catalog

The first of these is that there is no meaningful metadata, by which I mean 'business names' and descriptions for tables and attributes and no table relationships defined in the database system catalog.

The result is, that even if a data catalog scanner reads the schema of the application database, the results are of little value. Users would be faced with trying to decipher what the physical names for tables and attributes mean as well as how tables are related without any information about primary and foreign key constraints.

To illustrate the point, how would your typical data catalog user or data analyst know what this SAP table TF120 contains from this information?

SAP Table Name	***Field Names***
TF120	MANDT
	ITCLG
	ITLGH
	BLIND
	AREIND
	SETGENMODE

The size of the data model makes it difficult to navigate and find what you need

The second main challenge with this approach is that ERP and CRM packages have large or very large data models. For example, SAP has over 90,000 tables, a typical Oracle eBusiness Suite implementation has over 20,000 tables and even JD Edwards systems have in excess of 4,000 tables. In addition these are not fixed numbers because the great majority of implementations have had customisations made to the data model in terms of tables and attributes being added and/or amended.

It is unlikely that scanning and importing so many tables into a data catalog's metadata repository is necessary, valuable or even viable. For example it is quite possible for many thousands of tables to be unused which means that they are probably unnecessary for the catalog.

To compound that, adding thousands of tables to a data catalog with no business names or descriptions means that your target users will derive no benefit from them.

Over the past few years, we have noticed that even organisations with large Salesforce landscapes are challenged by a lack of mechanisms for identifying and sharing relevant metadata about their systems with solutions such as data catalogs.

In theory, it should be possible to reverse engineer the database system catalog using a data modeling tools such as SAP PowerDesigner, erwin or ER/Studio. However, these are still only able to access the physical names for tables and attributes and cannot cope with such large quantities of data.

Another method of discovering metadata, could be to search through whatever documentation is available. One challenge with this approach is how to be certain that the information is up to date and reflects any customisations.

Finding the valuable metadata

ERP and CRM packages typically store their useful metadata in a series of data dictionary tables in the application layer of the product. As one would expect there are differences between ERP and CRM vendors in the types and detail held, however, as a minimum they all contain business names and descriptions for tables and attributes. With the exception of JD Edwards they all contain the information to determine how tables are related. To find out how tables are joined in JD Edwards it is necessary to infer that information from the Business View layer in the application.

There is often other information about Views, Domains etc., or other data which for example allows for different kinds of application hierarchy to be constructed from the information in the Data Dictionary tables.

If it is possible to access and query the data dictionary tables then it is possible to construct SQL, or possibly ABAP, queries in the case of SAP, to extract their contents to populate a database or perhaps a spreadsheet.

But, what do you do with the metadata once you have done this?

This leads on to the second of the three problems associated with making effective use of ERP and CRM metadata through search, analysis and scoping.

2, Metadata analysis and scoping (how can I find what is relevant to me?)

The objective of this part of the process is to quickly and accurately identify which tables represent the topics you need for your data catalog and to ensure that there is sufficient business centric information included to make it of value to users.

What methods and tools are there for locating the tables and related tables needed?

Vendor tools

Each packaged application vendor (SAP, Oracle, Microsoft and Salesforce) provide some tools which can be used to find tables and related tables. They can also provide the business descriptions which are so vital to making the data catalog of value to the end user.

These tools however are not designed for use by data analysts or architect. They do not provide the global search, introspection and filtering capabilities necessary for true metadata exploration of such large and complex systems. They are commonly only used by technical application specialists and do not provide an intuitive interface into the metadata.

Unless the vendors partner with a specialist metadata discovery software provider, data catalog software products will struggle to provide you with access to a navigable set of accurate metadata which includes the customisations in your system. They may claim to be able to connect to some or all of these packages and to import their metadata. However, without having access to all the metadata for a package and easy yet comprehensive analysis and scoping facilities, it will be necessary to know what you are looking for before you start. This can put your data catalog at risk of delay or even failure.

Documentation and templates

As mentioned above, searching for documentation, assuming it is up to date and accurate would be a good option. However, there are some limitations with this approach. For example, finding relevant tables and related tables which are used by specific business topics from tens of thousands may take some time, especially if the documentation is not structured in a way that makes this feasible.

Using application or data catalog vendor templates may also appear to be a solution, however, these are unlikely to reflect the customisations you have had made and may not even be based on the same version of the software you are using. Identifying the differences would be a frustrating and time-consuming task.

In addition, these templates do not cover all aspects of the applications data model so there are likely to be significant gaps between what you need and what they provide.

Using external consultants

One common tactic is to employ external consultants to perform this work for you. It frees up your own staff and hopefully the resource you hire have specialist knowledge about the applications under scrutiny. Sometimes these consultants may be part of the data catalog vendor's implementation team or its partner organisation and sometimes they may be independent.

This can be an expensive and risky undertaking. It means that your own staff are not in control and that when changes or rework is required you need to hire consultants again.

It may also be necessary for the consultants to have access to some application vendor tools to perform this work.

Internet search or guesswork

It is possible to try to locate data models from your ERP or CRM application package using internet search.

This presents its own challenges as the results may not be as accurate as necessary, perhaps because the versions are different and obviously anything found will not reflect your customisations.

Anything found in this way is often part of documentation. This means that they are static and there is no way to make use of that information in your data catalog without rekeying information.

Finally, it is often necessary to ask a technical specialist to interpret the model and augment it with relevant information, perhaps about table joins.

Using metadata extracted into a spreadsheet or database

Even if you have managed to extract all the metadata from your application and load it into a database or spreadsheet you will initially be faced with task of trying to formalise the relationships between tables and also how tables are related to views and domains. You may also want to know, for example, which tables are accessed by specific transactions or programs or other components of the application which represent specific business concepts.

All of this represents a considerable amount of work and lost time on the project.

One other challenge you may encounter is if you have multiple instances of the same ERP such as SAP and you are not sure if their data models are exactly the same. The problem of analysing their metadata for the data catalog is much worse. Salesforce customers seem to experience this proliferation of instances more often. For example, we have worked with one organisation who has 50 separate Salesforce 'orgs'.

If you use any of the above methods and techniques for analysing and scoping the application metadata there is still the remaining task of provisioning it into the data catalog solution.

3, Delivering ERP and CRM metadata into a data catalog

It is critical that the metadata from these packages can be provisioned into the data catalog solution.

However you access and scope the metadata, if it is not in machine readable form then the only options you have are to hand key it or perhaps copy and paste the information into the data catalog or into an intermediate file or other staging area. This is time a consuming, costly and potentially inaccurate method.

Imagine for example trying to rekey the information from a single SAP table, MARA (General Material Data) into a data catalog. It has over 240 attributes (columns) each with a business description and is related to over 1500 other tables in the SAP system. We normally estimate that it takes about 1 day to rekey the metadata from 5 tables accurately into another system.

Clearly you will need metadata from more than a few tables in your data catalog so this is not really a viable option.

Obviously the quickest, most effective and accurate method is to import it using whatever mechanisms are available. Depending on which data catalog you are implementing, this may be via API or scanning or importing files in various formats.

However, this is not necessarily a simple task, especially when the metadata is complex as you may also need to understand the data structures of the data catalog in order to know how to create the source to target mappings necessary. If you have large quantities of metadata, there may be issues with speed of processing. Also, if you have not been able to access and analyse the rich metadata in the application then loading physical table and column names into the data catalog is of dubious value.

What would be better?

What is required to quickly and easily enable these three steps (Discovery, Analysis and Scoping, Delivery) is a single product which accesses and extracts the rich metadata from wherever it resides, allows for the selection of relevant metadata based on the needs of the business and delivers the selected metadata into the data catalog.

> *"We took what would have been months of work (and possibly a barrier to progress) and completed the activity within hours and nominal resource investment. We're pleased to have been able to achieve a high degree of efficiency with this collaborative effort."*
>
> ***Wellington Holbrook, ATB Chief Transformation Officer discussing the impact of using Safyr to provision Collibra with SAP metadata.***

Safyr: a single software solution for discovery, scoping and delivery of ERP and CRM metadata

This is exactly what our software product Safyr does. It is a unique, specialist self-service metadata discovery product for ERP and CRM packages. It ensures that your data catalog is a comprehensive and inclusive solution that is populated with rich metadata from these complex systems.

Referring back to the example of Table TF120 (actual name Consolidation Charts of Accounts) it makes much more sense when the contents are presented in your data catalog as in the illustration below. Data analysts and users will then be able to use this information more easily in the context of the business scenarios under review.

Table Name	***Short Description from SAP***	***Long Description from SAP***
TF120	Consolidation Charts of Accounts	
Field Names	***Short Description***	***Long Description***
MANDT	Client	A legally and organizationally independent unit which uses the system.
ITCLG	Cons chart of accts	Enter a consolidation chart of accounts.
ITLGH	Output item length	Maximum length of the financial statement items in the consolidation chart of accounts. Specify a length of 10 or fewer places.
BLIND	Lock indicator	Set this indicator if you want to specify that the master records of FS items in this chart of accounts can no longer be modified. In this way, you can ensure that no unwanted changes are made to item definitions. You can unlock the consolidation chart of accounts at any time.
AREIND	ARE bal. sheet	Appropriation of retained earnings in the balance sheet. If you do not select this flag, the appropriations are shown in the income statement.
SETGENMODE	Set generation	You determine when an automatic generation of the sets for cons groups/ totals items should take place, via this indicator.

Recommendations

If your organisation is seeking, or engaged in implementing, a data catalog and you have one or more ERP or CRM solutions then it is important to make sure you understand how their metadata will be accommodated in your data catalog.

- Ask prospective vendors how they can provide you with fast, accurate access to the rich metadata in your SAP, Oracle, Microsoft and Salesforce packages. Remember scanning the RDBMS System Catalog is not a solution because there is no metadata of value there.

- Ensure that they can show you how their tools will allow you to search, navigate, analyse and save relevant subsets of that metadata quickly and easily.

- Ensure that you have detailed discussions about how to provision the data catalog with metadata from your ERP and CRM packages with the vendors during the product selection phase. Ask them to demonstrate how this is achieved.

Sometimes the data catalog vendors may propose some of the approaches discussed above. These might include using the package vendors own tools, crowdsourcing, engaging internal specialists or external consultants, or even using prebuilt templates. These have some value, however, they all provide additional challenges in terms of the amount of time and cost this adds to the project, the level of accuracy as well as physical activity of mapping and importing the metadata into the catalog.

Some vendors already partner with Silwood and so they will be able to help you with this process. They have realised the value to their customers of providing a comprehensive metadata discovery capability which enables ERP and CRM packages to form part of a data catalog solution. The technical integration mechanisms may differ from vendor to vendor, however, the outcome in terms of speed, accuracy and reliability of metadata import is the same.

> "Silwood Technology is unique in having recognised, and acted on, the need for a better understanding of enterprise applications, notably the dominant ERP and CRM solutions provided by SAP and Oracle.
> Having this depth of understanding is critical for many enterprise projects, ranging from business intelligence to data governance, from data integration to master data management, and from data migration to application development"
>
> **Philip Howard,**
> **Bloor Research**