Collections Trust

Mapping digitised collections in England

Scoping report

Prepared for the Department for Digital, Culture, Media and Sport
February 2019 (v2.2)

Collections Trust
Rich Mix 35-47
Bethnal Green Road
London E1 6LA

www.collectionstrust.org.uk

Contents

Executive summary

DCMS has commissioned the present study to consider, from a technical point of view, how the Culture White Paper ambition to 'access particular collections in depth as well as search across all collections' might be realised.

> *'This project is a feasibility study, to develop and evaluate a practical framework for collecting relevant data in order to map cultural collections and consider what functionalities a tool based on this framework might possess given the state of existing technology. This project will provide the framework for carrying out this mapping exercise. It is not expected to produce the tool or mapping itself, but help us scope options for a technical solution.'*

The study has been carried out by Collections Trust, working with Knowledge Integration Ltd and with input from Culture24.

Note that the scope of the study is limited to developing a 'framework' and demonstrating its principles in test conditions. It presents neither the business case nor specification for any particular system to put the framework into practice.

This report confirms that the only viable model for the framework is an aggregator that can gather and deal with the data it needs from cultural heritage institutions of all sizes and levels of technical capacity, through all the aggregation methods currently used and likely to emerge in coming years.

Informed by surveys of relevant past initiatives (**appendix A**), and of relevant emerging technologies (**appendix B**) the report proposes an architecture for the required framework. This describes an aggregator that could gather data in whatever form, and by whatever means, contributing institutions can manage. The architecture also acknowledges that the incoming data will be messy, and that the aggregator would have to use various techniques and tools to mitigate this. Finally, the architecture proposes that, as well as being flexible about how data gets into the aggregator, it should also be possible for users (both human enquirers and other systems, such as third-party websites) to get the data they want through any of the means currently available ('dissemination and syndication').

Desk research into the online availability of digitised collections in a sample of 88 institutions was used to propose a smaller number of institutions whose data might be used to test the prototype tool that will be configured in the next phase of the project. The sample aims to give a good spread of institutions large and small, with a range of governance arrangements and collection types, and with varying degrees of digital sophistication, giving the opportunity to illustrate different methods of data-gathering and data-enhancement in the testing phase.

The report proposes five use-case scenarios for the testing phase:

- A curator looking for potential loans for a forthcoming exhibition about Charles Darwin's life and work.
- An academic researcher looking for information about ancient Egyptian ceramics in museums, with digitised images licensed for non-commercial use on her research

blog, and ideally with information about the people who collected and donated the material.

- A primary school teacher based in Essex looking for engaging, openly-licensed images of Anglo-Saxon objects, especially ones found in the county, as source material for a Key Stage 2 history project.
- A Subject Specialist Network seeking to combine collections data and produce thematic digital exhibitions on aspects of, for example, farming, with deeper diving into online collections where possible.
- A member of the public seeking information about the 'Monks Hall Hoard', discovered by an ancestor of his in 1864.

Finally, the following provisional success factors are suggested for evaluating the test, but it is recommended that they remain open to review and revision as the prototype tool is built.

- Be helpful to users in the agreed scenarios, leading them to whatever relevant data may be available more quickly and easily than is currently possible.
- Be helpful to the contributing institutions, by accepting and coping with their existing data as it is, no matter how inconsistent, incomplete or unstructured; and also allow them, if they wish, to retrospectively apply enhancements made at aggregator level.
- Show how resources from museums, libraries and archives might be brought together as seamlessly as possible.
- Show the potential and pitfalls of the agreed emerging technologies.
- Show how the framework would be scalable to involve and be useful to institutions of all sizes and levels of technical capacity.

1 Introduction

1.1 Policy context

In the 2016 Culture White Paper, the Department for Digital, Culture, Media & Sport (DCMS) set out its ambition to 'make the UK one of the world's leading countries for digitised public collections content. We want users to enjoy a seamless experience online, and have the chance to access particular collections in depth as well as search across all collections.'[1]

DCMS has commissioned the present study to consider how, from a technical point of view, the second part of this ambition might be realised: 'access particular collections in depth as well as search across all collections.'

> *'This project is a feasibility study, to develop and evaluate a practical framework for collecting relevant data in order to map cultural collections and consider what functionalities a tool based on this framework might possess given the state of existing technology. This project will provide the framework for carrying out this mapping exercise. It is not expected to produce the tool or mapping itself, but help us scope options for a technical solution.'* [2]

The study has been carried out by Collections Trust (CT),[3] working with Knowledge Integration Ltd (K-Int)[4] and with input from Culture24 (C24).[5] Note that the scope of the study is limited to developing a 'framework' and demonstrating its principles in test conditions. It presents neither the business case nor specification for any particular system to put the framework into practice.

1.2 Scope of this report

This report is the deliverable of the scoping phase of the project, the aims of which were:

- *'To research digitised cultural collections to develop common categorisations and terminology for a searchable database of cultural content. There is a wide sample encompassing fifteen national museum and gallery groups, over 100 collections designated as being of national or international significance, around 1800 accredited museums, and several thousand more non-Accredited museums.*
- *To research the current applications of emerging and innovative technologies, such as artificial intelligence, on collections and the readiness of such technologies for use in large-scale analysis of collections data in the cultural sector.*

---

[1] DCMS, *The Culture White Paper,* https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/510798/DCMS_The_Culture_White_Paper__3_.pdf, p39.

[2] DCMS, Invitation to tender (ITT) for: Contract for mapping digitised collections in the UK (ref 101343), 27 November 2018.

[3] www.collectionstrust.org.uk

[4] www.k-int.com

[5] www.weareculture24.org.uk

- *To research the viability, practicality and usefulness of a single framework that could be used to map the cultural content outlined above. And, if necessary, suggest alternatives.'* [6]

This scoping report also includes the proposed framework architecture and design, and the test cases recommended to evaluate the prototype tool. In the next phase of the project, the prototype will be configured and tested using data that illustrates the agreed use scenarios. The prototype is intended only to demonstrate key principles of the framework using a sample of real data. It will not have a user interface, nor will it offer any kind of ongoing service.

The final report will consider the results of the testing phase and draw lessons that should be kept in mind if and when the framework moves from theory to practical implementation.

1.3 Some key concepts

By its nature, this scoping report contains technical vocabulary and concepts. Where possible, these have been confined to the appendices. The following introductory notes unpack some of the terms used in the brief as a primer to the more technical sections of the report.

1.3.1 'Cultural collections'

The collections at the heart of this study comprise the artefacts, specimens and documents held by museums, archives and libraries. In this report, we use 'cultural heritage institution' as the general term for the various kinds of collections-based organisation. The vast majority of their holdings are physical items, but they increasingly acquire 'born digital' material, such as artworks and archival documents that have only ever existed in electronic form.

In this report we use 'collections' to mean the physical and born-digital items noted above, but note that it is *information* that gives these items meaning and significance. This, too, is a mix of digital and non-digital. The digital information includes the structured records of 'collections management systems' and other databases, but also electronic files of text created for websites, exhibitions, research and publications. Most institutions also have paper-based information about their holdings, such as accessions records, catalogue cards, files, and printed catalogues. These may go back decades, or even centuries, and are valuable historical documents in their own right.

1.3.2 'Digitised cultural collections'

Collections Trust's working definition of 'digitisation', which has been adopted by Arts Council England in its annual survey of National Portfolio Organisations, provides a useful starting point:

> *'Digitising' museum objects [etc] means making copies of physical originals in digital form – for example, by scanning or photographing 2D items or transferring the*

[6] Ibid.

> *contents of reels of film or audio tape into digital formats. It can also refer to 3D scanning of objects or, more loosely, any digital photography of collections.'* [7]

In this report, we refer to the digital reproductions resulting from such processes as 'digital assets'. Except where the context demands, we call the information about these assets 'data' (side-stepping the blurred meaning of 'metadata', which strictly means 'data about data' but is widely used just to mean 'data').

'Digitisation' can also include the process of transferring paper-based collections documentation to a digital format, either by scanning records (and perhaps using optical character recognition tools to create machine-readable text) or typing out transcriptions of their contents. For some institutions, 'digitising the collection' simply means using a computer system to catalogue it; for others, it means creating online content about key items or whole collections.

Therefore, although we use DCMS' term 'digitised collections', it is perhaps more helpful to think of 'digital collections' that may, or may not, include digitised copies of physical items, might include born-digital material, and will have varying levels of data.

1.3.3 'Mapping'

In this report we take ‘mapping’ to mean compiling an overview of digitised collections that can lead a user in the right direction towards resources that might be of interest and relevance.

1.3.4 'Collecting relevant data' into a 'searchable database of cultural content'

This study is all about how to ‘collect relevant data’ from many different cultural heritage organisations. The brief also specifies 'a searchable database of cultural content'. This implies something centralised, even if its data comes from lots of different cultural heritage institutions. Is that the right model? Why can't we just search all the online databases of individual institutions simultaneously in real time? And why not just use Google?

Simultaneous searching: the 'federated' model

Online tools such as flight comparison websites do indeed search many different databases simultaneously in real time: a process called 'federated' or 'broadcast' searching. In the cultural sector, from the 1990s onwards, libraries successfully shared bibliographic data through a number of 'virtual union catalogues' that used the federated searching model. These simultaneously searched the 'online public access catalogues' of many different library services in real time, and delivered the results to the user as a single 'hit list'.

The libraries' federated approach ensured that the search results were as up to date as possible and reduced the need for centralised data storage. However, the user experience could be poor, as the search speed was only as fast as the slowest response, and potentially relevant results would be missed if an individual catalogue was offline for any reason.

---

[7] www.collectionstrust.org.uk/digital-isnt-different/digitisation/

Moreover, the federated approach demands a high level of consistency between the data from different institutions; in a simultaneous search there is, with current technology, no time to analyse and tweak messy data. This is less of a problem with simple bibliographic records that follow rigorous standards, but would be a challenge with the more complex and variable data about the wider range of cultural heritage collections. Even assuming all 1,700 Accredited museums managed to get their collections online under their own steam - and keep the information up to date - the variability of the data is simply too great for the federated approach to be viable.

The 'aggregation' model

The technical term for 'collecting relevant data' into a 'searchable database' is 'aggregation', and the system that does it is an 'aggregator'. By themselves, these are fairly neutral terms and do not imply any specific solution beyond some kind of centralised database that is pre-loaded with 'cached' information gathered one way or another from other data sources. Note that not all the original source data need be cached; often only enough information for indexing purposes, and a link back to the original data or digital assets such as image files that would take up too much storage space if copied into the aggregator's own database.

That is how aggregators such as Google and other search engines work: they do not explore the entire World Wide Web in the few seconds after you hit the search button. Rather, they refer to the massive databases they have made earlier, which are updated regularly by the automated process of ‘crawling’ the Web. Having information to hand in this way speeds things up for the user, and means that potentially relevant content is less likely to be missed due to a website being temporarily offline.

Different aggregators currently gather their cached data in one or more of the following ways:

- By 'crawling' webpages using 'bots' in the manner of Google and other search engines; a generally blunt, free-text approach that can be refined if the webpages have machine-readable annotations (such as 'embedded ***microdata'***) that help the bot interpret the content.
- By 'harvesting' data that is exported from the source and imported into the aggregator using a defined standard template known as a 'protocol'. This process can either be automated or done manually using spreadsheets.
- Using 'Applications Programming Interfaces' (***API***s), which are tools that either proactively send ('push') data from the original source to the aggregator, or allow the aggregator to 'pull' data from the original source. A certain amount of configuration is needed to connect the aggregator to the specific API of a data source, so it is not quite as straightforward as harvesting.

There are also some other data-sharing and data-gathering methods used by cultural heritage institutions and aggregators. These include publishing information about collections as 'linked data' (or, when published with an ‘open’ licence for re-use, ‘***linked open data***’). In linked data, complex information (eg a catalogue record about a Turner watercolour) is broken down into a series of ‘semantic statements’; but instead of text (eg ‘JMW Turner’) to denote the painter, an ‘identifier’ such as http://vocab.getty.edu/ulan/500026846 is used to make a link to authoritative information about him published somewhere else (in this case,

the *Union List of Artist Names*).[8] If this sounds complicated, it is, and there are further complexities, too, that put this approach beyond the reach of all but the largest and most technically-sophisticated institutions.

If Google is an aggregator, why not just use that?

The practical limitations of using Google to find *all* the cultural heritage items that might be relevant to a search, and *only* relevant items, are best demonstrated by attempting to use it in the scenarios suggested in **section 4**.

As noted above, Google, and other search engines like it, is a general-purpose tool that treats most web content as a stream of free text. It therefore misses out on the potential benefits of structured metadata ('data about data') that could distinguish between, say, records about: things *created by* Charles Darwin; things *collected by* him; and things *about* him. Emerging developments such as embedded microdata might eventually go some way towards improving this situation, but they still require somebody, or some automated tool, to create and add meaningful annotations to each relevant webpage.

Google's custom search engine[9] allows developers to provide a search interface that is limited to a specified website or group of sites. This can be delivered for free, based on Google receiving revenue via Google AdWords. The main disadvantage of this approach, particularly for a framework intended to be an impartial resource on the nation's digitised cultural heritage, is that the 'relevance ranking' of a webpage is determined by Google's secret algorithms that, among other things, seek to boost advertising revenue. 'Search engine optimisation' is a huge business, and larger cultural heritage institutions would inevitably be better placed to boost the ranking of their resources than smaller ones.

Moreover, the 'just use Google' approach has the same major drawback as the federated searching model. In the museum sector, for example, in order for their collections to show up in search results, every single one of the country's 1,700 Accredited museums would have to have a crawl-able online collection as part of its own website. This is usually the complicated and expensive part of developing a new site and is currently beyond the means of many cultural heritage institutions, even larger local authority services, to judge from the research carried out for this study.[10]

Conclusion

For all the reasons set out above, in this report we assume that the language used in the brief - 'collecting relevant data' into a 'searchable database' - does indeed reflect the only viable model for the framework: an aggregator that can gather and deal with the data it needs from cultural heritage institutions of all sizes and levels of technical capacity, through all the aggregation methods currently used and likely to emerge in coming years.

[8] www.getty.edu/research/tools/vocabularies/ulan/
[9] https://cse.google.com/
[10] See the attached spreadsheet **table 2**.

1.3.5 Artificial intelligence

The brief calls for research into 'emerging and innovative technologies, such as artificial intelligence ... and the readiness of such technologies for use in large-scale analysis of collections data in the cultural sector.'

Artificial intelligence (AI) is a broad term which (strictly speaking) is used to describe systems and technologies that can essentially 'self-learn and correct' without human intervention. The definition, however, is often broadened to include the application of technologies that can be algorithmically 'trained' to recognize patterns in data but can only be improved by further human intervention.

In both cases, the key requirement is that the AI system has access to a representative 'training corpus' of material (such as words and/or images) for its initial programming. For both text and image-based techniques, this works best when the training corpus is large and homogenous. Text-based approaches have been particularly successful in sectors with such data, including the pharmaceutical industry,[11] and in certain specific fields of digital humanities (such as analysing the texts of Shakespeare's accepted canon of work to identify his stylistic traits.) [12]

The text-based resources of libraries and archives are already the subject of cutting-edge AI research such as the British Library and Turing Institute's *Living with machines* project.[13] According to Dr Mia Ridge of the British Library,[14] this £9.2 million collaboration with the Turing Institute, which runs for five years from 2019, will involve the Library digitising millions of pages from newspapers, including regional publications, published during and immediately following the industrial revolution. A multidisciplinary team of scientists will look to combine this data with other sources (such as geospatial data and census records) to develop new tools, including machine learning algorithms, which are capable of unearthing patterns in the data which will lead to new insights into the societal changes during that period. The availability of the Library's vast corpus of training material provides the perfect environment for the development of tools which have the potential, in future, to be of much wider use to researchers within the digital humanities.

Such large and homogenous sets of material are rarer in museums. One example is the Science Museum's collection of around 1,000 historic photographs of electricity pylons.[15] This would make an excellent corpus for an image-based AI tool that could then be trained to recognize pylons in other landscape images.

However, in order to achieve the numbers of similar things needed for a useful training corpus, digitised collections from many institutions need to be brought together. A typical art collection, for example, might have one or two oil paintings that include a particular historic fashion item. But if you bring together images of almost every oil painting in public ownership, as the aggregator Art UK has done, you have the raw material to pick out a

[11] Liu, Shengyu & Tang, Buzhou & Chen, Qingcai & Wang, Xiaolong. (2015). 'Drug Name Recognition: Approaches and Resources'. *Information*. 6. 790-810, www.mdpi.com/2078-2489/6/4/790/htm
[12] https://newatlas.com/algorithm-shakespeare-coauthor-marlowe/46130/
[13] www.turing.ac.uk/research/research-projects/living-machines
[14] Pers. comm.
[15] https://collection.sciencemuseum.org.uk/documents/aa110067037/albums-of-photographs-of-electricity-pylons-in-various-countries

training corpus large enough for training an AI tool to recognize that fashion item in any painting. Indeed, Art UK has already successfully collaborated with Oxford University's Visual Geometry Group to train image-recognition software to complement the work of human 'taggers'.[16]

Teaching an AI system to play 'snap' using a training corpus is just the start. To pursue the fashion example further, if the AI tool had not only been trained to recognize the fashion item in any digitised painting, but could also access data about when and where the artwork was painted, it could track the fashion across time and place.

Given that, with a few exceptions such as Art UK, it is not currently possible to aggregate digitised collections at the scale needed to train AI systems for any of the tasks we might want to set them, the question is not whether AI technologies are ready to be applied to cultural collections, but how we get digitised collections ready for the AI tools already used in other sectors. For this, we need to bring them together at scale.

1.3.6 'Common categorisations and terminology' and data enhancement

This aspect of the brief reflects perhaps the biggest opportunity for the proposed aggregation model to add value to the collections data needed by users – be they human or AI tools – to find the precise needles they are looking for in the digital haystack.

Cultural heritage data is messy. The same things, people, places and concepts can be recorded using quite different terms. Getting everyone to agree to use exactly the same ones consistently can work in some specific cases (eg the titles of published books) but is impractical across the hugely diverse range of material held by cultural heritage institutions.

This is not just a problem for the cultural sector, and the wider 'semantic web' has developed lots of useful resources that can be used by an aggregator to mitigate the inconsistencies within the data of a single institution, let alone across the country.

For example, a curator might call a certain spade not a 'spade' but a 'turfcutting iron'. Or perhaps a 'turf-cutting iron'. An aggregator's data-enhancement tools can be pointed at terminology sources that know these two terms are equivalent, and also that this is a type of 'spade'. Then, within reason, it does not matter what the object is called, nor whether it is part of a 'farming', 'rural life' or 'agricultural' collection.

[16] https://artuk.org/about/blog/the-art-of-computer-recognition

2. Research into online availability of digitised collections

In order to sketch the early outlines of a map of the country's digitised collections, desk research was carried out into the extent to which a sample of museums, and some Designated archives and libraries, had made information about their holdings available online. This section explains how the desk research presented in the accompanying spreadsheets (**tables 1-4**) was carried out.

The brief talks about a 'wide sample' including national museums, Designated collections, Accredited museums and several thousand non-Accredited museums. At the start-up meeting it was clarified that was the population from which a sample should be selected for the desk research.

There is currently no definitive list of non-Accredited museums (indeed, this is one of the goals of a £1m, AHRC-funded *Mapping museums* project, which is only halfway through its four-year work).[17] This category was therefore excluded from the study sample.

Terms in ***bold italics*** are defined and referenced in the glossary at **appendix C**.

2.1 Sampling

The starting points for the sampling process were three current lists published by Arts Council England (ACE):

- Accredited museums in England.[18] There are 1,319 of these, of which 1,189 are fully Accredited. This list also states the ACE region in which each museum operates, and also the nature of the governing body (eg 'National', 'Local Authority', 'University, etc).
- Museums that are National Portfolio Organisations (NPOs) within the ACE National Portfolio 2018-22.[19] There are 57 museum NPOs, but several of these (eg Birmingham Museums Trust) are services with more than one sites, and some (eg Wessex Museum Partnership) are consortia of more than one museum service. Depending on the amount of ACE funding they receive, NPOs fall into three funding bands; level 3 is the highest, with these museums receiving more than £1m annually.
- Designated collections.[20] There are currently 149 of these, the most significant non-national collections in England. Usefully for the purposes of this study, the Designation scheme includes archives and libraries as well as museums.

There is much overlap between these lists, but also a lack of consistency that suggested the top level of the proposed framework should be an authoritative source of museum names and the relationships between individual museum sites and their governing bodies (and, in some cases, NPO partnerships).

[17] http://blogs.bbk.ac.uk/mapping-museums/about/
[18] www.artscouncil.org.uk/sites/default/files/download-file/List_Accredited_Museums_UK_CI_IoM_28_Nov_2018.xlsx
[19] www.artscouncil.org.uk/sites/default/files/download-file/NPO_2018_22_Jan2019_0.xlsx
[20] www.artscouncil.org.uk/sites/default/files/download-file/Collections_List_Nov_2018_0.pdf

To whittle down these collections to a manageable number the legacy Cornucopia website (discussed in **appendix A**) was searched with a number of keywords. These, like the user scenarios described in **section 4,** were chosen because, from the experience of the consultants, they were thought likely to yield relevant collections across a wide range of institutions, as indeed they did, allowing connections to be demonstrated between different collection types.

- 'Egyptian'.
- 'Saxon'.
- 'Darwin'.
- 'Farming' (also checked against the Art UK website to find galleries with paintings of farming scenes, as these are unlikely to be noted in the Cornucopia data).

The results from searching by these keywords were tabulated and helped the process of eliminating smaller museums, particularly those that had none or only one of the categories of material in their collections. Fortunately, as a result of the National Trust's work over the past two decades, the collections of almost all its museums (around 10% of all Accredited museums) can be searched through a single website. The National Trust was therefore treated as a single collection.

Using this process, a longlist sample of 88 institutions was agreed with DCMS (see attached spreadsheet **table 1**). The sample aims to give a good spread of institutions large and small, with a range of governance arrangements and collection types, and with varying degrees of digital sophistication, giving the opportunity to illustrate different methods of data-gathering and data-enhancement. Note that the sample was not intended to be statistically representative of the total population of institutions.

2.2 Hierarchy of information

In order to focus the desk research and structure its findings across such a broad range of online availability, the consultants adopted the following hierarchy.

- **Level 1: institutions**. The names of cultural heritage institutions (which can be linked to information held elsewhere about their location, opening times, contact details, etc).
- **Level 2: collections**. Information about the analogue collections held by each institution, ranging from one or two keywords to descriptive summaries of the scope and highlights of collections (and, where appropriate, sub-collections reflecting departmental responsibilities, etc).
- **Level 3: item-level catalogues**. As a minimum, an indication of whether or not searchable, item-level catalogue information is available online (whether at the institution's own website or via an aggregator); where available, aggregated data allowing users to search across the holdings of participating institutions.
- **Level 4: digital assets**. where available, images and other digital assets (eg sound and video files) associated with item-level records.

2.3 Online research strategy

2.3.1 Level 1: institutions

It was immediately apparent that there was considerable variation in the way institutions named themselves, or were named by others, in the various sources consulted. This pointed to the need for the top level of the information hierarchy to be a definitive source of institution names, with preferred and alternate names of cultural heritage institutions, with the hierarchical relationship between their governing bodies (and, where needed, that between multi-site services and their individual venues).

2.3.2 Level 2: collections-level descriptions

The detail of collections-level description available across the sample ranges from single-word category keywords to lengthy statements of significance. The research compared the keywords used to categorise collections in the sample cultural heritage institutions in the following datasets (see **appendix A** for information about Culture 24 and Cornucopia):

- Culture24's venues database[21]
- Cornucopia[22]
- The Museums Association's directory of museums[23] (MA members/subscribers only).

2.3.3 Level 3: item-level catalogues

At the next level of the framework, catalogue information about individual items (whether digitised or not), the research asked the following questions of the sample institutions:

- Is there a searchable online catalogue of some, or all, of the institution's collections?

If the answer is 'yes':

- What is the URL of the landing page?
- Is it possible to tell the total number of records?
- Is there any indication of how the total number of records compares to the total number of items in the institution's holdings (even if the latter is a broad estimate)?
- What search strategies are available to the user?
- Do controlled lists of keywords seem to have been used, or are they compiled on-the-fly from data that is clearly inconsistent?
- Do individual catalogue records include a reference URL for citation purposes?
- Is there any evidence that a data-sharing ***API*** is available to allow collections information to be reused by others?
- Is there a sitemap?
- Do the pages contain any markup (eg ***Schema.org***)?

[21] www.culture24.org.uk
[22] http://cornucopia.orangeleaf.com/
[23] www.museumsassociation.org/find-a-museum

- Is the licensing status of the catalogue information clearly stated, either within individual records or in a general policy covering the site as a whole?

The research also noted the number of item-level records aggregated to the following sites, whether with digital assets associated or not. In some cases, institutions with no online catalogues themselves aggregated information to one or more of these platforms.

- Art UK[24]
- Global Biodiversity Information Facility[25]
- Culture Grid[26]
- Europeana[27]
- Archives Hub[28]

2.3.4 Level 4: digital assets

Finally, the research considered the following questions for each of the sample institutions with its own online catalogue:

- Does the online catalogue include digitised versions of some, or all, of the institution's collections?

If the answer is 'yes':

- Is it possible to tell the total number of records with associated digital assets?
- Is it possible to filter only records with associated digital assets?
- Does data specific to a digital asset include a reference URL for citation purposes?
- Is the licensing status of digital assets clearly stated, either within specific data or in a general policy covering the site as a whole?
- Can the user search for digital assets by licensing status (eg to find items available for re-use)?
- Is the user invited to download digital assets, where licensing permits?
- Does the download process record any data about the user or proposed use (beyond automated analytics)?
- Does any data-sharing API include data to allow digital assets to be used by others?

The research also noted the number of item-level records aggregated to the aggregator sites listed above.

[24] https://artuk.org/
[25] www.gbif.org
[26] www.culturegrid.org.uk
[27] www.europeana.eu
[28] https://archiveshub.jisc.ac.uk/

3 Framework architecture and design

This section of the report is a necessarily technical discussion of the proposed architecture for the required framework. It builds on the conclusion drawn in **section 1** that the 'aggregation' model is indeed the only viable approach for searching across such complex and variable data as the digitised collections of many hundreds of cultural heritage institutions. The architecture is also informed by the discussion of relevant technologies in **appendix B**.

In short, the proposed architecture describes an aggregator that could gather data ('ingest') in all the ways described in **section 1**, and also be flexible enough to respond to emerging approaches too. It therefore aims to take data in whatever form, and by whatever means, contributing institutions can manage. The architecture also acknowledges that the incoming data will be messy, and that the aggregator will have to use various techniques and tools to mitigate this ('processing, cleanup and enrichment'). Finally, the architecture proposes that, as well as being flexible about how data gets into the aggregator, it should also be possible for users (both human enquirers and other systems, such as third-party websites) to get the data they want through any of the means currently available ('dissemination and syndication').

Terms in ***bold italics*** are defined and referenced in the glossary at **appendix C**.

3.1 Overview

Above all, the architecture for the required framework must be flexible and extendable (eg, being able to plug in new tools and services). It is always difficult to predict the direction of future technologies and institutional preferences. For example, version 1.0 of the ***IIIF*** Image ***API*** was not published until 2012. The community now includes a large number of cultural heritage institutions and open source software companies including some of the most significant institutions worldwide.[29] An architecture which restricts itself to a set of pre-determined formats, and cannot react quickly to new developments, will soon be outdated.

Many aggregation frameworks place the burden of compatibility on the shoulders of the institutions supplying the data. An architecture flexible enough to accept data in whatever formats institutions can supply it will lower this potential barrier to participation.

It is also key that the architecture supports 'scaling on demand' from the hardware perspective and is itself scalable both horizontally (in terms of separating functionality across machines) and vertically in terms of adding more processing power to a specific partition. There may be significant bursts of activity where processing is required (for example image processing) and the ability to scale the resources accordingly is key to the ongoing cost.

All key points within the architecture, ingest, processing and syndication should be pluggable so that emerging technologies and services can be used as part of the standard processing routines and mechanisms.

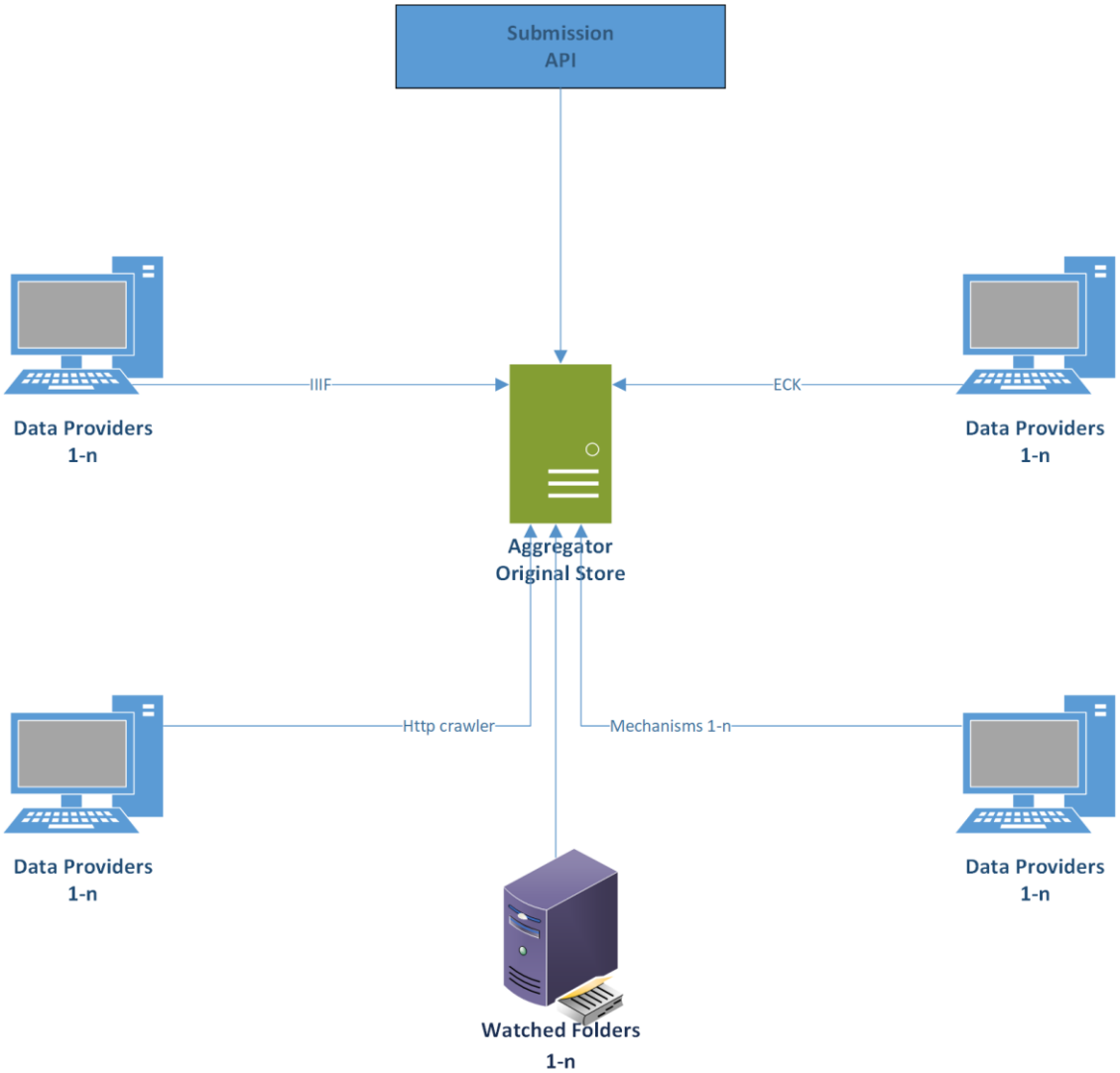
[29] https://iiif.io/community/#participating-institutions

In short, the architecture described below does three main things:

- Bring together data from a wide range of institutions, however they can supply it.
- Use a flexible selection of plug-in tools and services to process, clean, and enhance that data (making clear what has been done and keeping any changes separately from the original data).
- Make the data available in various ways for uses that are limited only by any licensing restrictions that contributing institutions might specify.

3.2 Ingest

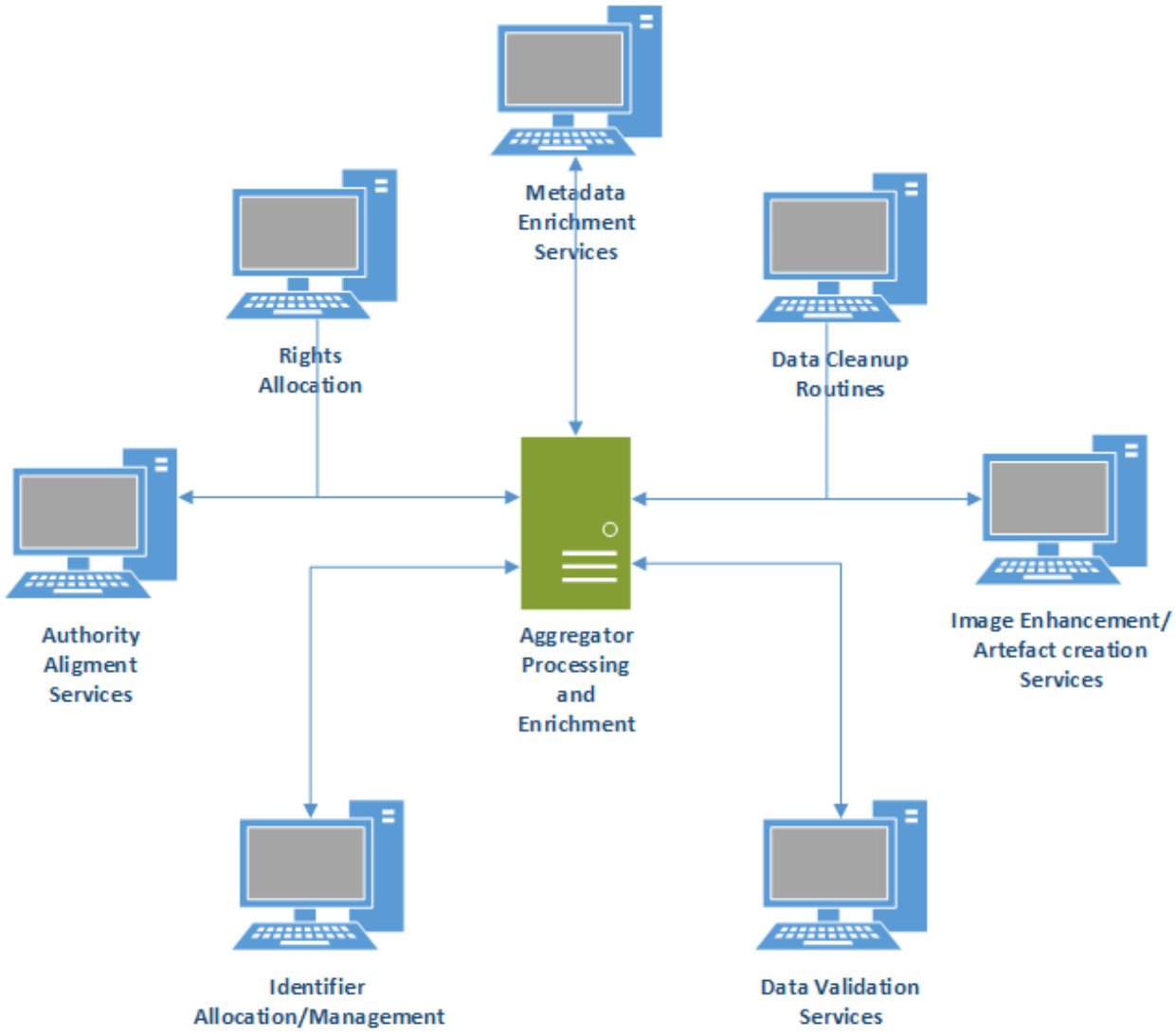
The most important aspect of the ingest architecture is that whatever mechanism for data supply/extraction, a copy of the raw source data and any other ingested assets (eg thumbnail images) must be retained within the system. This ensures that, when new services and enhancement routines are added going forward, they can always be added cumulatively to the original data. Any and all processing of the data should happen downstream of these 'ingest snapshots. Additionally, however the data is supplied, all data must pass through the same pipeline.



The aggregator must be able to support scheduled extraction from data sources to ensure that the data within remains up to date. It must also have a documented submission ***API*** so that providers (who are technically able to do so) can write push routines to deliver updated data at point of edit/creation. From an architectural design perspective, there should be no technical barriers placed on the mechanism and format of data supply to the aggregator.

3.3 Processing, cleanup and enrichment

Once ingested into the aggregator, the required architecture will allow the flexible application and coordination of services which act on the data. This can range from the allocation of ***persistent identifiers*** to data source specific cleaning and validation, along with alignment and enhancement. The most important aspect is that the fields within the data that should be acted on can be specified and the target location for the enhanced output given together with the order of the applied transformations. It is also important that any machine enhanced data is indicated as such within the metadata.



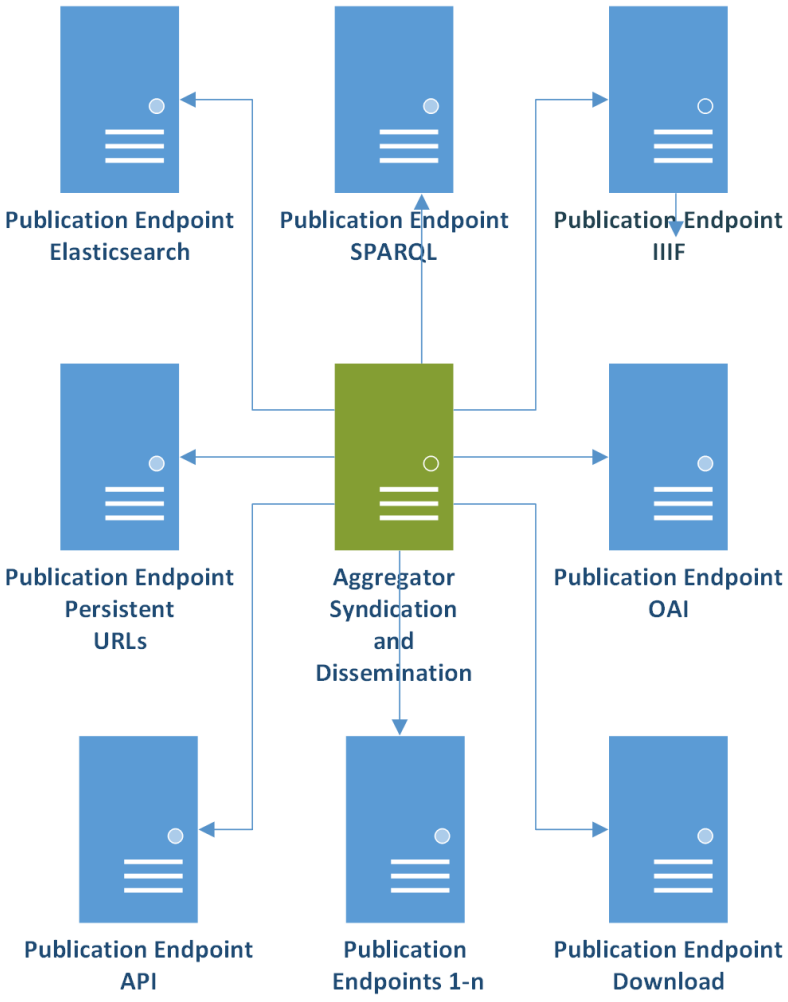
As an example, a place entity recognition service could be applied to the textual content of all descriptions. This entity recognition output could then be submitted (along with additional place data held in the records) to a gazetteer to apply coordinates to the textual data.

The most important part of this aspect of the architecture is that it is asynchronous and iterative. ie once a new service is identified as a possible candidate for data enrichment, it can be linked into the aggregator and added to the required pipeline(s). This service can then be retrospectively applied to all previously processed data and automatically applied to all new data. The asynchronous nature of the processing ensures that there is not a cumulative delay on record ingest into the system as more complex processing is added.

3.4 Dissemination and syndication

The main architectural requirement for dissemination and syndication is that one size does not fit all. The delivery mechanisms for the aggregated data must be the most flexible part of the architecture. Delivery requirements can range from providing an ***API*** onto data for small institutions that cannot host their own, to downloading and sharing subsets of aggregated data for researchers who wish to import it into their specific research platform.

There are many and diverse requirements for this aspect of the service and the ability of the system to enable this diversity is key, both directly (where the specific format and protocol is supported) and indirectly by providing an API that a developer can use to support the required function.



4 Proposed testing phase of the project

This section of the report proposes the collections to be used in the test phase of the project, the data collection/enhancement approaches to be prototyped, and the criteria and success factors that might be used to evaluate the test. Within the timescale and resources available, we suggest that the testing process should focus more on data enhancement and augmentation than on data collection, as this is where there is greatest potential for innovations that might improve the discoverability and usefulness of the available information. For example, AI services such as Amazon Rekognition,[30] Google Place[31] or IBM Watson[32] could be plugged into the prototype and applied to the ingested test data.

Terms in ***bold italics*** are defined and referenced in the glossary at **appendix C**.

4.1 Proposed use scenarios

As noted in **section 2**, to help whittle down a manageable sample of institutions for the desk research presented in **tables 1-4**, four keywords ('Egyptian', 'Saxon', 'Darwin' and 'farming') were used as search terms in the legacy Cornucopia database of collections-level descriptions. These four keywords were also used as the basis for some typical user scenarios that could be run in the prototype-testing phase of the project. A fifth user scenario was added at the request of DCMS to reflect the kind of enquiry a member of the general public might make.

The specific details of the use scenarios, and notional users, are less important than the fact that, between them, they present the opportunity to explore how the framework might succeed or fail to provide useful answers to these typical questions that might be asked of it.

4.1.1 Finding items relating to a specific person for a forthcoming project

This scenario is loosely based on one described by the Wessex Museum Partnership, which is planning a major exhibition about Thomas Hardy.[33] The curators want to find out what relevant material there might be in other cultural heritage institutions, potentially anywhere in the country.

For the purposes of this study, we have changed the scenario to a curator looking for potential loans for a forthcoming exhibition about Charles Darwin's life and work. The curator is interested in: natural history specimens collected by Darwin; archival documents written by him or relating to him; personal belongings; portraits; first editions of his books and articles; and items that reflect how his ideas were portrayed in contemporary popular culture. Although digitised images would be helpful, they are not essential. The curator is happy to make further enquiries if it seems as though an institution might have something relevant. The copyright status of digitised content is not important at this initial stage of the research.

---

[30] https://aws.amazon.com/rekognition/
[31] https://developers.google.com/places/web-service/intro
[32] https://www.ibm.com/watson
[33] David Dawson, 'Just give us the data', Collections Trust website, https://collectionstrust.org.uk/blog/just-give-us-the-data

4.1.2 Researcher interested in collectors of Egyptian ceramics

A researcher specialising in ancient Cypriot ceramics wrote on the CT blog that, if she could search seamlessly across the nation's collections, she would like 'search by collector, it would be really valuable in mapping intellectual networks and the itineraries of ancient objects.'[34] For this study, we imagine an academic researcher looking for information about ancient Egyptian ceramics, with digitised images licensed for non-commercial use on her research blog, and ideally with information about the people who collected and donated the material. As she is aware that not everything is likely to be catalogued, let alone digitised, she also wants to get a sense of the extent to which the resources available online represent the total number of items held.

4.1.3 Teacher looking for images of Anglo-Saxon objects

For this scenario, we imagine a primary school teacher based in Essex looking for engaging, openly-licensed images of Anglo-Saxon objects, especially ones found in the county, as source material for a Key Stage 2 history project.

4.1.4 A Subject Specialist Network digital exhibition

With Subject Specialist Networks (SSNs) becoming increasingly active and funded to deliver collaborative projects, our fourth scenario is an SSN (eg Rural Museums Network) seeking to combine collections data and produce digital exhibitions on aspects of, for example, farming, with deeper diving into online collections where possible.

4.1.5 A member of the public with a specific enquiry

This is a genuine enquiry emailed recently to Collections Trust:

> *I wonder if you could point me in a direction whereby I might discover the whereabouts of the Monks Hall Hoard, discovered in 1864. A newspaper clip from my father suggests that the British Museum and also the Manchester museum may hold some. The hoard was found by [an ancestor] ... I would be delighted to view some of his discovery. It would make me extremely proud to view some of the findings knowing that I have an association with it. Any help would be welcomed.*

Monks Hall is in Eccles, but the enquiry does not specify this. There is also the potential false lead that, until a few years ago, the building was itself a museum, now closed after a fire. The fifth scenario is, therefore, to find information about this hoard just from the search terms 'Monks Hall Hoard', 'British Museum' and 'Manchester Museum'.

4.2 Proposed collections

We have looked closely at the longlisted sample institutions (see attached spreadsheets **tables 1-4**) and considered how best to demonstrate the use scenarios, using the technical approaches proposed, across institutions of varying digital sophistication. We suggest the following shortlist of institutions should be in the test at levels 1 and 2 (ie to collection-level information).

[34] https://twitter.com/cypriotartleeds/status/1062027954698547200

- Ashmolean Museum (Egyptian; Saxon)
- British Museum (Egyptian; Saxon)
- British Library (Saxon)
- Colchester Museum (Saxon)
- Cambridge University Library (Darwin)
- English Heritage Down House (Darwin)
- Horniman Museum (farming)
- Historic England Archive (Darwin)
- Leicester New Walk Museum and Art Gallery (Egyptian)
- Museum of London (Saxon)
- National Gallery (farming)
- National Portrait Gallery (Darwin)
- National Trust (farming)
- Natural History Museum (Darwin)
- Norwich Castle Museum (Darwin)
- Nottingham Museums (farming)
- Oxfordshire Museum (farming)
- Petrie Museum (Egyptian)
- Portable Antiquities Scheme (Saxon)
- Royal Albert Memorial Museum (Egyptian)
- Royal Botanic Gardens Kew (Egyptian; Darwin)
- Saffron Walden Museum (Saxon)
- Science Museum (farming)
- The National Archives (Egyptian)
- Torquay Museum (Egyptian)
- Worcestershire Museum (farming)

For the level 1 authority file, we will prepare a thesaurus, in SKOS format, that will show the hierarchical relationship between governing bodies, multi-venue services (where appropriate) and individual venues within the shortlisted sample. The thesaurus will show preferred names for all these entities, along with any alternate names found in:

- Institutions' own websites
- ACE list of Accredited museums[35]
- ACE list of Designated collections[36]
- Culture24 database[37]
- Cornucopia database[38]
- Museums Association directory
- A few Virtual International Authority File[39] URLs to demonstrate potential.
- A few Wikipedia URLs to demonstrate potential.

Level 2 collection-level information will be aggregated from Cornucopia and Culture24.

[35] www.artscouncil.org.uk/sites/default/files/download-file/List_Accredited_Museums_UK_CI_IoM_28_Nov_2018.xlsx
[36] www.artscouncil.org.uk/sites/default/files/download-file/Collections_List_Nov_2018_0.pdf
[37] www.culture24.org.uk/home
[38] http://cornucopia.orangeleaf.com/
[39] https://viaf.org

For some of the shortlisted sample, it will not be possible to give any further information except to say that the institution does not currently have any searchable data online:

- Colchester Museum
- English Heritage Down House
- Leicester New Walk Museum and Art Gallery
- Saffron Walden Museum
- Worcestershire Museum

In other cases, it will be possible to direct the user to an online collection viewable at the institution's own website:

- Ashmolean Museum
- British Museum
- Horniman Museum
- Museum of London
- National Portrait Gallery
- National Trust (aggregated online collection on own website)
- Norwich Castle Museum
- Oxfordshire Museum (via Oxfordshire CC Heritage Search)
- Petrie Museum
- Royal Botanic Gardens Kew (online economic botany and herbarium databases)
- Torquay Museum (online collection at own website via MODES)

In some (but not necessarily all) of the remaining cases, it may be possible to aggregate data to item level (level 3) and any digital assets (level 4) using various approaches.

- British Library (subset via ***API***/download)
- Cambridge University Library (via Archives Hub)
- Historic England Archive (via Culture Grid, under English Heritage Viewfinder)
- National Gallery (tbc)
- Natural History Museum (via AP***I***/downloaded data subset)
- Nottingham Museums (via Culture Grid)
- Portable Antiquities Scheme (via ***OAI***)
- Royal Albert Memorial Museum (via API of regional aggregator South West Collections Explorer)
- Science Museum (API or ***IIIF***)
- The National Archives (via online Discovery portal; via API?)

4.3 Data collection/enhancement approaches

4.3.1 Data collection

To demonstrate different methods of collecting data, the prototype tool will use:

- ***OAI*** harvesting (eg Cornucopia from Culture Grid)
- ***API*** (eg from Culture24, Science Museum (***IIIF***?), National Gallery - subject to agreement (IIIF? - just from one), and Nottingham via CultureGrid)

- Web-crawling using ***JSON-LD*** (tbc)

Although the British Museum offers them, ***SPARQL*** endpoints are not yet common enough within the sector to be viable for most cultural heritage institutions or for demonstration within this project.

4.3.2 Data enhancement

We will demonstrate the potential of text-mining and terminological alignment to allow finding of resources using search terms not included in the source records.

- Place names
- Using equivalences between controlled terms to improve discoverability (eg ‘Farming’ = ‘Agriculture’)
- People (eg the Egyptologist Sir WM Flinders Petrie)

4.4 Evaluation criteria and success factors

We suggest the following provisional success factors for evaluating the test phase, but recommend that they remain open to review and revision as we build the prototype tool.

- Be helpful to users in the agreed scenarios, leading them to whatever relevant data may be available more quickly and easily than is currently possible.
- Be helpful to the contributing institutions, by accepting and coping with their existing data as it is, no matter how inconsistent, incomplete or unstructured; and also allow them, if they wish, to retrospectively apply enhancements made at aggregator level.
- Show how resources from museums, libraries and archives might be brought together as seamlessly as possible.
- Show the potential and pitfalls of the agreed emerging technologies.
- Show how the framework would be scalable to involve and be useful to institutions of all sizes and levels of technical capacity.

Appendix A

Summary of relevant past initiatives

The present project is by no means the first to address the issues noted in the project brief. In considering a future framework for mapping digitised collections, it is useful to be aware of past attempts and draw lessons from them. Relevant initiatives over the last 25 years are briefly considered here.

The technical legacy of these and other past projects includes innovations such as data models, schemas and related software. These often went on to be developed iteratively, forming (or, at least, informing) the technical foundations of current systems 'owned' by an engaged community of technical strategists and practitioners.

1 The Digest of Museum Statistics (DOMUS)

The Digest of Museum Statistics (DOMUS) database was launched in 1994 by the Museums and Galleries Commission (MCG, which later became the Museums Libraries and Archives Council (MLA)) and was active until 2000. DOMUS gathered information from 'Registered' (later termed 'accredited') museums via an annual survey. The DOMUS survey and datasets evolved over its lifetime, becoming ever-more layered and complex. DOMUS gathered data including:

- 'Core' museum information including: name, address, museum type and status, charity number, collection types, venue facilities and services and more.
- 'Annual' museum information relating to finance and operations including: policies, programmes, visitor numbers, staffing levels, charges, governance structure.
- 'Supplementary' information: this grew to include detailed questions on collections care; collections type, display, access and more; public services and information technology.
- Registration scheme information.
- Designation scheme information.

DOMUS relied upon a labour-intensive process of museums submitting information via paper questionnaires which was then entered into the database manually. As a result the data gathered was never reliably comprehensive, accurate or up to date across all institutions.

Archival versions of the datasets and more detail on the DOMUS project are available online at The National Archives.[40]

One of the projects that went on to use records from DOMUS was Cornucopia.

[40] http://discovery.nationalarchives.gov.uk/browse/r/h/C6787

2 Cornucopia

Cornucopia is an online database of collections held by cultural heritage collections across the UK, still live at the time of this report.[41] It was established in 1998 as a pilot project by the Museums & Galleries Commission (MCG, which later became the Museums Libraries and Archives Council, MLA.)

Cornucopia was initially created to make information on the 'Designated' collections – those deemed as being of outstanding importance, held by 62 English museums, accessible online. Following an evaluation in 1999, MLA expanded the project remit to include all Registered museums.

An article from 2004 by Chris Turner in Ariadne magazine[42] gives an account of Cornucopia's development from 1998 to 2004. Turner describes the information included:

> *'Key to the design of the original schema was the representation of the relationship between collections and the institutions that hold them. The database reflected this structure by preserving the concept of three 'levels' of information:*
>
> - *Institutional data (eg. address, access, Web site & institutional (or 'overall') collection)*
> - *Collections data (eg. title of collection, subject area, object type, geographical and temporal coverage)*
> - *Collection strengths (eg. objects of particular importance which would not otherwise be retrieved by a collections-level description).'*

The institutional data was originally derived from DOMUS but as the project scope widened, so too did the range of data sources and methods of contributing information.

An associated project that went on to use data drawn from Cornucopia was the 24 Hour Museum, described in section 4 below.

3 EnrichUK

Between 1999 and 2004 the UK saw the first major strand of public investment in digitisation of cultural collections through the New Opportunities Digitisation programme (NOF-Digi). Fifty million pounds of Lottery funding was distributed across 148 grants. Within this programme an online 'learning portal and search facility' called EnrichUK was developed by UKOLN[43], launching in 2003. EnrichUK aimed to be a one-stop-shop for audiences, giving them access to all of the content created with NOF-Digi funding.

The vision for EnrichUK was never fully realised and the site was short-lived. Indeed an evaluation report[44] published in 2006 states: 'EnrichUK was conceived as the 'learning portal' but must be considered a failure'. Very little citable information remains available on

---

[41] http://cornucopia.orangeleaf.com/
[42] www.ariadne.ac.uk/issue/40/turner/ - Cornucopia: An Open Collection Description Service; Chris Turner
[43] www.webarchive.org.uk/wayback/en/archive/20180630234548/http://www.ukoln.ac.uk/ or https://en.wikipedia.org/wiki/UKOLN
[44] 'The Fund's ICT Content Programme Final Evaluation Report', Education for Change, March 2006. www.biglotteryfund.org.uk/-/media/Files/Research%20Documents/er_eval_ict_final_rep.pdf

why EnrichUK wasn't a success. However, as the NOF-Digi programme overall was so wide-ranging in its diverse aims, reach, subject matter, delivery partnerships and target audiences it is clear with hindsight that the original idea of EnrichUK as a 'one-stop-shop' in which to aggregate and present everything created, to everyone, was deeply flawed.

In the years following the 2006 evaluation report MLA and BECTA[45] (British Educational Communications & Technology Agency) both looked into the potential of rejuvenating EnrichUK in some way but by that time a great many of the materials and data sets digitised within NOF-Digi were no longer available online as sustainability was not built into their original plans.

4 The 24 Hour Museum/Culture24

The 24 Hour Museum website launched in 1999,[46] created by the Campaign for Museums[47] and the Museums Documentation Association (MDA, which later became Collections Trust). It was branded as the first 'national virtual museum' with the aim of being a public-facing portal to all of the UK's museums - an enticing, open-all-hours gateway to museum visiting information, events, exhibitions and collections. The database behind the site was populated in the first instance with data drawn from DOMUS/Cornucopia - the institutional data, collection strengths and collection overviews.

Initially the site contained information on approx. 2000 Registered museums but quickly grew as the remit was expanded from 2000 to include all non-commercial visual arts and heritage venues. Over time data sets such as National Trust and English Heritage property records were uploaded manually and thousands of venues joined individually as the CMS offered an online interface for museums to sign up and enter their own information.

24 Hour Museum's key development was to collect and publish exhibition and events listings alongside the institutional and collections data. The transient nature of these events encouraged museums to sign up for the online data entry and update their venue's records but take-up and usage by venues was never consistent, requiring a significant level of support and encouragement from the 24 Hour Museum team.

24 Hour Museum became 'Culture24' in 2007 and launched a new public-facing website, retiring 24 Hour Museum. That website is still live at the time of writing, albeit not updated or maintained.[48] The website's third iteration, launched in 2017, is called Museum Crush.[49]

The venue database has expanded and has grown to be the most comprehensive visual arts and heritage listings dataset in the UK, holding over eleven thousand UK venue and organisational records and sharing all data via API for republishing by third parties. The database is still community-generated, relying on staff and volunteers from the venues inputting their own information, with nurturing, support and extra uploading where required by Culture24 staff.

[45] https://en.wikipedia.org/wiki/Becta
[46] http://news.bbc.co.uk/1/hi/sci/tech/342954.stm
[47] www.culture24.org.uk/home/art68003
[48] www.Culture24.org.uk
[49] www.museumcrush.org.uk

5 Culture Grid/Europeana

Culture Grid[50] is the *de facto* national aggregator for UK museums. It began life in 2005 as the People's Network Discovery Service (PNDS) set up by MLA.[51] As Culture Grid, under Collections Trust, the aggregator received start-up project funding in 2009-11 from MLA and other bodies, then from 2011 it became part of the core service for which Collections Trust received grant support from MLA. When ACE took over as the lead body for museums in 2012, Culture Grid continued to be part of CT's grant-funded activity until March 2015. In April 2015, CT's then chief executive, Nick Poole, decided to close Culture Grid to new additions,[52] and in 2016 the CT board decided to seek a futureproof alternative in the hands of a more-securely funded institution.[53]

Culture Grid was, and for the time being remains, the pipeline through which UK museums can be part of the Europeana ecosystem.[54] Culture Grid currently has around 3 million records from around 100 institutions, many of which have been further aggregated into Europeana's 58 million records. Through its API, Culture Grid delivers content to third-party sites such as *Exploring 20th Century London*[55] and the Scottish University Museums Group portal *Revealing the hidden collections*.[56]

6 Other relevant initiatives

There has been a range of initiatives mapping and aggregating digitised collections in medium or subject-specific domains. Some of these are small-scale, some bring together hundreds of thousands of records across multiple collection holders. They include:

- Art UK[57]. This project began with digitising all oil paintings in public ownership in the UK, publishing them online as 'Your Paintings' in partnership with the BBC in 2011. The current website, launched in 2016,features artworks from over 3250 collection holders.[58] The charity is currently working on a major initiative to digitise the nation's sculpture. Of particular interest to this report is their Collections Portal, the interface for collection-holders.

- FENSCORE (Federation for Natural Sciences Collections Research) was formed in 1980 to coordinate the work of natural sciences (geology, botany and zoology) curators working across the country to map and aggregate data about their work and collections. The FENSCORE database was active through the 1980s but hasn't been updated since 1999. It is still available online,[59] published by NatSca,[60] the Natural Sciences Collections Association, a membership organisation for natural science

[50] www.culturegrid.org.uk
[51] http://museum-api.pbworks.com/w/page/31386355/Culture%20Grid%20Profile
[52] www.museumscomputergroup.org.uk/culture-grid/
[53] www.collectionstrust.org.uk/news/collections-trust-asks-dcms-to-futureproof-the-way-museums-share-their-collections-online/
[54] www.europeana.eu/portal/en
[55] www.20thcenturylondon.org.uk
[56] www.revealing.umis.ac.uk
[57] https://artuk.org/about/history
[58] www.artuk.org
[59] http://fenscore.natsca.org/
[60] https://www.natsca.org/

curators and one of Arts Council England's Subject Specialist Networks. A full technical history is available on the website.[61]

- In 2014 NatSca launched the next generation of FENSCORE - a crowdsourcing project 'Natural History Near You' – gathering information about natural history collections in Britain and Ireland through a simple online form and publishing it online using a map interface to present results.[62] To date the project has gathered information on several hundred collections. As with FENSCORE, the project gathers high-level information including collection name, institution, institution, type, size, display status, time periods, and associated collector names. Neither Natural History Near You nor FENSCORE appear to gather information about whether or not collections have been digitised or accessible online, though there is a field to display a single URL.

[61] http://fenscore.natsca.org/fensdbexplan.php
[62] http://www.natsca.org/NHNearYou

Appendix B

Summary of relevant emerging technologies

Within the constraints of a short timescale the project team set out, using the research and architectural issues described in **section 2** and **section 3** as a starting point, to identify the technologies and approaches that could be used to support the mapping framework, including emerging and innovative developments. In completing this task, the team drew on their experience of working on research and development projects within the cultural heritage sector, contacts within the technical community, and a focused scan of literature published on the open web. What follows is not intended to be a detailed, structured academic literature review, but an informed opinion on what might be useful technical approaches both now and for the future.

1 Introduction

Historically, a number of approaches to acquiring, analysing, augmenting and disseminating data about collections have been utilised within the sector. These established technologies (which are outlined below) have a reasonable level of adoption and still have a place in any framework and can serve as a benchmark against which the utility of alternatives might be considered. However, it must be recognised that the barriers to adoption of some of these are high which has often meant that smaller institutions have been unable to participate in past activities. If emerging technologies can lower the cost of participation, they may have an important role to play in any future framework, even if some aspects of them are less optimal and efficient as existing technologies, by having the potential to improve the breadth, if not the depth, of data available.

Terms in ***bold italics*** are defined and referenced in the glossary at **appendix C**.

2 Current approaches

2.1 Data acquisition/discovery

In order to discover, compare and interrogate data derived from distributed sources the data needs to be accessible from a single point. This can either be achieved in real time, through 'federated' or 'broadcast' searching of multiple data sources, or achieved by first assembling a central index or cache of data from the remote search and searching that single source instead. This latter approach, often referred to as 'aggregation', is the approach taken by all major internet search engines, such as Google, Bing, etc. it is worth noting that aggregation does not necessarily mean that the full content of the original data source is replicated centrally. A common strategy is to aggregate key metadata from the data source and refer users back to the original data source to retrieve detailed information.

Within the cultural sector, the library domain was one of the first to explore options for data sharing. A number of successful 'virtual union catalogues' (eg as part of JISC's electronic libraries 'clumps' programme) were set up based on simultaneous searching of multiple ***online public access catalogues (OPACs)*** with the results being returned to the user as a

single ‘hit list’. Whilst this approach has many advantages, such as guaranteeing that the information being returned is as up to date as possible and reducing the requirement for centralised storage of data (including, in the present day, any data protection concerns that might arise from centralised storage of third-party data), it also has several disadvantages. These include:

- A reliance on third party data sources being available and performant.
- A need for a high level of consistency between data sources as there is little scope data analysis and adaptation ‘on the fly’.
- Potentially poor user experience, as the speed with which a complete result set us available to users is always determined by the performance of the slowest of the remote sources.

In general, whilst federated searching does have a role to play in certain applications, it is best suited to scenarios where there are a relatively small number of data sources which are robust and homogeneous in terms of both content search interface. By way of contrast the cultural sector can be characterised as potentially having a huge number of disparate data sources of unknown stability. For this reason, the only practicable basis for the mapping framework is aggregation.

In constructing an aggregation infrastructure, there are a number of existing options. The low cost and initially appealing approach is to build on an existing, general purpose, aggregator such as Google. Google’s custom search engine[63] allows developers to provide a search interface that is limited to a subset of the web pages indexed by Google (typically a single site or group of sites). This can be delivered for free, based of Google receiving revenue via Google AdWords. The main disadvantages of this approach, particularly for a national framework intended to be an impartial resource for researchers and professionals, include:

- Relevance ranking is determined by secret algorithms that, amongst other things, seek to boost advertising revenue. Search Engine Optimisation (SEO) is a huge business in its own right and it is certain that larger institutions will be better placed to ‘play the game’ and boost the ranking of their resources than smaller ones.
- Google is a general purpose tool and treats all web content as being pretty much a stream of free text. This means that this approach misses out on the potential benefits of structured metadata, such as limiting searches to a particular creator, material type, subject, etc.
- The approach only works for web pages indexed by Google.

Another option would be to base an approach on a commercial enterprise search tool such as HP Autonomy. These tools, which were extremely popular in the 2000s, are designed to provide ‘big data’ searching capabilities across a range of document types, not only web pages. Despite being based on unstructured data, many claim the ability to ‘learn’ about specific domains and evolve better relevance ranking as a result. However, like Google PageRank, these algorithms are proprietary and this, combined with the high cost of

[63] https://cse.google.com/

purchase, means that this approach in unlikely to be the best basis for a mapping tool for the sector.

There are a number of more specialised tools currently in use within the sector. These tend to be based on data push, data pull, or a combination of both.

Data push refers to the situation where the data source makes an active decision to make its data available (ie publish it) to a third-party system. This can include activities such as uploading a data set to data.gov.uk or filing a tax return. Many systems in the cultural sector have mechanisms, including ***API***s, which allow data to submitted in this way, for example the WorldCat metadata ***API***. Jisc has promoted ***SWORD*** a standardised protocol for submitting data to a repository but take-up outside of higher education has been limited and progress on version 3 of the protocol seems to be slow or have stalled. Culture Grid included support for ***SWORD*** v2 along with an ***FTP***-based data submission form and Europeana experimented with Operation Direct – a 'write' method implemented on the Europeana ***REST API*** – although this does not currently appear to be supported. In principle, a ***REST*** based data submission ***API*** would be a useful component of the mapping framework. However, it is important to be aware that 'zero touch' processing of data submitted in this way is reliant on strict adherence to data standards from the submitting organisation. Dealing with issues in the data submission and processing workflow has the potential to incur an administrative (and therefore cost) overhead that could affect the ongoing viability of the framework.

Data pull refers to the process of 'harvesting' or 'crawling' data from data sources. Here the data source is largely passive (although there are sometimes methods by which sources can signal that they wish to initiate a pull) and the timing and frequency of data acquisition is under the control of aggregating system. Web Crawlers, or 'bots' are ubiquitous on the modern internet and are the main method by which internet search engines acquire and index content. Because of their nature as general-purpose tools, search engines tend to treat the content of a web page as relatively unstructured, although certain tags, or other forms of mark-up result in special treatment. Also, the analysis and ranking algorithms used by popular search engines are highly commercially sensitive and difficult to replicate. However, because of the extremely low barriers to entry and widespread adoption of the underlying mechanism, some form of content crawling is, we believe, an essential component of the mapping framework. In the next section we examine ways in which emerging approaches to improving the machine readability of web content can enhance the utility of this approach.

The other main approach data pull is the harvesting of more structured content. Europeana, and most of the national and thematic aggregators which feed it (including Culture Grid) adopts the ***OAI-PMH*** protocol for harvesting. Version 2 of the protocol was released in 2002 and was adopted widely in digital archives and repository repository software such as EPrints, DSpace, MetaLib, etc. Google initially supported the protocol but withdrew this in 2008. Since then it's general use has declined. The ***OAI ORE*** specification was launched in 2008 in a bid to update this in particular in the area of aggregations of web resources based on a resource map. However, this never gained much traction. More recently the Open Archives Initiative has worked with NISO to produce the ResourceSync Framework Specification. The specification extends the ***sitemap*** protocol originally developed by Google as an alternative to ***OAI-PMH*** and used to guide search engines and web crawlers through a site's contents. The extensions include methods for describing resources drawn

from ***OAI ORE*** as well as methods for describing changes to promote synchronisation. Despite v1 being published in 2002 and v1.1 in 2007 it has yet to be incorporated into many sites and repositories in the cultural heritage sector.

As well as standards-based protocols such as ***OAI-PMH***, a number of institutions and repositories have developed their own Applications Programming Interfaces (***API***s) which allow access to content. Although these ***API***s differ in terms of the specific commands and syntax required, the majority share a common underlying access method (eg a ***REST***ful architecture). This means that, whilst adding a new institution with its own ***API*** is not quite a straightforward as adding an institution using ***OAI-PMH***, a configurable mechanism can be developed to limit the effort (and cost) involved.

In defining the scope of tools to support the mapping framework, the above summary demonstrates that there are a large number of potential mechanisms for acquiring content and that a single approach will not be sufficient to achieve the level of coverage desired. Any tools developed should not only be able to support the main mechanisms outlined above but should also be extensible and adaptable to be able to cope with new approaches that develop over time.

2.2 Analysis, normalisation and augmentation

Many data aggregators use techniques to improve the consistency and searchability of data acquired from disparate sources. These approaches include:

- Terminological alignment – using crosswalks between controlled vocabularies to allow a common method for information retrieval.
- Data transformation to ensure common data structures.
- Normalisation of data types (eg using common date formats).
- Assignment of identifiers (including ***URI***s).
- Extraction of 'embedded' metadata (eg ***Exif*** data embedded in a JPEG).
- Extraction of structured data from unstructured or semi-structured data elements. A common example of this is looking for common structures (e.g. subjects) within a 'description' data element. The Europeana Food & Drink Semantic Demonstrator[64] used this approach to extract subject classifications (wikipedia) from free text data. The approach went beyond simple string matching, using natural language processing techniques to reduce 'false hits' and expert curators to help train the machine learning algorithm.

Archives Hub[65] in the UK has developed a comprehensive processing 'pipeline' which allows for multiple data processing routines to be applied sequentially to data from a given source to provide an output format (in this case ***EAD*** – encoded archival descriptions) which conform to the Hub's profile and are suitable for onward syndication to Archives Portal Europe. Processing steps have been defined which can be applied to identified groups of contributors (eg. providers using CALM will go through a different processing routine to

[64] https://foodanddrinkeurope.eu/professional-applications/semantic-demonstrator
[65] https://archiveshub.jisc.ac.uk

those using ADLIB). Whilst this requires effort to set up initially, it allows the prospect of virtually zero touch processing in 'steady-state' use and for transformations and pipelines to be applied and monitored by non-technical staff.

Europeana offers a variety of ***API***s[66] which can be used to enrich data supplied by cultural heritage organisations. These include named entity extraction, semantic enrichment and annotations (user and machine generated). The 'enriched' data is then made available to the supplying organisation with the potential for this to be re-imported into the data source. Although these have been around in experimental form for a number of years, they are still variously referred to as 'alpha' or 'beta' releases on the Europeana Pro website. We will examine these in more detail in the emerging technologies section.

2.3 Dissemination

The main purpose of bringing together data from a variety of sources to make it possible for third parties (human and machine) to retrieve and explore the data. User interface design is constantly evolving and data indexing tools (such as ***Lucene*** and products based on it) allow data to be surfaced in new ways. Any modern data aggregation system should be able to conform to best practice in areas such as:

- Providing a simple 'google-like' search interface to meet the expectations of the majority of non-specialist users.
- Faceted searching to allow users to explore data based on an initial search.
- Browsing of data based on key dimensions.
- Visualisations of data (***Ngrams***, etc).
- Providing an ***API*** (or ***API***s) which allow machine to machine interaction.
- Allowing subsets of data to be downloaded in commonly understood formats (eg ***CSV***, ***JSON***, ***XML***) for further off-line analysis (eg by academic researchers).
- Share records, or groups of records, via an ***API***, eg. to construct a virtual exhibition containing items from several institutions.

3 Emerging technologies

3.1 Data acquisition/ discovery

Web Crawling

The blueprint report 'Towards a UK Digital Public Space'[67] prepared for the Strategic Content Alliance by SERO Consulting in 2014 identified three potential levels (of data richness) of web crawling that had potential for use in the aggregation process:

- Crawl HTML web pages (and associated objects such as images) and index the unstructured content.

[66] https://pro.europeana.eu/resources/apis
[67] https://digitisation.jiscinvolve.org/wp/2014/12/08/towards-a-uk-digital-public-space-a-blueprint-report

- Crawl HTML web pages and use any embedded ***microdata*** (specifically including ***Schema.org microdata***).
- Crawl Linked Data published on the web and index the structured data retrieved.

For the purposes of the mapping framework, we recommend that the last of these should be explicitly supported.

As stated in previous section, web crawling can be augmented and improved by content providers adding markup to their comments to aid machine readability. A number of different markup structures have been proposed over the years including ***RDFa***, ***HTML5 Microdata*** and ***JSON-LD***. However, whilst these provide the structural elements necessary for machines to recognise and extract marked up data, they do not, on their own, specify a common grammar for use within a community of interest. Most current activity is centred around ***Schema.org***, a community initiative which allows schemas, or vocabularies, to be created and shared. The resulting markup can be embedded within a website in any of the above formats. https://schema.org/Museum shows a simple example of how machine-readable data about a museum can be embedded in a web page in each format.

Mappings have been produced between ***Schema.org*** types and the Europeana Data Model[68]. The new Europeana Common Culture project, which started in Jan 2019, will follow this up with a trial of aggregating content via ***Schema.org*** mark-up. It is recommended that a route to acquire content through crawling is incorporated into the mapping framework and that the prototype data collection tools implements this if a suitable collection with a web presence can be identified that has a sitemap and incorporates ***Schema.org*** mark-up.

Whilst we agree that crawling of linked data has potential, and is also being investigated by Europeana, we think that this is less important to demonstrate in the prototype as, at the moment, any organisation capable of publishing collections information as ***LOD*** is likely to be at the technologically sophisticated end of the spectrum and already be making its content available by other means. Therefore, including this data acquisition method is unlikely to improve the scope of data being aggregated (although it does have potential for improving the quality).

IIIF Manifests

Research involving National Library of Wales and others for Europeana has identified the potential for ***IIIF manifests*** to be used not only to locate images within image collections but also to source data about the images, and potentially the objects that the image represents. As we are aware that a large number of cultural heritage institutions are implementing ***IIIF*** for their image collections, we feel that this approach needs to be incorporated within the framework and, if a suitable collection can be identified, investigated further within the prototype.

[68] https://journal.code4lib.org/articles/12330

APIs

A number of institutions have provided ***API***s to allow their collections to be searched. Some of these also support harvesting of data. Whilst a dedicated harvest ***API*** is preferable for the purposes of aggregating content, in practice it is often possible to use structured queries to achieve the same purpose. Experience of working with Culture Grid and CCIM middleware is that, for most ***REST API***s, the process of harvesting data is not dissimilar to harvesting using ***OAI-PMH***. We therefore recommend that this approach to data acquisition is included within the mapping framework and as part of the prototype.

3.2 Analysis, normalisation and augmentation

Whilst approaches to data acquisition based on highly structured data sources can greatly help to ensure consistency of data, it does not guarantee this. For example, institutions will use a variety of different controlled vocabularies and classification systems for their internal purposes. In order to search across these, some form of mapping or crosswalking between vocabularies is required. The problem is exacerbated when acquiring data from web pages with little or no structure. The sections below builds on the approaches described in 3.2.2 above to identify the key features that the mapping framework should include to maximise the potential users to gain a consistent view across data sources.

Text mining/natural language processing

As mentioned previously, Europeana have implemented prototype services based on these technologies for a number of years, with the EFD Semantic Demonstrator being an excellent case study for the auto-classification of objects or collections based on free text descriptions. Ontotext, the lead partner in the demonstrator, has also been actively involved in the development of GATE toolkit[69], hosted at the University of Sheffield. For demonstration purposes, GATE offers a number of advantages. Firstly, it is available as open source software. Secondly, Knowledge Integration already have experience in using the toolkit to assist with named entity extraction for a number of institutions including the Horniman Museum & Gardens.

A potential use of this within the demonstrator would be to use a similar process to that used in the Europeana EFD demonstrator (subject, of course, to time and resource constraints) to attempt to derive subject classifications, where these do not already exist, from data sources including Cornucopia collection level descriptions.

Image analysis and processing

There is widespread interest and activity in image recognition and analysis tools. Much of the academic research has been focussed of scientific analysis around provenance for art works (eg. analysis of canvas weave, pigments, brushstrokes, etc. which, whilst of interest to the community, is not directly applicable to the mapping project. Mass market tools such as Magnus[70] and Smartify[71] have an important role to play in democratising and demystifying the world of fine art but are also not directly applicable, although the underlying algorithms, if made available, would be of real potential use.

[69] https://gate.ac.uk
[70] www.magnus.net/about
[71] https://smartify.org

Useful types of data that could potentially be extracted from images of cultural heritage objects, and particularly of art works, include:

- Title.
- Creator (Artist).
- Genre/Subject.
- Dates.

Most of the tools capable of performing this sort of analysis are either commercial or highly experimental. As tools need to be trained on data sets. Commercial organisations are keen to gain access to themed collections for development purposes, although the terms associated with access might not be appropriate to a national framework. Of the open source tools available, the VIC engine[72] from Oxford University appears to be promising. As this has already been trained on samples from Google Images and Art UK[73], there may be potential for including a trial of this using a suitable art collection during the prototype. In any case, any framework for aggregation should demonstrate the potential for this or similar tools to be integrated as they emerge and mature into robust products.

There has also been work carried out on image recognition at Science Museum, using the Google Vision ***API***[74] to apply classification tags to image content.

3.3 Dissemination

As well as a modern end user interface and the ability to download data in specific formats, indicative requirements for accessing and using the data have been collected during the consultation process. However, we do expect all the elements of a modern system to be included, including end user (human) interfaces and machine interfaces (***API***s). We also note the growing importance within the wider public sector of publishing data sets for use in 'mash-ups' and other forms of data combination. Data.Gov.UK[75] contains a large number of data sets of relevance to the cultural heritage sector and the mapping framework should be capable of publishing relevant data sets alongside these and updating them on a regular basis. Care should be taken when publishing data sets to ensure findability.

Due to time and resource constraints, we do not envisage the prototype including any user interface or ***API***s (except, perhaps, native access to a prototype ***Elasticsearch*** index).

[72] www.robots.ox.ac.uk/~vgg/software/vic
[73] https://artuk.org
[74] https://cloud.google.com/vision
[75] https://data.gov.uk

Appendix C

Glossary

API

Application Programming Interface. Sets of rules that allows computers to communicate and exchange data. For example a web browser using an API can retrieve and display data from a server.

See: https://en.wikipedia.org/wiki/Application_programming_interface

CSV

Comma Separated Values. Data separated by commas (eg for an object it might include: object name,title,material,date). Each line represents a different record. It is possible to use a different delimiting character such a tab.

See: https://en.wikipedia.org/wiki/Comma-separated_values

EAD

Encoded Archival Description. A standard for tagging archival finding aids that can be processed by computers.

See: https://www.loc.gov/ead

Elasticsearch

Search engine software that can be used to search many kinds of document. It is based on ***Lucene***.

See: https://www.elastic.co/products/elasticsearch

Exif

Exchangeable image file format. A file format that allows data about the image to be embedded in the file itself.

See: https://en.wikipedia.org/wiki/Exif

FTP

File Transfer Protocol. A standard for the creation of software that allows the transfer of data between a client computer and a server computer.

See: https://en.wikipedia.org/wiki/File_Transfer_Protocol

IIIF

International Image Interoperability Framework. A standard defining APIs for searching and presenting images over the web. An aim is to support interoperability between different image repositories. A IIIF manifest is the data that allows an image to be viewed.

See: https://iiif.io/

JSON

JavaScript Object Notation. An open standard for creating human-readable text to transmit data objects consisting of attribute–value pairs. ***JSON-LD*** (JavaScript Object Notation for Linked Data) is one method for doing this.

See: https://en.wikipedia.org/wiki/JSON

Linked open data (LOD)

Structured data that is published on the web which allows links to other structured data. When the the data is published with a licence that allows it to be reused it is said to be 'open'.

See: https://en.wikipedia.org/wiki/Linked_data

Lucene

Software for the searching of text documents for the extraction of indexes from them.

See: http://lucene.apache.org

Microdata (HTML)

A standard for embedding relevant data into web pages that can be extracted to give users better search results.

See: https://en.wikipedia.org/wiki/Microdata_(HTML)

Ngram

In this context a method of visualising the frequency of subset of controlled terms in a set of documents.

See: https://en.wikipedia.org/wiki/N-gram

OAI ORE

Open Archives Initiative Object Reuse and Exchange. Standard defining the description and aggregation of web resources.

See: https://www.openarchives.org/ore

OAI-PMH

Open Archives Initiative Protocol for Metadata Harvesting. Protocol developed for harvesting metadata descriptions of records in an archive so that services can be built using metadata from many archives.

See: https://www.openarchives.org/OAI/openarchivesprotocol.html

OPAC

Online Public Access Catalogue. Used especially by libraries to give users access to their materials on the web. Replaced card catalogues at the library.

See: https://en.wikipedia.org/wiki/Online_public_access_catalog

Persistent identifier

In this context, a long-lasting ***URI***, or 'permalink', that should not end up as a broken link in the future.

RDF

Resource Description Framework. A standard based on the idea of making statements about things (in particular web resources) in the form of 'subject–predicate–object', known as triples.

See: https://en.wikipedia.org/wiki/Resource_Description_Framework

REST

REpresentational State Transfer. A standard for developing services on the web based on those standards already existing..

See: https://en.wikipedia.org/wiki/Representational_state_transfer

Schema.org

A collaboration to create small pieces of structured data describing the content of web pages. These allow useful services to be made from that data.

See: https://schema.org/Museum

SPARQL

SPARQL Protocol and ***RDF*** Query Language. Pronounced 'sparkle', it is the standard way of querying ***linked open data*** on the web or for databases containing ***RDF***.

See: https://en.wikipedia.org/wiki/SPARQL

SWORD

Simple Web-service Offering Repository Deposit. A standard for depositing content from one place to another

See: http://swordapp.org

URI

Uniform Resource Identifier. A string of characters which uniquely identifies a resource (eg a web page). The URL (Uniform Resource Locator) is the most common type of URI.

See: https://en.wikipedia.org/wiki/Uniform_Resource_Identifier

XML

eXtensible Markup Language. A standard for marking up (tagging) documents in order to give meaning to parts (elements) of the document. A set of rules for the marking up is defined by a 'schema'.

See: https://en.wikipedia.org/wiki/XML