Metadata Development and Documentation for a Research Data Repository

Huang-Sin Syu, Cheng-Jen Lee, Yao-Hsien Yeh,
Tyng-Ruey Chuang

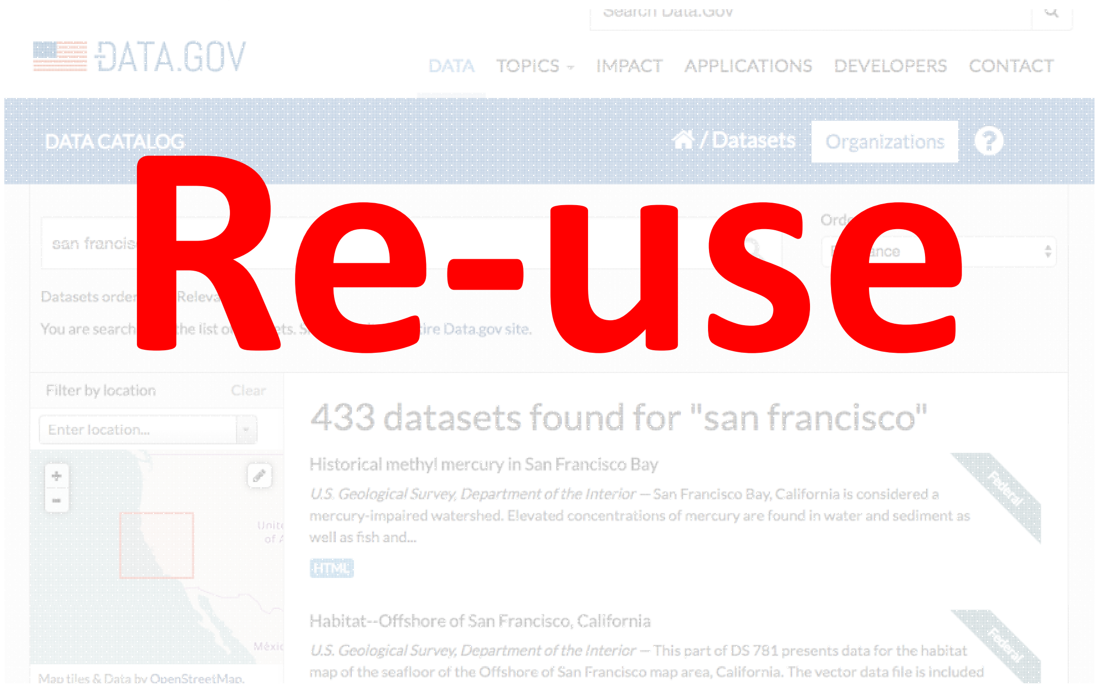
Institute of Information Science, Academia Sinica, Taiwan

Outline

- Research Data Repository
- Why we need metadata?
- Metadata schema and the quality of metadata
- Creating metadata for research datasets
- Documentation for Research Data Repository
- Documentation workflow for metadata schema
- Current results and future works

Research Data Repository (RDR)

- Active management, adding value and maintaining access to research data
- Accurate, complete, retrievable



Re-use

Datasets order Releva

You are search the list o ts. S ire Data.gov site.

Map tiles & Data by OpenStreetMap.

433 datasets found for "san francisco"

Historical methyl mercury in San Francisco Bay

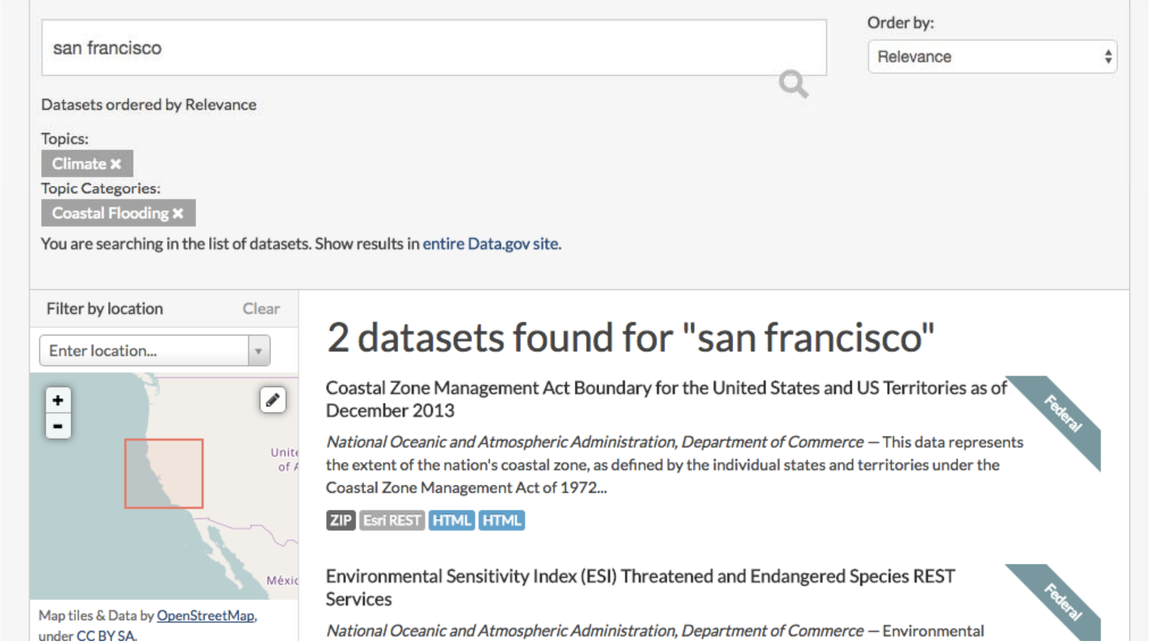
U.S. Geological Survey, Department of the Interior — San Francisco Bay, California is considered a mercury-impaired watershed. Elevated concentrations of mercury are found in water and sediment as well as fish and...

Habitat--Offshore of San Francisco, California

U.S. Geological Survey, Department of the Interior — This part of DS 781 presents data for the habitat map of the seafloor of the Offshore of San Francisco map area, California. The vector data file is included

Why we need metadata?

- Reduce dataset duplication
- Use datasets more effectively
- Make research data searchable



san francisco

Order by:

Relevance

Datasets ordered by Relevance

Topics:

Climate ✖

Topic Categories:

Coastal Flooding ✖

You are searching in the list of datasets. Show results in entire Data.gov site.

Filter by location Clear

Enter location...

Map tiles & Data by OpenStreetMap, under CC BY SA.

2 datasets found for "san francisco"

Coastal Zone Management Act Boundary for the United States and US Territories as of December 2013

National Oceanic and Atmospheric Administration, Department of Commerce — This data represents the extent of the nation's coastal zone, as defined by the individual states and territories under the Coastal Zone Management Act of 1972...

Environmental Sensitivity Index (ESI) Threatened and Endangered Species REST Services

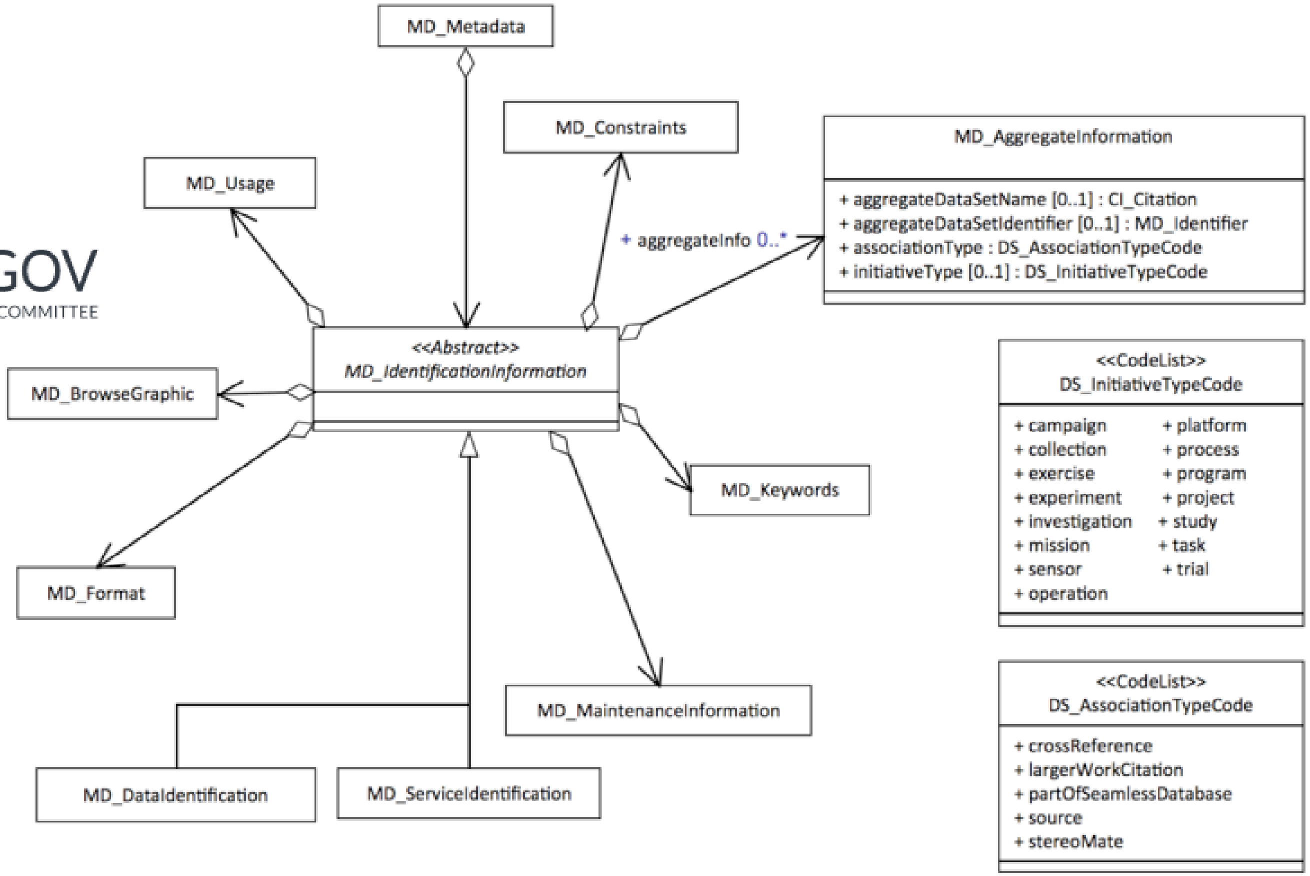
National Oceanic and Atmospheric Administration, Department of Commerce — Environmental

Metadata schema and the quality of metadata

- General standard
  - Dublin Core Metadata Initiative(DCMI)
- Geospatial
  - ISO 19115:2003 Geographic information metadata

Metadata schema and the quality of metadata

FGDC.GOV
FEDERAL GEOGRAPHIC DATA COMMITTEE



MD_Metadata

MD_Constraints

MD_Usage

+ aggregateInfo 0..*

MD_AggregateInformation

+ aggregateDataSetName [0..1] : CI_Citation
+ aggregateDataSetIdentifier [0..1] : MD_Identifier
+ associationType : DS_AssociationTypeCode
+ initiativeType [0..1] : DS_InitiativeTypeCode

<<Abstract>>
MD_IdentificationInformation

MD_BrowseGraphic

MD_Keywords

MD_Format

MD_MaintenanceInformation

MD_DataIdentification

MD_ServiceIdentification

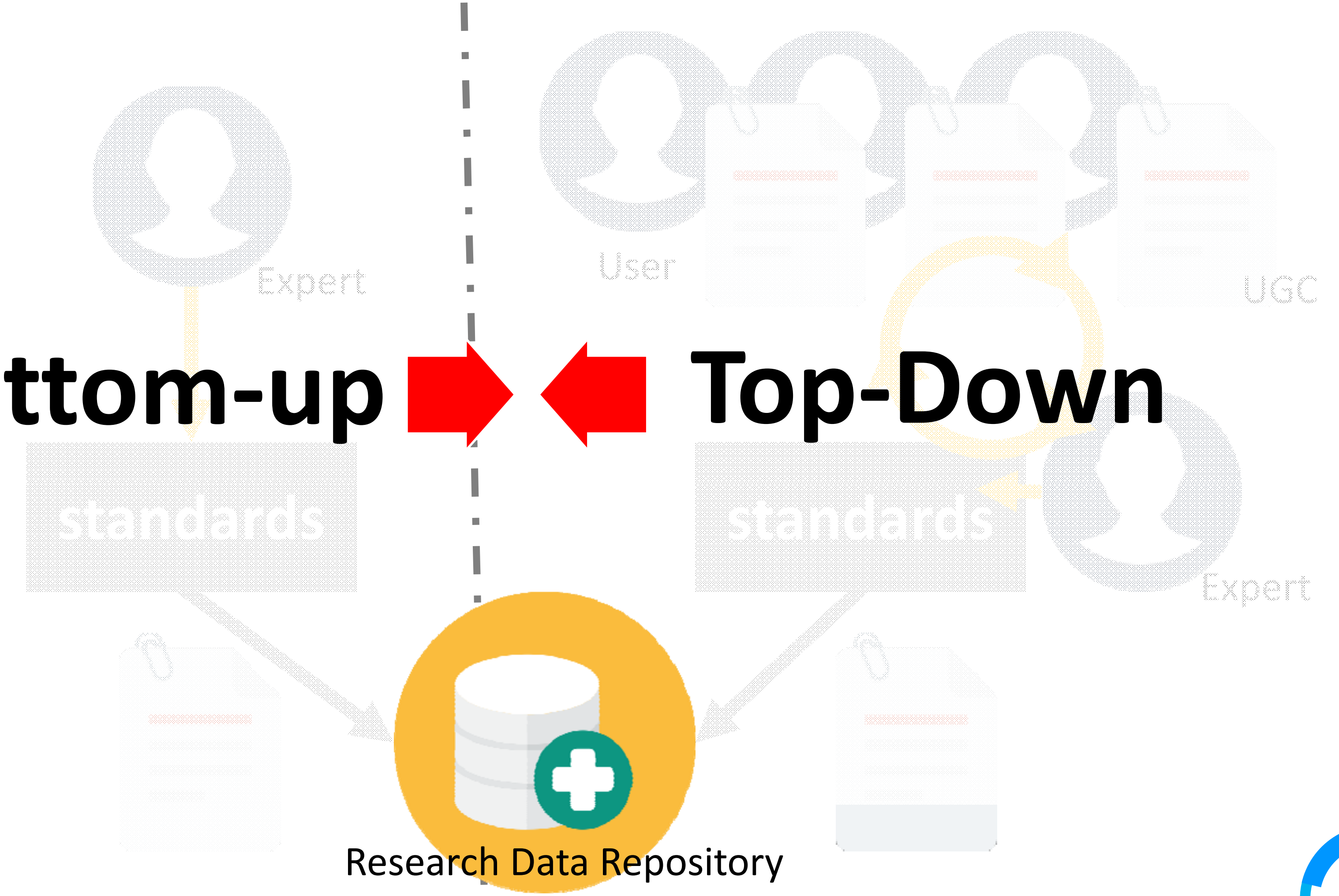
<<CodeList>>
DS_InitiativeTypeCode

+ campaign
+ collection
+ exercise
+ experiment
+ investigation
+ mission
+ sensor
+ operation
+ platform
+ process
+ program
+ project
+ study
+ task
+ trial

<<CodeList>>
DS_AssociationTypeCode

+ crossReference
+ largerWorkCitation
+ partOfSeamlessDatabase
+ source
+ stereoMate

Metadata schema and the quality of metadata



Buttom-up ➡ ⬅ Top-Down

Research Data Repository

Metadata schema and the quality of metadata

- **Continuous revisions of metadata schema**

- **Customized metadata schema**

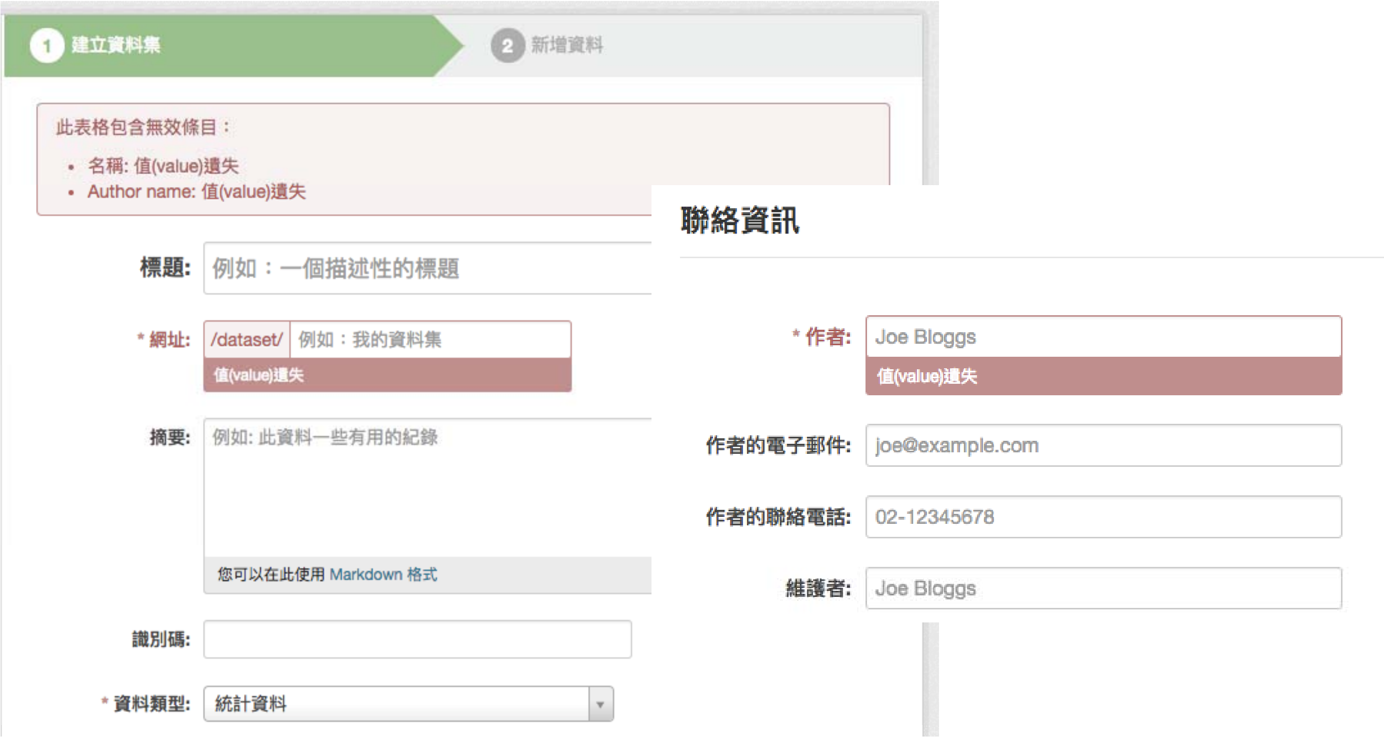
Creating metadata for research datasets

- Use CKAN to build Research Data Repositories
  - Taijiang Project
  - Ka-lam Project
  - More

- Metadata authoring via CKAN webpage
- Dataset/Metadata Bulk Upload (Customized process for Taijiang project)
- Dataset/Metadata Harvest API

Creating metadata for research datasets

- Metadata authoring via CKAN webpage



此表格包含無效條目：

- 名稱: 值(value)遺失
- Author name: 值(value)遺失

標題: 例如：一個描述性的標題

* 網址: /dataset/ 例如：我的資料集

值(value)遺失

摘要: 例如: 此資料一些有用的紀錄

您可以在此使用 Markdown 格式

識別碼:

* 資料類型: 統計資料

聯絡資訊

* 作者: Joe Bloggs

值(value)遺失

作者的電子郵件: joe@example.com

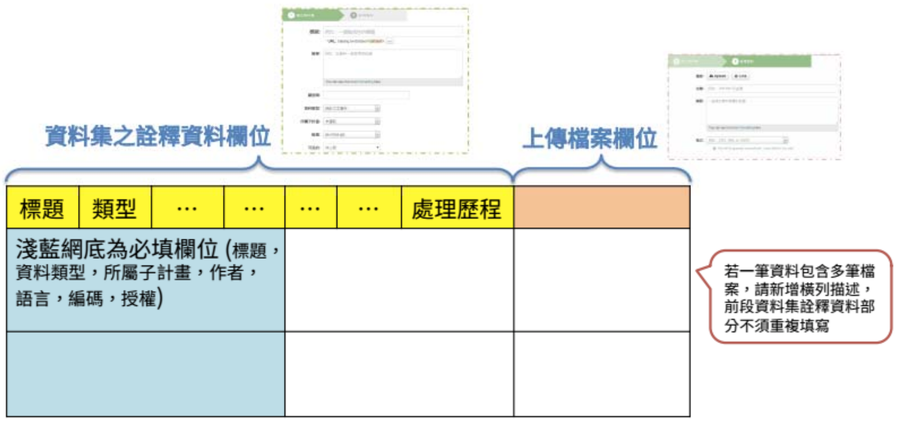
作者的聯絡電話: 02-12345678

維護者: Joe Bloggs

Creating metadata for research datasets

- Dataset/Metadata Bulk Upload (Customized process for Taijiang project) https://taijiang.tw/en/help

詮釋檔案說明



標題	類型	…	…	…	…	處理歷程	
淺藍網底為必填欄位 (標題，資料類型，所屬子計畫，作者，語言，編碼，授權)							

若一筆資料包含多筆檔案，請新增橫列描述，前段資料集詮釋資料部分不須重複填寫

- 特定格式欄位 (欄位文字以藍色表示)：參考下頁特定格式欄位說明
- 特定選項欄位 (欄位文字以紅色表示)：選項請參考範例 excel 檔之「清單選項」頁籤
- 接受多重值之欄位 (主題關鍵字、空間範圍關鍵字、使用史料、群組、標籤、參考來源，後兩者可自由填寫)，請以「;」做區隔

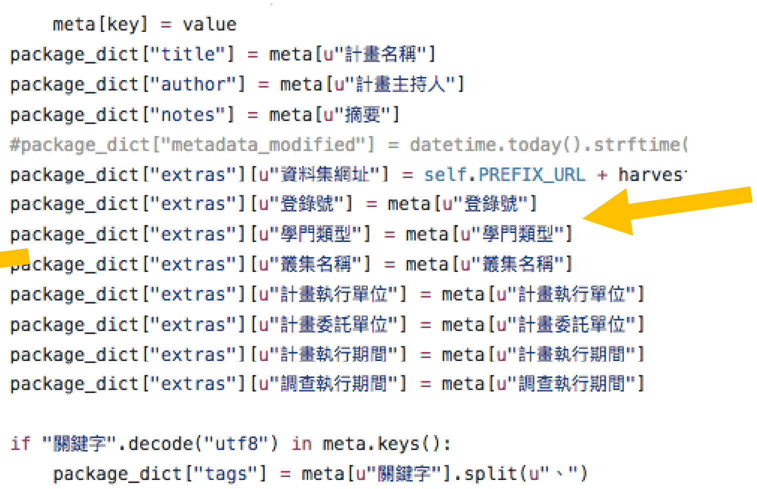
上傳檔案欄位說明

欄位名稱	說明	備註
檔案標題	顯示於系統之檔案標題	
檔案摘要	檔案之摘要描述	
檔案名稱	欲上傳檔名 (含副檔名)	各筆檔案此兩欄位請擇一填寫 (不可同時填寫)
檔案網址	檔案於網際網路上之位置	
坐標參考系統	若檔案格式為 shp 請務必填寫，否則無法正常預覽	填寫格式請參考上頁說明
檔案格式	檔案之副檔名	若為打包為 zip 檔案之 shp 檔案，請填寫 “shp”

含所有批次上傳之檔案

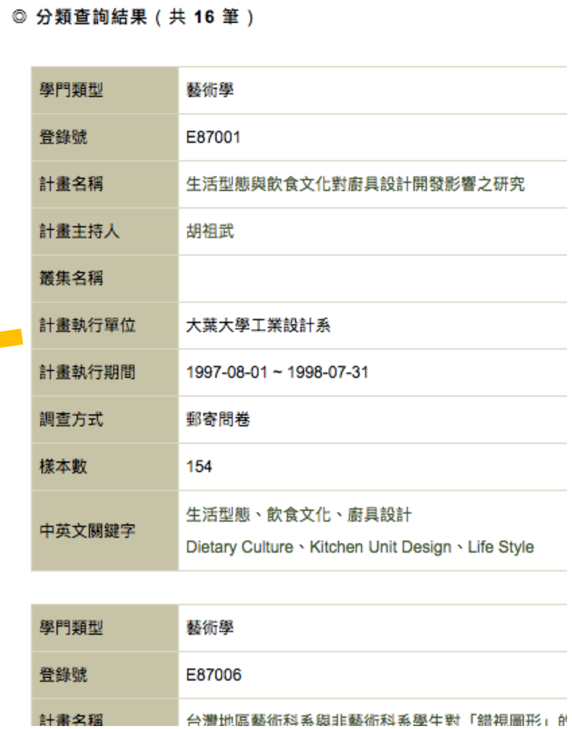
Creating metadata for research datasets

- Dataset/Metadata Harvest API



```
    meta[key] = value
package_dict["title"] = meta[u"計畫名稱"]
package_dict["author"] = meta[u"計畫主持人"]
package_dict["notes"] = meta[u"摘要"]
#package_dict["metadata_modified"] = datetime.today().strftime(
package_dict["extras"][u"資料集網址"] = self.PREFIX_URL + harves
package_dict["extras"][u"登錄號"] = meta[u"登錄號"]
package_dict["extras"][u"學門類型"] = meta[u"學門類型"]
package_dict["extras"][u"叢集名稱"] = meta[u"叢集名稱"]
package_dict["extras"][u"計畫執行單位"] = meta[u"計畫執行單位"]
package_dict["extras"][u"計畫委託單位"] = meta[u"計畫委託單位"]
package_dict["extras"][u"計畫執行期間"] = meta[u"計畫執行期間"]
package_dict["extras"][u"調查執行期間"] = meta[u"調查執行期間"]

if "關鍵字".decode("utf8") in meta.keys():
    package_dict["tags"] = meta[u"關鍵字"].split(u"、")
```

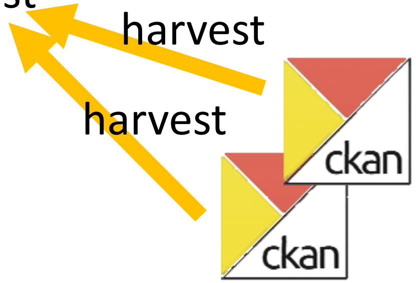


◎ 分類查詢結果（共 16 筆）

學門類型	藝術學
登錄號	E87001
計畫名稱	生活型態與飲食文化對廚具設計開發影響之研究
計畫主持人	胡祖武
叢集名稱	
計畫執行單位	大葉大學工業設計系
計畫執行期間	1997-08-01 ~ 1998-07-31
調查方式	郵寄問卷
樣本數	154
中英文關鍵字	生活型態、飲食文化、廚具設計 Dietary Culture、Kitchen Unit Design、Life Style

學門類型	藝術學
登錄號	E87006
計畫名稱	台灣地區藝術科系與非藝術科系學生對「錯視圖形」的

ckanext-harvest



https://srda.sinica.edu.tw/search/field/2

Institute of Information Science, Academia Sinica, Taiwan

Documentation for RDR

- Explain metadata schema
- Aid to interpretation

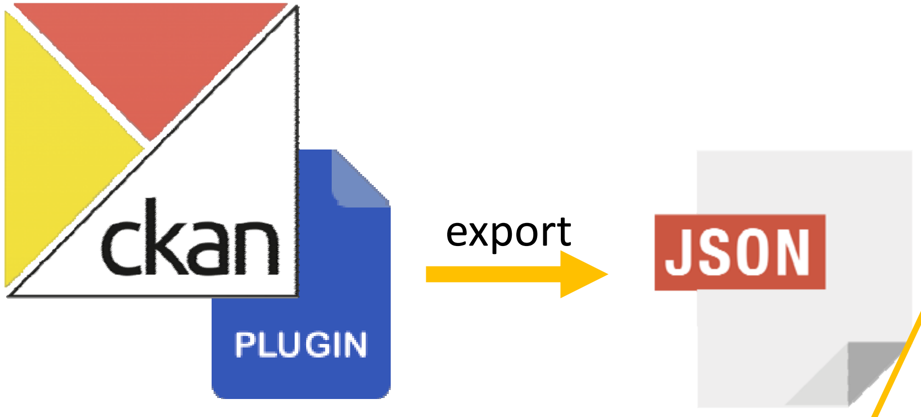
- Workshop
- User guide
- Documentation for different usage scenarios

Documentation for RDR

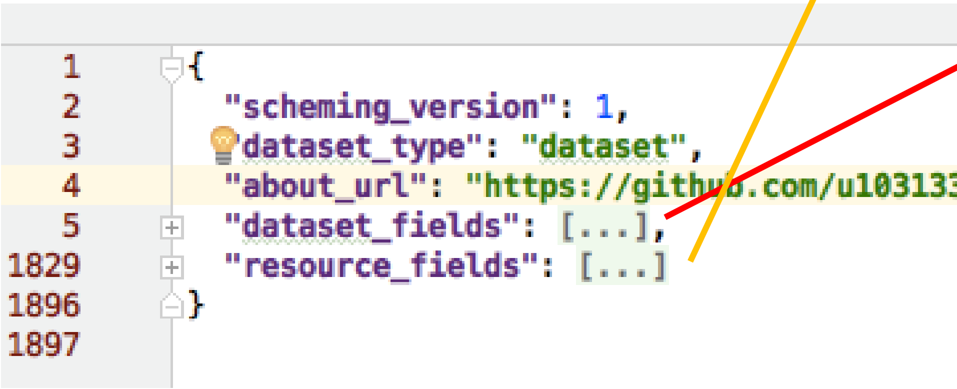
- **One source document for all metadata schema**
- **Rapid revision to metadata schema**

Documentation workflow for metadata schema

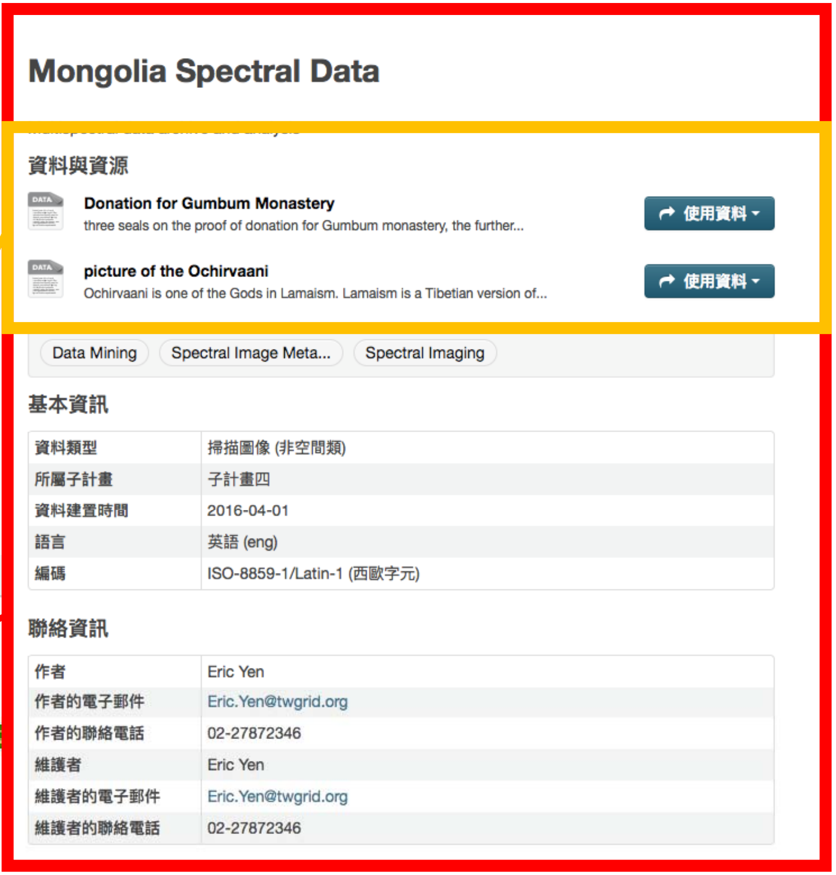
https://taijiang.tw/dataset/mongolia-spectral-data



ckanext-scheming



```
{
  "scheming_version": 1,
  "dataset_type": "dataset",
  "about_url": "https://github.com/u103133
  "dataset_fields": [...],
  "resource_fields": [...]
}
```



Mongolia Spectral Data

資料與資源

Donation for Gumbum Monastery
three seals on the proof of donation for Gumbum monastery, the further...

picture of the Ochirvaani
Ochirvaani is one of the Gods in Lamaism. Lamaism is a Tibetian version of...

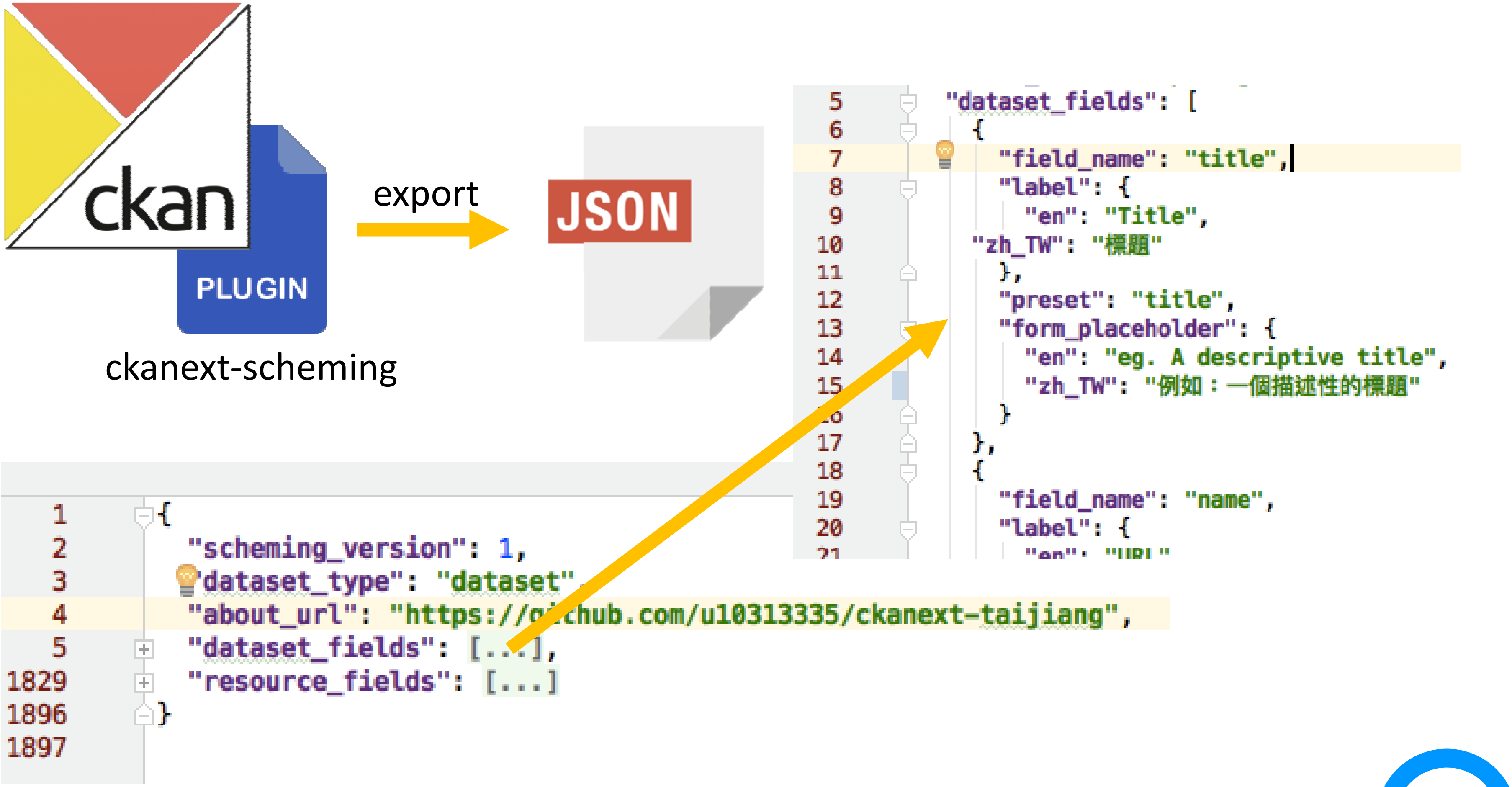
基本資訊

資料類型	掃描圖像 (非空間類)
所屬子計畫	子計畫四
資料建置時間	2016-04-01
語言	英語 (eng)
編碼	ISO-8859-1/Latin-1 (西歐字元)

聯絡資訊

作者	Eric Yen
作者的電子郵件	Eric.Yen@twgrid.org
作者的聯絡電話	02-27872346
維護者	Eric Yen
維護者的電子郵件	Eric.Yen@twgrid.org
維護者的聯絡電話	02-27872346

Documentation workflow for metadata schema



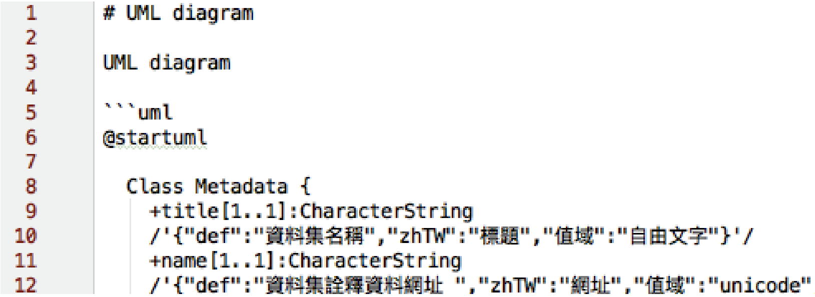
```
{
  "scheming_version": 1,
  "dataset_type": "dataset",
  "about_url": "https://github.com/u10313335/ckanext-taijiang",
  "dataset_fields": [...],
  "resource_fields": [...]
}
```

```
"dataset_fields": [
  {
    "field_name": "title",
    "label": {
      "en": "Title",
  "zh_TW": "標題"
    },
    "preset": "title",
    "form_placeholder": {
      "en": "eg. A descriptive title",
      "zh_TW": "例如：一個描述性的標題"
    }
  },
  {
    "field_name": "name",
    "label": {
      "en": "URL"
```

Documentation workflow for metadata schema

MD

UML

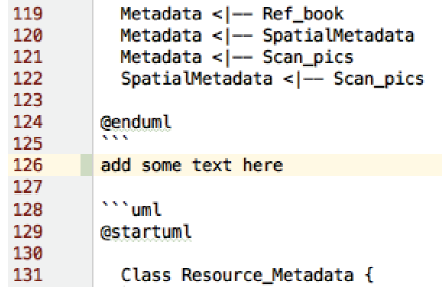


~~~
# UML diagram

UML diagram

```uml
@startuml

  Class Metadata {
    +title[1..1]:CharacterString
    /'{"def":"資料集名稱","zhTW":"標題","值域":"自由文字"}'/
    +name[1..1]:CharacterString
    /'{"def":"資料集詮釋資料網址 ","zhTW":"網址","值域":"unicode","not
~~~

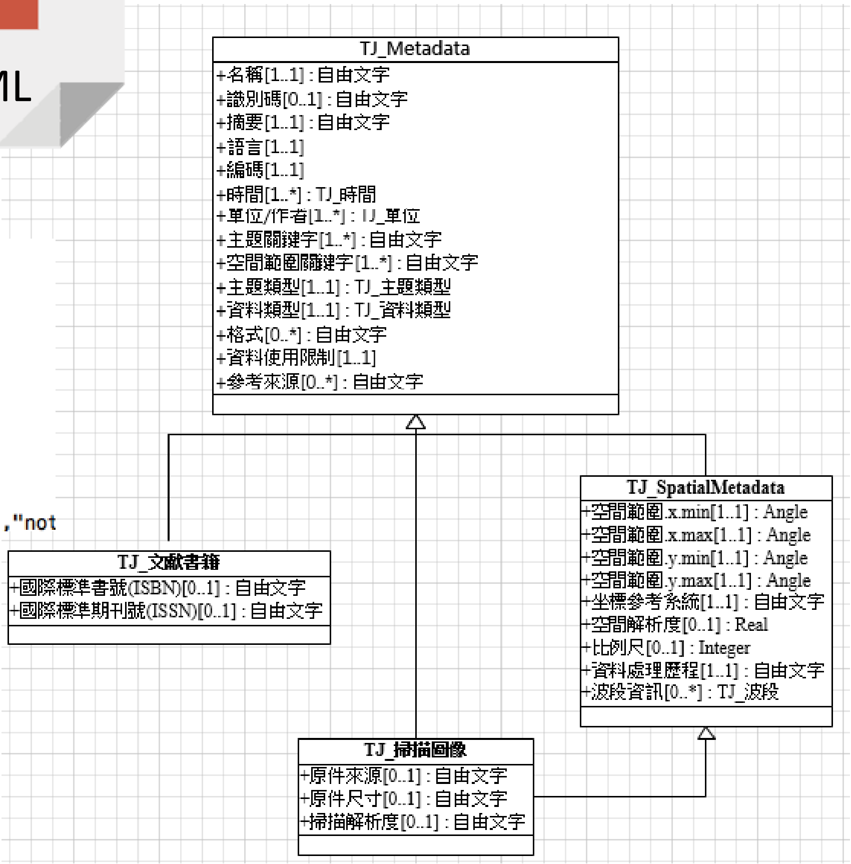


```
  Metadata <|-- Ref_book
  Metadata <|-- SpatialMetadata
  Metadata <|-- Scan_pics
  SpatialMetadata <|-- Scan_pics

@enduml
```
add some text here

```uml
@startuml

  Class Resource_Metadata {
```



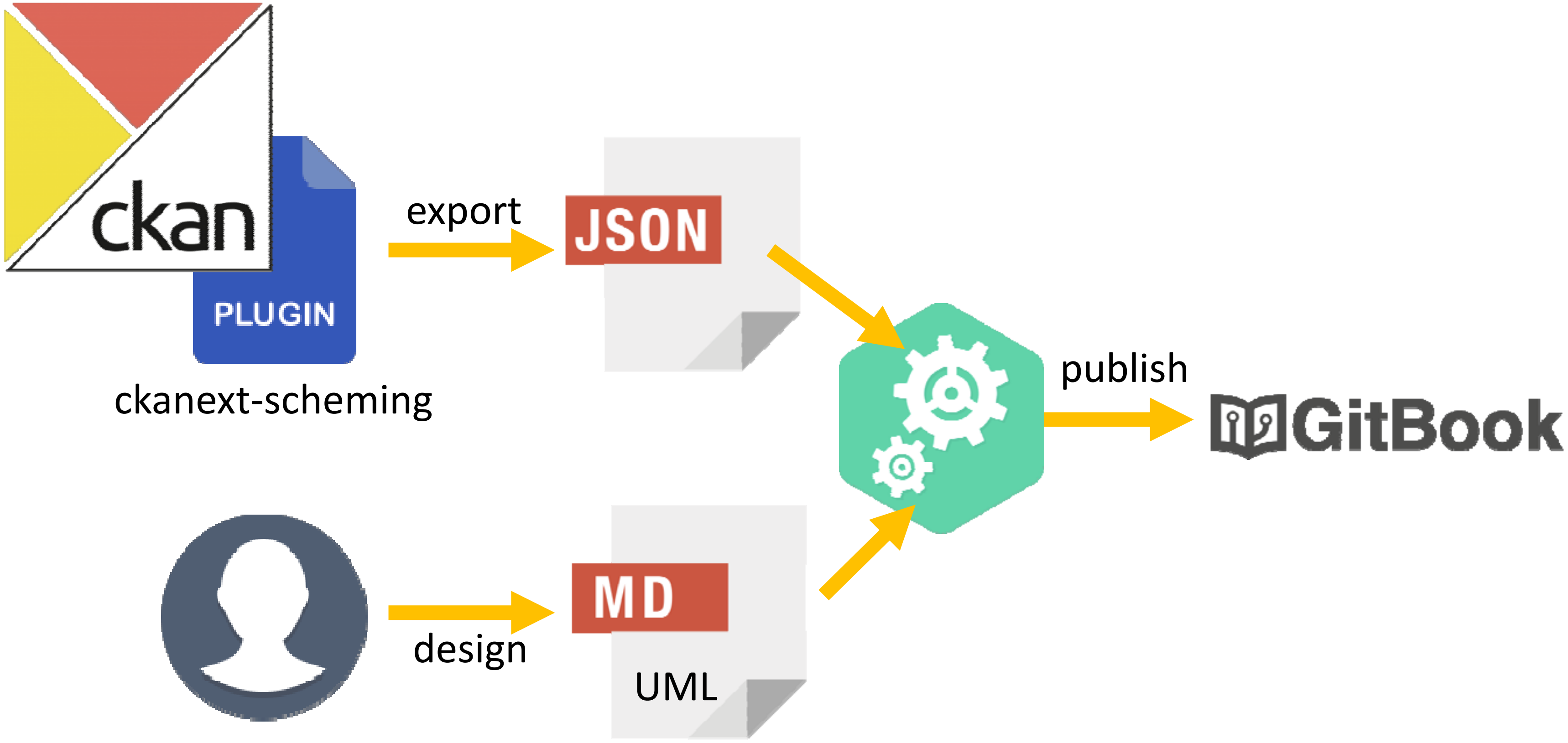
TJ_Metadata
+名稱[1..1] : 自由文字
+識別碼[0..1] : 自由文字
+摘要[1..1] : 自由文字
+語言[1..1]
+編碼[1..1]
+時間[1..*] : TJ_時間
+單位/作者[1..*] : TJ_單位
+主題關鍵字[1..*] : 自由文字
+空間範圍關鍵字[1..*] : 自由文字
+主題類型[1..1] : TJ_主題類型
+資料類型[1..1] : TJ_資料類型
+格式[0..*] : 自由文字
+資料使用限制[1..1]
+參考來源[0..*] : 自由文字

TJ_文獻書籍
+國際標準書號(ISBN)[0..1] : 自由文字
+國際標準期刊號(ISSN)[0..1] : 自由文字

TJ_SpatialMetadata
+空間範圍.x.min[1..1] : Angle
+空間範圍.x.max[1..1] : Angle
+空間範圍.y.min[1..1] : Angle
+空間範圍.y.max[1..1] : Angle
+坐標參考系統[1..1] : 自由文字
+空間解析度[0..1] : Real
+比例尺[0..1] : Integer
+資料處理歷程[1..1] : 自由文字
+波段資訊[0..*] : TJ_波段

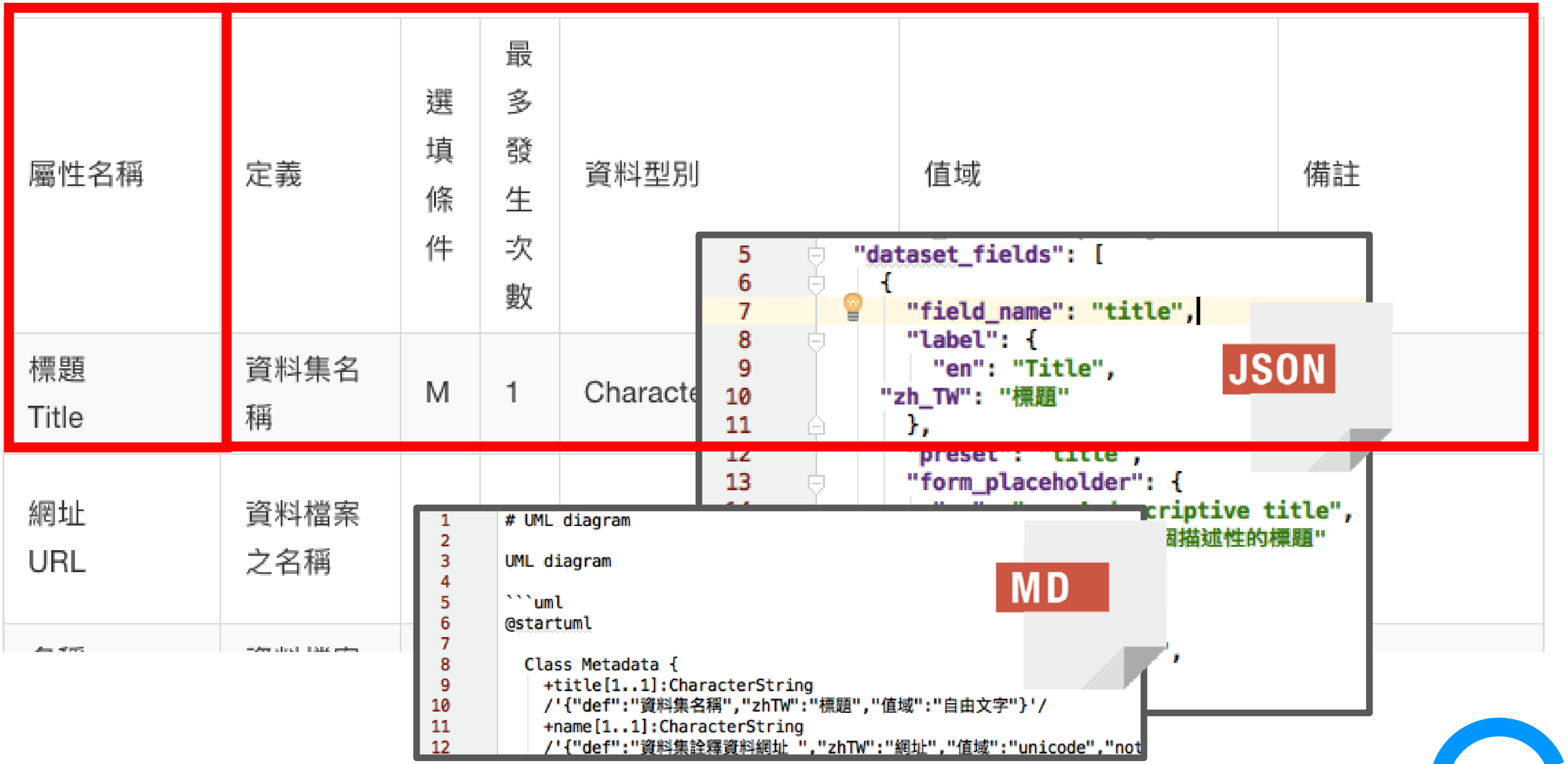
TJ_掃描圖像
+原件來源[0..1] : 自由文字
+原件尺寸[0..1] : 自由文字
+掃描解析度[0..1] : 自由文字

Documentation workflow for metadata schema



Documentation workflow for metadata schema

Schema + Description



屬性名稱	定義	選填條件	最多發生次數	資料型別	值域	備註
標題 Title	資料集名稱	M	1	Characte		
網址 URL	資料檔案之名稱					

```
  "dataset_fields": [
    {
      "field_name": "title",
      "label": {
        "en": "Title",
    "zh_TW": "標題"
      },
      "preset": "title",
      "form_placeholder": {
                      criptive title",
                      描述性的標題"
```

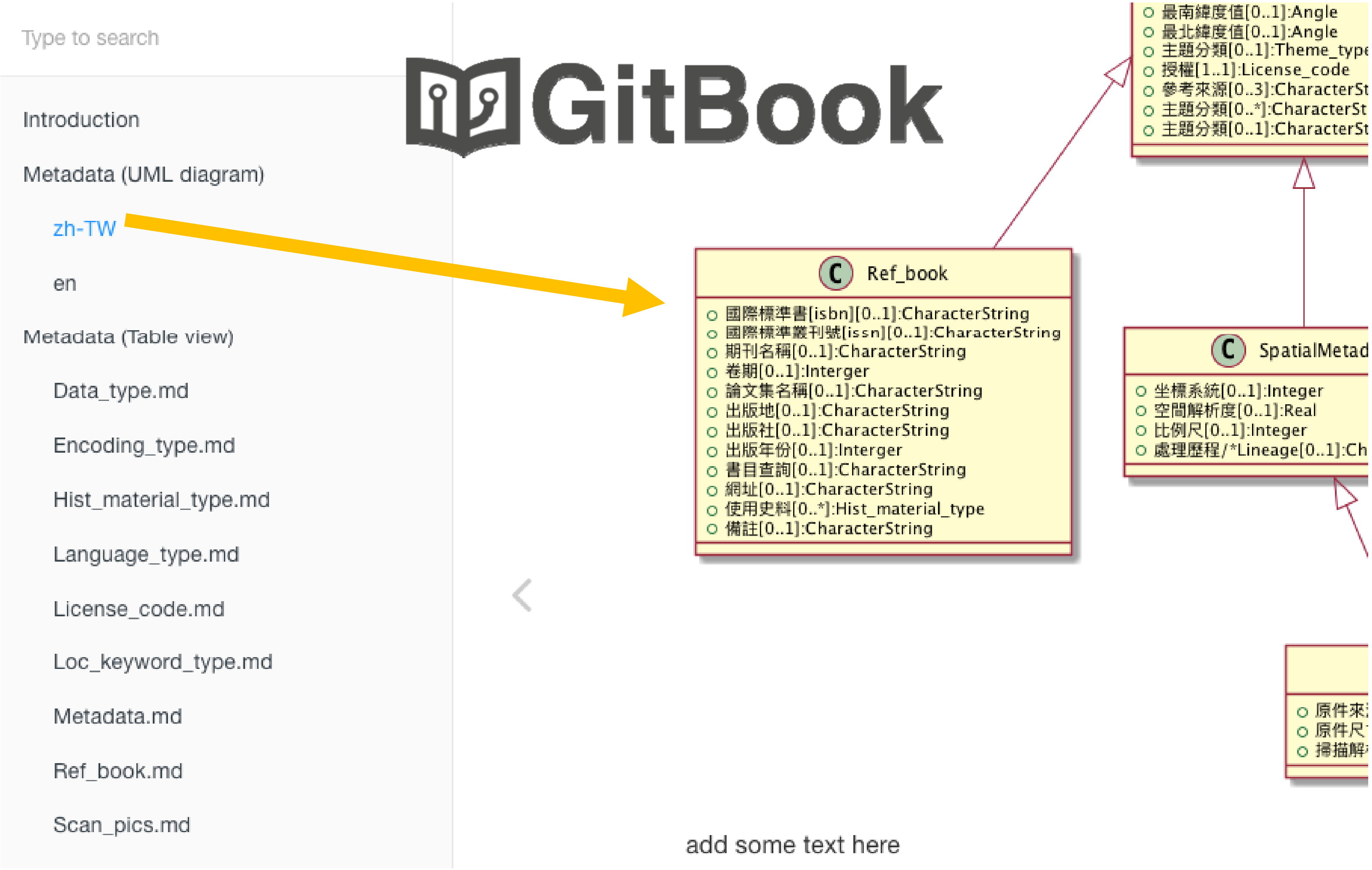
~~~
# UML diagram

UML diagram

```uml
@startuml

  Class Metadata {
    +title[1..1]:CharacterString
    /'{"def":"資料集名稱","zhTW":"標題","值域":"自由文字"}'/
    +name[1..1]:CharacterString
    /'{"def":"資料集詮釋資料網址 ","zhTW":"網址","值域":"unicode","not
~~~

Documentation workflow for metadata schema



Type to search

- Introduction
- Metadata (UML diagram)
  - zh-TW
  - en
- Metadata (Table view)
  - Data_type.md
  - Encoding_type.md
  - Hist_material_type.md
  - Language_type.md
  - License_code.md
  - Loc_keyword_type.md
  - Metadata.md
  - Ref_book.md
  - Scan_pics.md

GitBook

```
○ 最南緯度值[0..1]:Angle
○ 最北緯度值[0..1]:Angle
○ 主題分類[0..1]:Theme_type
○ 授權[1..1]:License_code
○ 參考來源[0..3]:CharacterSt
○ 主題分類[0..*]:CharacterSt
○ 主題分類[0..1]:CharacterSt
```

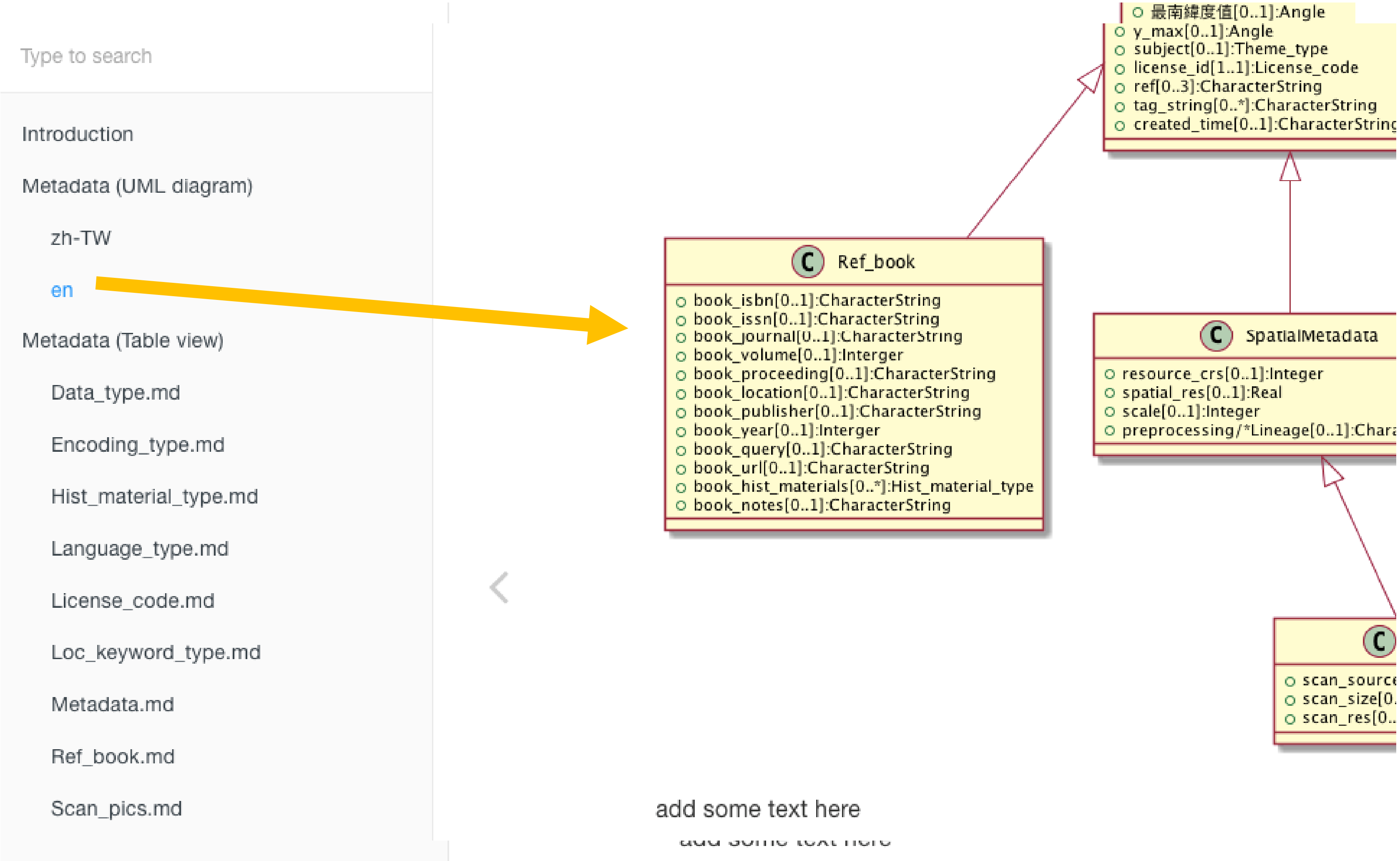
```
Ⓒ Ref_book
○ 國際標準書[isbn][0..1]:CharacterString
○ 國際標準叢刊號[issn][0..1]:CharacterString
○ 期刊名稱[0..1]:CharacterString
○ 卷期[0..1]:Interger
○ 論文集名稱[0..1]:CharacterString
○ 出版地[0..1]:CharacterString
○ 出版社[0..1]:CharacterString
○ 出版年份[0..1]:Interger
○ 書目查詢[0..1]:CharacterString
○ 網址[0..1]:CharacterString
○ 使用史料[0..*]:Hist_material_type
○ 備註[0..1]:CharacterString
```

```
Ⓒ SpatialMetad
○ 坐標系統[0..1]:Integer
○ 空間解析度[0..1]:Real
○ 比例尺[0..1]:Integer
○ 處理歷程/*Lineage[0..1]:Ch
```

```
○ 原件來
○ 原件尺
○ 掃描解
```

add some text here

Documentation workflow for metadata schema



Type to search

Introduction

Metadata (UML diagram)

zh-TW

en

Metadata (Table view)

Data_type.md

Encoding_type.md

Hist_material_type.md

Language_type.md

License_code.md

Loc_keyword_type.md

Metadata.md

Ref_book.md

Scan_pics.md

○ 最南緯度值[0..1]:Angle
○ y_max[0..1]:Angle
○ subject[0..1]:Theme_type
○ license_id[1..1]:License_code
○ ref[0..3]:CharacterString
○ tag_string[0..*]:CharacterString
○ created_time[0..1]:CharacterString

Ref_book

○ book_isbn[0..1]:CharacterString
○ book_issn[0..1]:CharacterString
○ book_journal[0..1]:CharacterString
○ book_volume[0..1]:Interger
○ book_proceeding[0..1]:CharacterString
○ book_location[0..1]:CharacterString
○ book_publisher[0..1]:CharacterString
○ book_year[0..1]:Interger
○ book_query[0..1]:CharacterString
○ book_url[0..1]:CharacterString
○ book_hist_materials[0..*]:Hist_material_type
○ book_notes[0..1]:CharacterString

SpatialMetadata

○ resource_crs[0..1]:Integer
○ spatial_res[0..1]:Real
○ scale[0..1]:Integer
○ preprocessing/*Lineage[0..1]:Chara

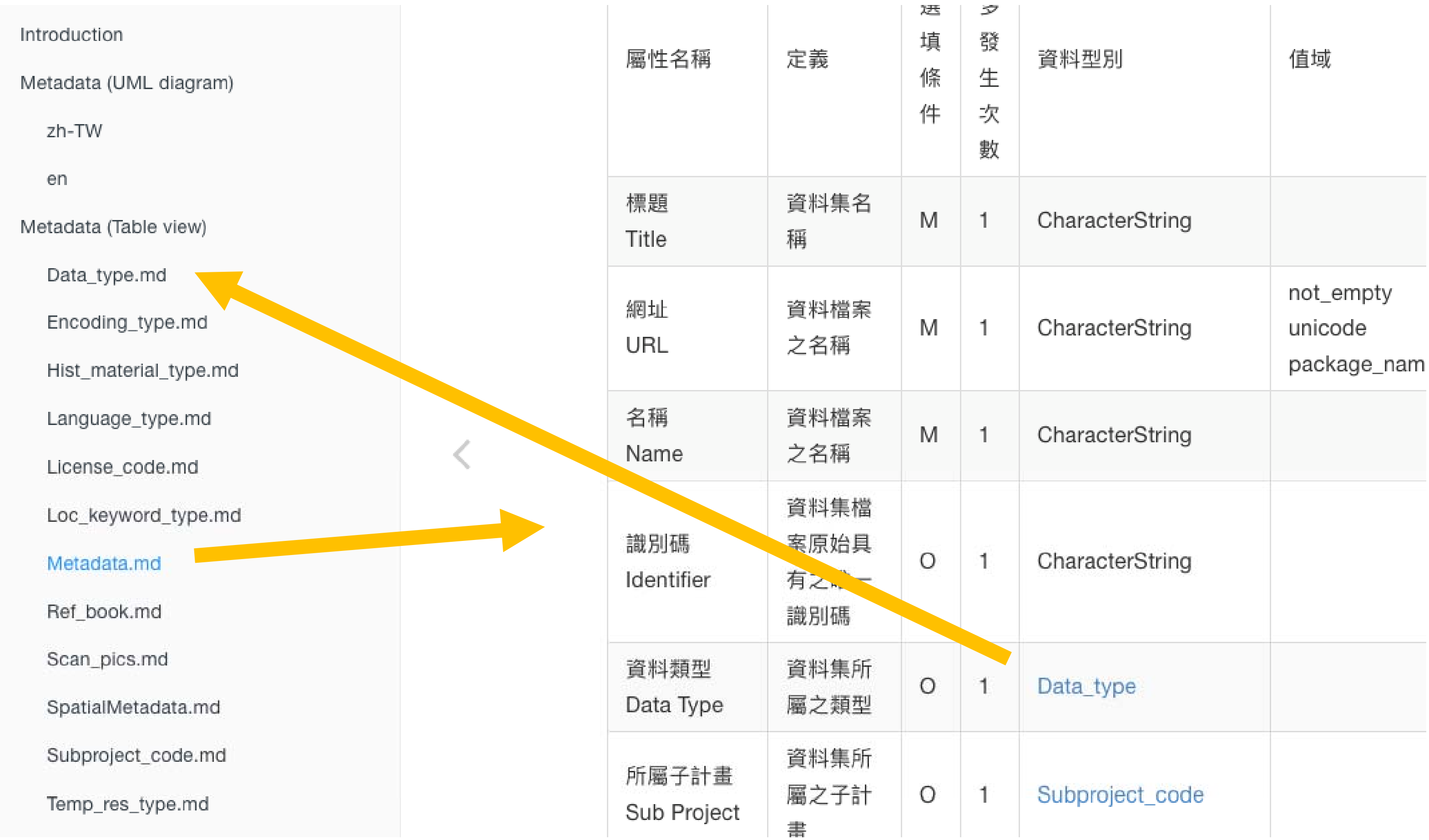
○ scan_source
○ scan_size[0.
○ scan_res[0..

add some text here

add some text here

Institute of Information Science, Academia Sinica, Taiwan

Documentation workflow for metadata schema



- Introduction
- Metadata (UML diagram)
  - zh-TW
  - en
- Metadata (Table view)
  - Data_type.md
  - Encoding_type.md
  - Hist_material_type.md
  - Language_type.md
  - License_code.md
  - Loc_keyword_type.md
  - Metadata.md
  - Ref_book.md
  - Scan_pics.md
  - SpatialMetadata.md
  - Subproject_code.md
  - Temp_res_type.md

屬性名稱	定義	選填條件	多發生次數	資料型別	值域
標題 Title	資料集名稱	M	1	CharacterString	
網址 URL	資料檔案之名稱	M	1	CharacterString	not_empty unicode package_nam
名稱 Name	資料檔案之名稱	M	1	CharacterString	
識別碼 Identifier	資料集檔案原始具有之唯一識別碼	O	1	CharacterString	
資料類型 Data Type	資料集所屬之類型	O	1	Data_type	
所屬子計畫 Sub Project	資料集所屬之子計畫	O	1	Subproject_code	

Current results and future works

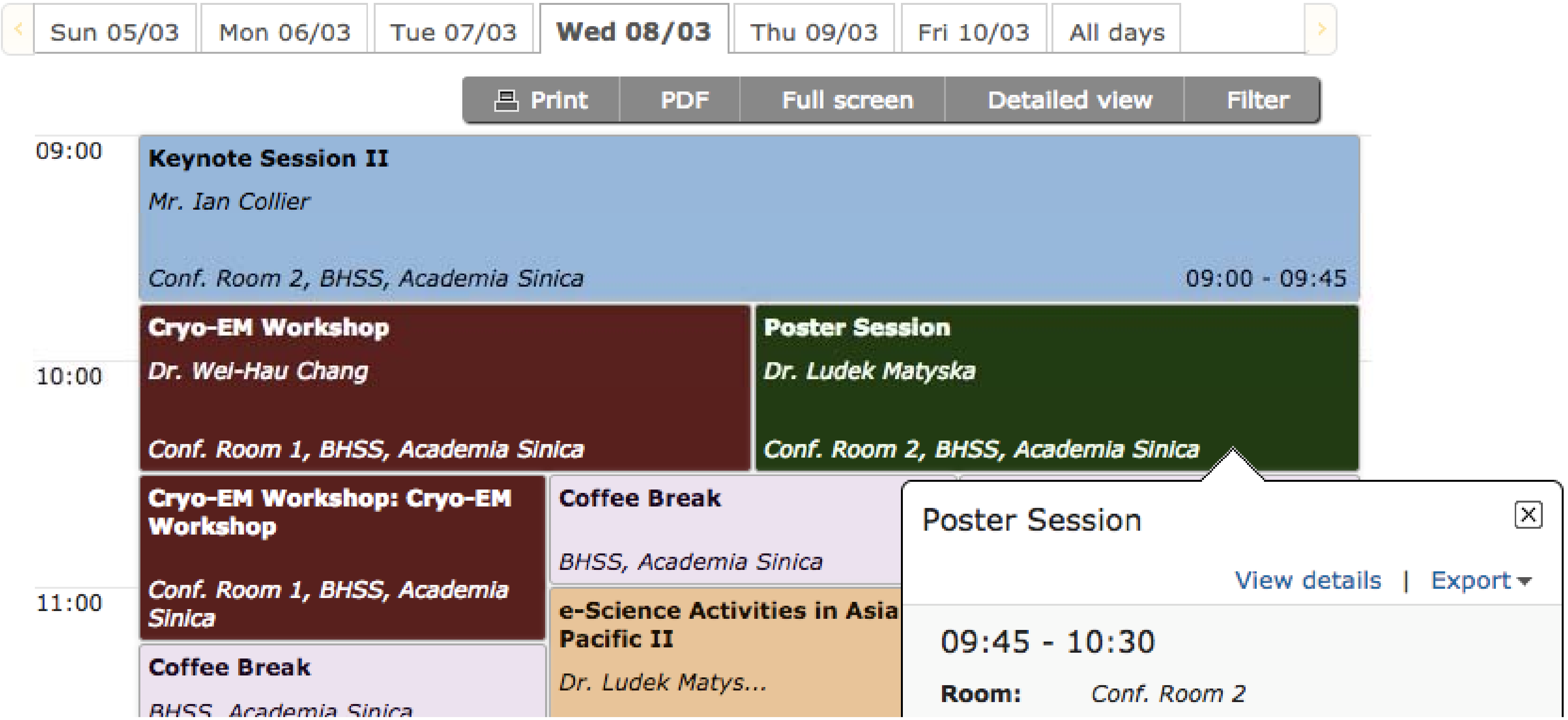
- **Rapid documentation for** CKAN-based repository and publish to GitBook.

- Metadata schema designers **only need to maintain UML files.**

- Publishing to GitBook **makes schema and documents searchable.**

Current results and future works

Metadata as Linked Data for Research Data Repositories (Mr. Cheng-Jen LEE)



09:00 **Keynote Session II**
Mr. Ian Collier
Conf. Room 2, BHSS, Academia Sinica 09:00 - 09:45

10:00

Cryo-EM Workshop
Dr. Wei-Hau Chang
Conf. Room 1, BHSS, Academia Sinica

Poster Session
Dr. Ludek Matyska
Conf. Room 2, BHSS, Academia Sinica

Cryo-EM Workshop: Cryo-EM Workshop
Conf. Room 1, BHSS, Academia Sinica

Coffee Break
BHSS, Academia Sinica

11:00

e-Science Activities in Asia Pacific II
Dr. Ludek Matys...

Coffee Break
BHSS, Academia Sinica

Poster Session
View details | Export
09:45 - 10:30
Room: *Conf. Room 2*