西安电子科技大学
XIDIAN UNIVERSITY

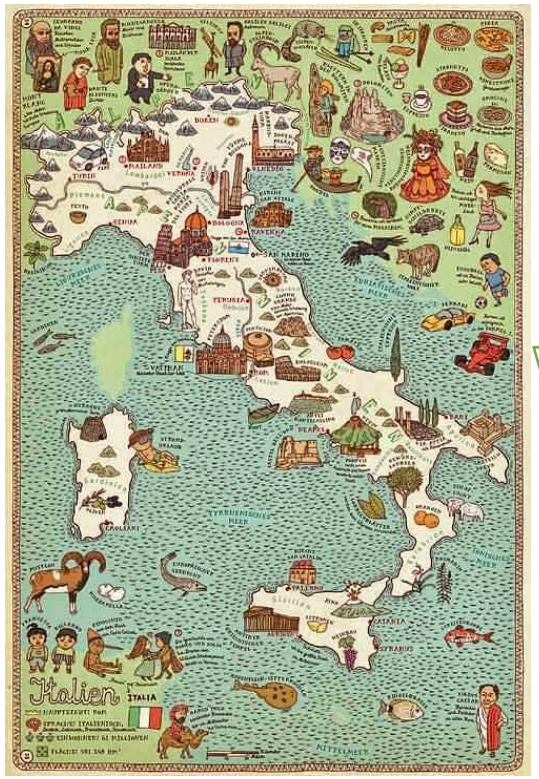
第三章 元数据管理

马 晶

经济与管理学院 信息管理系

Email: majing@xidian.edu.cn

元数据（meta data）：描述数据的数据



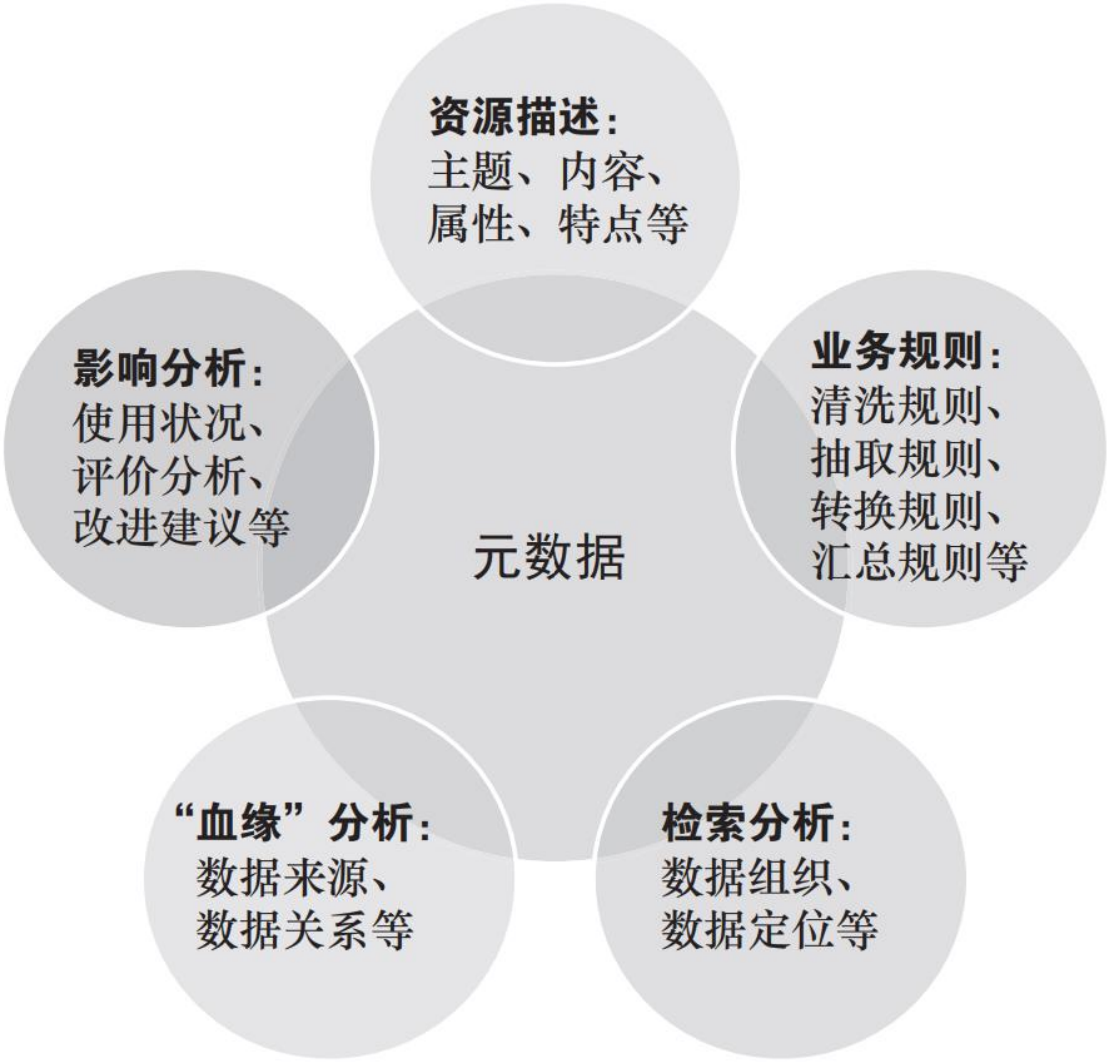
这是一张地图

地图类型、地图图例，包括地图名、空间参照系统和坐标、地图内容说明、比例尺、精度、出版单位、发布日期、销售信息等

元数据是描述**数据属性**的信息，其主要目的是帮助组织、整理和查找相关的信息和资源。

元数据描述的不是特定的实例或记录，IT部门和业务部门都需要高质量的元数据来**理解现有数据**。元数据不仅表示数据的类型、名称、值等信息，还提供**数据的上下文**描述，比如数据的所属业务域、取值范围、数据间的关系、业务规则、数据来源等。

元数据可能涉及的范围


资源描述：
主题、内容、
属性、特点等
业务规则：
清洗规则、
抽取规则、
转换规则、
汇总规则等
元数据
影响分析：
使用状况、
评价分析、
改进建议等
“血缘”分析：
数据来源、
数据关系等
检索分析：
数据组织、
数据定位等

元数据的组织方式

- **从数据集的视角出发，为每个数据集存储对应的元数据文件，文件中包含与对应的数据集相关的全部元数据。**

不足：当数据集非常多时，需要存储大量的元数据文件，这给元数据的索引、管理和关联操作分析带来不便；当数据不断演变时，元数据可能发生变化，多个独立的元数据文件会给数据演变时的维护工作造成困难。

- **从元数据管理和可扩展性的角度出发，建立元数据库，专门存储和管理元数据。**
- 线路1：将元数据构建成结构化数据库表，例如，数据库表的一列代表元数据的一个要素，行代表一条元数据内容。

属性名	业务描述	精确度
每股收益	每股收益即每股盈利（EPS），又称每股税后利润、每股盈余，指税后利润与股本总数的比率。	0.01

不足：与简单的元数据文件相比，建立元数据库的代价通常更高，包括前期设计、建立，以及后续的维护和更新；灵活性受限，如元数据异构性较强（不同数据集的元数据要素类型差别较大）或结构比较复杂，构建元数据表比较困难；元数据本质上是结构化数据库表的一行，也需要用户能够接受和使用这种形式。

元数据的组织方式

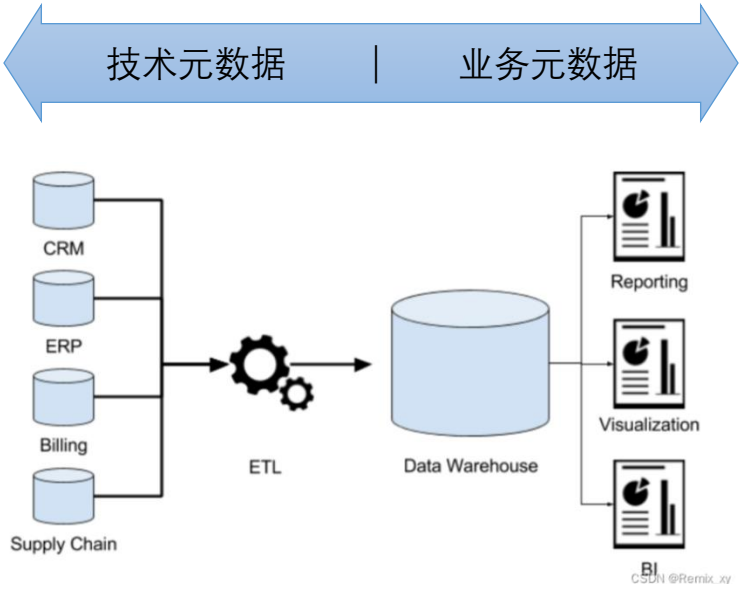
- 线路2：为解决结构化数据库表灵活性受限的问题，使用资源描述框架（Resource Description Framework, RDF）。RDF是万维网联盟（W3C）提出的一组标记语言的规范，采用“主-谓-宾”的形式表达资源描述和网络资源的内容与结构。经常使用XML文件的形式表示。

```
<? xml version="1.0"? >
<RDF>
    <Description about="http://www.xxxx.com/">
        <标题>大数据治理：理论与方法</标题>
        <作者>王宏志，李默涵</作者>
        <出版年份>2021</出版年份>
</Description>
</RDF>
```

- 当前企业和组织要处理的数据量越来越大，元数据库易于管理和高可扩展性的优势更加凸显。
- 元数据库既可以通过统一的格式和规则提供对数据的重要信息的一致描述方便统计和管理；也可以作为一类数据索引，实现数据集成，以及关联分析。

作用和意义—描述、定位、检索、管理、评估、交互

- 对**底层的数据集成**，元数据定义了多元异构集成所必需的关键信息。
- 对**模型设计构建者**，元数据提供了沟通上下层次的关键信息。
- 对**数据使用者**，元数据提供了帮助理解数据的关键信息。
- 对于**随时间演变的数据**，元数据有助于维护良好的数据质量。



从企业和组织层面，元数据管理应重视以下几点：

1. 确立清晰完善的元数据管理和维护策略，基于科学的方法和专业的工具来管理元数据。
2. 维护清晰一致的元数据指向关系，将元数据与正确的对象链接在一起。
3. 建立增量式的元数据构建和维护体系，可以基于多种方式采集和更新元数据。

元数据分类

- **业务元数据：**存储业务角度的数据描述，提供了沟通终端用户和实际系统之间的语义层，帮助不了解系统技术细节的业务人员顺利使用数据，包括业务规则、业务术语、指标定义、系统使用业务语言等。
- **技术元数据：**存储技术角度的数据描述，提供数据的关键技术细节，存储数据的系统以及在系统内和系统间数据流转的过程，包括结构信息描述（如接口信息、表信息等），好数据处理的相关描述（如程序信息、存储过程、函数等）。
- **操作元数据：**描述处理和访问数据的细节，是数据处理日志及运营情况的数据，描述数据的操作属性，包括管理部门、责任人等，是数据安全管理的基础。

业务元数据

- 数据集、表和字段的定义和描述
- 业务规则、转换规则、计算公式和推导公式
- 数据模型
- 数据质量规则和核检结果
- 数据的更新计划
- 数据溯源和数据血缘
- 数据标准
- 有效值约束
- 利益相关方联系信息（所有者、管理员）
- 数据的安全/隐私级别
- 数据问题
- 使用说明

- 业务元数据必须真正代表业务层对数据的理解，其关系到对数据的需求是否准确地被实现，以及数据架构是否产生漂移。
- 业务元数据的术语来源可能比技术元数据更加多样化，包括ERP系统中的业务逻辑和规则，各类业务报表，数据模式（计算公式、填写说明等）产品说明书（型号、等级、价格、成分等）。系统的需求文档中往往也包含了大量元数据。

例：关于点击率的元数据

名称： 点击率

定义： 网站页面上某一内容被点击的次数与被显示次数之比

应用场景： 在网络广告中，点击率是在网页上的一条广告打开后被点击的次数百分比，例如，如果该网页被打开了1000次，而该网页上某一广告被点击了10次，那么该广告的点击率为：1%。

相关数据源： S135、S136、S137

业务元数据

实践要点

1. **自上而下的建立业务术语体系**

包括企业数据模型的高层信息、整个企业的业务概念、不同粒度的概念的层级从属关系、不同类别信息间的相互联系等。

2. **在自动化抽取元数据的同时也需要人工保证元数据的准确性**

业务元数据往往存在于非结构化数据中，包含的语义信息往往比较复杂专业，自动化抽取很难保证完全准确，因此，需要专业人员予以修正和确认。

3. **为业务元数据做好恰当的标记**

标记业务元数据和技术元数据的关联关系，可以将业务元数据链接到相关的技术元数据和原始数据中去，达到索引数据的目的。

4. **必要时设立专门的管理者**

在必要时，企业和组织应当专门设置管理者，负责识别和管理业务元数据。

技术元数据

- 物理数据库表名和字段名
- 字段属性
- 数据库对象的属性
- 访问权限
- 数据CRUD规则
- 物理数据模型，包括数据表名、键、索引
- 记录数据模型与实物资产之间的关系
- ETL作业详细信息
- 文件格式模式定义
- 源到目标的映射文档
- 数据更新的调度计划和依赖
- 恢复和备份规则

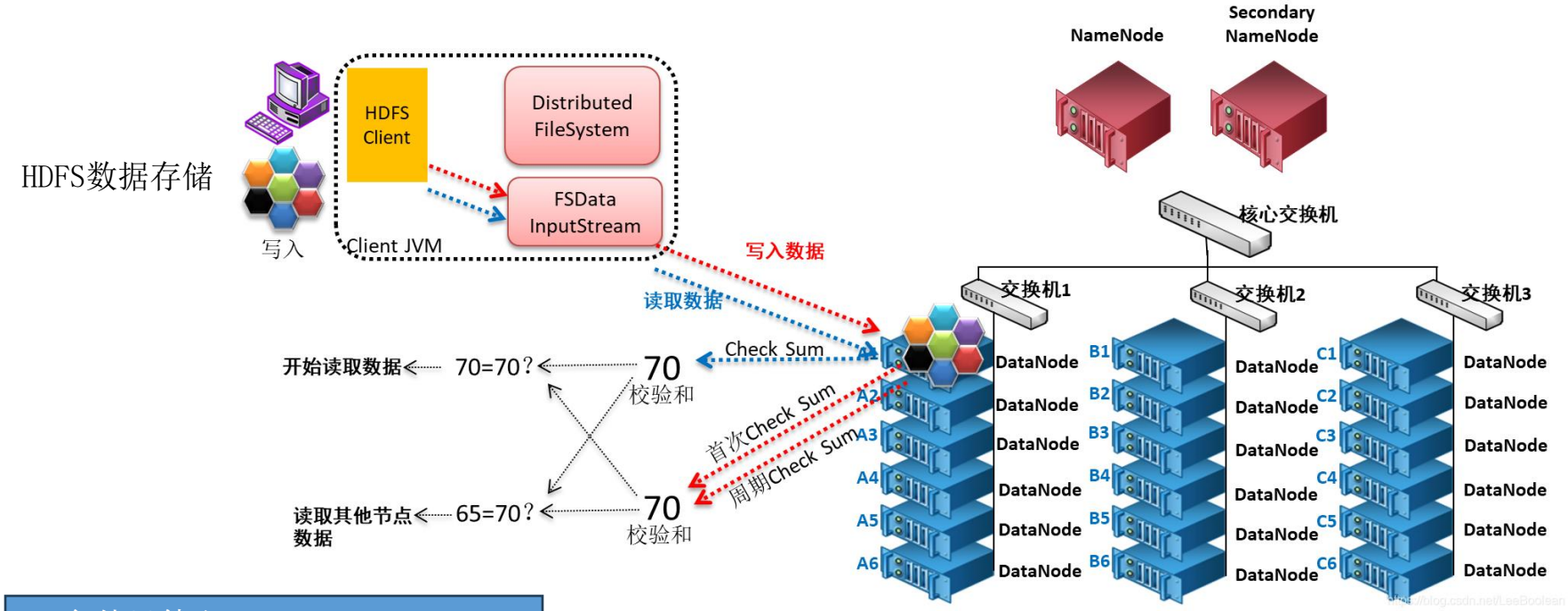
技术元数据是系统设计与管理人员在开发和日常管理数据仓库时需要的元数据，其既包括一些静态的技术参数，也包括一些动态参数。技术元数据可能以各种样式各异的结构存在于系统各处。

(a) 系统分层结构

```
if((wouldToggleZenMode(ringermode) && checkCallerSystemOrSamePackage(caller) &&
checkAccessPolicy(caller)){
        throw new SecurityException(…)
}
```

(b) 安全性检查策略

技术元数据



- 文件是什么
- 文件被分成多少块
- 每个块和文件是怎么映射的
- 每个块被存储在哪个服务器上面

操作元数据

- 批处理程序的作业执行日志
- 抽取历史和结果
- 异常处理
- 错误日志
- 补丁和版本的维护计划和执行情况
- 备份、保留、创建日期和灾备恢复预案
- 服务水平协议要求和规定
- 容量和使用模式
- 数据归档、保留规则和相关归档文件
- 数据共享规则和协议
- 技术人员的角色、职责和联系信息

根据使用场景或资源类型，对元数据还有其他分类方式。如

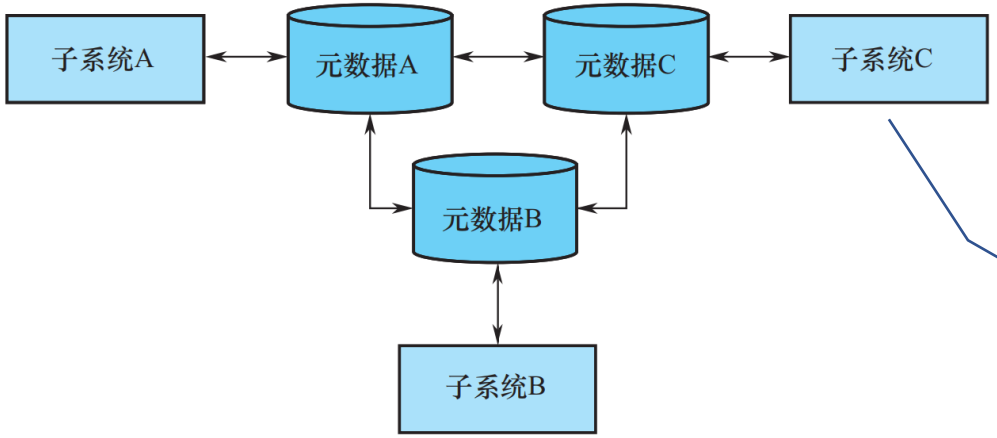
- 描述性元数据、管理性元数据、和应用性元数据
- 网络资源元数据方案、数字图书馆元数据方案、政府信息元数据方案等。

元数据管理

根据元数据的管理方式，可以将元数据的管理方案分成3类：**分布式、 集中式、联邦式。**

分布式

早期各系统相互独立，具有较高的异构性、自治性、冗余性。连接各子系统时，将元数据集成到一起会导致很高的开销且容易出错，因此企业或组织会选择依旧由各子系统各自维护自身的元数据，并提供必要的共享机制。

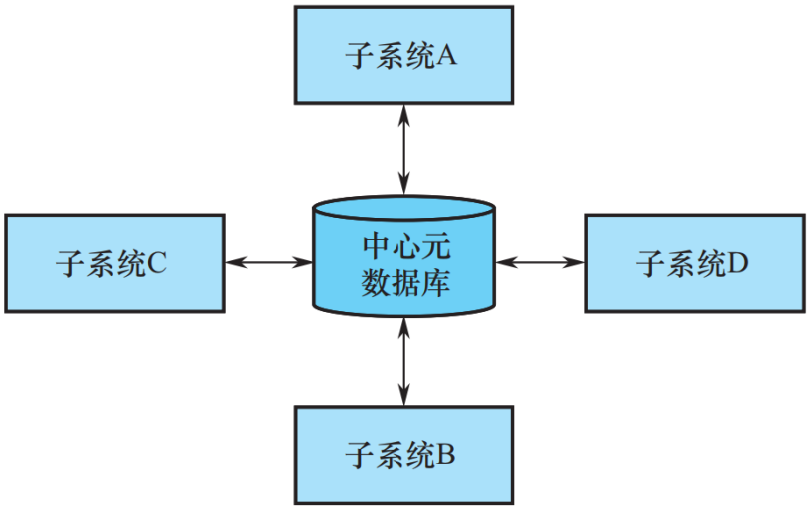


优点：开销相对较小；可以以比较自然的方式管理元数据，元数据能够和其关联的数据源或数据集一起管理；不易因元数据集成而导致错误。

缺点：元数据异构性强，元数据共享通道设计较为复杂；元数据冗余性高，更新时易导致元数据不一致；需要很多的元数据交换接口，维护不易。

集中式

集中式元数据管理可以克服分布式元数据管理的一些缺点，其设计了一个集中式的数据管理组件，专门用于管理元数据。该组件通常称为中心元数据库，负责管理和发布所有子系统的相关元数据，提供元数据的统一表达，每个子系统直接从中心元数据库获取元数据。

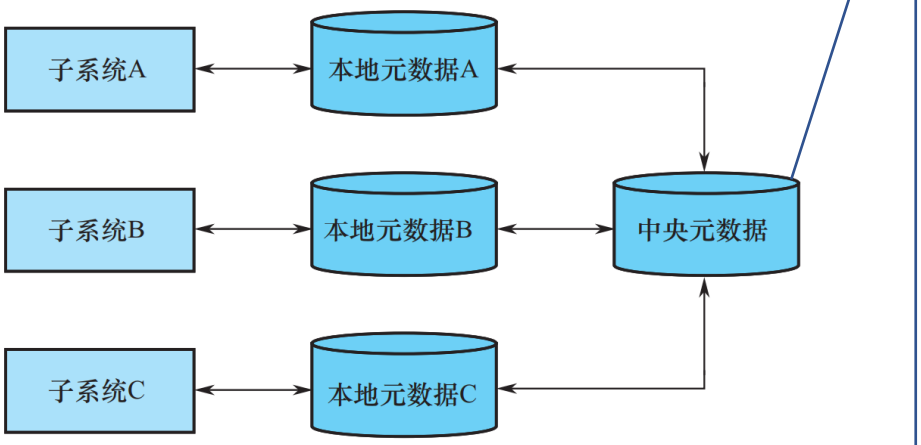


优点：数据通道的数量较分布式大为减少，只需要维护每个子系统到中心元数据库的数据通道，元数据本身的共享在元数据库内部即可完成；降低了元数据的冗余性；可以提供元数据的标准的、一致的表达。

缺点：对于中等规模的企业而言，中心元数据库足够维护和管理其涉及到的所有元数据，但对规模较小的企业和规模较大的企业而言，集中式则存在一些问题。

联邦式

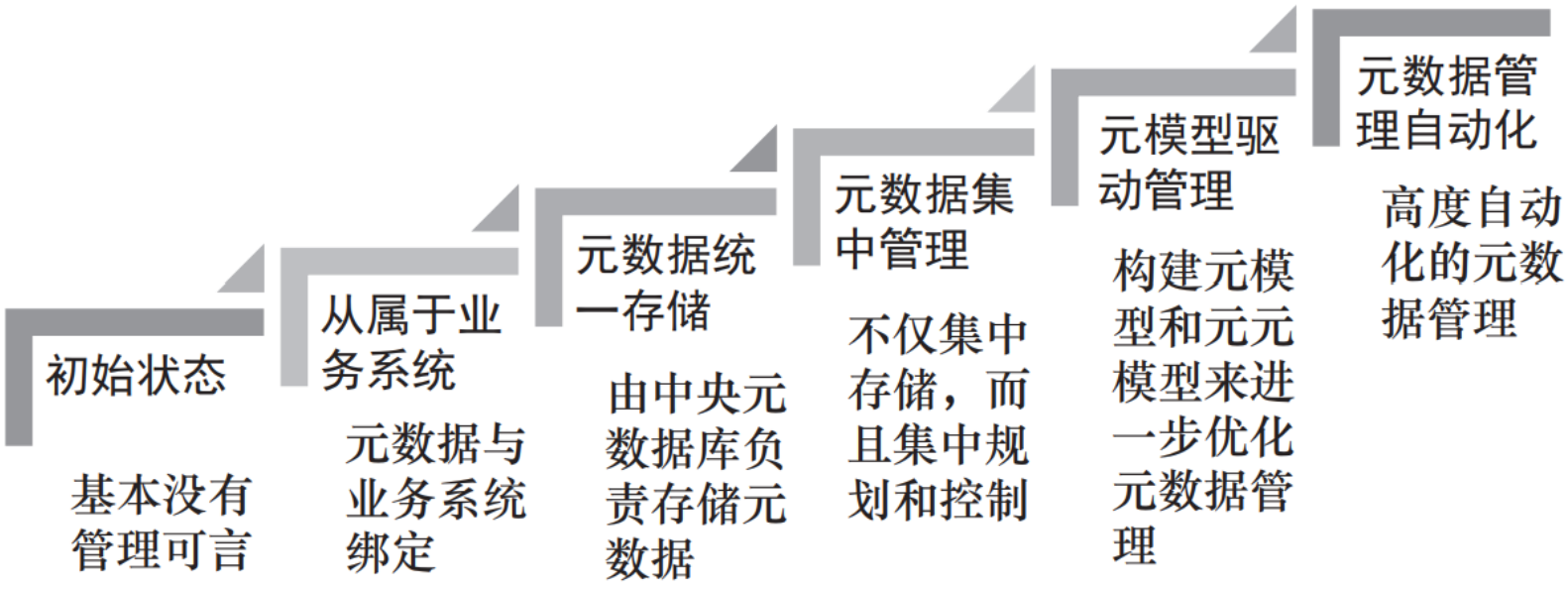
联邦式元数据管理尝试综合这两类方案的优点，同时保留了本地元数据和共享（中心）元数据机制，建立了层级式的元数据管理方案。



优点：保留了分布式元数据管理方案的一部分特点，使得不同类别的元数据可以相对独立、自主的被管理；共享元数据提供了元数据的规范的、一致的表示；元数据通道相对比较简单（相对分布式而言）。

缺点：系统相对前两种方案更为复杂，需要考虑的细节也更多，因此对系统的设计质量要求更高，成本也相应会有所增加。

IBM元数据管理的成熟度模型



数字公共图书馆的元数据

元素分类	元素名称
源资源	替代标题、集合、贡献者、创建者、日期、描述、范围、格式、识别符、语言、地点、发行者、关系、替代品、代替、版权所有者、主题、子类型、时间范围、标题、类型
网络资源	文件格式、版权声明、IIIF 清单、IIIF 基 本URL
集合	聚合的源资源、数据提供者、数字资源原始记录、观点、中间提供者、显示、对象、预览、提供者、版权说明
版权说明	版权说明、定义、注释

美国政府资源索引服务定位记录

美国政府资源索引服务（government information locator service，GILS）是一种支持公众搜寻、获取和使用政府公开信息资源的分布式信息资源利用体系。借助该索引服务可以定义元数据、获取政务信息的描述信息，并且能够获取到使用政务资源的方式。GILS定位记录是该索引服务使用的元数据标准，由若干核心元素组成，并含有应用系统自定义元素与具体应用系统相容的其他Z39.50属性规范中定义的元素，主要用来描述信息资源的内容、位置、服务方式、存取方法等。

相关维度	元素名称
资源内容信息	题名、资源语言、摘要、规范主题索引、非规范主题索引、空间域、数据来源、方法、补充信息、目的、机构项目、参照、记录语言
资源管理信息	目录号、控制标识符、原始控制标识符、记录源、最后修改日期、记录审查日期
资源发布信息	出版日期、出版地、获取方式、时间
资源的责任者	创始者、贡献者、获取限制、使用限制、联系点

我国政务信息资源目录体系GBT 21063.3- 2007

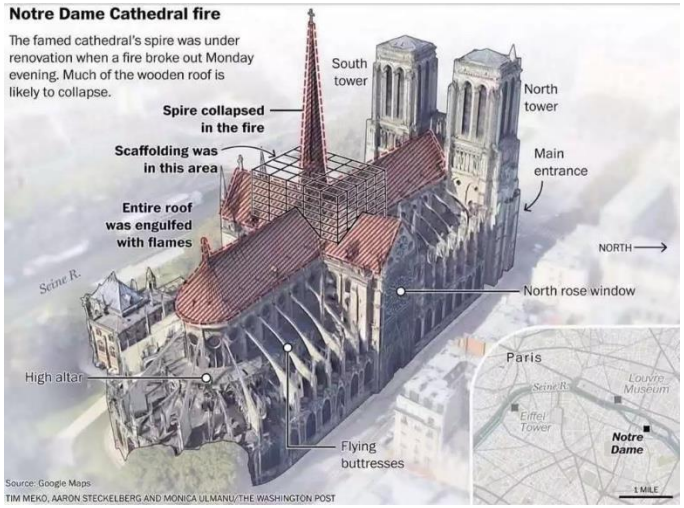
政务信息资源目录体系GBT 21063.3-2007由国务院信息化办公室在2007年提出，其中第3部分的主题是核心元数据。GBT 21063.3-2007的第3部分规定了描述政务信息资源特征所需的核心元数据及其表示方式，给出了各核心元数据的定义和著录规则，用以描述政务信息资源的标识、内容、管理等信息，并给出了核心元数据的扩展原则和方法，适用于政务信息资源目录的编目、建库、发布和查询。

元数据	说明
	成部门：外交部、发展改革委等；“目”下设置“细目”，由政务部门自行编制部门信息资源分类，可根据需要设置多级分类。地方层面，“项”之下按省（自治区、直辖市）和计划单列市展开。
2.信息资源名称	定　　义：缩略描述政务信息资源内容的标题。 数据类型：字符型。 注　　解：必选项；最大出现次数为1。 说　　明：缩略描述对应政务信息资源具体内容的标题。
3.信息资源代码	定　　义：政务信息资源的唯一不变的标识代码。 数据类型：字符型。 注　　解：必选项；最大出现次数为1。 说　　明：信息资源代码规则详见6.3。按照6.3.1分类码、6.3.2顺序码和附录3政务信息资源代码结构的规则进行编码。
4.信息资源提供方	定　　义：提供政务信息资源的政务部门。 数据类型：字符型。 注　　解：必选项；最大出现次数为1。 说　　明：具体提供信息资源的部门和单位，原则上中央政务部门细化到司局或所属行政事业单位，地方政务部门细化到内设机构和所辖政务部门。
5.信息资源提供方代码	定　　义：提供政务信息资源的政务部门代码。 数据类型：字符型。 注　　解：必选项；最大出现次数为1。 说　　明：代码采用《国务院关于批转发展改革委等部门法人和其他组织统一社会信用代码制度建设总体方案的通知》中规定的法人和其他组织统一社会信用代码。信息资源提供方代码采用资源分类“项”中的政务部门代码，而非部门内设机构和直属单位代码。
6.信息资源摘要	定　　义：对信息资源内容（或关键字段）的概要描述。 数据类型：字符型。 注　　解：必选项；最大出现次数为1。 说　　明：对资源内容进行概要说明（或关键字段）的描述。

借助数据让文物遗产重生

巴黎圣母院火灾过后，法国政府宣布将对它进行重建。其中一份重建的希望，就躺在美国瓦萨学院已故建筑历史学家安德鲁·塔隆的硬盘里。塔隆从2011年就已经开始精确测量这座教堂。他记录的数据点超过10亿个，生成的模型能够描述出巴黎圣母院最微小的细节，包括它的缺陷，误差只有大约5毫米。 这些数据最终超过70T。

思考：元数据管理在这其中的作用？



本章要点

- 理解元数据、业务元数据和技术元数据的概念与意义
- 熟悉元数据管理方案
- 理解元数据在不同行业和领域中的应用