digirati

# SM20C Project

## Post digitisation metadata enrichment tools evaluation report

| | |
|---|---|
| Date: | 13th September 2019 |
| Version: | 1.1 |

# Contents

**Contents** 2

**1. Summary** 4

**2. Introduction** 5

2.1 Background 5

2.2 Approach 5

**3. Workflows** 6

3.1 User journeys and experience maps for existing workflows 6

3.2 Revised workflows for exploiting identified ‘digital first/early’ opportunities 7

**4. An assessment of tools and methods** 10

4.1 Methodology 10

4.2 Grouping and assessment 12

4.2.1 Grouping 13

4.2.2 Stakeholder evaluation 14

4.3 Opportunities for evaluation 15

4.3.1 Text 15

4.3.2 Subject and topic description 15

4.3.3 Named entity recognition 15

4.3.4 Controlled vocabularies and taxonomies 15

**5. Solution Blueprint** 16

5.1 Workflow Overview 17

5.1.1 Page or image centric steps 17

5.1.2 Object level steps 18

5.2 Implementation 18

5.2.1 DLCS Gap Analysis (Page level workflow) 19

Natural language processing (NLP) 20

Handwriting recognition (HTR) and OCR 20

Management interface/dashboards 21

Workflow segregation 21

5.2.2 DLCS Gap Analysis (Object level workflow) 21

Sorting Room 22

Keywording 22

Subject and topic tagging 23

Taxonomies and controlled vocabularies 24

Catalogue metadata, metadata reformatting, and metadata convergence 24
Metadata reformatting and metadata convergence 26
5.2.3 What this means in the context of IIIF 27
Machine readable metadata 27
Linking properties 27
IIIF Discovery 28
5.3 What this means in the context of SM20C Digital First workflows 28
5.3.1 LSE 29
5.3.2 UCL 29
5.3.3 SM20C Content 30
5.4 Potential roadmap 31
Page level workflow steps 31
Object level workflow steps 33
5.5 Critical recommendations 35
Appendix 1: Jisc Open research hub 37

# 1. Summary

The aim of this evaluation is to explore and evaluate potential solutions to current issues with the cataloguing of special collections which are candidates for digitisation, and to explore technological approaches to easing bottlenecks in digitisation imposed by cataloguing.

This investigation was conducted as part of Jisc's Social Movements of the 20th Century (SM20C) project. Through a number of collaborative workshops with Jisc, UCL and LSE (see Section 3), and stakeholder interviews with researchers and other Jisc stakeholders, Digirati explored requirements for both researchers and archivists/special collections librarians.

First, we identified the current workflows that are in use at UCL and LSE, before mapping the experiences and user journeys through these workflows. These experience maps enabled us to identify areas where there were pain points, or bottlenecks, and consider how these areas might be assisted through technology. In these same workshops Digirati presented a number of technological approaches—using IIIF and machine processing—for review and assessment by stakeholders.

Second, we identified a number of key opportunities for improvements, which were ranked by stakeholders, and the highest ranked opportunities were:

- IIIF for digital surrogates and as an enabling/disseminating technology
- Searchable text via handwriting recognition and OCR
- Named entity recognition of people, places, dates and organisations.
- Keywording (by machine methods)
- Subject classification (by machine methods)
- Discovery and the creation of convergent/common metadata for use by: Internet Archive, Jisc portals, and IIIF Discovery.

Digirati provided some key recommendations around workflow steps that could be built to provide solutions for these opportunities. See 5. Solution Blueprint below.

Finally, given time and budget constraints, we consider what the minimum viable product might be. See 5.5 Critical recommendations below.

# 2. Introduction

## 2.1 Background

The aim of this evaluation is to explore and evaluate potential solutions to current issues with the cataloguing of special collections which are candidates for digitisation. This includes both low-quality cataloguing and inconsistencies with cataloguing languages which Jisc has experienced on a number of digitisation projects. Furthermore the time taken to manually catalogue collections in order that they can be digitised has proved a major bottleneck.

This investigation was conducted as part of Jisc's Social Movements of the 20th Century (SM20C) project in partnership with UCL and the London School of Economics, another 11 HEI institutions are also involved in the project. The initial aim of the SM20C project is to focus on pamphlets that were produced between the first and second world wars by women's rights organisations at that time.

## 2.2 Approach

Through a number of collaborative workshops with Jisc, UCL and LSE (see Section 3), and stakeholder interviews with researchers and other Jisc stakeholders, Digirati explored requirements for both researchers and archivists/special collections librarians. We looked to understand the context of SM20C project and wider Jisc digitisation goals as well as the potential and merits of integrating with Jisc Open Research Hub.

We examined how these requirements and current issues could potentially be addressed via a 'digital first' approach to digitisation that utilises semantic tools and the latest open standards such as the International Image Interoperability Framework (IIIF) and W3C annotations. Potential 'Digital first' workflows, and the underlying digital methods and tools that would be required, were assessed and prioritised by UCL and LSE stakeholders.

Further desk research was then completed to collate and evaluate the findings from the workshops and interviews and create a set of recommendations in the form of a Solution Blueprint and High Level Roadmap (see Section 5). The work required for the various roadmap items has been estimated based on using the Digital Libraries Cloud Service (DLCS) as an underlying platform which provides a number of the required services out-of-the-box and is built to be extended for these use cases. The DLCS was also used to provide example digital surrogates and data for illustrative purposes during the workshops and interviews.

# 3. Workflows

## 3.1 User journeys and experience maps for existing workflows

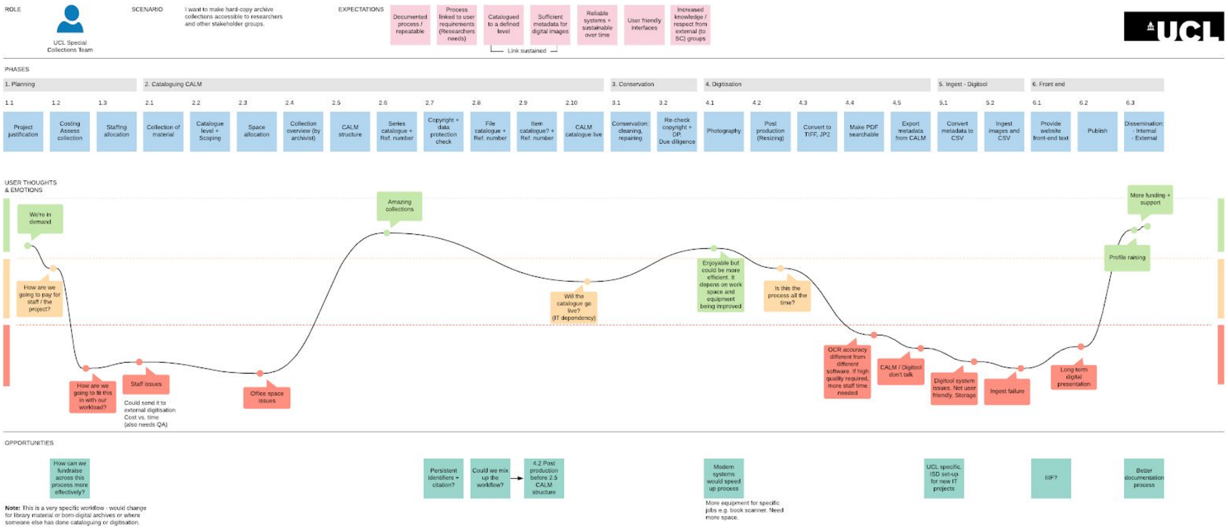
A key activity during the workshops was to map the current workflows and user experience for the metadata enrichment of digitized collections at UCL and LSE. Archivists and librarian teams from both institutions, supported by the Jisc and Digirati teams, produced their own journey maps to visualise their workflows.

User actions, thoughts and emotions were captured in the visualisation to help uncover moments of both frustration and delight throughout the interactions with their current processes.

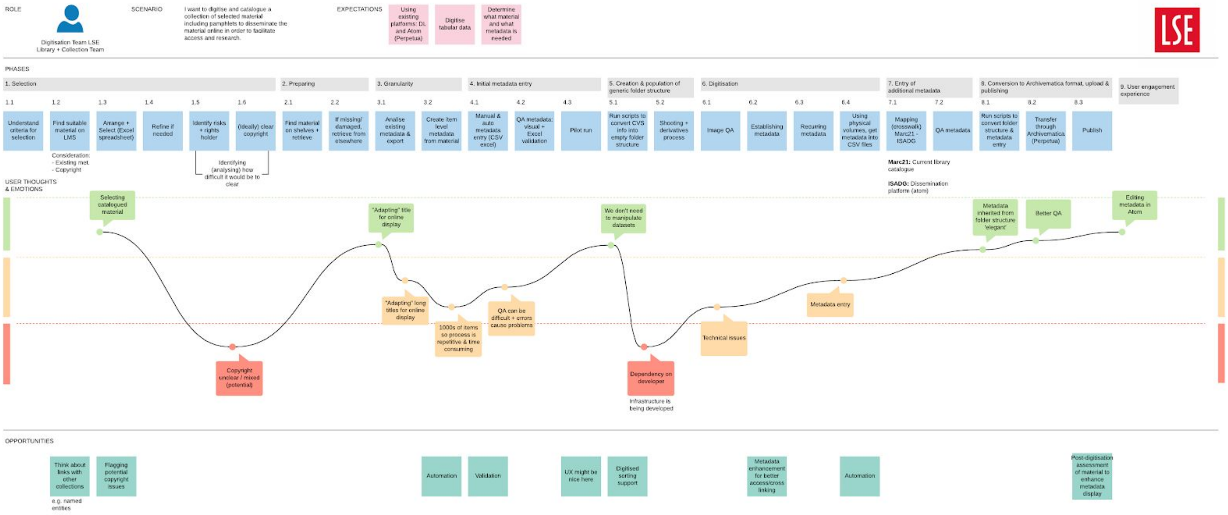
The mapping activity also allowed all parties to start discussing insights and opportunities that could improve these workflows.

*UCL and LSE discussing the archivist teams user journey maps during the workshops facilitated by Jisc and Digirati at Jisc's offices.*

The experience maps for both UCL and LSE can be found below:


ROLE
UCL Special Collections Team
SCENARIO
I want to make hard-copy archive collections accessible to researchers and other stakeholder groups.
EXPECTATIONS
Documented process / repeatable
Process linked to user requirements (Researchers needs)
Catalogued to a defined level
Sufficient metadata for digital images
Link sustained
Reliable systems + sustainable over time
User friendly interfaces
Increased knowledge / respect from external (to SC) groups
UCL
PHASES
1. Planning
2. Cataloguing CALM
3. Conservation
4. Digitisation
5. Ingest - Digitool
6. Front end
1.1 Project justification
1.2 Costing Assess collection
1.3 Staffing allocation
2.1 Collection of material
2.2 Catalogue level + Scoping
2.3 Space allocation
2.4 Collection overview (by archivist)
2.5 CALM structure
2.6 Series catalogue + Ref. number
2.7 Copyright + data protection check
2.8 File catalogue + Ref. number
2.9 Item catalogue? + Ref. number
2.10 CALM catalogue live
3.1 Conservation: cleaning, repairing
3.2 Re-check copyright + DP. Due diligence
4.1 Photography
4.2 Post production (Resizing)
4.3 Convert to TIFF, JP2
4.4 Make PDF searchable
4.5 Export metadata from CALM
5.1 Convert metadata to CSV
5.2 Ingest images and CSV
6.1 Provide website front-end text
6.2 Publish
6.3 Dissemination: - Internal - External
USER THOUGHTS & EMOTIONS
We're in demand
How are we going to pay for staff / the project?
How are we going to fit this in with our workload?
Staff issues
Could send it to external digitisation Cost vs. time (also needs QA)
Office space issues
Amazing collections
Will the catalogue go live? (IT dependency)
Enjoyable but could be more efficient. It depens on work space and equipment being improved
Is this the process all the time?
OCR accuracy different from different software. If high quality required, more staff time needed
CALM / Digitool don't talk
Digitool system issues. Not user friendly. Storage
Ingest failure
Long term digital presentation
Profile raising
More funding + support
OPPORTUNITIES
How can we fundraise across this process more effectively?
Note: This is a very specific workflow - would change for library material or born-digital archives or where someone else has done cataloguing or digitisation.
Persistent identifiers + citation?
Could we mix up the workflow?
4.2 Post production before 2.5 CALM structure
Modern systems would speed up process
More equipment for specific jobs e.g. book scanner. Need more space.
UCL specific. ISD set-up for new IT projects
IIIF?
Better documentation process

UCL experience map is also available online

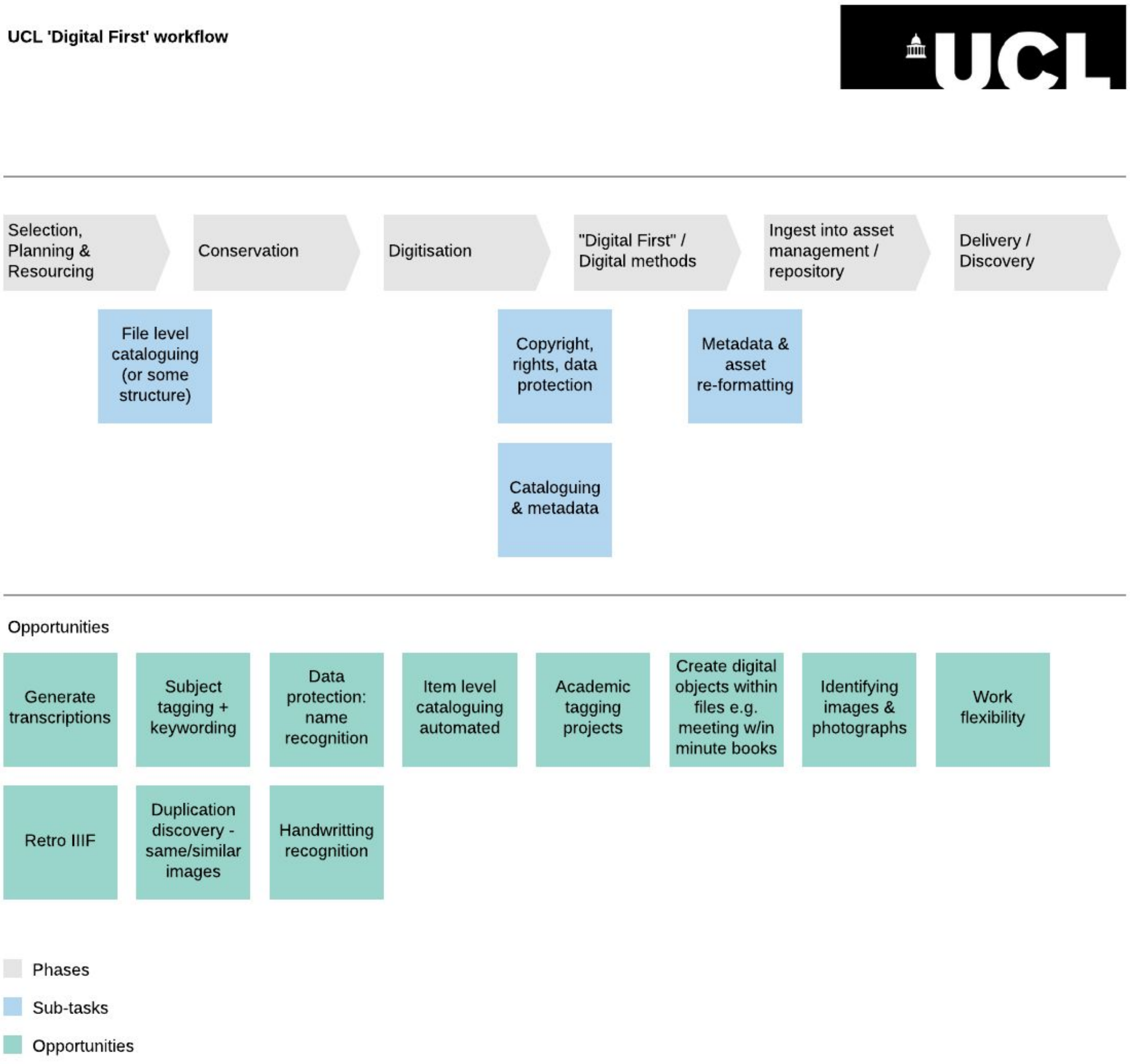

ROLE
Digitisation Team LSE Library + Collection Team
SCENARIO
I want to digitise and catalogue a collection of selected material including pamphlets to disseminate the material online in order to facilitate access and research.
EXPECTATIONS
Using existing platforms: DL and Atom (Perpetua)
Digitise tabular data
Determine what material and what metadata is needed
LSE
PHASES
1. Selection
2. Preparing
3. Granularity
4. Initial metadata entry
5. Creation & population of generic folder structure
6. Digitisation
7. Entry of additional metadata
8. Conversion to Archivematica format, upload & publishing
9. User engagement experience
1.1 Understand criteria for selection
1.2 Find suitable material on LMS
Consideration: - Existing met. - Copyright
1.3 Arrange + Select (Excel spreadsheet)
1.4 Refine if needed
1.5 Identify risks + rights holder
1.6 (Ideally) clear copyright
Identifying (analysing) how difficult it would be to clear
2.1 Find material on shelves + retrieve
2.2 If missing/ damaged, retrieve from elsewhere
3.1 Analise existing metadata & export
3.2 Create item level metadata from material
4.1 Manual & auto metadata entry (CSV excel)
4.2 QA metadata: visual + Excel validation
4.3 Pilot run
5.1 Run scripts to convert CVS info into empty folder structure
5.2 Shooting + derivatives process
6.1 Image QA
6.2 Establishing metadata
6.3 Recurring metadata
6.4 Using physical volumes, get metadata into CSV files
7.1 Mapping (crosswalk) Marc21 - ISADG
7.2 QA metadata
Marc21: Current library catalogue
ISADG: Dissemination platform (atom)
8.1 Run scripts to convert folder structure & metadata entry
8.2 Transfer through Archivematica (Perpetua)
8.3 Publish
USER THOUGHTS & EMOTIONS
Selecting catalogued material
Copyright unclear / mixed (potential)
"Adapting" title for online display
"Adapting" long titles for online display
1000s of items so process is repetitive & time consuming
QA can be difficult + errors cause problems
We don't need to manipulate datasets
Dependency on developer
Infrastructure is being developed
Technical issues
Metadata entry
Metadata inherited from folder structure 'elegant'
Better QA
Editing metadata in Atom
OPPORTUNITIES
Think about links with other collections
e.g. named entities
Flagging potential copyright issues
Automation
Validation
UX might be nice here
Digitised sorting support
Metadata enhancement for better access/cross linking
Automation
Post-digitisation assessment of material to enhance metadata display

LSE experience map is also available online

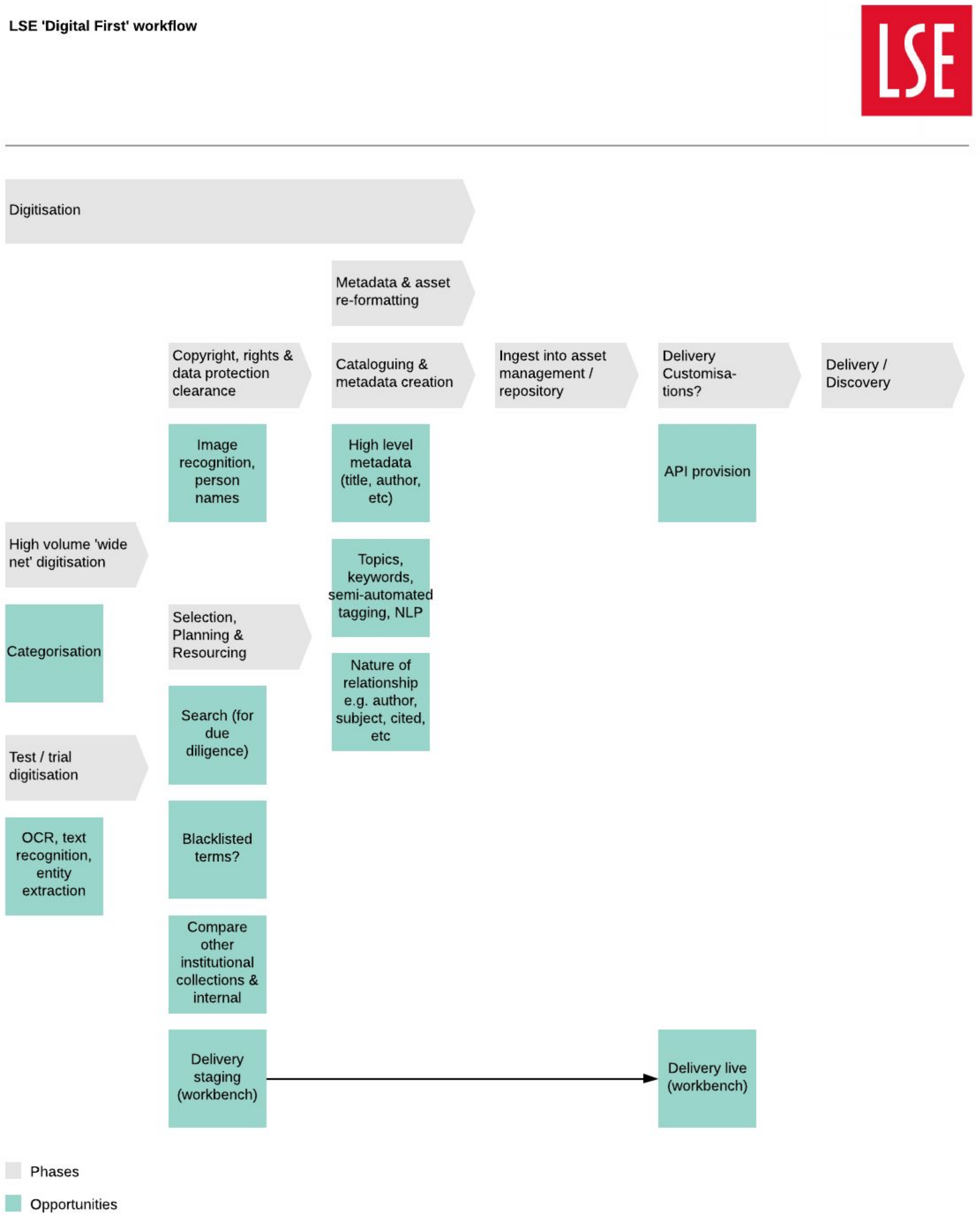
## 3.2 Revised workflows for exploiting identified 'digital first/early' opportunities

The experience maps for the existing workflows allowed us to identify shared opportunities that could inform the development of potential 'digital first/early' workflow solutions. In workshop 2, Digirati proposed a draft 'digital first/early' workflow based on these opportunities. The task that followed in the workshop involved UCL and LSE to individually iterate over this draft workflow so it could be integrated into their own existing practices.

The diagrams for the resulting integrated versions of the ‘digital first/early’ workflows for both UCL and LSE are displayed below. These diagrams reflect slightly different approaches to organising this information that were taken during the workshops. LSE phases often overlap and so it was useful to reflect that and map the ‘opportunities’ to the different phases to help clarify. UCL’s workflow was more straightforward and the opportunities were simply collated rather than mapped directly to workflow steps during the exercise (although this was discussed  later).


UCL 'Digital First' workflow
UCL
Selection, Planning & Resourcing
Conservation
Digitisation
"Digital First" / Digital methods
Ingest into asset management / repository
Delivery / Discovery
File level cataloguing (or some structure)
Copyright, rights, data protection
Metadata & asset re-formatting
Cataloguing & metadata
Opportunities
Generate transcriptions
Subject tagging + keywording
Data protection: name recognition
Item level cataloguing automated
Academic tagging projects
Create digital objects within files e.g. meeting w/in minute books
Identifying images & photographs
Work flexibility
Retro IIIF
Duplication discovery - same/similar images
Handwritting recognition
Phases
Sub-tasks
Opportunities

UCL ‘digital first/early’ workflow diagram is also available online


LSE 'Digital First' workflow
LSE
Digitisation
Metadata & asset re-formatting
Copyright, rights & data protection clearance
Cataloguing & metadata creation
Ingest into asset management / repository
Delivery Customisa-tions?
Delivery / Discovery
Image recognition, person names
High level metadata (title, author, etc)
API provision
High volume 'wide net' digitisation
Topics, keywords, semi-automated tagging, NLP
Categorisation
Selection, Planning & Resourcing
Nature of relationship e.g. author, subject, cited, etc
Search (for due diligence)
Test / trial digitisation
OCR, text recognition, entity extraction
Blacklisted terms?
Compare other institutional collections & internal
Delivery staging (workbench)
Delivery live (workbench)
Phases
Opportunities

LSE ‘digital first/early’ workflow diagram is also available online

The opportunities that emerged from fleshing out and discussing these ‘digital first/early’ workflows translated into key tools and methods which were then categorised and voted by UCL and LSE for their potential for impact as discussed in section 4.2.

# 4. An assessment of tools and methods

## 4.1 Methodology

In preparation for the two workshops, Digirati produced some IIIF digital surrogates of UCL archival objects digitised for the SM20C project.

These surrogates were provided with:

1. *IIIF based interoperable images*
2. *OCR text*
3. *Machine generated extractive summaries*
4. *Keywords*
5. *Topic modeling or subject tagging* (done via simple fuzzy matching rather than more advanced approaches)
6. *Named entity extraction:* identifying people, places, dates, and organisations.
7. *Parts of speech tagging and other natural language processing outputs:* such as lemmatization, sentence boundary detection, etc.

This data, along with explanatory text, was provided to the LSE and UCL stakeholders in advance of the second workshop for review and assessment. This data can be found here: https://digirati-co-uk.github.io/sm20c-workshop-samples/summary.html

In the workshop, UCL and LSE teams were asked to revisit their experience maps (see 3.1 User journeys and experience maps for existing workflows above) after having had a chance to review Digirati's initial interim report.

Following this, UCL and LSE teams were asked to consider what their workflows might look like if they employed digital methods earlier in the workflow to assist with pain points in their existing process, such as:

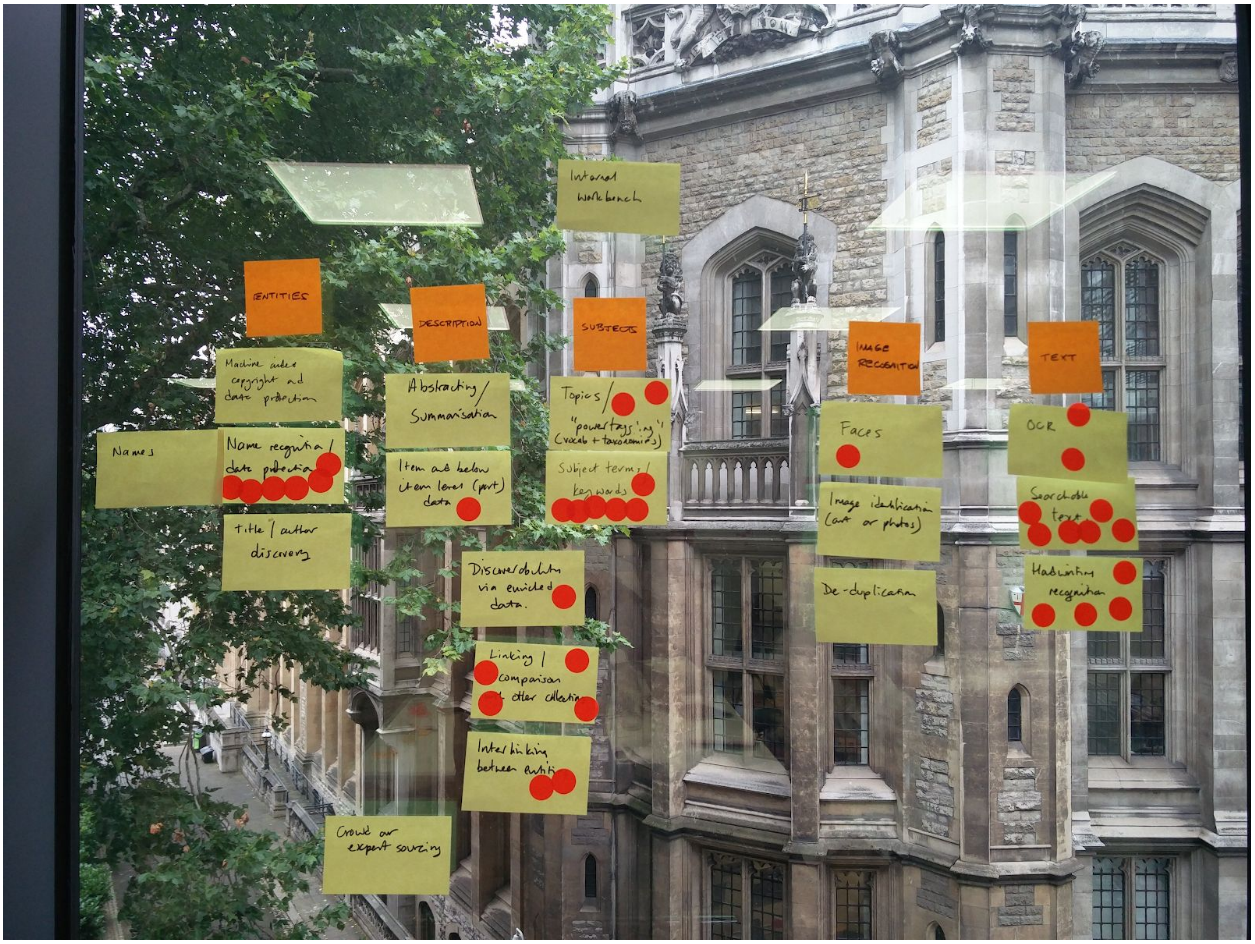
- Enabling more efficient series and item level cataloguing, by:
    - Making more data available earlier to the cataloguer
    - Reducing the amount of manual data entry required
- Reducing space and staffing constraints, because:
    - objects are off the shelves for less time
    - Item level cataloguing can happen after digitisation, removing a potential bottleneck
    - Cataloguing can potentially happen from anywhere, even off-site

- Easing metadata transformation and convergence through the production of machine readable metadata earlier in the workflow.

See 3.2 Revised workflows for exploiting identified ‘digital first/early’ opportunities above for details.

## 4.2 Grouping and assessment

Reviewing these “digital first”, or “digital earlier” workflows, UCL and LSE stakeholders identified a number of key tools or methods which might be useful, and then clustered these into broad categories.


Internal workbench
ENTITIES
DESCRIPTION
SUBJECTS
IMAGE RECOGNITION
TEXT
Machine aided copyright and data protection
Names
Name recognition / data protection
title / author discovery
Abstracting / Summarisation
Item and below item level (part) data
Topics / "powertags'ing'" (vocab + taxonomies)
Subject terms / keywords
Discoverability via enriched data.
Linking / comparison with other collections
Interlinking between entities
Crowd or expert sourcing
Faces
Image identification (art or photos)
De-duplication
OCR
Searchable text
Handwriting recognition

After clustering, stakeholders were asked to vote for those tools or methods they thought would be most useful in their potential “digital first” workflow.

### 4.2.1 Grouping

| Category | Tool or method |
|---|---|
| Text | OCR |
| | Handwriting recognition |
| | Searchable text |
| Entities | Named entity recognition, especially of people |
| | Title / author discovery |
| | Machine aided copyright and data protection assessment |
| Description | Abstracting / summarisation |
| | Item level and below item level data |
| Subjects | Subject terms / keywords |
| | “Power-tagging” with controlled vocabularies and taxonomies |
| Image Recognition | Face recognition |
| | Identification of photographs or art works |
| | Deduplication |
| Discovery and interlinking | Discoverability via enriched data |
| | Linking and comparison with other collections or objects |
| | Interlinking between entities |
| | Crowd or expert sourcing |

### 4.2.2 Stakeholder evaluation

Stakeholders were clear that they were assessing these tools both for their usefulness, and for their likely effectiveness.

For example, while copyright clearance is a major consumer of staff time and resources, it was felt that machine assisted copyright clearance was unlikely to be effective given the complexity of this task and the risk attached to error.

Ordering those methods which received votes, by votes received:

| Method | Category | Votes |
|---|---|---|
| Searchable text | Text | 6 |
| Named entity recognition | Entities | 6 |
| Subject terms and keywords | Subjects | 6 |
| Handwriting recognition | Text | 4 |
| Linking and comparison with other collections or objects | Discovery and interlinking | 4 |
| OCR | Text | 2 |
| Interlinking between entities | Discovery and interlinking | 2 |
| “Powertagging” | Subjects | 2 |
| Face recognition | Image recognition | 1 |
| Item and below item level data | Description | 1 |

If we further cluster these, we can see three key opportunities for improved digital first workflows, with a fourth requirement that spans across these.

## 4.3 Opportunities for evaluation

### 4.3.1 Text

There are interdependencies between: searchable text, handwriting recognition, and OCR, which collectively received 12 votes. Searchable text implies the existence of either: human-created transcription, OCR for typescript, or handwriting recognition (HTR) for handwritten text.

In addition, machine-readable text is a requirement for machine assisted subject description, and named entity recognition. Getting machine-readable textual data, as early as possible in the workflow, is one of the key takeaways from the workshop sessions.

### 4.3.2 Subject and topic description

Identification of subjects and keywords for documents, and potentially, the integration of these with the ability to employ and flexibly extend a taxonomy ("powertagging") received 6 votes.

### 4.3.3 Named entity recognition

Identification of named entities, and in particular, person names scored highly with 6 votes, and is also an implicit requirement or aid to linking and comparison tasks, which also received (across two items) 6 votes.

### 4.3.4 Controlled vocabularies and taxonomies

There are a number of requirements:

1. Linking and comparison with other collections or objects
2. Interlinking between entities
3. "Powertagging"
4. Subject terms and keywords

Which received a high number of votes, and which either explicitly, or implicitly, imply the normalisation of terms and the use of taxonomies or controlled vocabularies.

Subject identification and "powertagging" with topics imply the existence of a controlled vocabulary and/or taxonomy. This might be either a fixed source, such as LCSH subject headings, or a flexible/editable taxonomy shared by LSE and UCL, or employed across other SM20C digitisation projects.

Similarly, while “raw” named entity recognition, and recognition of person names is useful in itself, if those names can be matched to a controlled vocabulary -- again, this could be an existing dataset, or one managed for the SM20C project -- those named entities can be used more efficiently for cross-linking between collections, and linking between objects within a collection.

There are a number of other implicit technical requirements that span the highly voted tools and methods and, in particular:

1. Natural language processing:
    a. Tokenization
    b. Lemmatisation
    c. Parts of speech tagging
    d. Sentence boundaries
2. Document segmentation, in particular, the breaking up of file or series level digitisation outputs into object/item level digital surrogates.
3. IIIF APIs as enablers.

We will discuss these requirements below.

# 5. Solution Blueprint

The digital first workflows for LSE and UCL (see 3.2 Revised workflows for exploiting identified ‘digital first/early’ opportunities above) suggest a potential architecture and technical approach which has two main stages:

1. Page/Image level steps: fully automated steps which can run early in the workflow, whether or not the sequence of images has been broken up into individual documents.
2. Object level steps: which depend on the existence of document boundaries.

Where these two stages are able to be run at different points in the workflow, and flexibly integrated into the specific workflows of each institution.

These two main stages can potentially bookend manual cataloguing or data entry tasks, or hybrid tasks, in which cataloguers and archivists/special collections librarians act on the outputs of digital methods to filter or select outputs, or to assist in their manual tasks.

## 5.1 Workflow Overview

### 5.1.1 Page or image centric steps

Digital first workflows require the creation of digital surrogates early in the workflow, where early in the workflow may mean:

- Before potentially any item level identifiers or item level metadata records exist; or,
- where those identifiers may be references to a catalogue record which is sparse or just a stub record at the time of digitisation.

There are, many digital methods/tools which can be applied:

- in the absence of catalogue data, or
- in the absence of known page boundaries that indicate where items begin and end.[1]

At this stage, we can:

- *Create IIIF resources* at file or series level, via ingest APIs that allow for the uploading of images with a series or file level identifier, and running numbers. These IIIF resources provide the *enabling* APIs for the rest of the workflow, and a point of *integration* for data created in subsequent workflow steps.
- *Create OCR text* at the page or image level.
- *Create HTR* (handwritten text recognition) text at the page or image level.
- Carry out *initial natural language processing* at the page or image level for:
    a. Tokenization
    b. Lemmatisation
    c. Parts of speech tagging
    d. Sentence boundaries
- Initial *named entity recognition*:
    a. People
    b. Places
    c. Dates
    d. Organisations

These steps can be carried out before item level cataloguing has taken place, and are agnostic about the future structure of the archival hierarchy, or the relationship between images and item level catalogued objects, as these workflow steps work at a per-page or per-image level.

[1] For example, where an entire file or folder has been digitised with running numbers, and item level cataloguing has not yet been done.

These steps will enable an already enriched set of images to be interacted with by staff and users in order to carry out segmentation (into item level digital surrogates) and further cataloguing.

### 5.1.2 Object level steps

Both LSE and UCL's workflows could also then involve the creation of more detailed object level data, once the digital surrogates created earlier in the workflow have been segmented into item level digital objects and identifiers assigned.

Some types of data are *intrinsically* object level data, rather than page level data.

For example, the subject or topic of a document, or summary text for a document requires the object boundaries to have been established, as the subjects and keywords will change depending on which pages are included or excluded.

These steps cannot be taken until object level surrogates have been created and assigned identifiers. The second workflow stage should enable this segmentation to take place:

- *IIIF driven graphical sorting/segmentation tools* to enable digitisation staff, archivists or cataloguers to break up larger level digital surrogates into objects at the appropriate level of granularity for the digitisation project involved.
- *Keywording*: most important terms/bigrams/trigrams. This could take place relative to a general corpus of English language material, relative to a specific corpus, or relative to the corpus of SM20C material (or some combination of all of these).
- *Subject and topic tagging* based on controlled vocabularies and/or taxonomies
- *Generation of derivative metadata* for ingest into:
    a. Internet Archive
    b. Jisc Historical Texts or Journal Archives.
    c. IIIF based discovery environments, e.g. via producing *seeAlso* metadata files for use in IIIF Presentation API manifests or via IIIF Change Discovery API activity streams.
    d. Jisc Open Research Hub.

## 5.2 Implementation

In the sections below, we provide a high level summary of how these workflow steps can be carried out, and provide a gap analysis vis a vis the current capabilities of Digirati's Digital Library Cloud Service (DLCS).

It would be possible to provide this workflow without using the DLCS, and the existing DLCS architecture can provide a guide to the services that would need to be developed in order to do this. However, using the DLCS, or a derivative of it, is likely to be a much quicker and more cost effective approach.

### 5.2.1 DLCS Gap Analysis (Page level workflow)

In preparing the demonstration material (https://digirati-co-uk.github.io/sm20c-workshop-samples/summary.html), for the SM20C Workshops, Digirati made use of the Digital Library Cloud Service (DLCS), an existing IIIF based SaaS workflow that Digirati have used on a number of cultural heritage projects.

In describing one possible architecture for implementing the digital first workflow steps identified above (5.1 Workflow Overview) we have assumed the DLCS, or something very similar, is in place.

The DLCS can currently provide the following core services, which are linked together in a loosely coupled event driven workflow:

1. *IIIF Image hosting*: with load balancing, authentication, format conversion, scalable tiled image and thumbnail delivery.
2. *IIIF Presentation API* hosting: a database driven service with APIs for storing, updating and delivering IIIF Presentation API manifests. This service acts as in integration hub for the data produced by additional services in the workflow.
3. A *message bus*: to allow communication between services, and to trigger services to act when the appropriate condition is met, for example, to trigger named entity recognition when a IIIF image has associated OCR text.
4. *OCR/HTR text* creation and normalisation: a lightweight service which wraps Google Vision for creating text from images, and which normalises the OCR format to a common model, and provides APIs for fetching text, and for fetching coordinates for text.[2]
5. *Named entity recognition*: a service which uses NLP software, and which can be configured with optional sources of controlled vocabulary -- prepared by technical staff on a case by case basis -- for tagging IIIF resources with named entities.
6. *Search*: which indexes annotations, including named entities, and OCR/HTR text for exposure via the IIIF Content Search API.

If we compare these services to the workflow steps, in the digital first page level workflow above:

[2] Google Vision's HTR (handwriting recognition) is not as good as dedicated solutions, but is quick and easy to implement.

- *Create IIIF resources*
- *Create OCR text*
- *Create HTR* (handwritten text recognition)
- *Initial natural language processing*
- *Initial named entity recognition*

The DLCS can already provide most of these functions. With a number of small gaps.

### Natural language processing (NLP)

This is currently carried out internal to the named entity recognition process. The service used in the DLCS uses spaCy (a fast Python/Cython based NLP engine) to carry out basic tokenization, lemmatization and parts of speech tagging.[3] This data is treated as ephemeral data in the processing pipeline, and is not persisted.

However, for the SM20C workflow, some of the workflow steps and digital first methods can reuse this data.

It would make sense to break this out as a separate workflow step, which persisted lemmatization, tokenization, and parts of speech data as database records, or as static JSON files.

These files would not necessarily have to be in spaCy's data model, and could be transformed to a simple data structure.

### Handwriting recognition (HTR) and OCR

Currently, Digirati's OCR service makes use of Google Vision Document Text Detection for recognising text. It can also be configured to use Tesseract, which does not incur usage charges,[4] but which our experience (see the extended discussion in this Medium.com article) has shown to fall a long way behind the cloud based OCR services, and Abbyy.

Google Vision is quite good at handwriting recognition, but falls down on older historic hands, and may not be as good as HTR specific tools, especially those such as Transkribus which have been trained on a particular hand. The existing text creation service may be suitable as is, as an MVP (minimum viable product)—Google Vision output may be good enough—but if there is high value and high demand for handwriting recognition, we could explore the creation of a dedicated service or workflow, which might integrate with specialist tools.

[3] The DLCS service wraps spaCy with additional pipeline steps that integrate with OCR data, understand IIIF annotation formats, can accept controlled vocabularies or lists of terms as JSON or CSV, and integrates with the DLCS message bus.

[4] Google Vision's usage charges are $1.50 per 1000 images.

This might be something to be postponed for a future phase of development, however, this could definitely be done, and Digirati have already done some technical work on documenting what an HTR architecture might look like. A discussion document which outlines this can be found here.

### Management interface/dashboards

The DLCS has a good portal interface/dashboard for IIIF Image API based services, but does not have a good user facing web interface or dashboards for other DLCS services, which tend to be run in an automated workflow without end user intervention.

It may be worth exploring whether surfacing these workflow outputs in a user friendly dashboard for SM20C stakeholders is a useful deliverable, or whether the current lightweight information—which is intended for developers, not for general users—is sufficient. A simple command line interface might be a low cost option.

### Workflow segregation

Currently, the DLCS text services—OCR/HTR, annotations, named entity recognition, and search—are run as a “single tenant” workflow. That is, there is no concept of segregation between data, anyone who has access to the workflow, has access to all of the data in that workflow.

That would mean, for example, as the system currently stands, if we were to deploy the existing DLCS stack on behalf of UCL and LSE, in common as part of SM20C, both institutions would be able to see each other’s data. This may well be acceptable, in terms of a convergent workflow, with convergent outputs, and a shared understanding of process.

However, if it is not, we would need to do some work on multi-tenancy, or deploy multiple workflows

Deploying multiple workflows, i.e. using a multi-instance architecture, would be more cost effective, and faster to produce than developing a single multi-tenant workflow. However, these workflows could share common configuration, to ensure convergent data.

## 5.2.2 DLCS Gap Analysis (Object level workflow)

For the object level workflow, the workflow would need to provide:

- *IIIF driven graphical sorting/segmentation tools*

- *Keywording*
- *Subject and topic tagging*
- *Generation of derivative metadata*

### Sorting Room

The DLCS can already provide a IIIF driven graphical sorting/segmentation tool (*Sorting Room*), which was demonstrated in the workshops.

The user experience for this tool might need to be revisited, and, in particular, the structure of the data captured from the end user at the point of object creation might potentially need to be changed to incorporate:

- Appropriate file/series level identifier (the identifier for the parent in the archival hierarchy)
- Appropriate item level identifier (the identifier for the object being created)
- Any relevant links to catalogue records, CALM structure, etc.
- Additional label or description information that might be captured at this stage from the end user
- In-line display of machine created data (created at the page level workflow stage), if this is of value to the user "minting" new items from digitisation outputs.

These would all represent incremental changes to an existing tool, rather than completely new requirements. The potential change might involve replacing the current form, which is a simple label form, with a multi-field form. We would also potentially need to explore how much data validation and interlinking would happen at this stage.

### Keywording

The DLCS does not currently have a keywording service. For the demo sessions in the two workshops, we used term-frequency inverse-document-frequency analysis (TF-IDF) to extract keywords from documents. The set of keywords is dependent on the structure of the document, so this would need to happen after a document or item level object had been created.

This is a relatively straightforward process, so could be provided as a discrete DLCS service with a dependency on an NLP service (see Natural language processing (NLP) above) and existing OCR text (see Handwriting recognition (HTR) and OCR above).

One suggestion that emerged from the discussion with James Baker (Senior Lecturer in Digital History and Archives, University of Sussex), regarding researcher use of this information, was that it might be useful to measure the *keyness* of a given word relative to a particular known English language corpus. If we employ corpora here we can potentially ground the keywords in a way that make the keywords useful to the right kind of user. If we grounded them, say, in just the SM20C corpus we may underestimate the importance of common terms that regularly appear in the SM20C material—women, equal pay, etc.—where these are the terms we actually want to pull out as keywords. This corpus might be one of the large "global" corpora, or it could be a specifically chosen smaller corpus drawn from Jisc's data.[5]

The complexity of the service might change, depending on how the integration with corpus data is carried out. There are no major technical problems to solve, however.

### Subject and topic tagging

Subject tagging, especially vis a vis a controlled vocabulary, is a harder problem to solve, although the analysis of needs above suggests it is one of the most useful workflow steps for the SM20C stakeholders. The approach used for the demos presented in the workshop was a relatively crude approach using fuzzy "levenshtein" matching between keywords against LSE's existing subject headings.

The data used to drive this was a list of LSE headings their associated keywords, and an inverse list of keyword sets, with their associated LSE subject headings. UCL documents were "tagged" with whichever subject sets were *fuzzily* closest to the list of keywords pulled from the document by TF-IDF.

This approach would probably not be adequate for a production service, as its parasitic on existing catalogue data, and limited by how extensive that existing catalogue data already is.

There are more sophisticated approaches that could be adopted here. For example, existing Python based machine-learning tools such as Gensim can provide topic modeling based on extracted text. In order to this, we would need to:

1. Transform the text into a vector representation (this could be done by the Natural language processing (NLP) service above).
2. Use techniques such as Latent Semantic Indexing, or Latent Dirichlet allocation to infer a topic model.
3. Associate this topic model with existing subject headings. These could be assisted by, or constrained to, the existing subject tags taken from LSE or UCL's catalogue or CALM data, or drawn from some wider data set or standard set of subject headings.

The key step here is step 3, since topic modeling is good at clustering documents, but we also need a way of relating these clusters to a specific human and machine readable subject term or set of subject terms. Typically this might involve a combination of manual assignment, assignment inferred from existing data sets, and training (machine learning).

[5] 19th century pamphlets, for example.

Our estimation is that this might be the largest task in terms of new software development, although, based on feedback already provided, this may be the most useful.

### Taxonomies and controlled vocabularies

Named entity recognition, and subject/topic tagging, are both tools that would benefit significantly from association with controlled vocabularies or taxonomies.

Digirati have worked on two different tools that provide formal taxonomy management (*Taxman*) or a lightweight topic management service (*Tacsi*, used on the Indigenous Digital Archive backend). Neither of these services are fully integrated, as yet, into a workflow of the type envisaged for the SM20C project.

We would expect to be able to make use of *Taxman* in a project like this, but there would need to be development work to:

1. integrate SKOS based taxonomies into topic clustering
2. Integrate SKOS based taxonomies into named entity recognition[6]
3. Convert SM20C specific sources of vocabulary into SKOS format
4. Potentially, extend these tools to incorporate feedback loops so the output of human cataloguers and archivists/special collections librarians can improve the outputs of the machine generated data.[7]

This might be a reasonable amount of development work, but offers a great deal of benefit in terms of data convergence, reusable data, discoverable digital assets, and interlinking between objects and collections.

### Catalogue metadata, metadata reformatting, and metadata convergence

In the scope of a project with a relatively moderate overall budget, it is unlikely we would have much success generating catalogue metadata—other than subjects, keywords, and named entities—using predominantly machine methods, with the level of accuracy that is required for formal catalogue records.

There are a number of reasons for this:

[6] Digirati's existing named entity recognition software — *Montague* – can integrate with external sources of vocabulary, but currently uses a lightweight JSON or CSV based format. We could convert this to use *SKOS*.

[7] *Taxman* does not currently work in this way, so we would need to potentially extend the software.

- *Heterogeneity of the source material*: It would be very hard to train a computer vision algorithm to identify, for example:

    - *Title*. Although we can potentially flag phrases or sentences that have “title like” shape (e.g. all caps, or title case).
    - *Date of publication*. Although we can use named entity recognition to produce a list of possible dates in the publication, and potentially flag dates that appear on the first or second page, or dates that appear early or late within a page, since these are more likely to be publication dates appearing in a header or footer, or on a title page.
    - *Author*. Although we can use named entity recognition to flag a list of names, and, again, like dates, potentially flag those names that appear on the first or second page, or names that appear early or late within a page, since these are more likely to be authors appearing in a header or footer, or on a title page.
    - *Publisher*. Although as per names and dates, we may be able to flag those organisation names that appear in headers or footers, or on early pages in a document. Additionally, organisations have very heterogeneous names, and named entity recognition of organisations can be hit or miss.
    - *Place of publication*. As per the above, we may be able to flag places that appear in appropriate locations in documents, but place names are also very heterogeneous. Use of controlled vocabularies—GeoNames, or Viaf, or Getty Thesaurus of Geographic Names, for example—may help, but is not a magic bullet.
- *Accuracy*: Keywords and subject terms are inherently somewhat subjective, and are still extremely useful for search and navigation even where they are not 100% accurate. Bibliographic metadata, however, is generally held (for good reason) to a much higher level of accuracy, and machine generated catalogue data is unlikely to meet this level without a significant investment of time, and expertise, with concomitant cost implications.

A more cost effective use of SM20C development time would be to produce a user friendly document summary page for each document, including:

- Subjects, linked to controlled vocabularies where possible.
- Keywords, linked to controlled vocabularies where possible.
- Named entities: people, places, dates, etc. linked to controlled vocabularies where possible.
- Text
- Extractive summaries (for quick visual scanning by cataloguers)

Where this information is provided in a form that is easy to cut and paste to transfer into whichever cataloguing tool—CALM, for example, or AtoM, or a library management system—is in use in the institutional workflow.

There will be two potential user-facing sources of the machine generated data.

1. Catalogue/archive management system
2. The IIIF surrogate.

Manual data entry into the catalogue/archival management system ensures that only accurate data finds its way into the catalogue. We may want to additionally, having a simple click-to-approve process where machine generated data can be removed from the IIIF manifest. This could be implemented as a simple UI component, perhaps associated with the *Sorting Room* like segmentation tool described above.

### Metadata reformatting and metadata convergence

We can assume that there are a small number of possible metadata formats in use internal to the workflows.

The workflow should provide a tool that can:

1. Export this metadata from the catalogue (or archival management system) using whichever API or format is available.
2. Normalise to a simple common denominator format.
3. Transform to a number of derivative formats, e.g.
    a. Schema.org
    b. DublinCore
    c. Jisc Journal Archives or Historical Texts
    d. Internet Archive
    e. Formats required by the Jisc Research Data Hub—see Appendix 1: Jisc Open Research Hub below—:
        i. Discovery and aggregation
        ii. Creation of DOIs via DataCite or similar services.
4. Store these derivative formats, with their metadata profiles as:
    a. XML, or
    b. JSON/JSON-LD
    c. Link these to IIIF Presentation API manifests, via the seeAlso process.
5. Push, via APIs, to Jisc Open Research Hub, or Internet Archive.

Transforming to various derivative formats may seem like a relatively complex task, but, each of a)-e) has a minimum metadata requirement that involves only a relatively small number of core fields:

- Author
- Title

- Date of publication
- Place of publication
- Publisher

Which can be enhanced or enriched by additional subject or keyword fields where this data is available.

Additionally, for upload to Internet Archive, it may be required to provide additional exports which export images (TIFFs or JPEG2000s) and OCR text (as XML or plaintext) as well as machine readable metadata.

Although conceptually straightforward, this may be a relatively large chunk of work, because of the number of integrations that may be required.

## 5.2.3 What this means in the context of IIIF

### Machine readable metadata

IIIF provides a mechanism for linking machine readable metadata with profiles (which can tell machines what to expect) to the IIIF manifest (the file that describes the digital surrogate).

This seeAlso mechanism will enable the workflow to associate the convergent metadata, in potentially multiple formats, with the IIIF manifest.

For example, this might include:

1. An SM20C discovery profile (the common denominator format) as JSON-LD or RDF/XML
2. Dublin Core as XML
3. Schema.org
4. Internet Archive metadata as XML

### Linking properties

IIIF also provides a mechanism for providing vertical links—for example, between an archival file or fond and the items within it—and horizontal links—for example, to related items. Via linking properties. These linking properties can situate an item in its wider archival and library context.

This metadata and these linking properties can be used to generate IIIF Collections, which can represent the archival or library context for that document.

The Wellcome Library, for example, have a IIIF Collection for their archive which can be found here: https://wellcomelibrary.org/service/collections/archives

This information can be used in context, to show the relationship between an object and hierarchy, as in this page: https://wellcomelibrary.org/item/b18184303

### IIIF Discovery

The IIIF Discovery Technical Specification group has been working on specifications for publishing sets of objects in a way that is useful for crawling and indexing by aggregators, and which can provide a way to surface IIIF Presentation API manifests and their machine readable metadata (the data linked via the *seeAlso* property) *en masse* and with their create, update, and delete events exposed.

The Change Discovery specification is currently at version 0.3, with a new version 0.4 revision imminent.

The convergent metadata produced by SM20C can be used to expose the SM20C objects for crawling and aggregation via this mechanism.

## 5.3 What this means in the context of SM20C Digital First workflows

In 3.2 Revised workflows for exploiting identified 'digital first/early' opportunities above, we looked, in the workshop sessions, at potential opportunities for integrating digital methods into LSE and UCL's workflows.

In terms of the software architecture described in outline above, how might this look?

After workshop one, we identified a few key workflow/process steps common to both institution's workflows:

- Selection, Planning, and Resourcing
- Copyright, rights, and data protection clearance
- Cataloguing and metadata creation
- Conservation
- Digitisation
- Metadata and asset re-formatting
- Ingest into Asset Management or Repository
- Delivery/discovery

### 5.3.1 LSE

In a digital first workflow, LSE's process (using the above process stages as labels), would look something like:

- Digitisation (high volume "wide net" digitisation, or trial sample digitisation)
- Page level digital workflow:
    - IIIF Image and Presentation APIs
    - OCR and handwritten text creation
    - Natural Language processing
    - Named entity recognition
- Copyright, rights, and data protection clearance (with assistance from text, topics, and entities, and an efficient cross-object search API produced by digital methods)
- Selection, Planning, and Resourcing
- Object level digital workflow:
    - Segmentation into item level digital objects
    - Subject and topic tagging
    - Keywording
- Cataloguing and metadata creation at the item level (with assistance from text, topics, and entities produced by digital methods)
- Metadata and asset re-formatting
- Ingest into Asset Management or Repository (with a full suite of existing IIIF resources and APIs produced by the digital first workflow)
- Delivery customisation (assisted by APIs)
- Delivery/discovery (assisted by IIIF resources, keywords, subject and topic tags, etc.)

Digitisation, based on existing catalogue records, could happen first, throwing a wide net across collections. Further refinement and selection could happen after this stage, with the digital tools assisting in copyright and data protection clearance, reformatting of assets, and the creation of item level rich catalogue records. APIs provided by the digital tools can assist with custom delivery, and with ingest into the discovery environment(s).

### 5.3.2 UCL

In a digital first workflow, UCL's process (using the above process stages as labels), might look something like:

- Selection, Planning, and Resourcing
- Cataloguing and metadata creation (at the file or series level)
- Conservation
- Digitisation
- Page level workflow:
    - IIIF Image and Presentation APIs
    - OCR and handwritten text creation

    - Natural Language processing
    - Named entity recognition
- Object level workflow:
    - Segmentation into item level digital objects
    - Subject and topic tagging
    - Keywording
- Copyright, rights, and data protection clearance (with assistance from text, topics, and entities produced by digital methods)
- Cataloguing and metadata creation at the item level (with assistance from text, topics, and entities produced by digital methods)
- Metadata and asset re-formatting
- Ingest into Asset Management or Repository (with a full suite of existing IIIF resources and APIs produced by the digital first workflow)
- Delivery/discovery (assisted by IIIF resources, keywords, subject and topic tags, etc.)

After initial item selection, high level cataloguing (the creation of file or series level CALM records), digitisation would take place. Further item level cataloguing, rights clearance and data protection assessment, reformatting, and ingest would take place with enriched digital assets.

### 5.3.3 SM20C Content

For the SM20C content, which will comprise pamphlets and similar publications, the outputs from this process would include.

- Named entities (using a shared vocabulary)
- Subject and topic tags (using a shared shared vocabulary)
- IIIF (for common delivery)
- Common derivative metadata formats, for:
    - Internet Archive
    - Jisc portals
    - Open research hub
    - Some potential future discovery portal for SM20C material

The key thing being that the same services, or similar services, will produce the data for both institutions, and share the same set of controlled vocabularies, and the same subject/keyword tagging model.

## 5.4 Potential roadmap

All Page level items described in the table below are required foundational components for subsequent Object level components (apart from Dashboards if a “black box” pipeline is sufficient). Of the Object level items the Named Entity Recognition and Subject Terms and Keywords scored most highly with stakeholders.  Searchable text also scored highly but should be achievable with the existing DLCS functionality.

### Page level workflow steps

| Deliverable | Relative size |
|---|---|
| Dashboards<br>*This requirement could potentially be dropped if the SM20C stakeholders were happy with the pipeline being essentially a black box. There may be a small increase to the “Sorting Room” development time, in that case, to provide more information in the Sorting Room UI.*<br>● UX and design work<br>● Track progress of content through the processing pipeline<br>● Direct access to the output of services<br>● Error reporting and logs | *Medium to large.*<br>Depending on UI/UX and level of integration. |
| Ingest<br>*The exact size of this package of work will depend on how SM20C stakeholders want to ingest content, e.g. from a simple command line tool run against a folder, or from a sophisticated web UI, which would fold into the Dashboard application.*<br>● APIs and tooling for ingest of digitisation outputs into the digital first workflow | *Small to medium.*<br>Depends on further investigation. |
| NLP service<br>● Persist NLP data as JSON<br>● APIs for access to NLP data<br>○ Tokens<br>○ Lemmas<br>○ Parts of speech<br>○ Sentences | *Small to medium.* |

| | |
|---|---|
| ● MVP might be a single JSON file per image identifier with no granular APIs | |
| Named entities<br>*The DLCS has a service to do this. Adding more sophisticated use of controlled vocabularies would add some development time.*<br>● People<br>● Places<br>● Dates<br>● Organisations | *Zero to medium.* |
| Infrastructure and architecture<br>● MVP might be a single shared pipeline used by both UCL and LSE<br>● Options (in cost order):<br>○ Multi-instance infrastructure with shared config but separate data stores and API keys<br>○ Full multi-tenant infrastructure | *Small to x-large.*<br>Depending on options. A shared pipeline would be the most cost efficient. |

### Object level workflow steps

| Deliverable | Relative size |
|---|---|
| “Sorting Room” interface for breaking up sequences of images into item level digital surrogates.<br>● UI/UX review and refresh<br>● Improved data entry interface to capture item level identifiers and additional catalogue data<br>● Surface digital first outputs for review/deletion<br>● User friendly interface for use by cataloguers | *Medium.*<br>Could be larger, depending on the sophistication of the data entry interface required. |
| Keywording service.<br>● TF-IDF based keyword extraction<br>● Simple JSON model for linking as a service to a canvas<br>● Keywords as Open Annotation or W3C Web Annotations for surfacing in a IIIF viewer<br>● Options:<br>○ Integration with corpora for “keyness” scoring | *Small to medium, extending to large.*<br>Depending on the level of integration with corpus data. |
| Subject or topic tagging.<br>● Clustering of documents by inferred subject<br>● Using Gensim or a similar library<br>● Options:<br>○ Dynamically retag as the corpus of SM20C documents increases.<br>○ Use of controlled vocabularies or formal subjects headings based on<br>■ Manual identification of tagged clusters<br>■ Machine learning<br>■ Inference from existing subject classification systems like LCSH/FAST. | *Large to x-large.*<br>Depending on how sophisticated the topic clustering is, and how controlled vocabularies and subject classification schemes are used. |
| Shared vocabulary/taxonomy service<br>*If a single existing controlled vocabulary is identified, this required might be dropped and replaced with a few extra days of development for the Named Entity Recognition.*<br>● Using Taxman or a similar SKOS based taxonomy manager<br>● Options:<br>○ Integration with other sources of controlled vocabulary, e.g. | *Large.*<br>Exact estimation will depend on further user research and technical review. |

| | |
|---|---|
| ■ GeoNames<br>■ Getty TGN<br>■ VIAF<br>■ etc. | |
| Metadata conversion<br>● Extraction of catalogue data<br>● Conversion to common denominator format<br>● Generation of derivative formats<br>● Persistence of derivative formats<br>● Linking to IIIF resources via *seeAlso* | *Medium to Large*<br>Depending on complexity of catalogue integration. |
| Onward integration (with Internet Archive, Jisc portals, or open research hub) | *Medium to Large* |

The relative sizes are given using a “T-Shirt Size” estimate. These don’t translate exactly into precise cost estimates, without further technical decision making, which would be done during the project.

However, as a rough rule of thumb, we’ve assumed something like the following:

- Small: can be done by a single developer, within a few days
- Medium: can be done within a two week sprint or less, but may need more than one person.
- Large: may require more than one sprint, or more than one person across sprints.
- X-Large: will definitely require multiple “person sprints” of effort.

Estimates are high-level, and have some contingency included. Once a detailed technical requirements gathering and UX process has taken place, these estimates can be provided more firmly, and could go down as well as up.

## 5.5 Critical recommendations

Budgetary and time constraints may mean that not all of the features described above can be built for the SM20C project. We would recommend that a minimum viable product, which builds on at least some of these features would comprise:

| Deliverable | Relative size |
|---|---|
| Ingest<br>*A simple command line tool run against a folder, or set of folders, with no associated web dashboard*<br>● APIs and tooling for ingest of digitisation outputs into the digital first workflow | *Small* |
| NLP service<br>● Persist NLP data as JSON<br>● APIs for access to NLP data<br>○ Tokens<br>○ Lemmas<br>○ Parts of speech<br>○ Sentences<br>● MVP might be a single JSON file per image identifier with no granular APIs | *Small to medium.* |
| Named entities<br>*The DLCS has a service to do this. Adding more sophisticated use of controlled vocabularies would add some development time.*<br>● People<br>● Places<br>● Dates<br>● Organisations<br>This may have reduced value without corresponding investment in controlled vocabularies, but may still be a very useful input for raw discovery and to assist human cataloguers. | *Zero to medium.* |
| Infrastructure and architecture<br>● A single shared pipeline used by both UCL and LSE | *Small to medium* |

| | |
|---|---|
| “Sorting Room” interface for breaking up sequences of images into item level digital surrogates.<br>● UI/UX review and refresh<br>● Improved data entry interface to capture item level identifiers and additional catalogue data<br>● Surface digital first outputs for review/deletion<br>● User friendly interface for use by cataloguers | *Medium.*<br>Could be larger, depending on the sophistication of the data entry interface required. |
| Keywording service.<br>● TF-IDF based keyword extraction<br>● Simple JSON model for linking as a service to a canvas<br>● Keywords as Open Annotation or W3C Web Annotations for surfacing in a IIIF viewer<br>● Options:<br>○ Integration with corpora for “keyness” scoring | *Small to medium, extending to large.*<br>Depending on the level of integration with corpus data. |
| Metadata conversion<br>● Extraction of catalogue data<br>● Conversion to common denominator format<br>● Generation of derivative formats<br>● Persistence of derivative formats<br>● Linking to IIIF resources via *seeAlso* | *Medium to Large*<br>Depending on complexity of catalogue integration. |
| Onward integration (with Internet Archive, Jisc portals, or open research hub) | *Medium to Large* |

This minimum viable product would:

1. drop the subject tagging, or could potentially incorporate a very lightweight version of it (similar to the fuzzy matching approach used for the workshop).
2. Workflows would be kept simple, and drive by a command line interface rather than a sophisticated dashboard.
3. Keywording might make limited use of corpora
4. Onward integration and metadata conversion would be timeboxed, rather than open ended.

## Appendix 1: Jisc Open research hub

Digirati, and Jisc had a call regarding the research hub on Tuesday 13th of August. The suggested process, identified in this call (although not yet approved by Jisc research hub staff) is:

1. IIIF objects enriched with machine generated data are not a perfect fit for deposit in a research data repository which has a store and share model.
    a. This is because certain elements within a manifest are services with APIs rather than documents.
        i. Images
        ii. Content Search
        iii. otherContent (annotation lists - may be either static or dynamic)
    b. Manifests are often dynamically generated from a repository solution or document store, rather than static resources in their own right, where, for example, an API call might fuse data from a catalogue, with data from a repository, to make a Presentation API manifest.
2. Although it may, in principle, be possible to flatten some of these datasets into static resources which could be ingested into a repository for future reference, even if these are not the resources actually linked from the manifest (I don't think this is a goal for MVP).
3. Metadata about these objects, however, are a good candidate for a store and share model, where this metadata will have been produced by a combination of:
    a. Machine generation
    b. Human approval of machine generated elements
    c. Human cataloguing
4. Metadata generated for SM20C should be transformable, in principle and in practice, into the content level data model required by the Jisc open research hub
5. Similar levels of minimal metadata are required to generate DOIs (via DataCite or a similar service)
6. The DOIs associated with the object can resolve back to the object (the IIIF representation), and the metadata associated with that object, via the open research hub, can be provided to aggregators to assist with discovery of those objects
7. Upload of metadata and DOI creation should be automated via API calls, and do-able en masse for large volumes of material.
8. IIIF objects provide a way to associate multiple machine readable metadata profiles with an object via the seeAlso mechanism, which can provide multiple linked profiles for metadata.
9. So, as well as creating the metadata profiles needed by local catalogues:

a. ISAD(G) compliant archival metadata,
b. MarcXML,
c. MODS,
d. BibFrame,
e. or whatever that might be

10. And whichever metadata profiles might be best consumed by aggregators  (IIIF or other):
    a. DC or DC terms,
    b. or Schema.org
    c. or whatever
11. We can also generate and associate with the object whatever metadata profile will be required to:
    a. Ingest into the open research hub
    b. Generate a DOI
12. The open research hub will, at this stage, not be the source for the IIIF resources.
13. The identifiers in the IIIF manifest are likely to resolve to either institutional IIIF services at the source institutions, or to a dedicated IIIF hosted solution, like Digirati's DLCS (or some similar successor).