# Short commentary on **Museum online catalogues-as-data investigation** | data-focused work (January – March 2023)

Dr Anna-Maria Sichani, Jamie Unwin
(Critical reader: John Stack
Assistance with data/code: Kunika Kono)

The investigation started with an interest to explore the museum (online) catalogues from a technical and data-point of view, in order to assess potential obstacles and weaknesses to the catalogues' underlying data that limit their linking with other museums' collections and records. In order to do a first assessment, we use the SMG catalogue dataset as a case-study.

## Why the SMG catalogue dataset ?

The SMG catalogue dataset (objects) has been chosen as a case study for this investigation for the following reasons:

- We are able to get immediate access to the catalogue dataset itself and we're lucky to have also accessible 'pockets of knowledge' behind it (via SMG colleagues and stakeholders), so we can better assess its structure and content
- It is a rich dataset both in terms of history /legacy and in terms of length and variety of content , allowing us to explore various layers and levels of data

## Creating and working with a "functional" dataset

**data import**

We collaborated with Kunika on the data wrangling aspect of the SMG dataset JSON file using Jupyter Notebook. The SMG's objects dataset was downsized to a more easily workable subset, based on the key fields for exploration identified by Anna-Maria, and converted from JSON Lines to regular JSON format. For this task we used a command line JSON processor called jq. The subset was then flattened and loaded into Pandas DataFrame.

A valid JSON format of the file is now ready and checked for accuracy of data. We then loaded the file into Pandas framework for further exploration. We tried a couple of Py libraries to structure data around different fields.

# Initial remarks on specific data fields - points

**1.<date>**

There are different fields describing multiple date-related pieces of information per record, not always easy to demystify and to play around without the necessary contextual (curatorial and technical) knowledge :

- Date **object** created
    - it refers to when the object in the SMG collection was made
    - **Lifecycle.creation.date.value**
    - The other (sub)fields are variations on it, we support earliest/latest as some objects may have an earliest and latest date range in Mimsy ie. created between 1890 and 1895. Object can also have multiple “dates” ie. commissioned date, re-built date etc. Although these are not the norm.
    - Given the complex legacy history of the catalogue, it not always the creation date the date the first index card was written
    - It may be worth talking to documentation (Lawrence / Thomas) to see if we can extract the **TRUE** acquisition date of an object. My memory of past attempts to get hold of this data from Mimsy was that it was tricky for some reason. Although happy to reopen that conversation as it would be a useful data point.


```
"creation_date": [
   {
    "earliest": 1976,
    "latest": 1976,
    "primary": true,
    "value": "1976"
   }
  ]
```


- Date **record** created (**data.created**) or modified (**data.modified**)
    - It refers to when the record was created in Mimsy, which could be for a variety of reasons including the merging of museum collections (Science and Industry Museum Manchester becoming part of SMG in 2012 and the recent One Collection decant).
    - In short, this field isn’t used as people expect (e.g. when **the object was accessioned**) and so any findings drawn from it will be flawed.
    - this information is stored as second since the Epoch ie. 826588800000 (different – not ISO – format from the date object created field above) – you should be able to input those seconds into (https://www.programiz.com/python-programming/time) and get out a year
    - Both these fields are highly problematic in terms of accuracy within the wider legacy history of the catalogue: if I make a minor edit today that record will have a modification date (**data.modified**) of today, even if it was substantially edited 10 years ago.
    - If the question is, can we plot the historical ‘spikes’ in SMG accession activity, which I expect is already known to curatorial, there are better ways to find that out. Asking the documentation team to create an export direct from Mimsy of the ‘Accession table’ might be one route, although I know that team is

extremely busy and unfortunately, I have no experience of that area of Mimsy myself.


```
 "data": {
    "type": "objects",
    "id": "co8084947",
    "attributes": {
        },
        "created": 1172066776000,
        "modified": 1681926102000,
      },
```


**Interesting question:** When were there peaks in modification or creation dates (spoiler: the biggest trends fall in line with the mergers of museums or systems realignment; although below those changes there may be more interesting trends.).

**2. <Description>**

- Free-text description of the record
- **Description.value**
- We have **no historical audit logs (in Mimsy)**, so no means of showing when substantial edits were made.
- The description field is expected to provide context for linking among collection's records and across different collections using text mining and entity extraction approaches, e.g. looking for and linking via specific key terms across the collections description fields or investigating the historical use of 'problematic words'. At collection level this could be done by simply doing a wildcard search against the description field. However, given the free-text nature of the field and the error-proning, obscure legacy history of the field, this won't paint a full and accurate historical picture.
- It would be interesting to investigate the evolution over time for the 'description length' re. date **record** modified, with ambiguous conclusions from this (NB If I make a minor edit today that record will have a modification date of today, even if it was substantially edited ten years ago).
- We could look at 'description length' against SMG category like "Locomotives and Rolling Stock" (categories.name) — although again, not sure how useful that is.


```
"summary_title": "Stephenson's Rocket (steam locomotive)",
    "description": [
      {
        "date": [
          {
            "earliest": 2013,
            "latest": 2013,
            "primary": true,
            "value": "2013-10-31"
          }
        ],
        "primary": true,
        "type": "description",
```

```
        "value": "Steam locomotive, remains of Robert Stephenson's 0-2-2
locomotive 'Rocket', designed by Robert Stephenson and made by Robert
Stephenson & Co., Newcastle upon Tyne, England, 1829"
      },
      {
        "type": "web description",
        "value": "The Rocket, designed by Robert Stephenson (1803-1859), was
the clear winner in the locomotive trials held at Rainhill in 1829 to decide
the motive power for the Liverpool & Manchester Railway. \n\nFive locomotives
competed at the Rainhill trials – Cycloped, built by Thomas Shaw Brandreth;
Novelty, built by John Ericsson and John Braithwaite; Perseverance, built by
Timothy Burstall; Sans Pareil, built by Timothy Hackworth; and the
Rocket.\nThe Rocket was the only locomotive to successfully complete the
trials, averaging 12 miles per hour and achieving a top speed of 30 miles per
hour. The Stephensons won the £500 prize and were awarded the contract to
produce locomotives for the Liverpool & Manchester Railway.\n\nWhen the
Liverpool & Manchester Railway was opened in 1830, Rocket, driven by John
Locke, was approaching Parkside Station where dignitaries were gathered on
the track after disembarking from their own special train. William Huskisson
MP was unable to get off the track safely and Rocket collided with his leg as
he tried to get out of its path. His injury proved to be
fatal.\n\nStephenson's Rocket represented a substantial technical advance
over previous designs, bringing together in one machine such developments as
the multi-tube boiler and the blast-pipe. The speed of progress, however,
meant that the Rocket was substantially rebuilt within 18 months and laid
aside within 10 years. It was preserved in 1862 incomplete and semi-derelict
but it had set the pattern for the future and remains the most significant of
all early railway locomotives.\nIn 1862, Rocket was donated to the Patent
Office Museum in London, which became the Science Museum.\nIn 1998, Rocket
went on display outside the UK for the first time, travelling to Japan for
the Festival UK 98, a festival of British arts and science, opened by the
then-Prime Minister Tony Blair."
      }
    ],
```


**3. <categories>**

- a flex-field (a separate table connected via a JOIN relationship in SQL) corresponding to **an external taxonomy in value**
- **Categories.museum.name.value**
- This field isn't commonly used across most records, but it does get used.
- These values tend to be from a legacy import of the Getty ATT thesaurus or the Mda Railway Object Name Thesaurus maintained by the Collections Trust. Although even when a object is linked to a Thesaurus term in Mimsy (at a database level) we do not hold a UID or URI for those terms in Getty ATT or elsewhere (although in some case it is possible to work out the Getty UID/URI by using the parent term; but this involved extra work/processing).
- **The full list on Categories (merged across museums) can be viewed here** https://collection.sciencemuseumgroup.org.uk/categories


```
    "name": [
      {
        "primary": true,
        "value": "steam locomotive"
      }
    ],
```

```
  "categories": [
    {
      "museum": "NRM",
      "name": "Locomotives and Rolling Stock",
      "value": "NRM - Locomotives and Rolling Stock"
    }
  ],
```

# An attempt to map fields at the online collections site to mimsy catalogue fields

| Collections site | Mimsy Catalogue | description | comments |
|---|---|---|---|
| description | description": [<br>{<br>"primary": true,<br>"value": "Steam locomotive and tender, British Railways, 9F class 2-10-0 No 92220 \"Evening Star\", designed by R.A.Riddles, built at Swindon in 1960, withdrawn in 1965."<br>},<br>{<br>"type": "web description",<br>"value": "Evening Star is historically significant as the 999th British Rail (BR) Standard, and indeed the last steam locomotive to be built by BR. \nEvening Star was the only 9F to be painted in BR's express passenger service livery of lined green. The name Evening Star was chosen following a competition held by the BR Western Region Staff Magazine. There were three winners, who all suggested Evening Star – a fitting name given that one of the first locomotives to run on the Great Western Railway was named Morning Star. \n\nEvening Star had an extremely short life span for a steam locomotive and was unexpectedly withdrawn from service in 1965. The locomotive was claimed for the National Collection in 1975.\n\nhttps://en.wikipedia.org/wiki/BR_Standard_Class_9F_92220_Evening_Star"<br>} | | |

| | | | |
|---|---|---|---|
| | ], | | |
| title | "summary_title": "Steam locomotive, entitled Evening Star (steam locomotive)",<br>"title": [<br>{<br>"primary": true,<br>"type": "display title",<br>"value": "Steam locomotive, entitled Evening Star"<br>}<br>], | | |
| type | "name": [<br>{<br>"primary": true,<br>"type": "object type",<br>"value": "steam locomotive"<br>}<br>], | | In most cases this is simply a free text value/ string (although it auto-completes in Mimsy so you get repeated usage of the same term), because it is free text you also get mis-spelling and alternative spellings.<br><br>Importantly, these object names are also not linked to any formal taxonomy. Although many of them do indirectly derive indirectly from Getty ATT (https://www.getty.edu/research/tools/vocabularies/aat/) this is not enforced in Mimsy nor do we have a ID or URI for those Getty terms in Mimsy. This is a hindrance to linking across collections. |
| Taxonomy | "name": [<br>{<br>"primary": true,<br>"type": "preferred",<br>"value": "VEHICLES AND VEHICLE COMPONENTS"<br>}<br>], | | A record can also be related to an external taxonomy in value. In which case the relationship is held in a flex-field (a separate table connected via a JOIN relationship in SQL). |

| | | | |
|---|---|---|---|
| Catalogue - Category | "categories": [<br>{<br>"museum": "NRM",<br>"name": "Locomotives and Rolling Stock",<br>"value": "NRM - Locomotives and Rolling Stock"<br>}<br>], | | This relates to the SMG curatorial department.The full list on Categories (merged across museums) can be viewed here. https://collection.sciencemuseumgroup.org.uk/categories |
| materials | "materials": [<br>"brass (copper, zinc alloy)",<br>"copper (alloy)",<br>"copper plated",<br>"enamel",<br>"glass",<br>"gunmetal",<br>"paint",<br>"rubber (unidentified)",<br>"steel",<br>"tin (metal)",<br>"vulcanised rubber",<br>"zinc plate"<br>], | | These are simply free text strings, but often indirectly come from Getty ATT (but are not linked or associated to Getty or any UID or URI) |
| <creation date> | "lifecycle": {<br>"creation": [<br>{<br>"date": [<br>{<br>"earliest": 1960,<br>"latest": 1960,<br>"primary": true,<br>"value": "1960"<br>}<br>], | | |

## Code and documentation

The code and data used and/or produced as part of this investigation can be found in the CE GitHub repo:

- the SMG objects dataset as uncompressed/unzipped file can be found here : smg_objects_06_06_2022.json
- SMG dataset mapping.xlsx: The key fields identified by AMS for analysis.
- [smg_objects_06_06_2022_extracted.json]: Dataset containing only the key fields extracted from the SMG objects dataset.
- [smg_objects_06_06_2022_extracted_converted.json]: Copy of smg_objects_06_06_2022_extracted.json as converted to regular JSON.

- smg_objects.ipynb: Jupiter Notebook containing the code used for data extraction. Data was extracted and exported as provided, no cleaning nor transformation was applied.

The following are CSVs of each key fields (see mapping, with nested fields flattened and extracted into separate columns and multi-value (array) field expanded into individual rows:

- smg_objects_categories.csv
- smg_objects_description.csv
- smg_objects_lifecycle_creation_date.csv
- smg_objects_materials.csv
- smg_objects_name.csv
- [smg_objects_title.csv]

N.B. Column names are in dot notation format, based on field data structure e.g. smg_objects_categories.csv:

| JSON | CSV |
|---|---|
| "categories": [ | |
| { | |
| "museum": "SCM" | category.museum |
| "name": "Veterinary Medicine" | category.name |
| "value": "SCM - Veterinary Medicine" | category.value |
| } | |
| ] | |

All the CSVs include object ids, which can be used to reconstruct the full extracted dataset e.g. by importing each CSV into a Pandas dataframe and joining/merging the dataframes on the [id] field.

## Comments - Remarks

- The actual dataset file (json), used as an entry point to the collections' data, is a JSON Lines file, which, unless indicated (e.g. with .jsonl extension), is not immediately obvious to a non-specialist.
- Cataloguing standards have grown organically and legacy cataloguing data is, therefore, inconsistent.
- (online) catalogues' data fields have been used inconsistently throughout the years, with limited or without documentation at all. Inconsistency or messy data refers to
    - different uses and qualities of information employed by different people through the years
    - incorrect use of fields - human-in-the-loop and error-prone
    - information format (eg dates) difficult to manage
- Data fields contain layers of information that is difficult – if not impossible – to reveal, assess and use.
- Catalogue records are thin, often containing very few fields and data in them, and many objects are similarly catalogued, thus limiting from the outset any external linkage, including between them, with other collections or related material.
- The dataset has only internal IDs for its records, often with duplications of records in place (ie the known 'whole-part' issue) with no links to external sources, further complicating disambiguation of entries.

Based on these findings and limitations, we were wondering how feasible and/or useful for this investigation and the project at large a closer analysis and visualisation of the SMG catalogue data would be, exploring information such as 'date object created over time' and 'description length over time', might be. Our focus here is not just to investigate or assess the SMG collection per se, but mainly to assess the collection's data, its structure and infrastructural/technical characteristics, as per its potential linkage to other collections and resources.

Having this in mind, we embarked on developing a data-informed, comprehensive report on the obstacles and weaknesses of the catalogues' underlying data, also from an infrastructure point of view, that limits their linking potential and proposed an assessment framework and matrix, before focusing on innovative technical solutions and prototypes. We believe that it is important to understand what and why is not working from a data and infrastructure point of view in museum collections' datasets before deciding to fix it.

A thorough assessment of museums' catalogues and datasets should be among the first steps, alongside a needs-based analysis and resources assessment, towards the development of technical solutions that will enable their access, creative (re)use and linking with other collections for various audiences.