上海图书馆
上海科学技术情报研究所
Shanghai Library

浅析第三代LSP的元数据管理
以Folio的Codex为例

上图 Codex & MM SIG | 夏翠娟

2019年6月6日

主要内容

图书馆元数据管理的需求与现状

01

- 元数据管理的需求
- 标准规范的沿革
- 元数据管理的现状

Folio的元数据管理

02

- 微服务架构
- 最小化原则
- Codex与元数据管理

上图Codex & MM SIG的工作

03

- 任务
- 进展
- 计划

01

图书馆的元数据管理

需求与现状

图书馆元数据管理的需求

A

完整的生命周期管理

采集（采购、捐赠）-创建-编辑-删除-更新（众包）-交换传输-长期保存

B

互联网环境下的书目控制与规范控制

FRBR（LRM），Linked Data,Knowledge Graph

C

资源类型兼顾

纸质图书报刊，购买的数据库，数字化特藏资源

D

标准规范兼容

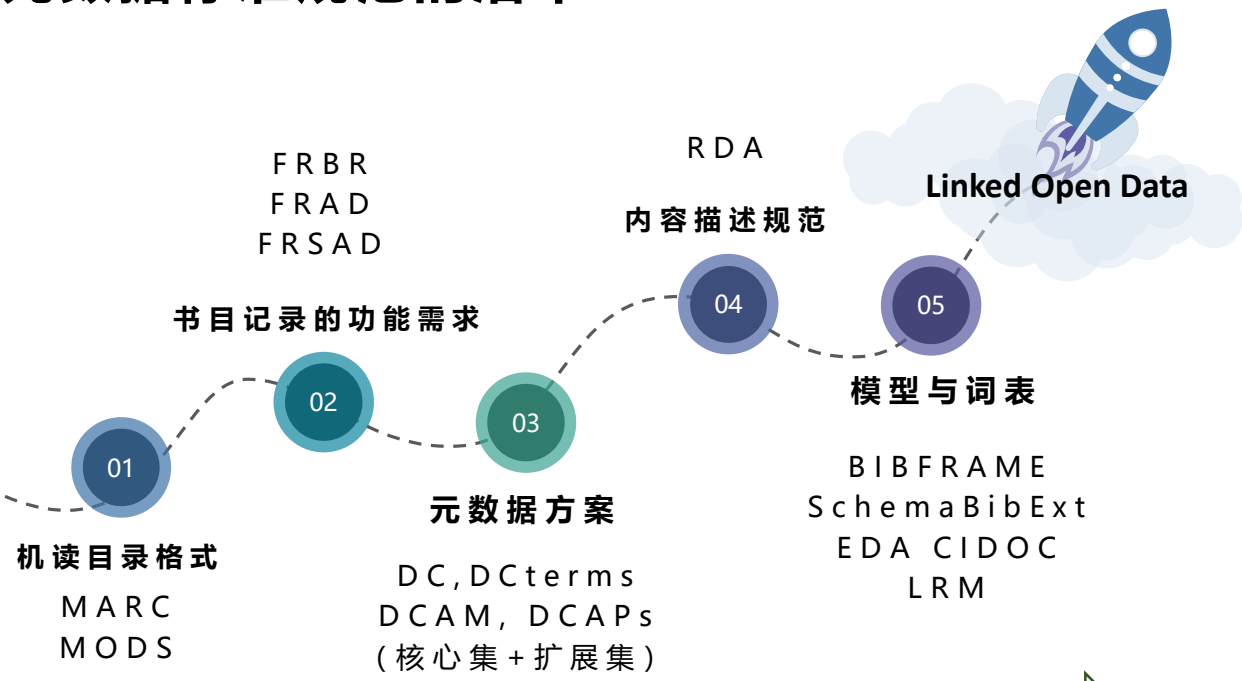
MARC/2709、MARC/XML、DC/XML、DC AP/XML、BIBFRAME/RDF

E

知识组织

描述文献的物理特征，揭示文献间的关系，基于内容的知识组织

元数据标准规范的沿革



从印本到多媒体，从单一标准到应用纲要，从文献描述到知识组织，
从机器可读到语义互操作，从专业领域到开放互联… …

元数据管理的现状

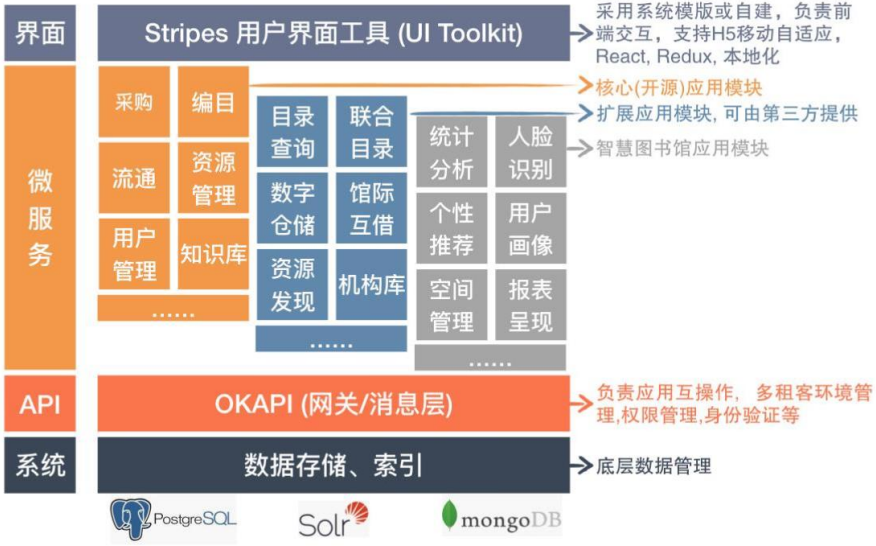
馆藏书目	购买的数据库	自建特色库	未编文献
MARC格式在ILS中	格式未知 由厂商掌握	格式多样 散落各处	?

02

Folio的元数据管理

Codex的愿景和现状

Folio的微服务架构

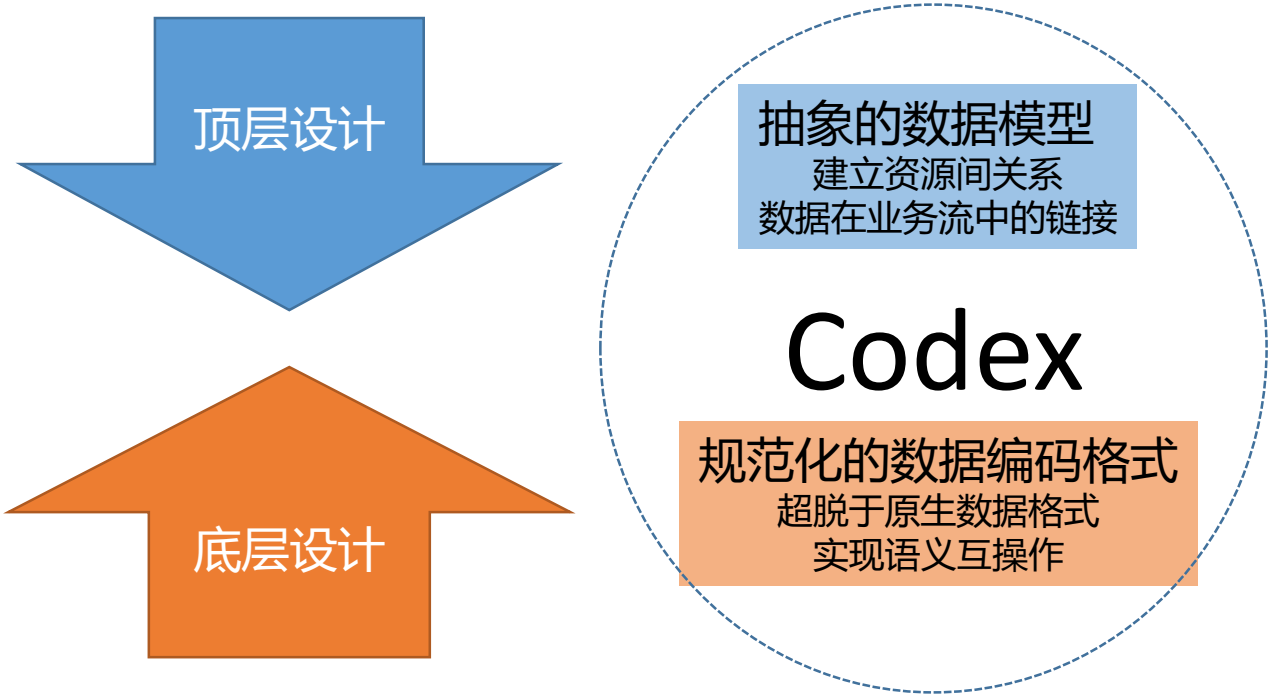


数据存储：如何解决跨库检索问题？

API：如何解决语义互操作的问题？

APPs：如何建立App间数据关联关系？

用户界面：如何实现知识导引和知识融合？


顶层设计
底层设计
抽象的数据模型
建立资源间关系
数据在业务流中的链接
Codex
规范化的数据编码格式
超脱于原生数据格式
实现语义互操作

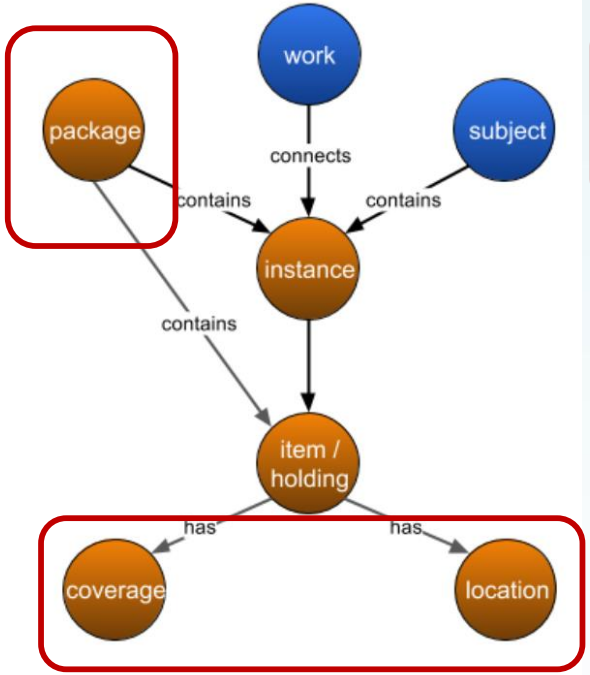
Codex是什么?

- Codex的本义
  - Codex起源于拉丁文，原意是“木块”、“书”，后引申为由若干纸、牛皮纸、纸莎草或类似材料构成的书，现专指古代手抄本。虽然，从技术上来讲，codex一种书籍装订方式。此外，Codex还可表示 “法典”、“药典”之意，**两者都是对所属行业内容的规范性汇编，具有指导与约束作用。**
- Codex在Folio中的作用
  - 是抽象的统一的数据模型（Data Model）
  - 是元数据方案 （Metadata Schema）
  - 有一套最小的元素集（Metadata Elements）

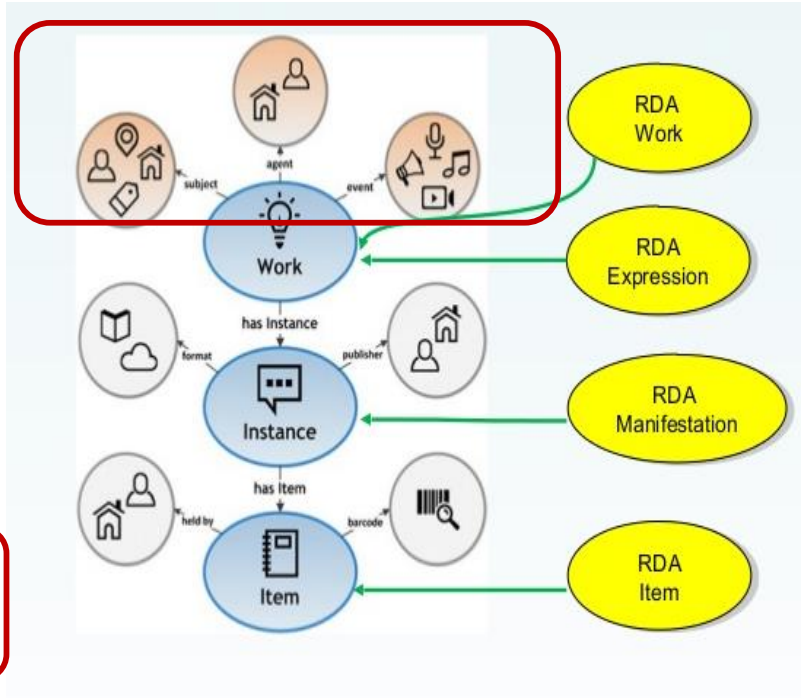
翻译：元数据规范、典范、法典，简称元典

Codex作为抽象数据模型

Folio Codex



BIBFRAME 2.0



Codex与BIBFRAME2的关系

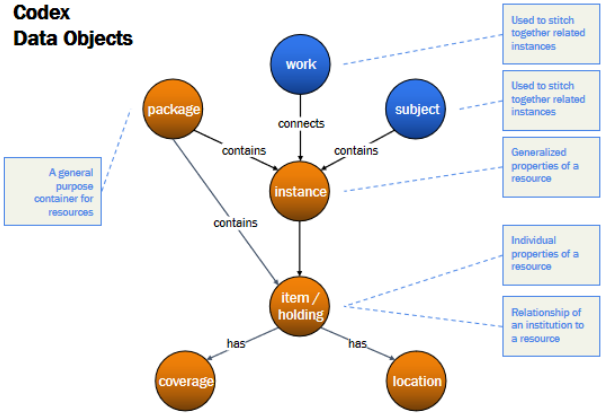
The data format used by Folio is highly compatible with the design of BIBFRAME2.
Folio的数据格式与BF2高度兼容
While Folio does not directly implement BIBFRAME,
Folio并非直接实施BF2
we can see that it is simple to create a crosswalk between the two.
Folio可便捷地与BF2建立映射

Though similar to BIBFRAME, this is not BIBRAME. It is only *inspired* by BIBFRAME

BIBFRAME2	Folio Object (RM)
Item	Holding
Instance	Instance
Work	Work
physicalLocation electronicLocator	Location

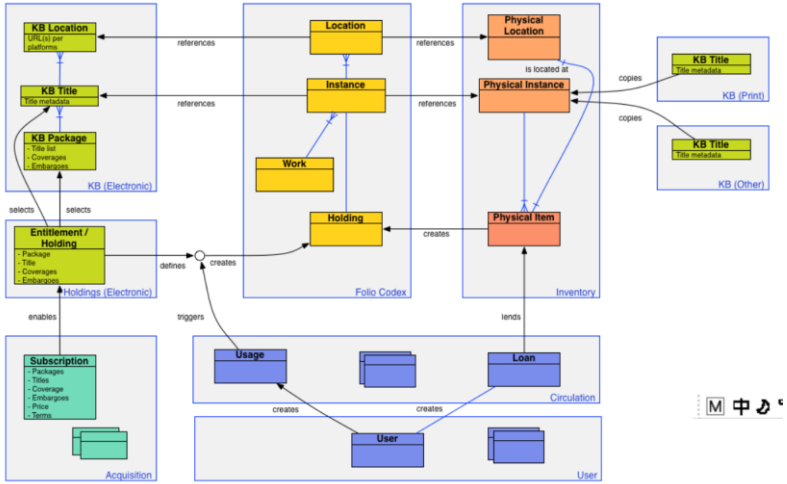
Codex 作为元数据方案（Metadata Schema）

- Instance　15+2个元素，必备2个
- Item/Holding 11+(12+2 from Instance)必备2+1个
- Package　13个　必备2个
  - Location　7+2个　必备1个
  - Coverage　4个　必备1个



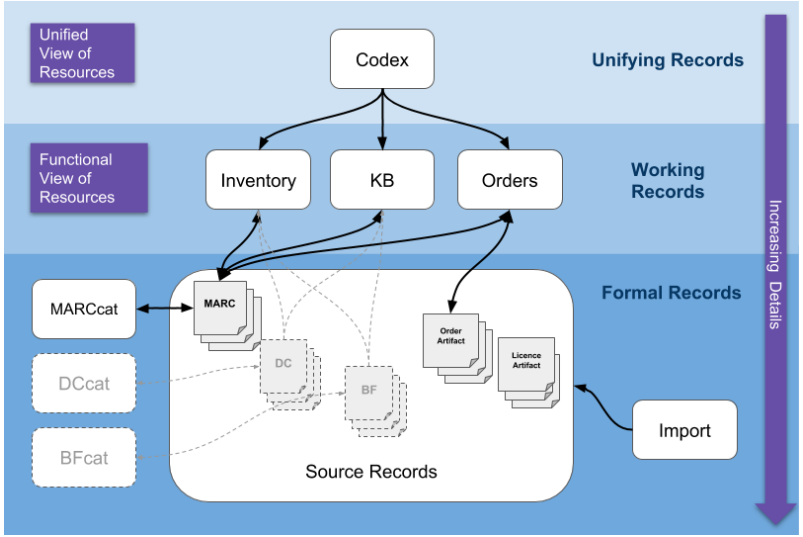
最小化原则（域的概念）

- Folio Codex Domain
- Knowledge base (KB) Domain
- Holdings Domain
- Acquisition Domain
- Inventory Domain
- Circulation Domain

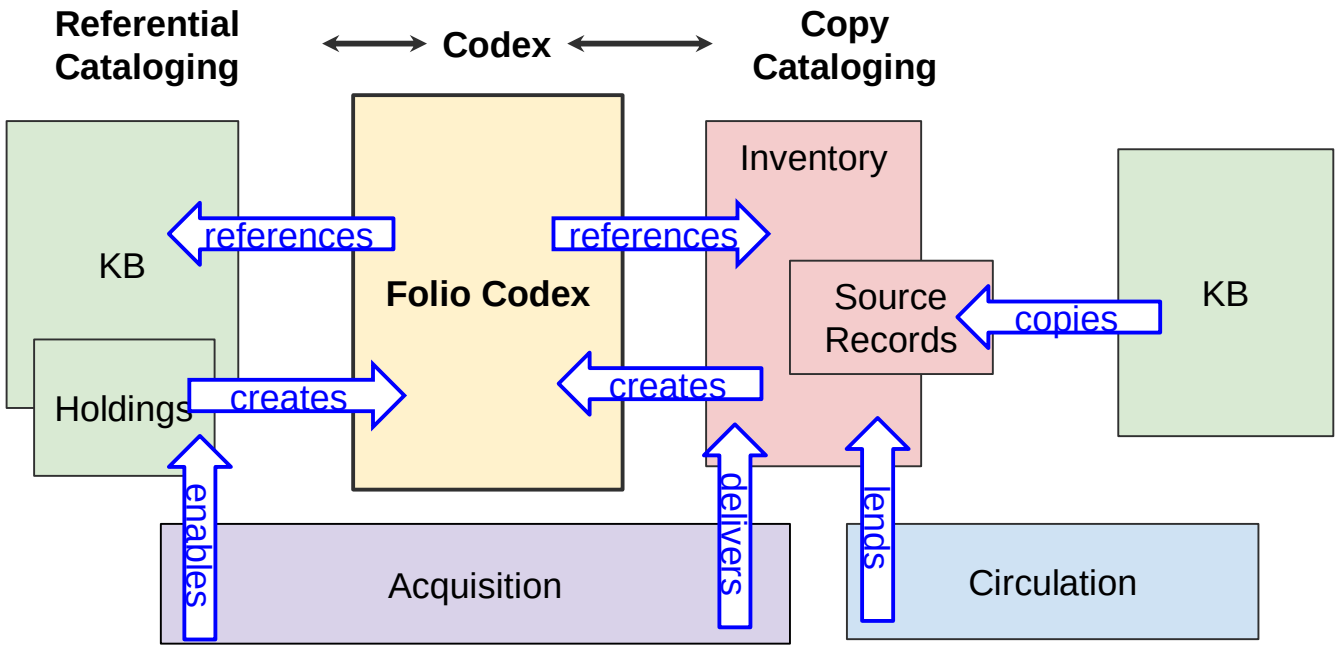


Codex作为Folio资源管理域中的数据整合层

层次化的元数据记录模型



- **Formal Records.** These are the most detailed records available in the Folio system.
- **Working Records.** These are the app-specific versions of records that are necessary for the apps to deliver their functionality.
- **Unifying Records.** This top layer is the one in which Codex and its modules operate. The purpose of this layer is to provide a uniform and high-level view across all of Folio.



KB 域和 Inventory 域与Codex有密切的数据互动关系
Codex的数据来源于KB和Inventory的自动推送

Folio中的MARC

- The scope of MARC records is mostly limited to the Inventory domain.

MARC记录的适用范围主要是**典藏域**。

- Folio will interpret the MARC records and extract out the relevant fields, mapping them to the very limited set of core metadata that is kept in the Codex domain and used to describe the resource.

Folio将从MARC记录中抽取**部分相关**字段，将其映射到Codex有限的核心元素集。

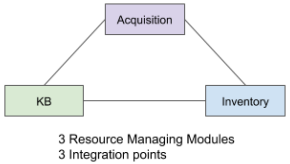
- The original MARC record is not discarded but is retained as an attachment tied to the new record - no information is lost.

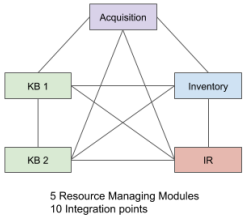
原始MARC记录**作为附件**链接到新的Codex记录中，以保证信息不损失。

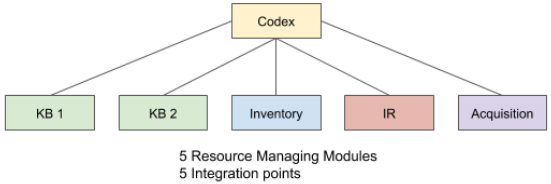
- Folio does not considers MARC records to be a data format. Rather, it considers them to be an interchange format - as was their original intent.

MARC记录仅作为**数据交换格式**。

Codex的愿景







Codex addresses the Entanglement Problem（跨域的数据链接中心）

Codex is the Entry Point for Resource Management（资源管理的入口）

Codex is a normalizing Data Model（规范化的数据模型）

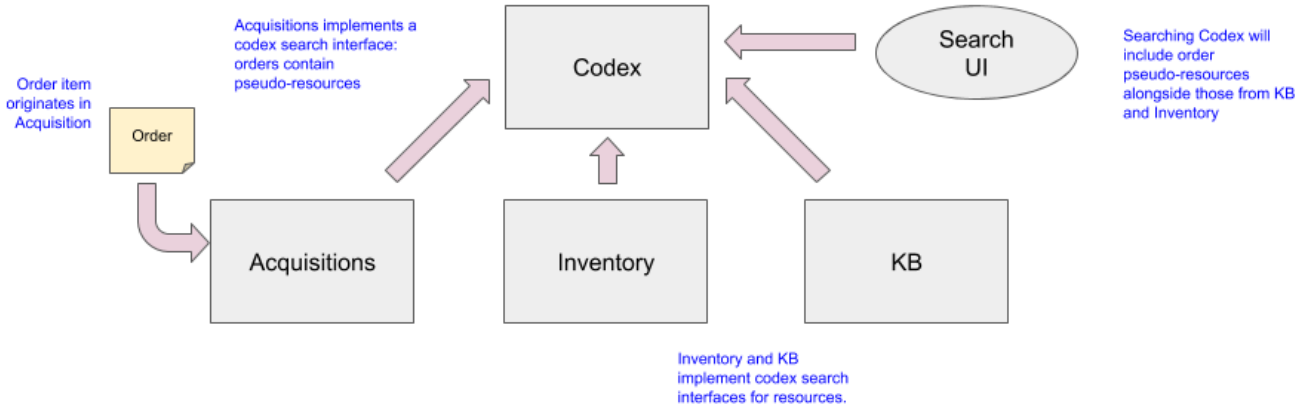
Codex manages Relationships between Resources（管理资源间的关系）

Codex is Resource Central（以资源为中心）

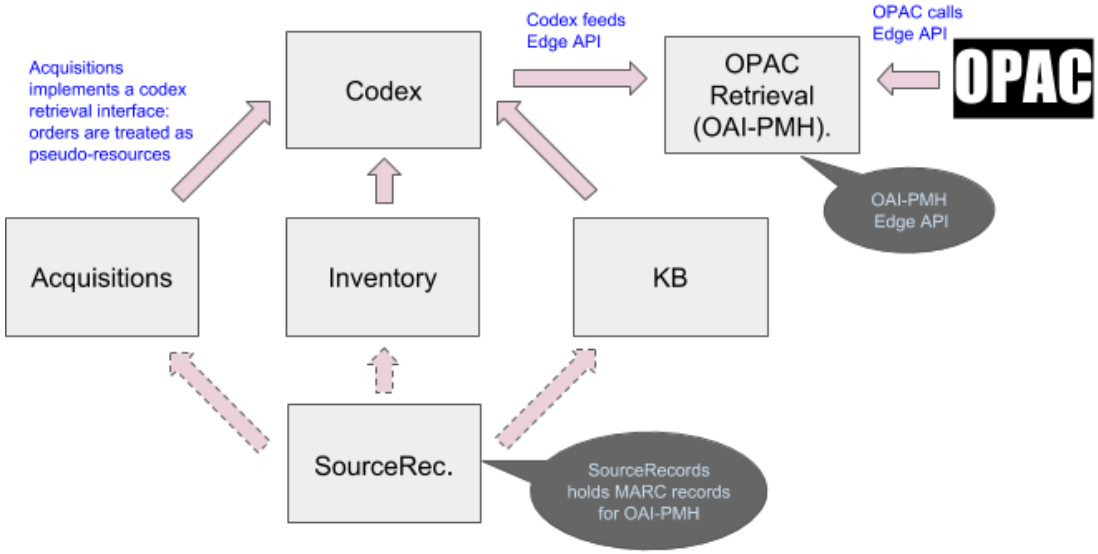
Codex includes Codex Search.Codex is the starting point for locating a resource.（仅用于定位资源，而不是揭示资源）

https://wiki.folio.org/display/PLATFORM/The+Codex+Vision

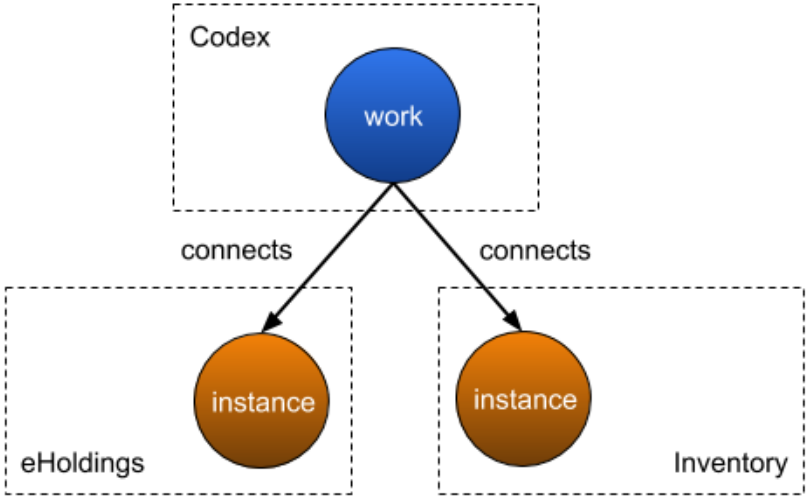
Use Case 1: Locating Resources
定位资源



Use Case 2: Exporting comprehensive Folio catalogs to an OPAC or Discovery system支持OPAC检索

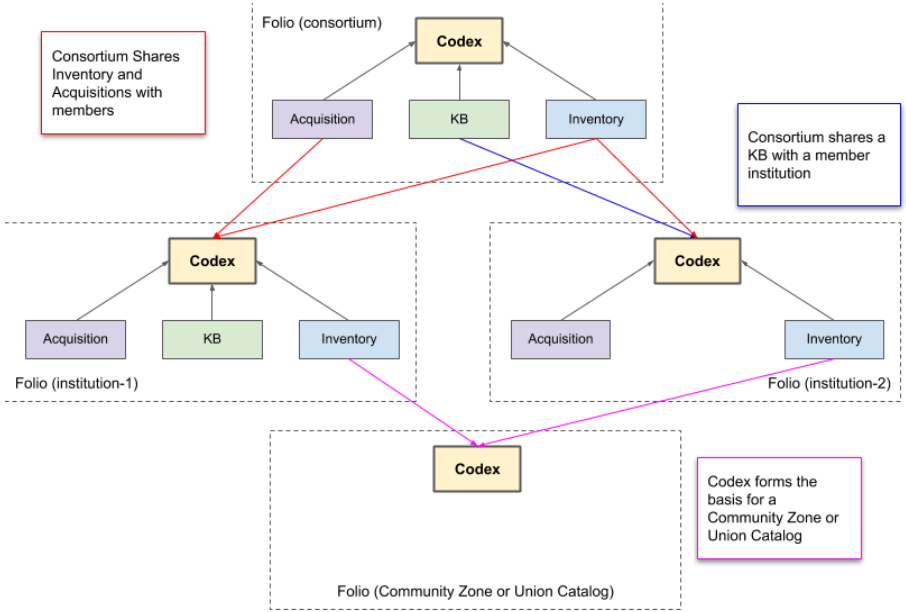


Use Case 3: Creating Associations between Resources建立资源间的关联关系

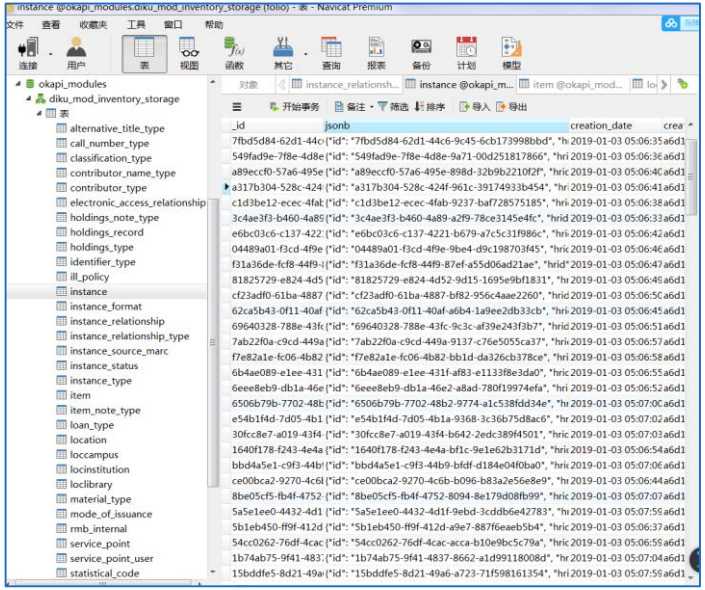


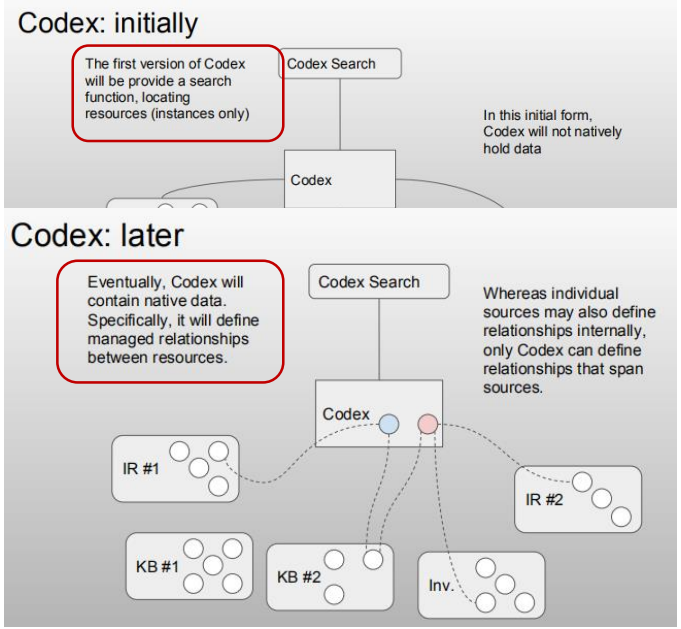
In keeping with the concepts of Linked Data we would want to create a Work object and link it to the different instance records that represent manifestations of that work.

Use Case 4: Inter-Folio Functionality
联盟或社区的资源共享和联合目录



Folio Codex 的现状

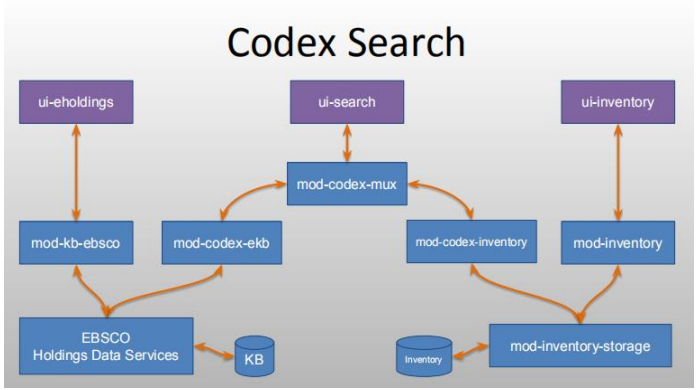




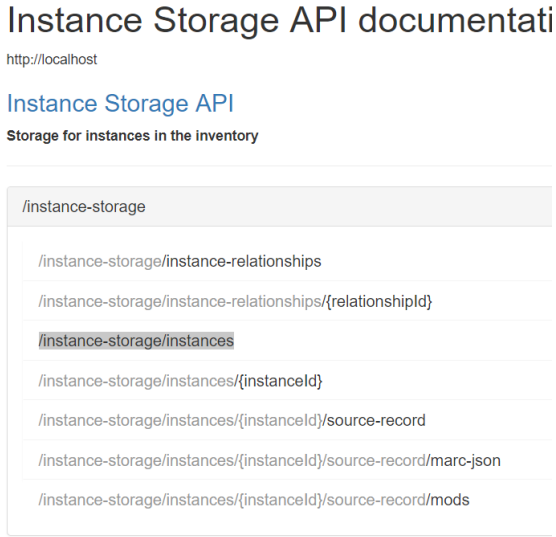
在Folio目前的版本中Codex只包含Codex Search
没有独立的数据存储

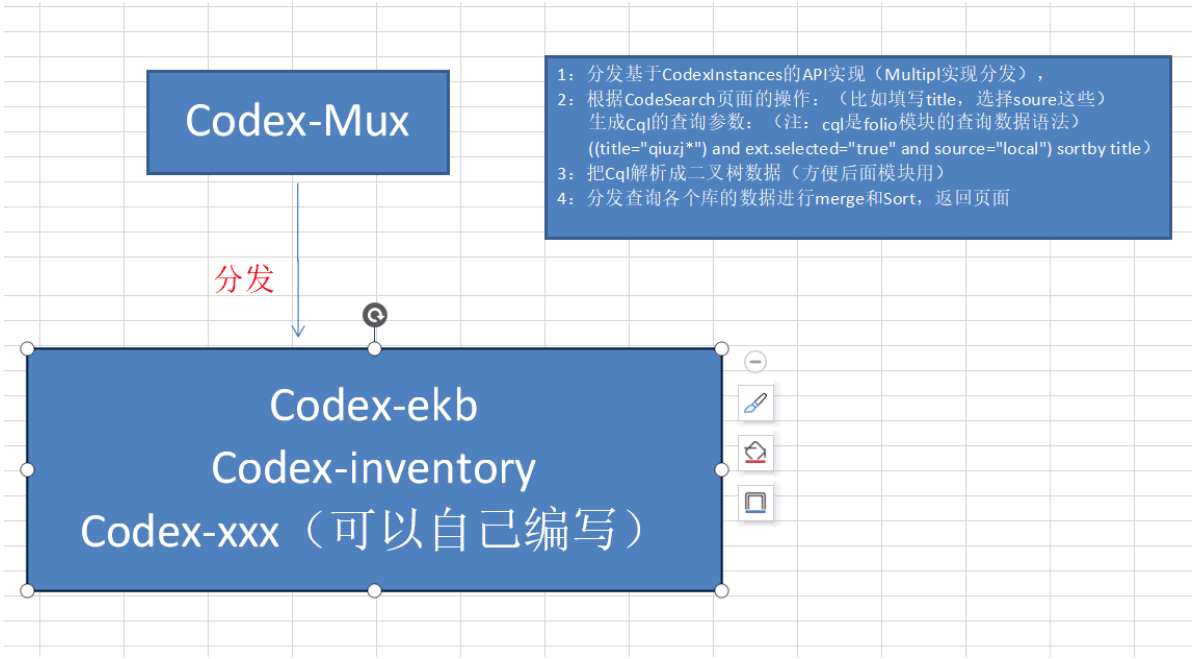
Codex Search 的业务逻辑

(搜索的过程中需要调用哪些数据？数据流如何走向？)



location数据获取流程 mod-codex-mux---》mod-codex-inventory--》 mod-inventory-storage，最后调用 /instance-storage/instances 这个方法





by 邱忠杰 of SHL Codex & MM SIG

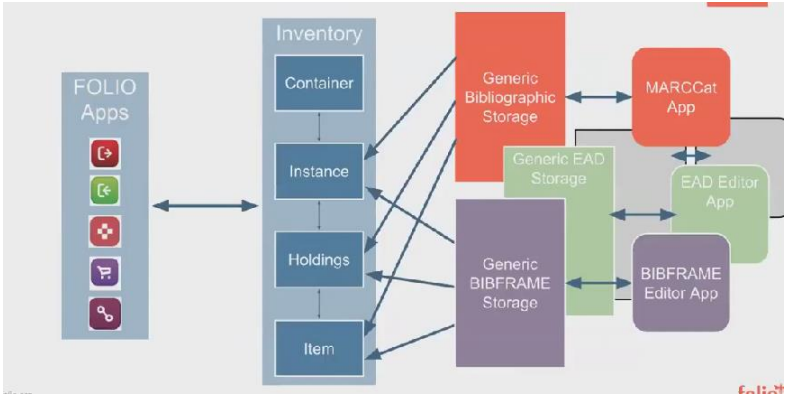
03

上图Cedex&MM SIG的工作

任务、进展、问题、计划

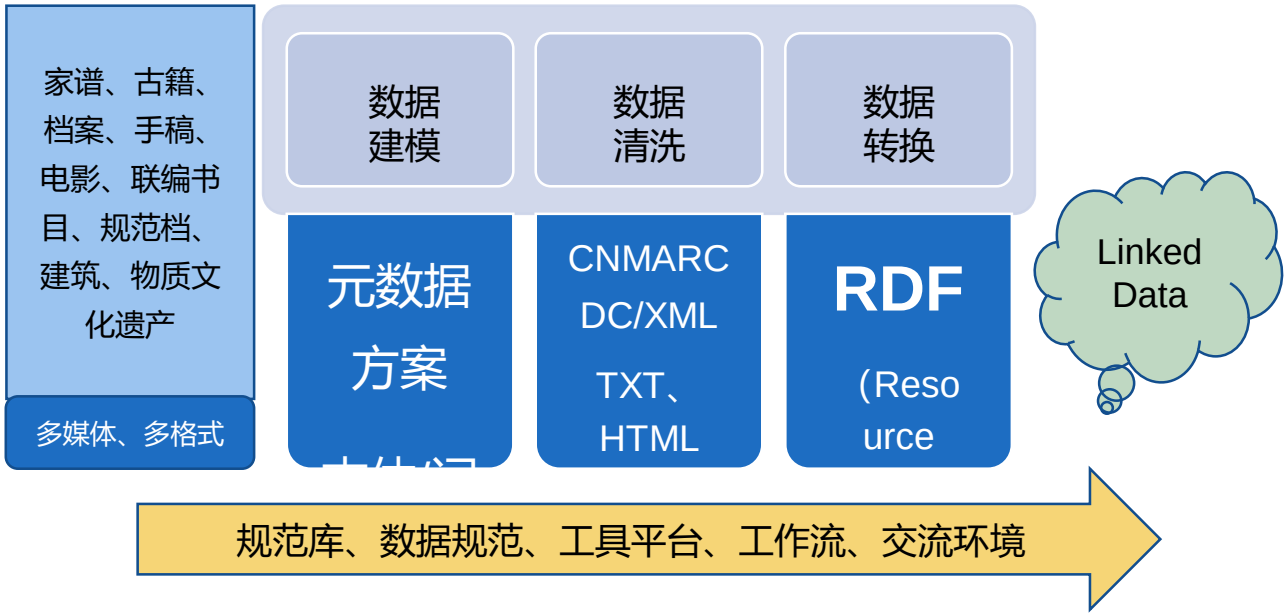
成立Codex & MM SIG (2019.1.17)
确立当前主要任务

The Metadata Management SIG works with developers to define essential bibliographic management functions: creating, editing, suppressing, deleting, importing, exporting, replacing, overlaying, and reporting. Defines essential data elements of a bibliographic control module. Explores various formats/schema that should be incorporated into FOLIO (MARC, RDA, BIBFRAME, Dublin Core, etc.) Considers metadata storage and harmonization .....

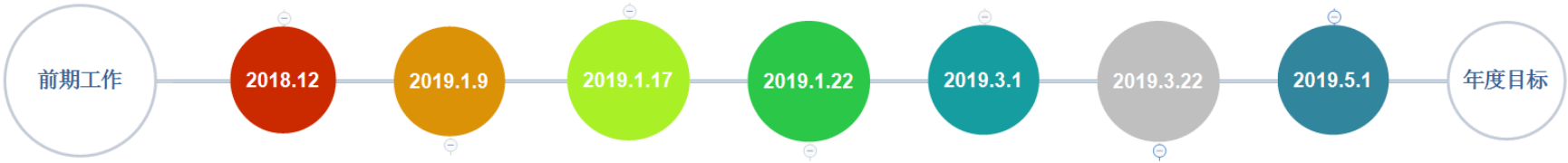


- 定义典型的元数据管理功能
- 定义支持书目控制的典型元数据方案（元素）
- 探索不同的数据模型/格式如何融入Folio的数据架构
- 探索Folio如何支持Linked Data

前期工作基础


家谱、古籍、档案、手稿、电影、联编书目、规范档、建筑、物质文化遗产
多媒体、多格式
数据建模
数据清洗
数据转换
元数据方案
CNMARC
DC/XML
TXT、HTML
RDF
(Resource
Linked Data
规范库、数据规范、工具平台、工作流、交流环境

上图Codex & MM SIG
2019年第一、二季度工作时间轴



上图Codex SIG成立

数据建模、元数据
编目、开发人员

调研：认识Codex

Codex在Folio中的地位和角色

调研：Codex现状

将元数据管理（MM）作为
SIG现阶段的主要任务
制定计划、分配任务

搭建共享空间
管理任务进度

提出问题、认领任务
分享文档、记录过程

试验与实施

查看数据库结构
查阅代码文档
理清流程逻辑
准备导入数据

数据结构映射

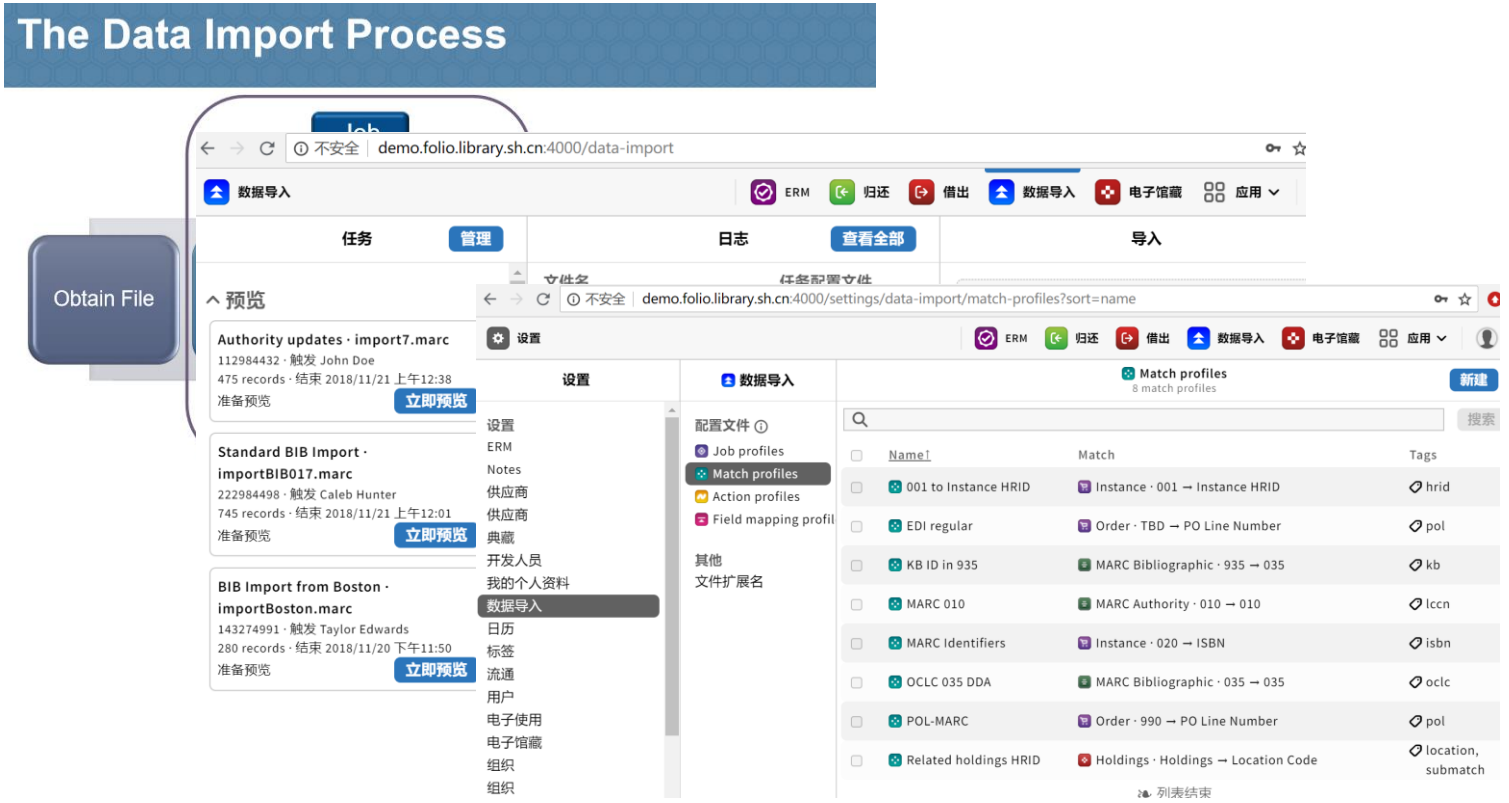
Codex/Inventory Metadata Set->
CNMARC的映射
Inventory 数据库数据结构->
上图联编中心BF2.0/RDF映射

数据导入技术路线
馆藏流通数据导入

制定映射规范、导入数据

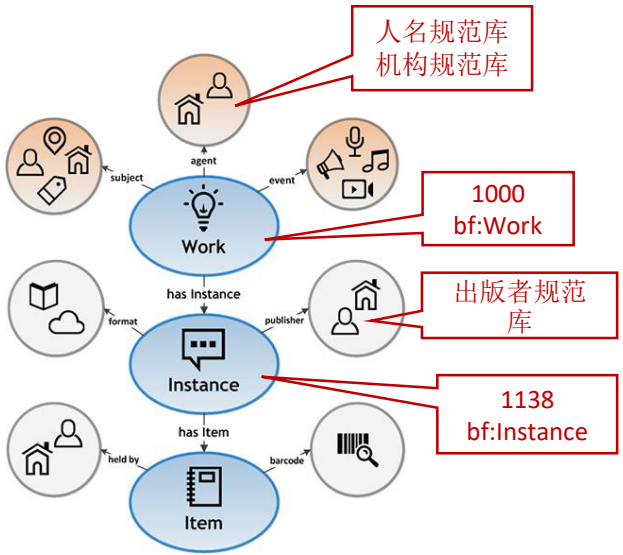
- 映射规范
  - Inventory 数据库数据结构->上图联编中心BF2.0/RDF映射
  - Codex/Inventory Metadata Set->CNMARC的映射
  - Circulation数据库数据结构->上图Horizon馆藏流通数据映射
- 试验数据导入
  - 上海联编中心1000种图书书目数据，BF2/RDF格式
  - 300余种CNMARC格式的书目数据
  - 与CNMARC格式数据相对应的馆藏和流通数据
- 全量数据导入
  - 提供API、局域网运行

Folio的数据导入策略

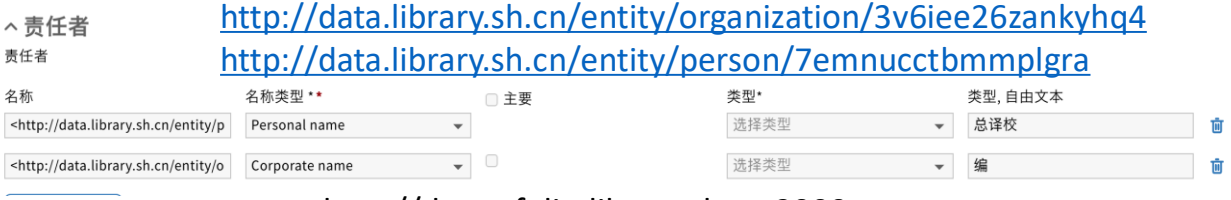


http://demo.folio.library.sh.cn:4000

数据导入试验结果



- 非侵入式（不直接操作数据库）
- 不能增加元数据元素
- 可增加取值词表的取值
- 无法将原始BF数据附在Inventory数据库的source表中
- 不支持基于HTTP URI的名称规范控制



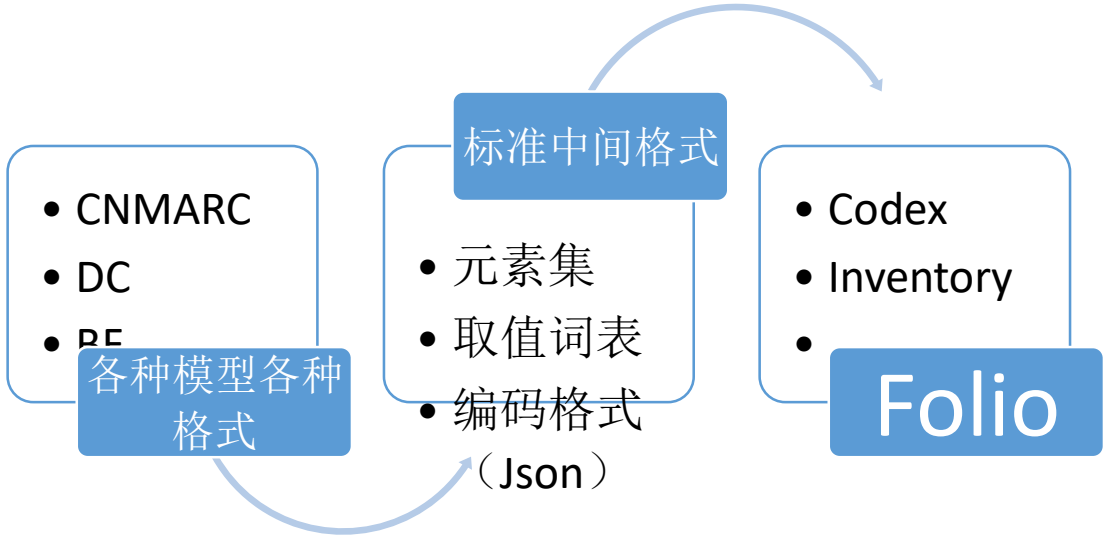
http://data.library.sh.cn/entity/organization/3v6iee26zankyhq4
http://data.library.sh.cn/entity/person/7emnucctbmmplgra

http://demo.folio.library.sh.cn:3000

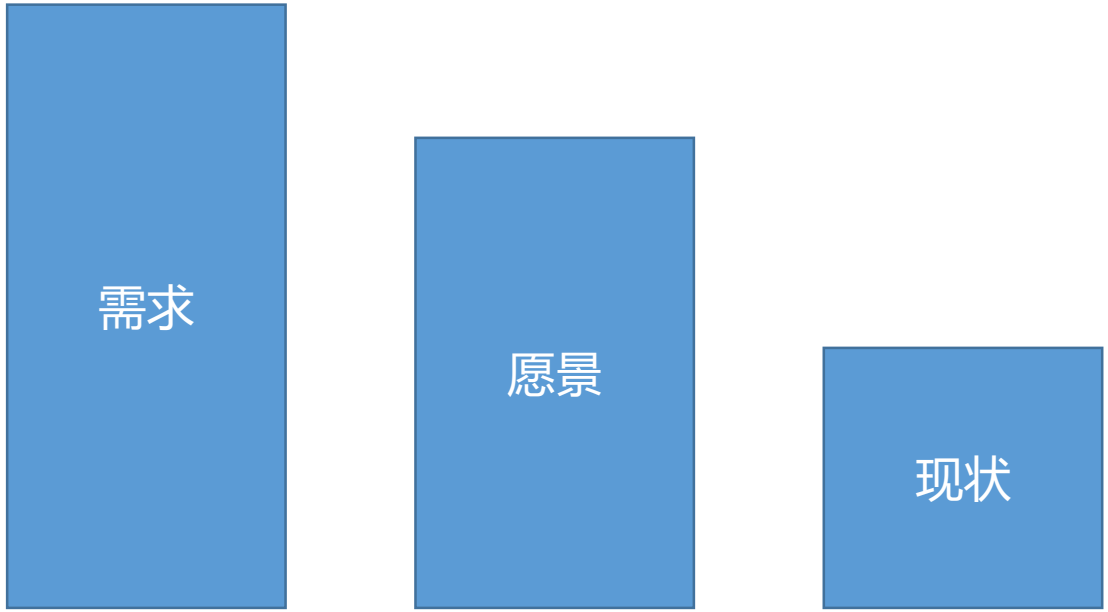
制定Folio的数据准入规范（中间格式标准）

按照原子字段顺序字符串拼接。拼接时，$v前加空格分号空格。$h前加句点空格。$i前如果有$h则加逗号空格、如果没有$h则加句点空格。$e前加空

系统自动生成。以in开头。
原始MARC的001在导入FOLIO的SRS时，001移到035，035=(801$b)001。SRS的001由FOLIO自动生成但省略开头in的HRID。
SRS中存储的源MARC字段，新增999字段。$i[Instance UUIR]$m[MARCcat Bib UUIR]$s[SRS Bib UUIR]



小结：Codex是否能实现元数据管理的需求？



年度目标

- CNMARC和BF格式的书目数据导入Inventory，建立书目数据的映射规范。导入馆藏和流通数据，建立馆藏及流通数据的映射规范。实现iPAC和流通模块的运行。
- 制订数据进入Folio的中间格式标准。

- 探索Folio整体元数据架构，定义各域的元数据管理需求，制订Codex域以及其他域的异源异构数据导入、映射、交换格式标准。

THANKS

上海图书馆Codex & MM SIG
2019.6.6 @大连