

【数据治理实践】第九期：数据时代的导航——数据目录

前言

[上期文章](#)，我们分享了银行数据质量管理的相关实践。本期，我们将分享讨论**数据资产目录**。说到数据资产目录必然需要提到元数据，而一旦提及元数据，可能大家会认为这是一项技术性的工作，但**其实这是一个业务与技术共同参与的管理性工作**，只是因为数据存储的特点，会让人联想到系统、数据库等，容易理解成IT的工作。下面我们将从监管要求，元数据概念，数据资产目录与元数据的关系，以及数据资产目录的价值体现来进行分析介绍。

监管要求

在《银行业金融机构数据治理指引》（下称“指引”）第三章的数据管理中有下列的要求：

第二十一条 银行业金融机构应当持续完善信息系统，覆盖各项业务和管理数据。**信息系统应当有完备的数据字典和维护流程，并具有可拓展性。**

第二十三条 银行业金融机构应当加强数据采集的统一管理，明确系统间数据交换流程和标准，实现各类数据有效共享。

——《银行业金融机构数据治理指引》

监管明确要求，信息系统应当有完备的数据字典和维护流程，并具有可拓展性。但是并没有对“数据字典”进行明确的定义，**狭义上讲，数据字典包括数据库，表结构与码表码值的定义；广义上讲，数据字典包括银行所有数据的定义、流向、分布，也就是银行的元数据管理。**

什么是元数据

元数据一般解释为“**描述数据的数据**”(catalog or schema)，所以其本质也是一种数据。元数据与非元数据之间的边界其实并非那么明确，在概念上，这个边界与数据所代表的抽象级别有关。例如一则客户的投诉录音是数据，而描述这份数据的元数据就会包括“电话号码”、“录音时长”等信息，而在结构化信息处理的时候我们通常也将“电话号码”和“录音时长”作为数据进行分析。所以哪些是数据，哪些是元数据，需要从企业管理的数据对象来看。

一般来讲，元数据可以分为**业务元数据、技术元数据**和**操作元数据**（也有分类为管理元数

据）。

那么这些分类分别是什么呢？

- **业务元数据**一般包括报表指标的定义说明、业务规则、数据标准等；
- **技术元数据**一般包括物理数据库的表和字段，数据库的ETL（抽取、转换和装载）作业信息；
- **操作元数据**则包括审计结果、错误日志、调度异常日志等。

为了理解元数据在数据管理中的重要作用，我们可以想象一下在一个大型图书馆中有成千上万的书籍。如果没有一个目录索引，或者主题的分类，读者很难找到一本特定的书籍或者是某一类型的书籍。同样对于银行来说，我们将数据类比为书籍，业务与技术人员类比为读者，要想在海量的数据中快速的找到自己需要的数据，那么自然离不开对元数据的管理。

没有可靠的元数据，银行就不知道自己有什么数据，数据表示什么，数据来自何处，数据如何在系统中流转，谁可以访问这些数据，更不用说要将数据作为资产进行管理。所以简单而言，开展元数据管理工作是开展数据资产管理的基础。

如何管理元数据

因为元数据的范围太广，而且银行管理元数据的边界并非都很明确，究竟要管理哪些元数据，如何去管理，会让不少人心生困惑。一般来讲，可以从元数据应用的角度来反向思考，例如需要构建数据目录，以支持未来的取数用数的查找和探索；需要根据数据质量问题追溯数据发生的源头；需要根据数据定义与数据标准进行对标等。明确具体的元数据应用以后，分析支撑这个应用需要管理和维护的元数据有哪些，再进行这些元数据的采集和存储。

元数据应用层	数据目录	数据标准对标	数据影响分析	敏感信息识别
	数据血缘分析	数据质量问题定位	数据地图	...
元数据管理存储层	元数据版本管理	数据标准存储	库表结构存储	敏感信息规则管理
	元模型管理	数据质量规则关系管理	信息系统架构存储	...
元数据采集层	业务术语	数据标准	库表结构	敏感信息规则
	ETL加工映射关系	数据质量规则	信息系统基础数据	...

只有从元数据的应用角度开展管理工作，才能避免为管理元数据而管理元数据，此外对于管理范围、管理目标也都比较明确，短期的成果也将切实可见，业务部门或者是相关的技术部门才能更有动力持续的维护和管理元数据。限于篇幅，本文以数据目录的应用为例，展开介绍元数据的管理。

数据目录

一般银行所构建的数据目录是服务业务部门与数据分析团队，此数据目录必须：

- 1) 是业务人员所熟悉的业务场景、表单与数据分析场景；
- 2) 是客观反映银行目前数据现状的；
- 3) 是可扩展的支持未来取数用数的；

我们可以归纳数据目录的一些特点：



以上特点任意一个没满足，都会影响用数人员的体验，导致元数据不能很好地支持应用，进而减少元数据管理的运维投入，导致进入恶性循环。

例如缺少灵活的多视角的数据查找，用户不能通过期望的路径或者方式，无法找到所需的数据在什么位置，那么自然而然就会弃用。虽然用户的用数场景不胜枚举，但一般可以结合现有的业务流程和数据场景进行分析提炼，此外提供相应的标签服务，维护元数据对应的业务标签，也可以快速通过标签，关联查找出用户所需的数据位置。

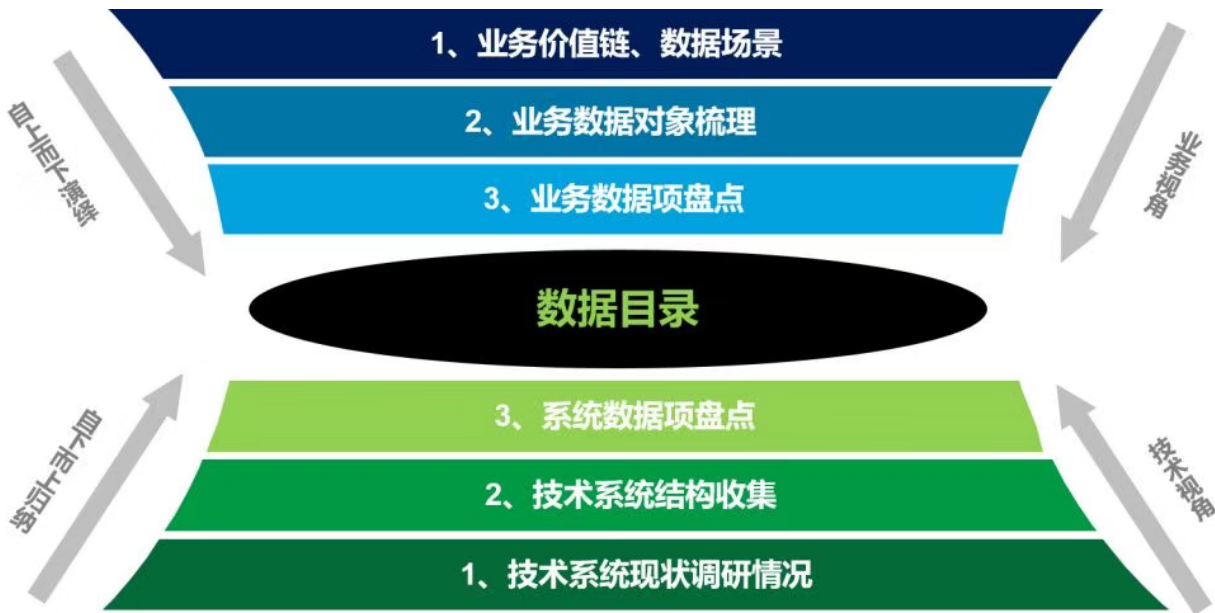
此外，互联网时代，用户已经习惯了关键词的组合搜索，如果一个数据目录的工具无法支持这种灵活的查询搜索，甚至是二次检索，自动推荐，也将逐渐被用户淘汰。

元数据的客观真实性，更不用多说，设想一个读者在图书馆，根据索引在某一个主题区的书架上找了半天，但是这本书根本不存在，无疑会很沮丧。再如根据地图搜索一个位置，而这个位置早已消亡，那么这种地图还会有人用么？所以对于元数据的管理也是如此，银行需要及时、真实、客观的反应信息系统实际的情况，维护和更新这个数据目录，保障用户可以找到自己需要的数据位置。

最后一个特点，支持用数，其实是元数据管理的“最后一公里”。一个流畅的取数用数的场景是，用户有一个想法，立即去检索这些数据，找到数据所在的位置以后，直接获取甚至适用数据。所以对于用户来讲，需要的是“一站式”解决。

如何构建数据目录

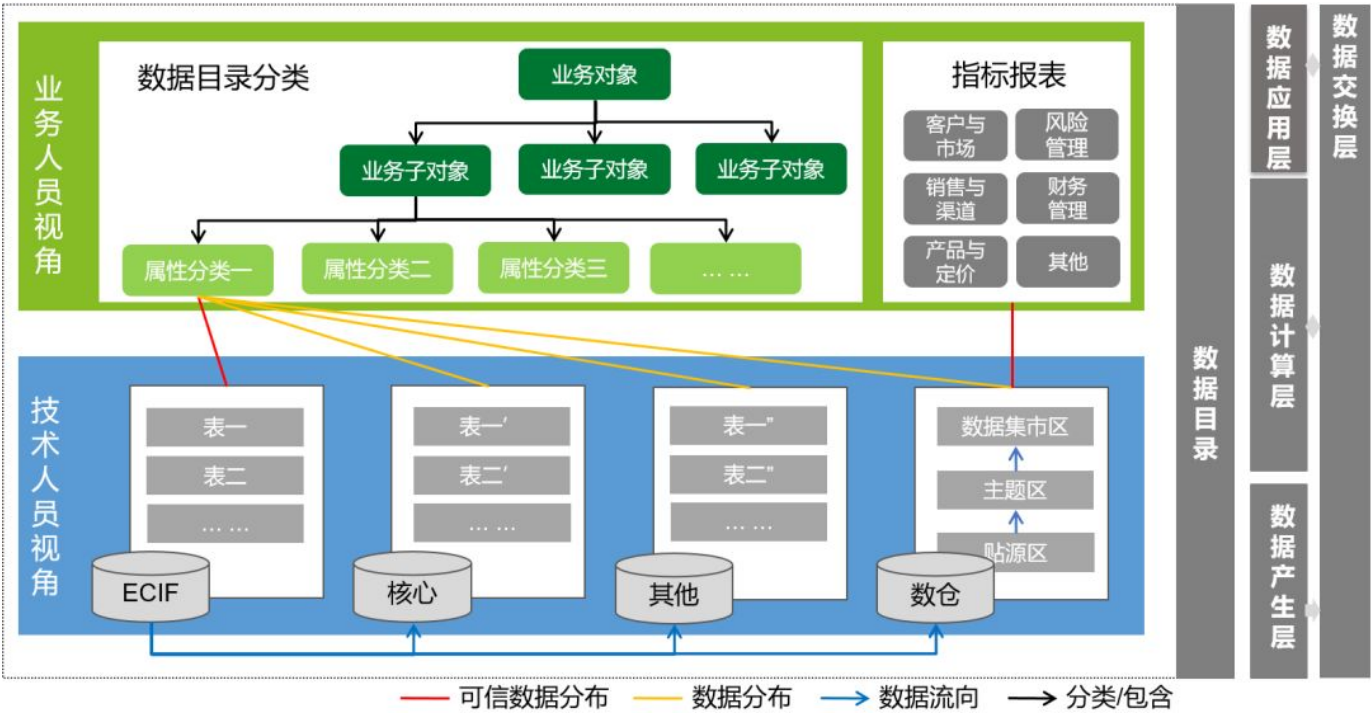
数据目录可以通过业务角度的自上而下演绎，以及结合数据角度自下而上归纳进行构建。



自上而下的演绎，是确保数据目录可以按照业务的视角进行组织的，可以关联业务的场景，包括业务流程中用到表单，术语等。自下而上的归纳，是为了满足数据目录真实客观，保证数据目录呈现的内容可以在真实的信息系统中查找到，可以最终映射关联到精确的信息项。

构建的数据目录可以是多视角的，一种方式就是通过概念对象组织呈现，如客户对象、个人客户对象、对公客户对象。每个对象拥有不同的属性分类，而这些分类与属性分别分布在不

同系统的表和字段中。此外数据目录还可以通过如业务域，安全分类等不同的视角或者标签进行呈现。



元数据管理工具

数据在各个重要系统中存储、流转、加工、应用，由于高度复杂的银行业务以及庞杂的信息系统，如果纯粹依靠人工的方式对各个库表结构，ETL关系等元数据进行采集建设，那么将投入极大量的人力和物力进行维护和管理，因此一个便捷的元数据管理工具是元数据成功有效管理的重要条件。元数据管理工具可包含元数据的采集，版本管理，元数据的应用如数据目录，血缘分析，影响分析等功能。具体元数据管理工具我们将在以后的篇章中着重分享。

结语

我们从信息时代进入数据时代，元数据的管理比任何时期都要来的更为迫切和重要。在海量的数据中，元数据的管理就如大型图书馆中的导航索引，也如一张城市的地图。管理好元数据，也就为数据资产管理打下坚实的基础。下期文章中，我们将分享数据安全实践，敬请期待。