

Mapping digitised collections in England

Final report

Prepared for the Department for Digital, Culture, Media and Sport
March 2019

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

www.collectionstrust.org.uk

Contents

Executive summary	2
Limitations.....	2
Project phases	2
Conclusions	3
1 Introduction.....	5
1.1 Policy context	5
1.2 About the consultants.....	5
1.3 Limitations.....	5
1.4 Project phases.....	6
1.5 Some key concepts	7
2. Summary of the scoping phase	13
2.1 Research into online availability of digitised collections	13
2.2 Framework architecture and design.....	16
3. Building the prototype	20
3.1 Base platform	20
3.2 'Pluggable' modules	21
4. Testing the prototype	22
4.1 Content ingested	22
5. Use scenarios.....	39
5.1 Material relating to Charles Darwin	39
5.2 Images of Egyptian ceramics	41
5.3 Anglo-Saxon material from Essex.....	42
5.4 Farming-related material	44
5.5 Monk's Hall Hoard	46
6. Conclusion.....	49
6.1 Evaluation against success factors.....	49
6.2 Benefits to the cultural heritage sector.....	52
6.3 Risks	56
Appendix A: Acknowledgements	58
Appendix B: Glossary	59

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 (CT), working with Knowledge Integration Ltd (K-Int) and with input from Culture24. Both CT and K-Int were involved in setting up the legacy aggregator Culture Grid and continue to maintain it on a *pro bono* basis.

Limitations

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.

Project phases

In the scoping phase, 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 scoping report also proposed an architecture for the framework that would achieve 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.

In the second phase of the project, the framework described above was turned into a prototype tool that aimed to illustrate the viability of the architecture and approaches proposed in the scoping phase of the project. The prototype integrated 'pluggable' modules to demonstrate the possibilities of various AI services, particularly in the area of content analysis and enhancement.

In the third phase, sample data was brought into the prototype from a range of sources and using several different technical approaches to demonstrate the flexibility of the framework.

A processing pipeline was created to allow the AI enhancement services to be applied to the test data. The results were considered against five hypothetical use-case scenarios.

Conclusions

The report concludes that:

- The test data demonstrated that, with a suitable user interface, the proposed framework would allow end users a single point of access to data at multiple levels from a wide range of institutions.
- The test demonstrated that the prototype could ingest data from a range of institutions using various technical means, and without the institutions having to format their contributions to any kind of set template.
- Although museums, libraries and archives follow different cataloguing standards (often even within the same institution), the flexible nature of the prototype means there is no technical reason why data from all three could not be harvested by an aggregator built using the same architecture, allowing searches to be made across the different collection types.
- The potential and pitfalls of agreed emerging technologies was demonstrated, as various generic AI services were applied to the test data had mixed success. Specific entities such as places were often recognised, and sometimes enhanced with additional information. However, there were enough mis-identifications and mis-classifications to illustrate the important principle that such AI enhancements should be clearly identified as auto-generated and kept apart from the original source data. Such problems, however, reflect the need to train such AI services with lots of data relevant to cultural heritage collections, rather than any inherent shortcomings of the technologies themselves.
- The test demonstrated that the framework is widely applicable and does not take the 'one size fits all' approach of previous aggregators. In particular, the hierarchy of information means that an institution could initially be represented with just collection-level records that would still be useful, and adding item-level records and digitised assets as and when they were ready, in whatever form and through whatever means they were able to provide.

The following potential benefits to the cultural heritage sector are noted:

- Enabling content curation to reach new audiences. It is important to stress that most of the public benefits likely to flow from the proposed data aggregation would be indirect rather than direct. The aggregator would not be a destination site for the wider public; rather it would be the tool behind limitless end-use scenarios that presented curated content to specific audiences.
- Supporting dynamic collections management, one of the priorities identified by the 2017 *Mendoza Review*. Addressing this priority would be a lot easier if those working

collaboratively across the sector could routinely search across the collections data that is currently siloed within individual museums.

- Strategic partnership with higher-education sector, starting with online access for researchers to the records in collections databases.
- Being part of international research and development, by providing the pipeline needed to connect UK collections data to European and global networks.
- A strategic, cross-sector approach to gathering audience data, such as applying digital fingerprinting at aggregator level to allow the onward journeys of downloaded or shared assets to be tracked with greater precision than currently attempted outside the commercial sector.
- Maintaining authoritative lists of cultural heritage institutions to allow more consistent data recording and management by funders and other sector bodies.

A number of potential risks are also identified:

- Confusion about potential audiences for aggregated data, most of which would not be the kind of curated content expected by audiences. Rather, the aggregator would allow a wide range of third parties to research, select and re-purpose the raw data. In framing the business case for any eventual aggregator built on the proposed framework, it will be important to keep this distinction in mind, and to value the behind-the-scenes use of aggregated data by curators for collections management purposes as much as the more obvious public-facing possibilities.
- Duplicate records if data is drawn from disparate sources, especially a mix of individual institutions and other aggregators.
- Mixing up original source data and versions processed at aggregator level, unless an original copy of the source data is kept, and the contributing institution is able to review the imported data.
- Broken links as contributing institutions rename their online content, move it around or otherwise fail to maintain it.
- Lack of long-term commitment, leaving cultural heritage institutions, software providers and developers of third-party applications that re-use aggregated data would be left high and dry if the service were not used, maintained and supported.

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:

'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.'

1.2 About the consultants

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

The mission of Collections Trust (CT) is to help museums capture and share the information that gives their objects meaning.² Founded in 1977 as the Museum Documentation Association, CT today is best known for its collection management standard, *Spectrum*. CT maintains, *pro bono*, the legacy aggregator CultureGrid³ referenced at several points in this report. CT is a sector support organisation within Arts Council England's national portfolio.

Knowledge Integration (K-Int) is a long-established software firm focusing on the publishing, aggregating, storing and retrieving of information by cultural heritage and educational organisations.⁴ K-Int's CIIM product, the middleware used to create the prototype for this assignment, is used by several DCMS-sponsored museums and galleries. K-Int developed the legacy aggregator, CultureGrid, and provides *pro bono* technical support as needed.

1.3 Limitations

The scope of this 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. It should also be stressed that the prototype was built, delivered and tested within just four weeks. As such, it was never intended to be a comprehensive demonstration of all possible routes through the proposed framework, nor to bring together all the digitised content found during desk research. It was

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 <https://collectionstrust.org.uk/what-we-do/>

³ <http://www.culturegrid.org.uk/>

⁴ <https://www.k-int.com/>

agreed that the prototype would not have a user interface, nor offer any kind of ongoing service. Instead, the prototype aimed to demonstrate key principles of the framework, using a sample of real data, in order to illustrate the viability of the architecture and approaches proposed.

1.4 Project phases

The study was carried out in three phases. The scoping phase began in early January 2019, with a draft scoping report submitted at the beginning of February. The aims of the scoping phase 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.*
- *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.’⁵*

The focus of the second phase, which ran through February 2019, was:

- *‘To develop and design a prototype data collection “framework” informed by the analysis and results of phase 1. This framework and design should be user friendly and allow for easy interrogation and connection of data.’⁶*

In the third phase, in March 2019, the prototype was ‘tested on a digitised collection(s) to demonstrate functionality.’⁷

Section 2 of this final report summarises the scoping phase report, which included the framework architecture and design. **Section 3** explains how the prototype tool was built. **Section 4** describes how test data was ingested into the prototype, and also how various artificial intelligence (AI) services were plugged into it and applied to the data. In **section 5** we consider to what extent the prototype – if loaded up with the kind of data found to exist in the scoping phase – might be helpful in five use scenarios. Finally, **section 6** considers lessons learned and the potential benefits and risks of moving beyond the prototype. Throughout the report, terms in ***bold italic*** are explained in the glossary at **appendix B**.

⁵ Ibid.

⁶ Ibid.

⁷ Ibid.

1.5 Some key concepts

The following introductory notes, which are slightly abridged from the earlier scoping report, unpack some of the terms used in the brief as a primer to some of the key concepts discussed in this final report.

1.5.1 'Cultural collections'

The collections at the heart of this study comprise the artefacts, specimens and documents held by museums, archives and libraries. 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.

It is *information* that gives these physical and born-digital 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.

1.5.2 '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 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.'*⁸

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.

⁸ www.collectionstrust.org.uk/digital-isnt-different/digitisation/

1.5.3 '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.

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' (*APIs*), 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*).⁹ 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 5**.

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 engine¹⁰ allows developers to provide a search interface that is limited to a specified website or group of sites. 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.

⁹ www.getty.edu/research/tools/vocabularies/ulan/

¹⁰ <https://cse.google.com/>

Moreover, the 'just use Google' approach has the same major drawback as the federated searching model. 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 crawlable 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.¹¹

Conclusion

For all the reasons set out above, in this report we assume 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.

1.5.4 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 recognise 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,¹² 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.)¹³

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.¹⁴ According to Dr Mia Ridge of the British Library,¹⁵ 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

¹¹ See spreadsheet **Table 2** attached to the scoping report.

¹² 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

¹³ <https://newatlas.com/algorithm-shakespeare-coauthor-marlowe/46130/>

¹⁴ www.turing.ac.uk/research/research-projects/living-machines

¹⁵ Pers. comm.

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.¹⁶ This would make an excellent corpus for an image-based AI tool that could then be trained to recognise 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 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'.¹⁷

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 recognise 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 connect digitised collections to the AI tools already available.

1.5.6 '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.

¹⁶ <https://collection.sciencemuseum.org.uk/documents/aa110067037/albums-of-photographs-of-electricity-pylons-in-various-countries>

¹⁷ <https://artuk.org/about/blog/the-art-of-computer-recognition>

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.

2. Summary of the scoping phase

2.1 Research into online availability of digitised collections

In the scoping phase, 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. The results of this research were presented in the form of spreadsheets submitted with the scoping report (**tables 1-4**). The process was meant to guide the sample selection and inform the use scenarios, not to be statistically representative. No attempt has therefore been made to draw any conclusions from the data that might suggest analytical rigour where none was intended.

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

- The 1,319 Accredited museums in England. ¹⁸
- The 57 museum National Portfolio Organisations within the ACE National Portfolio 2018-22. ¹⁹
- The 149 Designated collections in England, ²⁰ including archives and libraries as well as museums.

To whittle down these collections to a manageable number the legacy Cornucopia website ²¹ was searched using the following keywords: 'Egyptian', 'Saxon', 'Darwin', and 'farming'. These, like the user scenarios described in **section 5**, 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.

Using this process, a longlist sample of 88 institutions was agreed with DCMS. ²² 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.

2.1.2 Hierarchy of information

In order to focus the desk research, the consultants adopted the following information 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).

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

²¹ Discussed in **appendix A** of the scoping report.

²² See spreadsheet **table 1** submitted with the scoping report.

- **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.

2.1.3 Online research strategy

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).

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 ²³
- Cornucopia ²⁴
- The Museums Association's directory of museums ²⁵ (members/subscribers only).

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?

²³ www.culture24.org.uk

²⁴ <http://cornucopia.orangeleaf.com/>

²⁵ www.museumsassociation.org/find-a-museum

- 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**)?
- 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 ²⁶
- *Global Biodiversity Information Facility* ²⁷
- Culture Grid ²⁸
- Europeana ²⁹
- Archives Hub ³⁰

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)?

²⁶ <https://artuk.org/>

²⁷ www.gbif.org

²⁸ www.culturegrid.org.uk

²⁹ www.europeana.eu

³⁰ <https://archiveshub.jisc.ac.uk/>

- 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.

2.2 Framework architecture and design

This scoping report also proposed an architecture for the framework. This followed the conclusion drawn in **section 1** that the ‘aggregation’ model is the only viable approach for searching across such complex and variable data as digitised cultural heritage collections from many hundreds of institutions. The proposed architecture described 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 acknowledged 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 proposed 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’).

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.

2.2.1 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.

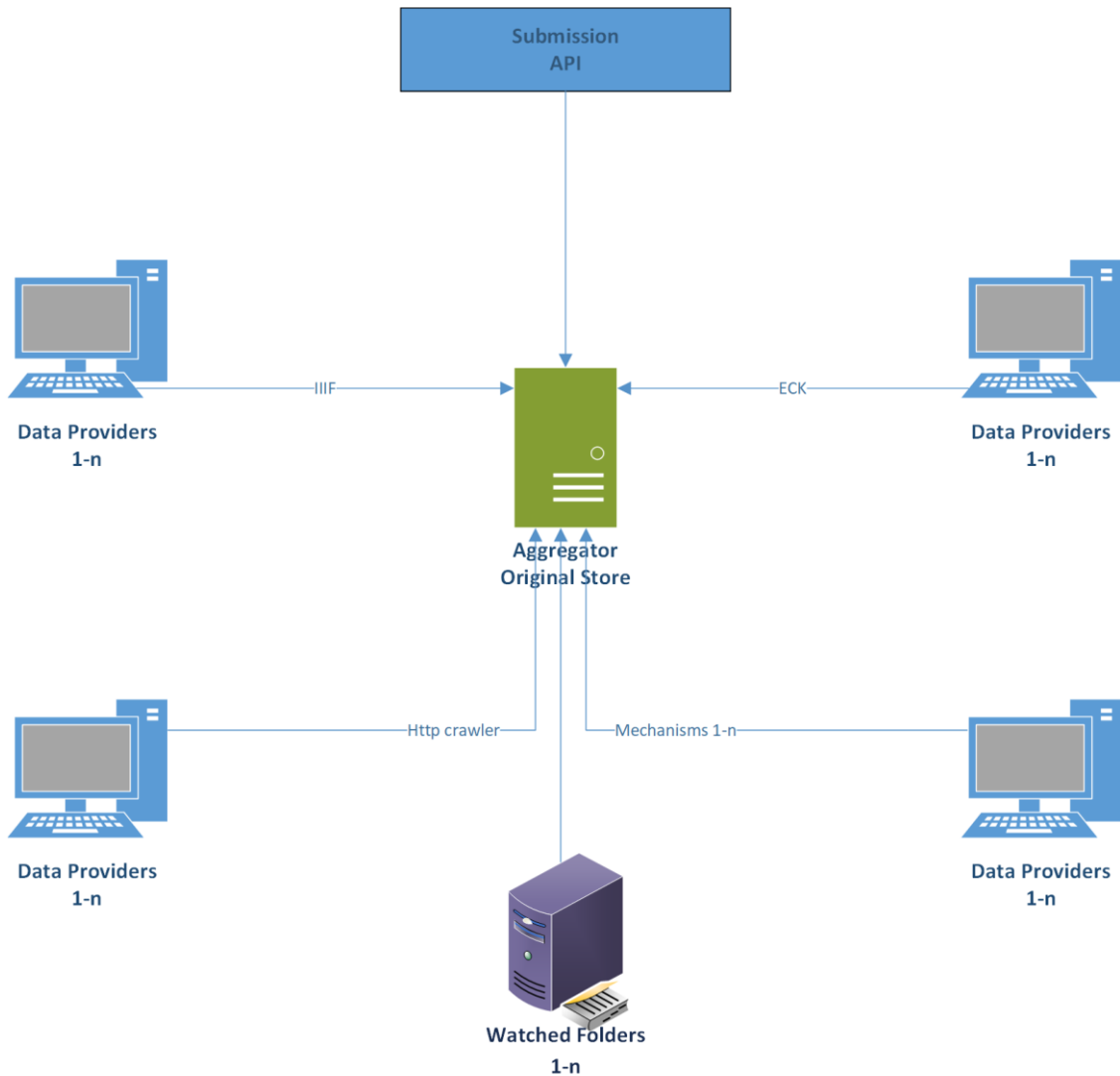


Figure 1 Ingest architecture. K-Int.

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.

2.2.2 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.

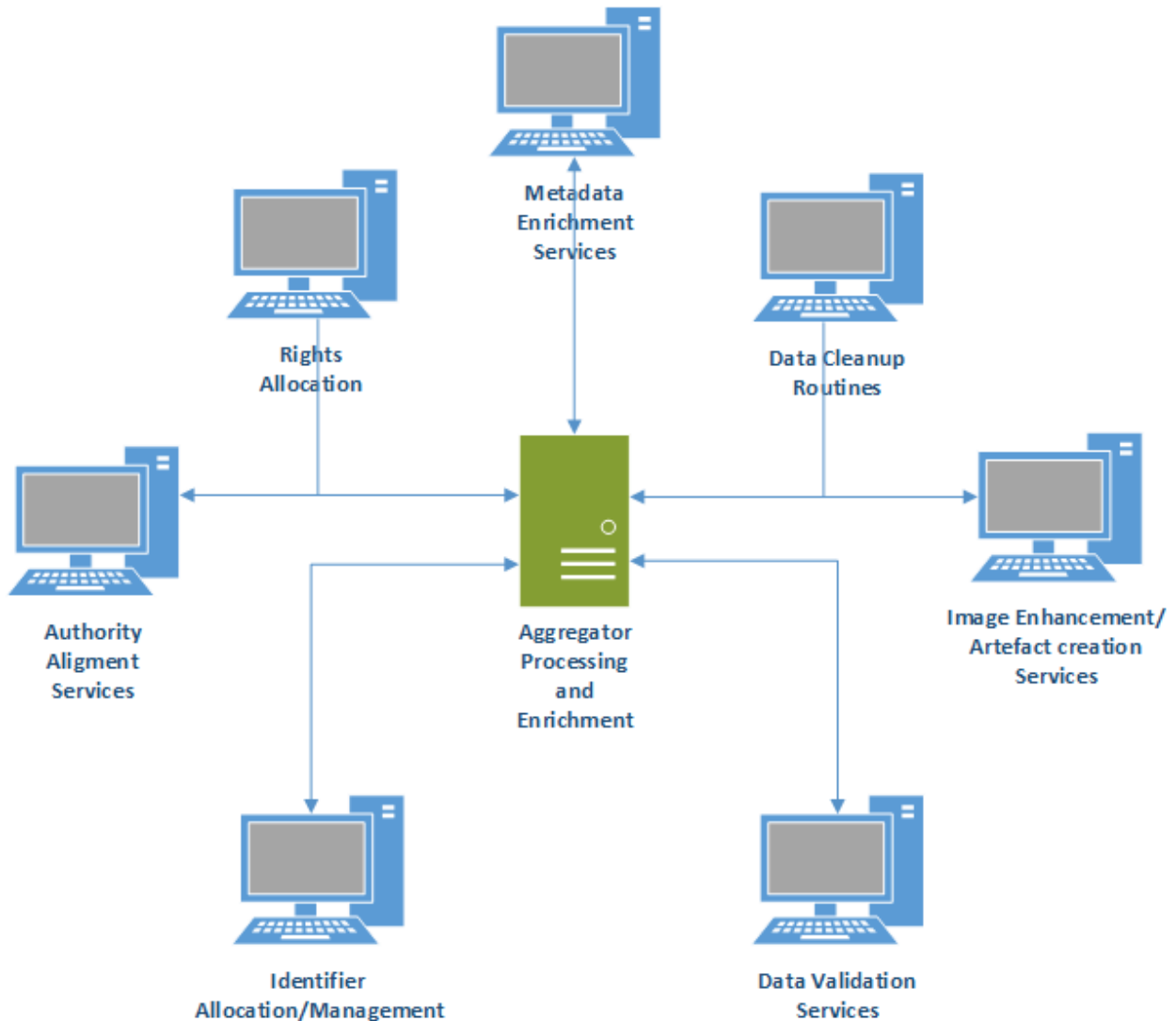


Figure 2 Processing and enrichment architecture. K-Int.

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. That is, 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.

2.2.3 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.

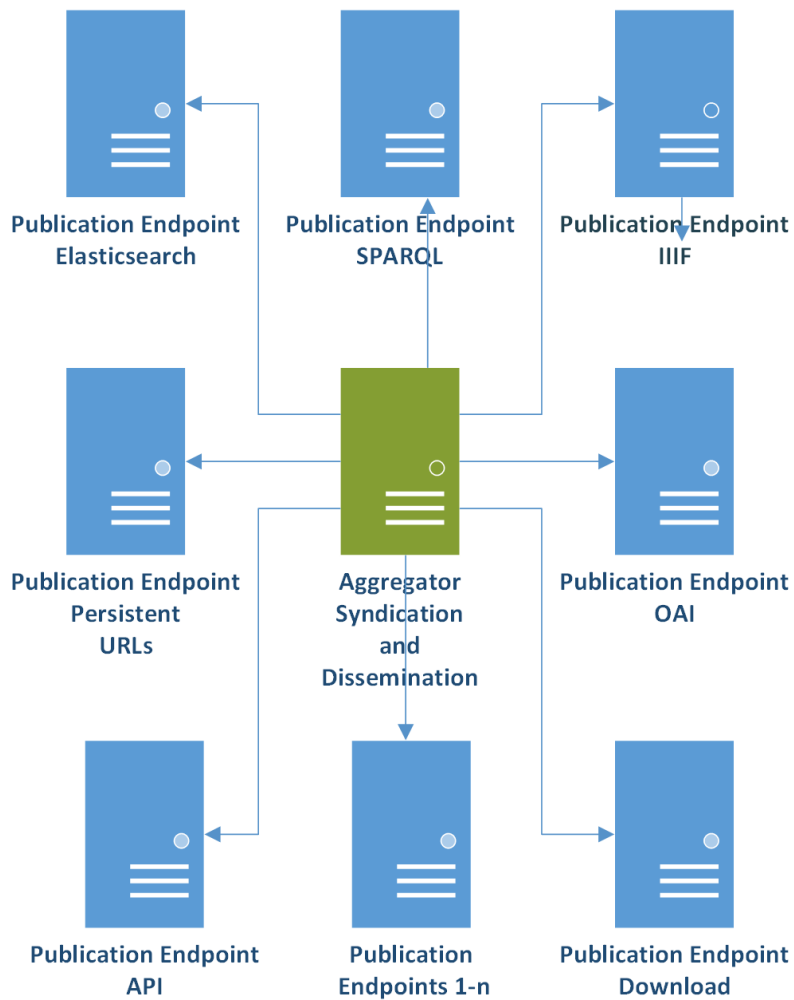


Figure 3 Syndication and dissemination architecture. K-Int.

3. Building the prototype

In the second phase of the project, K-Int turned the framework described above into a prototype tool. It bears repeating that the prototype does not have any front-end user interface. The screenshots included in this report therefore illustrate what might be going on under the bonnet of any eventual aggregator built according to the framework. This is not what the public would see in any of the limitless use scenarios that could draw data from the aggregator and re-purpose it for specific audiences.

To show how an aggregator could provide data for such use scenarios, the prototype published data to a number of endpoints, or APIs, that could be used to drive a user interface. One such endpoint is **Elasticsearch**,³¹ a powerful indexing tool used within the cultural sector³² and more widely.

3.1 Base platform

The prototype was based on K-Int's middleware and aggregation platform, CIIM (Collections Information Integration Middleware).³³ As well as being used by several national museums to combine data from various sources within their own individual systems, CIIM also powers Jisc's Archives Hub aggregator,³⁴ which brings together descriptions of thousands of archive collections from more than 330 institutions across the UK.

By basing the prototype on CIIM we were able to leverage a number of existing components for ingesting and processing data that have been proven to work with many of the data types and protocols used within the cultural heritage sector. Examples of different types of data, ingested through different technical routes, are given in **section 4**.

Figure 4 overleaf shows a view of the CIIM management interface in which the scheduled import of Culture24's data on venues (including the names and addresses of cultural heritage institutions) and events (at those venues) is in progress. Such scheduled tasks were repeated to improve the way incoming data is 'parsed' through re-processing of individual and groups of records from each of the data sources.

³¹ <https://www.elastic.co/products/elasticsearch>

³² Eg the Arches platform: <https://www.archesproject.org/standards/>

³³ <https://www.k-int.com/products/ciim/>

³⁴ <https://archiveshub.jisc.ac.uk/>

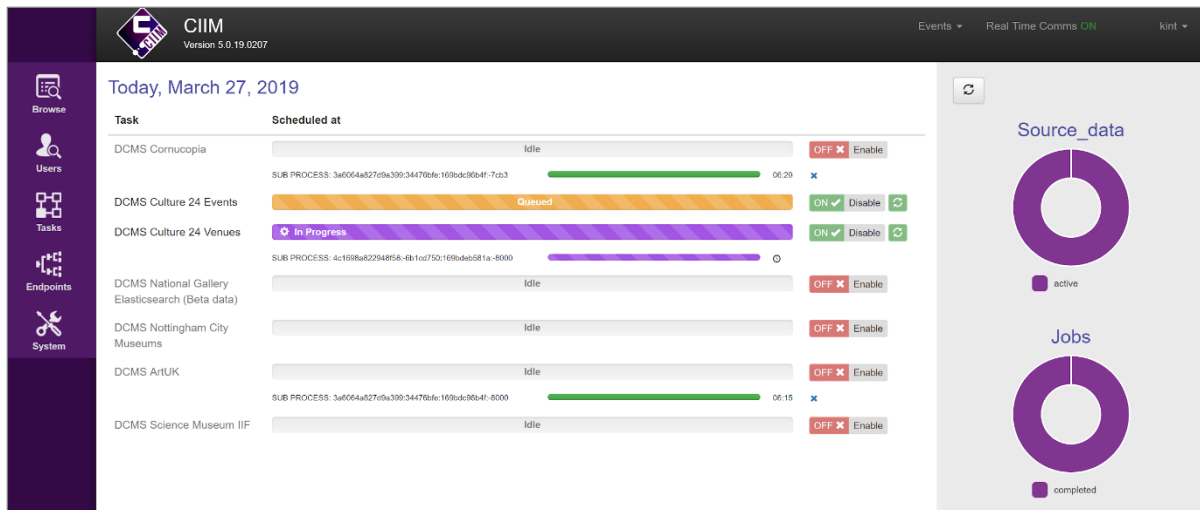


Figure 4 CIIM's scheduled task management interface. K-Int.

3.2 'Pluggable' modules

The prototype was designed to exploit CIIM's flexible architecture by integrating new 'pluggable' modules which demonstrate the possibilities of integrating with new services, particularly in the area of content analysis and enhancement – a key concern of the brief.

The following services were integrated. Screenshots showing how they were used with the test data are included in **section 4**.

- The University of Sheffield's GATE (General Architecture for Text Engineering) open-source toolkit that can be used for most types of text processing problems.³⁵ In particular, the 'entity recognition' plugin ANNIE (A Nearly-New Information Extraction System) was installed, which uses a list of internally configurable gazetteers to identify entities within data.
- Google's Places API,³⁶ a paid-for service (with a free quote each month) that returns information about places using **HTTP** requests. Places are defined within this API as 'establishments, geographic locations, or prominent points of interest'. Where available, the service supplies details such as opening times, address details, geo detection, names, etc.
- IBM Watson - Natural Language Understanding,³⁷ a paid-for HTTP service with a 'lite' development plan which is free for a capped number of calls per month. It provides multiple services for text analysis and enhancement.
- Amazon's Rekognition³⁸ service, which detects objects, scenes and faces, and its natural language processing service, Comprehend.³⁹ These are both paid-for service, with limited free plans for new customers.

³⁵ <https://gate.ac.uk/>

³⁶ <https://developers.google.com/places/web-service/intro>

³⁷ <https://www.ibm.com/watson/services/natural-language-understanding>

³⁸ <https://aws.amazon.com/rekognition/>

³⁹ <https://aws.amazon.com/comprehend/>

4. Testing the prototype

This section of the report describes the test data that was ingested into the prototype tool and the processing that was applied to it to illustrate the potential of available AI services to analyse and enhance the data. The discussion refers to the four-tier information hierarchy adopted throughout the project.

- **Level 1: institutions.** The names of cultural heritage institutions (linked to information held elsewhere about their location, opening times, contact details, etc).
- **Level 2: collections.** Information about analogue collections held by each institution, ranging from one or two keywords to descriptive summaries.
- **Level 3: item-level catalogues.** Where available, ingestible item-level catalogue information, either directly from an institution or via another aggregator.
- **Level 4: digital assets.** where available, images and other digital assets (eg sound and video files) associated with item-level records.

4.1 Content ingested

4.1.1 Data about institutions and collections

Sample data for the top two levels (institution names and collection-level keywords and descriptions) was brought into the prototype tool from a number of sources:

Cornucopia

Cornucopia is a legacy database commissioned by the former Museums & Galleries Commission in 1998, but not actively maintained for years.⁴⁰ It contains summary descriptions of collections held in a range of cultural heritage institutions, most of which are museums. In larger institutions collections are often described department by department. The descriptions vary considerably in length and detail, and the keywords used to tag them are also inconsistent. Nonetheless, although out-of-date and incomplete, the Cornucopia dataset remains the most comprehensive source of information about the holdings of the nation's museums. It is the museum equivalent of the collection-level 'finding aids' routinely compiled by archivists and aggregated through sites such the Archives Hub aggregator.⁴¹

A copy of the Cornucopia data exists in another legacy database, the aggregator Culture Grid.⁴² For the test, K-Int ingested the whole Cornucopia dataset into the prototype tool from there using Culture Grid's API.⁴³

Figure 5 below shows an example of an ingested Cornucopia record: a description of the archaeology collection in Cheltenham Art Gallery and Museum. Similar records describe the same institution's other collections: agriculture, archives, arms and armour, biology, costume and textiles, decorative and applied art, ethnography, fine art, geology, medals,

⁴⁰ For more background information, see **appendix B** of the scoping report.

⁴¹ <https://archiveshub.jisc.ac.uk/>

⁴² For more background information, see **appendix B** of the scoping report.

⁴³ Via an OAI-PMH harvest, with XML data elements conforming to the RSLP profile of the Dublin Core metadata element set.

numismatics, personalia, science and industry, social history, and transport. As is typical in such summaries, the level of detail varies from the specific ('iron age ... material from Salmonsbury Camp') to the more general ('Egyptian').

Collections	
Access	
AUDIENCE	Not Specific
Description	
VALUE	This collection covers the area of north Gloucestershire and spans the period from the palaeolithic to the industrial revolution. Particular strengths are finds and excavation archives from the neolithic long barrows of Belas Knap, Notgrove and West Tump. Significant iron age collections include material from Salmonsbury Camp (Bourton on the Water), Kings Beeches (Cleeve Hill), Oxenton Hill and Leckhampton Hill. Romano British settlement sites are represented by finds from Bourton on the Water, Andoversford, Syreford, Vineyards Farm (Charlton Kings), Haymes (Southam) and Wycomb and the villa sites by Compton Grove and Whittington Court. Anglo Saxon material from the cemetery at Bishop's Cleeve has recently been acquired. Medieval finds are so far the earliest artefacts attesting to permanent settlement on the site of Cheltenham itself. The collection also includes some Egyptian, Classical Greek and Roman artefacts. (less)
Subjects	
SUMMARY TITLE	Archaeology
Name	
VALUE	primary Archaeology
Title	
VALUE	primary Archaeology Collection
Contact	
Email	
VALUE	artgallery@cheltenham.gov.uk

Figure 5 Example of an ingested Cornucopia record. K-Int.

Thesaurus of institution names

The Cheltenham example shown above also illustrates the need for a dataset that was created specifically for this project and ingested into the prototype: a thesaurus of institution names. The institution known to Cornucopia as 'Cheltenham Art Gallery and Museum' now calls itself 'The Wilson'.⁴⁴ Since the prototype might have to deal with data containing either the old or new names, it needs to know that both refer to the same institution.

To address this problem, we created a test thesaurus of institution names using K-Int's Lexaurus terminology management tool.⁴⁵ Like any thesaurus that follows the standard for such resources⁴⁶ it allows a link to be made between a 'preferred term' (in this case, the new name: 'The Wilson') and any number of 'non-preferred' terms (such as 'Cheltenham Art Gallery and Museum', 'Cheltenham Art Gallery & Museum', 'Cheltenham Museum', etc.)

⁴⁴ www.cheltenhammuseum.org.uk

⁴⁵ www.k-int.com/products/lexaurus/

⁴⁶ ISO 25964, hosted at www.niso.org/schemas/iso25964

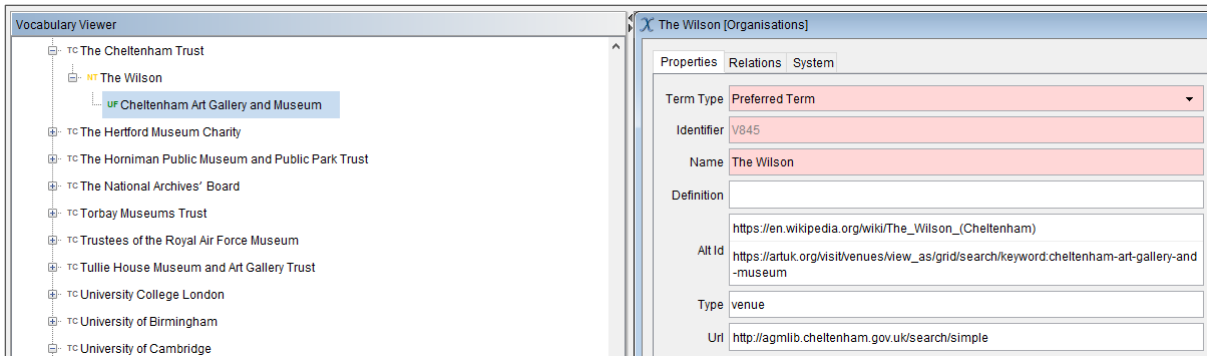


Figure 6 Lexaurus record for 'The Wilson' in the test thesaurus of institution names. K-Int.

A thesaurus also shows the hierarchical relationship between terms. In the test thesaurus, this feature was used to show the relationship between governing bodies and institutions (including the component parts of multi-venue services). The screenshot below shows the thesaurus record for the Science Museum Group.

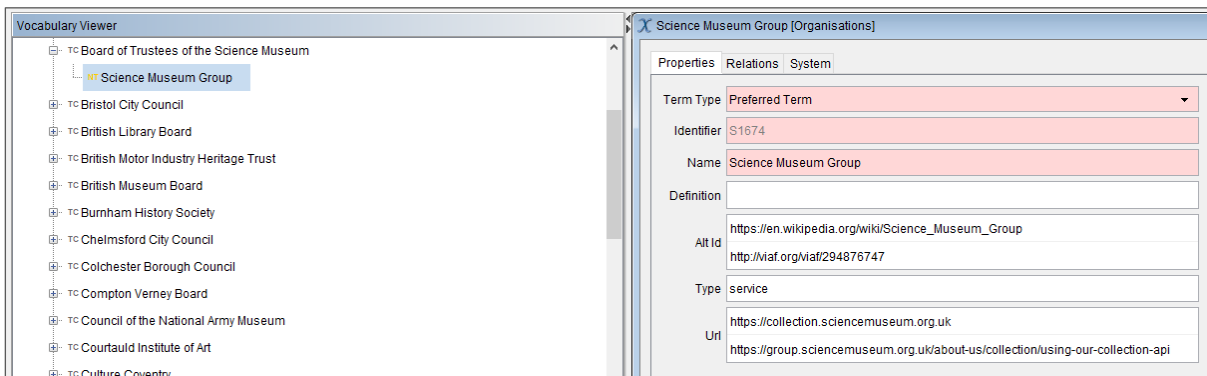


Figure 7 Lexaurus record for 'Science Museum Group' in the test thesaurus of institution names. K-Int.

In this example, the record for 'Science Museum Group' (SMG) sits below that of its governing body, the 'Board of Trustees of the Science Museum'. (With many former local authority services now run by independent trusts, changes in governing bodies can also be captured within the thesaurus structure, with explanatory notes where appropriate.) In a fuller thesaurus, the SMG hierarchy could be extended down to include the various branches of the group (eg 'National Railway Museum', etc).

The thesaurus created for the project covered around 80 institutions identified as a test sample in the scoping phase. The following reference sources were used to compile the test thesaurus:

- Institutions' own websites
- ACE list of Accredited museums ⁴⁷
- ACE list of Designated collections ⁴⁸
- Culture24's venues database (see below)
- Cornucopia database (see above)
- Museums Association 'Find a museum' directory ⁴⁹

⁴⁷ <https://www.artscouncil.org.uk/accreditation-scheme/about-accreditation#section-4>

⁴⁸ <https://www.artscouncil.org.uk/publication/designated-outstanding-collections>

⁴⁹ <https://www.museumsassociation.org/find-a-museum> (members/subscribers only)

- A few URLs from the Virtual International Authority File to demonstrate potential.⁵⁰
- A few Wikipedia URLs to demonstrate potential.


The data was exported from the Lexaurus tool and then ingested into the prototype.⁵¹

Culture24

Another source of information about collections-holding institutions are the datasets compiled and maintained by Culture24.⁵² Like Cornucopia, one of these datasets includes summary descriptions of collections held, sometimes quite detailed, but sometimes only one or two keywords. The collections-level information was not included in the test, but two other Culture24 datasets were: venues and events. The venues data includes details of institutions' address, website, contact details and accessibility, as shown in this screenshot from Culture24's current site, Museum Crush.⁵³

The Wilson Cheltenham Art Gallery and Museum

Cheltenham, Gloucestershire Gallery, Museum



Clarence Street
Cheltenham
Gloucestershire
GL50 3JT
England

Email: artgallery@cheltenhamtrust.org.uk
Phone: 01242 237431
Web Site: <http://www.cheltenham.artgallery.museum>

The Wilson, Cheltenham's newly extended Art Gallery & Museum reopened its doors to the public on 5 October 2013 with a new building housing expansive fine art and touring exhibition galleries for the first time.

Renewed gallery spaces allow visitors to explore highlights from the Museum's collections - including a new gallery space dedicated to the internationally renowned Arts & Crafts collection, open archives showing tales of local heroes, including the great Edward Wilson (one of Scott's key men on his 1912 expedition to Antarctica) and temporary exhibition spaces filled with varied programming including fun shows for families.

Further information

Opening Hours:

The Wilson is open daily, 9.30am - 5.15pm.
We are closed on 25 & 26 December, 1 January and Easter Sunday.

Figure 8 Venue record as displayed on Museum Crush. Culture24.

The events dataset contains details of events happening at the venues. The whole venues and events datasets were ingested into the prototype tool,⁵⁴ adding data that would help

⁵⁰ <https://viaf.org>

⁵¹ XML import; controlled vocabulary in SKOS format.

⁵² For more background information, see **appendix B** of the scoping report.

⁵³ https://museumcrush.org/todo/?item_id=sw000021

⁵⁴ Via Culture24's REST API; custom data elements in JSON format.

geolocate other records and opening the possibility for an eventual user interface to combine information about collections with details of the institutions holding them and forthcoming events that might be of interest.

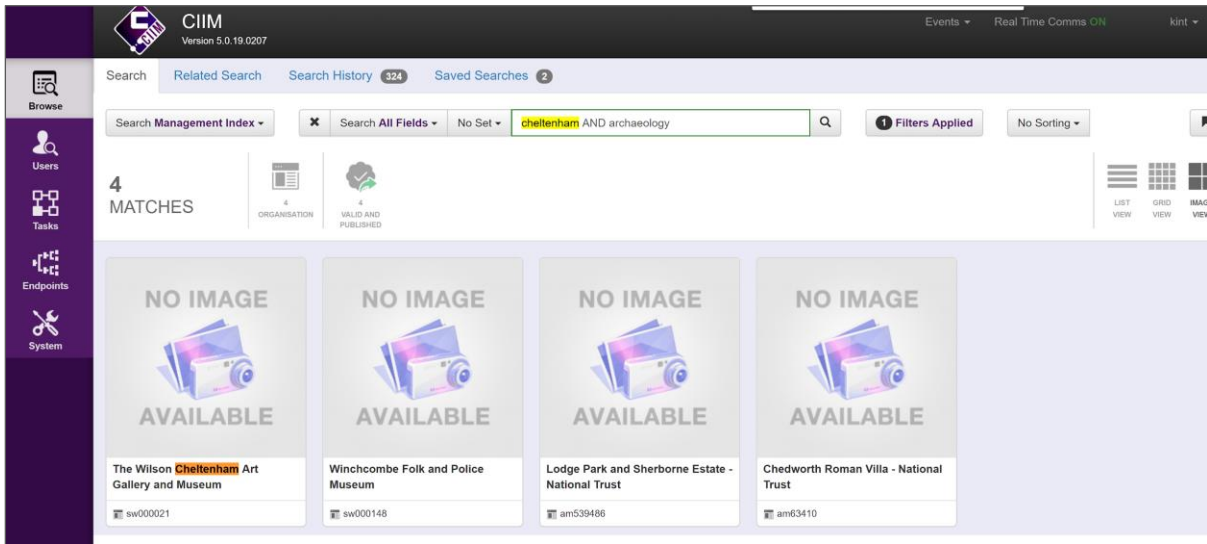


Figure 9 A search for ‘Cheltenham’ and ‘archaeology’ within the ingested Culture24 venues data. K-Int.

4.1.2 Data about item-level catalogues and digital assets

Sample data for the more detailed levels of the hierarchy (catalogue records for individual items and, where available, thumbnail images and metadata about digital assets) was brought into the prototype tool from the following sources identified in the scoping report:

Source	Ingest mechanism
Art UK (cross-institution images and metadata)	REST API Images and metadata in a custom format
Science Museum Group (images)	IIIF Image API (REST) ⁵⁵ Images and metadata in JSON format
Nottingham City Museums (images and metadata)	OAI-PMH harvest from Culture Grid API XML data elements conforming to the People’s Network Discovery Service profile of the Dublin Core metadata element set ⁵⁶
National Gallery beta (no images - as not yet live) Library and Object records	Elasticsearch search API ⁵⁷ Custom data elements in JSON formats (elements have been mapped to standard formats including LIDO and CIDOC-CRM)

⁵⁵ <https://iiif.io/api/image/2.1/>

⁵⁶ <https://www.webarchive.org.uk/wayback/en/archive/20150828225323/http://www.ukoln.ac.uk/metadata/pns/pndsdcap/>

⁵⁷ <https://www.elastic.co/guide/en/elasticsearch/reference/current/search-search.html>

The rationale behind these choices was to demonstrate an approach to data import that was not prescriptive, but flexible; working with the available formats and protocols rather than mandating specific technical mechanisms. The content was ingested and published into the **Elasticsearch** indexing tool to illustrate how the content could be used by an end-user interface. All fields within the data were indexed with common fields for common access points. For example, when searching across institutions and objects, the basic metadata model is not the same for the two entities. However, for the purposes of combined searching, it is reasonable to map the ‘title’ search access point to institution ‘name’, and object ‘name/title’. Where no equivalence is possible, eg institution ‘opening hours’, there is no unified search across all data types.

It should be noted that, as listed below, a number of other potential sources suggested in the scoping report did not, in the end, form part of the test. Rather than repeating the process of ingesting data from several sources using essentially the same technical mechanism, the decision was taken to prioritise the data enhancement side of the prototype, in order to demonstrate the AI tools of particular interest to DCMS. Consequently, the following sources were mentioned in the scoping report but not ingested during the test. In each case, an API would have been used to ingest data into the prototype in the same ways as the data from the four test sources listed above.

- British Library
- Cambridge University Library (via Archives Hub)
- Historic England Archive (via Culture Grid)
- Natural History Museum
- Portable Antiquities Scheme
- Royal Albert Memorial Museum (via the South West Collections Explorer) ⁵⁸
- The National Archives

The scoping report expressed the hope that it might be possible to demonstrate the use of ‘crawling’ embedded **microdata** (such as **JSON-LD**) in webpages as a way of ingesting data into the prototype. However, our desk research into the online availability of digitised material found that none of the sample institutions was currently embedding such microdata in its online collection in enough detail to be useful to an aggregator. ⁵⁹ One reason for this is likely to be that microdata standards in general are still embryonic. ⁶⁰ It is possible that the cultural heritage sector might in future develop the standards needed for this approach to be widely adopted, but it is unlikely ever to be the primary means for an institution to make its digitised collection discoverable online. It is more likely that web pages might be annotated with embedded microdata generated as a by-product of other data-sharing processes.

In theory, a further ingest mechanism, via **SPARQL** endpoints, could have been demonstrated within the test. However, although the British Museum offers one, SPARQL endpoints are not yet common enough within the sector to be viable for most cultural heritage institutions and were therefore not considered a priority for demonstration within the test.

⁵⁸ <https://swcollectionsexplorer.org.uk/website-api/>

⁵⁹ See **table 2** submitted as a separate spreadsheet with the scoping report.

⁶⁰ [https://en.wikipedia.org/wiki/Microdata_\(HTML\)](https://en.wikipedia.org/wiki/Microdata_(HTML))

4.2 AI processing pipeline

A processing pipeline was created to allow enhancement services to be applied to data from multiple sources, based on a path to content within the data sources. An original copy of the data is stored on ingest and this can be enhanced by the repeated application of AI services. These AI services can also be re-applied to the required fields when the content in these specific fields changes.

The pipeline allows multiple services to be applied to the same data and the location in which the enhanced data is added to the source record can also be specified, to allow the fact that this is automatically enhanced data (and should therefore not be treated as necessarily 'authoritative').

Figure 10 below shows an overview of the application of multiple external AI services to the ingested Cornucopia record about Cheltenham's agricultural collection.

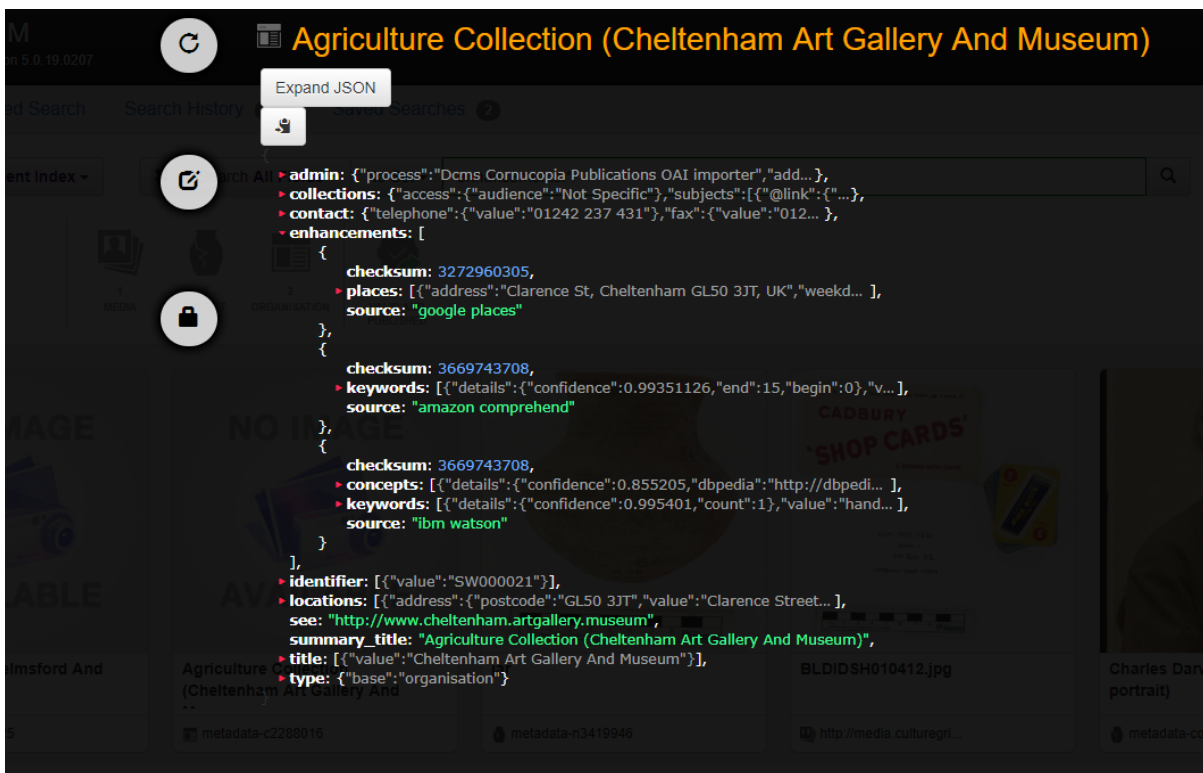


Figure 10 Multiple AI services applied to an ingested Cornucopia record. K-Int.

4.2.1 Recognising entities within the data

The screenshots below show how the IBM Watson and Amazon Comprehend services found ‘entities’ within the free text of the ingested Cornucopia record about Cheltenham’s archaeology collection (see **figure 5**). ‘Entities’ here means specific people, places or organizations, etc. The analysed text field is reproduced below for comparison.

Description	
VALUE	<p>This collection covers the area of north Gloucestershire and spans the period from the palaeolithic to the industrial revolution. Particular strengths are finds and excavation archives from the neolithic long barrows of Belas Knap, Notgrove and West Tump. Significant iron age collections include material from Salmonsbury Camp (Bourton on the Water), Kings Beeches (Cleeve Hill), Oxenton Hill and Leckhampton Hill. Romano British settlement sites are represented by finds from Bourton on the Water, Andoversford, Syreford, Vineyards Farm (Charlton Kings), Haymes (Southam) and Wycomb and the villa sites by Compton Grove and Whittington Court. Anglo Saxon material from the cemetery at Bishop’s Cleeve has recently been acquired. Medieval finds are so far the earliest artefacts attesting to permanent settlement on the site of Cheltenham itself. The collection also includes some Egyptian, Classical Greek and Roman artefacts.</p> <p>(less)</p>

Figure 11 Text from ingested Cornucopia record about Cheltenham’s archaeology collection. K-Int.

Figure 12 below shows the ‘entities’ found in this text by IBM Watson. Some of the very specific place names have not been recognised: the neolithic long barrow ‘Belas Knap’ has been interpreted as the name of a person, while the iron age site ‘Kings Beeches’ is thought to be a facility.

```

- entities: [
  {
    > details: {"confidence":0.924869,"count":2,"name":"Bourton-on-the-W..."},
    value: "Bourton"
  },
  {
    > details: {"confidence":0.778982,"count":1,"name":"Cleeve Hill","dbpe..."},
    value: "Cleeve Hill"
  },
  {
    > details: {"confidence":0.593164,"count":1,"type":"Facility"},
    value: "Kings Beeches"
  },
  {
    > details: {"confidence":0.587021,"count":1,"type":"Person"},
    value: "Cleeve"
  },
  {
    > details: {"confidence":0.585324,"count":1,"type":"GeographicFeature"},
    value: "Leckhampton Hill"
  },
  {
    > details: {"confidence":0.565849,"count":1,"type":"Person"},
    value: "Belas Knap"
  },
  {
    > details: {"confidence":0.534519,"count":1,"type":"Organization"},
    value: "Charlton Kings"
  }
]

```

Figure 12 Sample of entities found by IBM Watson in the Cornucopia record about Cheltenham’s archaeology collection. K-Int.

```

close
- entities: [
  {
    > details: {"confidence":0.8859572,"end":56,"type":"LOCATION","begin... },
    value: "north Gloucestershire"
  },
  {
    > details: {"confidence":0.74163264,"end":230,"type":"LOCATION","be... },
    value: "Belas Knap"
  },
  {
    > details: {"confidence":0.91459215,"end":240,"type":"LOCATION","be... },
    value: "Notgrove"
  },
  {
    > details: {"confidence":0.99497354,"end":254,"type":"LOCATION","be... },
    value: "West Tump"
  },
  {
    > details: {"confidence":0.79273975,"end":327,"type":"LOCATION","be... },
    value: "Salmonsbury Camp"
  },
  {
    > details: {"confidence":0.55591327,"end":336,"type":"LOCATION","be... },
    value: "Bourton"
  },
  {
    > details: {"confidence":0.8306473,"end":365,"type":"LOCATION","begi... },
    value: "Kings Beeches"
  },
  {
    > details: {"confidence":0.98383784,"end":378,"type":"LOCATION","be... },
    value: "Cleeve Hill"
  },
  {
    > details: {"confidence":0.9796833,"end":393,"type":"LOCATION","begi... },
    value: "Oxenton Hill"
  },
  {
    > details: {"confidence":0.9942878,"end":414,"type":"LOCATION","begi... },
    value: "Leckhampton Hill"
  },
]

```

Figure 13 Sample of entities found by Amazon Comprehend in the Cornucopia record about Cheltenham’s archaeology collection. K-Int.

In figure 13 above, Amazon Comprehend fares better in recognising most of these entities as locations.

4.2.2 Enhancing information about entities

In the example shown in figure 14 below, Google Places has analysed the ingested Cornucopia record about Cheltenham’s archaeological collection, found the entity ‘Cheltenham Art Gallery and Museum’, and enhanced it with additional information from its own databases. Google Places has made the connection between the old name of the museum and its current name, ‘The Wilson’, and also added the address and information about opening hours. It is worth repeating that these enhancements are clearly identified as such in the prototype tool and not mixed up with the original source data, since they might be wrong.

```

    admin: {"process": "Dcms Cornucopia Publications OAI importer", "add..."},
    collections: {"access": {"audience": "Not Specific"}, "subjects": [{"@link": "..."},
    contact: {"telephone": {"value": "01242 237 431"}, "fax": {"value": "012..."},
    enhancements: [
      {
        checksum: 3272960305,
        places: [
          {
            address: "Clarence St, Cheltenham GL50 3JT, UK",
            bounds: {"coordinates": {"system": "latitude/longitude", "latitude": 51.9... },
            details: {"hours": {"periods": [{"close": {"time": [0,16], "day": "SUNDAY..."},
            name: "The Wilson",
            shutdown: false,
            weekdays: ["Monday: Closed", "Tuesday: 9:30 AM - 5:15 PM", "Wednesda... ]
          }
        ],
        source: "google places"
    ]
  ]

```

Figure 14 Google Places service applied to an ingested Cornucopia record. K-Int.

4.2.3 Recognising keywords within the data

As well as recognising specific entities, AI services such as IBM Watson can also spot 'keywords', both individual words and significant phrases within text. **Figure 15** below shows a sample of the keywords it found in the summary of Cheltenham's archaeology collection. Unsurprisingly, this generic service has gone for 'British settlement sites' rather than 'Romano British settlement sites', not helped by the fact that the source record has not used the more usual, hyphenated 'Romano-British'.

```

- keywords: [
  {
    details: {"confidence": 0.909339, "count": 1},
    value: "Significant iron age collections"
  },
  {
    details: {"confidence": 0.69966, "count": 1},
    value: "Medieval finds"
  },
  {
    details: {"confidence": 0.653538, "count": 1},
    value: "Particular strengths"
  },
  {
    details: {"confidence": 0.644749, "count": 1},
    value: "British settlement sites"
  },
  {
    details: {"confidence": 0.599414, "count": 1},
    value: "permanent settlement"
  },
  {
    details: {"confidence": 0.581191, "count": 2},
    value: "collection"
  },
  {
    details: {"confidence": 0.580137, "count": 1},
    value: "Classical Greek"
  },
  {
    details: {"confidence": 0.57917, "count": 1},
    value: "Saxon material"
  }
]

```

Figure 15 Sample keywords found by IBM Watson in the Cornucopia record about Cheltenham's archaeology collection. K-Int.

By way of comparison, **figure 16** shows the keywords picked out from the same piece of text by the Amazon Comprehend service. This illustrates the value of applying several AI services to the same data, since there are some differences in the keywords identified by IBM Watson and Amazon Comprehend.



Figure 16 Sample keywords found by Amazon Comprehend in the Cheltenham archaeology collection record. K-Int.

4.2.4 Recognising concepts within the data

In **figure 17** below, IBM Watson has not only recognised words and phrases, but has enhanced the data by suggesting broader concepts and providing links to the relevant resources within DBpedia,⁶¹ a crowd-sourced database of structured content extracted from the information created in the various Wikimedia Foundation projects.⁶²

⁶¹ <https://wiki.dbpedia.org/about>

⁶² <https://wikimediafoundation.org/>

While ‘Gloucestershire’ and ‘Industrial Revolution’, for example, appear in the source record, the phrase ‘Villages in Gloucestershire’ does not. From the number of place names mentioned in the text that are indeed villages located in that county, the AI service has drawn an inference. Similarly, the phrase ‘Middle Ages’ is not in the source text, but ‘medieval finds’ is, and IBM Watson makes the connection.

```

    concepts: [
      {
        details: {"confidence":0.960593,"dbpedia":"http://dbpedia.org/resour..."},
        value: "Villages in Gloucestershire"
      },
      {
        details: {"confidence":0.711264,"dbpedia":"http://dbpedia.org/resour..."},
        value: "Europe"
      },
      {
        details: {
          confidence: 0.6352,
          dbpedia: "http://dbpedia.org/resource/Middle_Ages"
        },
        value: "Middle Ages"
      },
      {
        details: {"confidence":0.631844,"dbpedia":"http://dbpedia.org/resour..."},
        value: "Cheltenham"
      },
      {
        details: {
          confidence: 0.625852,
          dbpedia: "http://dbpedia.org/resource/Industrial_Revolution"
        },
        value: "Industrial Revolution"
      },
      {
        details: {"confidence":0.58677,"dbpedia":"http://dbpedia.org/resourc..."},
        value: "Gloucestershire"
      }
    ]
  
```

Figure 17 Sample concepts found by IBM Watson in the Cornucopia record about Cheltenham’s archaeology collection. K-Int.

4.2.5 Classification

IBM Watson can also be asked to classify content it finds against a five-level taxonomy.⁶³ However, this generic scheme is not well suited to the cultural heritage data sets used in the test, as **figure 18** illustrates below.

It shows text from a record about an Egyptian ceramic jar that was ingested into the prototype tool from Nottingham City Museums via Culture Grid.⁶⁴ Despite the many chronological dates sprinkled through the text, the AI service has decided that the sentence ‘The cemeteries at Harageh have a range of dates,’ means that this record should be classified under ‘society/dating’, as in ‘going out on a date’.

⁶³ <https://cloud.ibm.com/docs/services/natural-language-understanding?topic=natural-language-understanding-categories-hierarchy#categories-hierarchy>

⁶⁴ <http://www.culturegrid.org.uk/search/3419946.html>

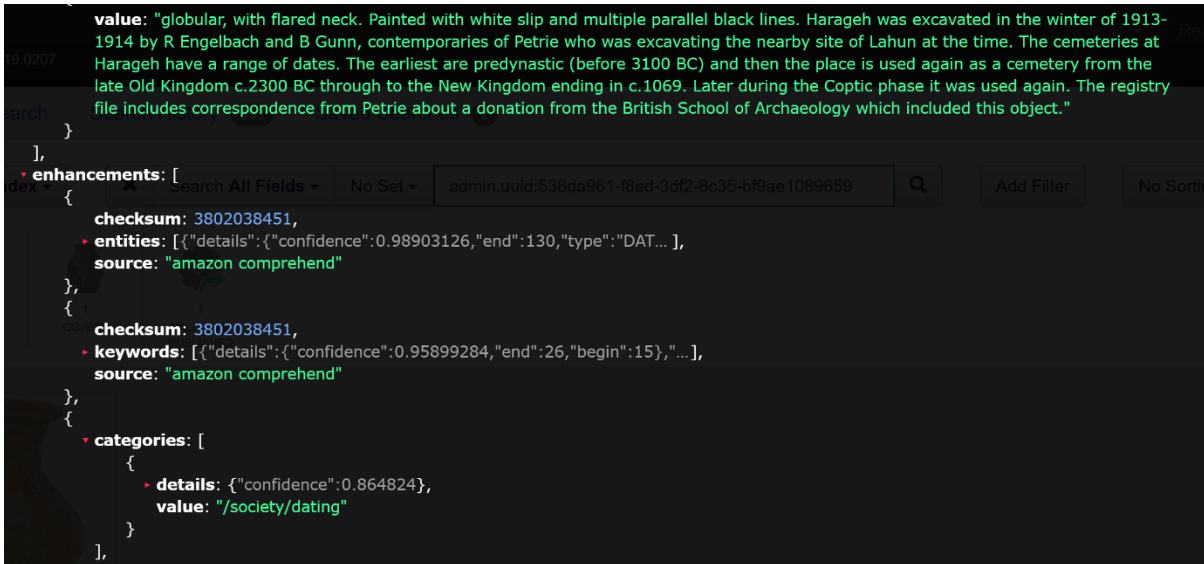


Figure 18 Misclassification of 'date' by IBM Watson. NB the record was also enhanced by Amazon Comprehend. K-Int.

However, standard cultural heritage terminology sources could be used to help train these generic AI services to understand the specific meanings of words likely in collections records. For example, the Art & Architecture Thesaurus has the sense of 'dating' we are after in the context of archaeological finds.⁶⁵



Figure 19 Example record from the Art & Architecture Thesaurus. Getty Research Institute.

4.2.6 Recognising images

As noted earlier, Art UK has already successfully collaborated with Oxford University's Visual Geometry Group to train image-recognition software to complement the work of its volunteer human 'taggers'. In the test for this project Amazon's Rekognition⁶⁶ service was applied to a selection of ingested images.

⁶⁵ <http://www.getty.edu/vow/AATFullDisplay?find=dating&logic=AND¬e=&page=1&subjectid=300054714>

⁶⁶ <https://aws.amazon.com/rekognition/>

Figure 20 shows a studio portrait of Charles Darwin ingested into the prototype from the Science Museum.⁶⁷

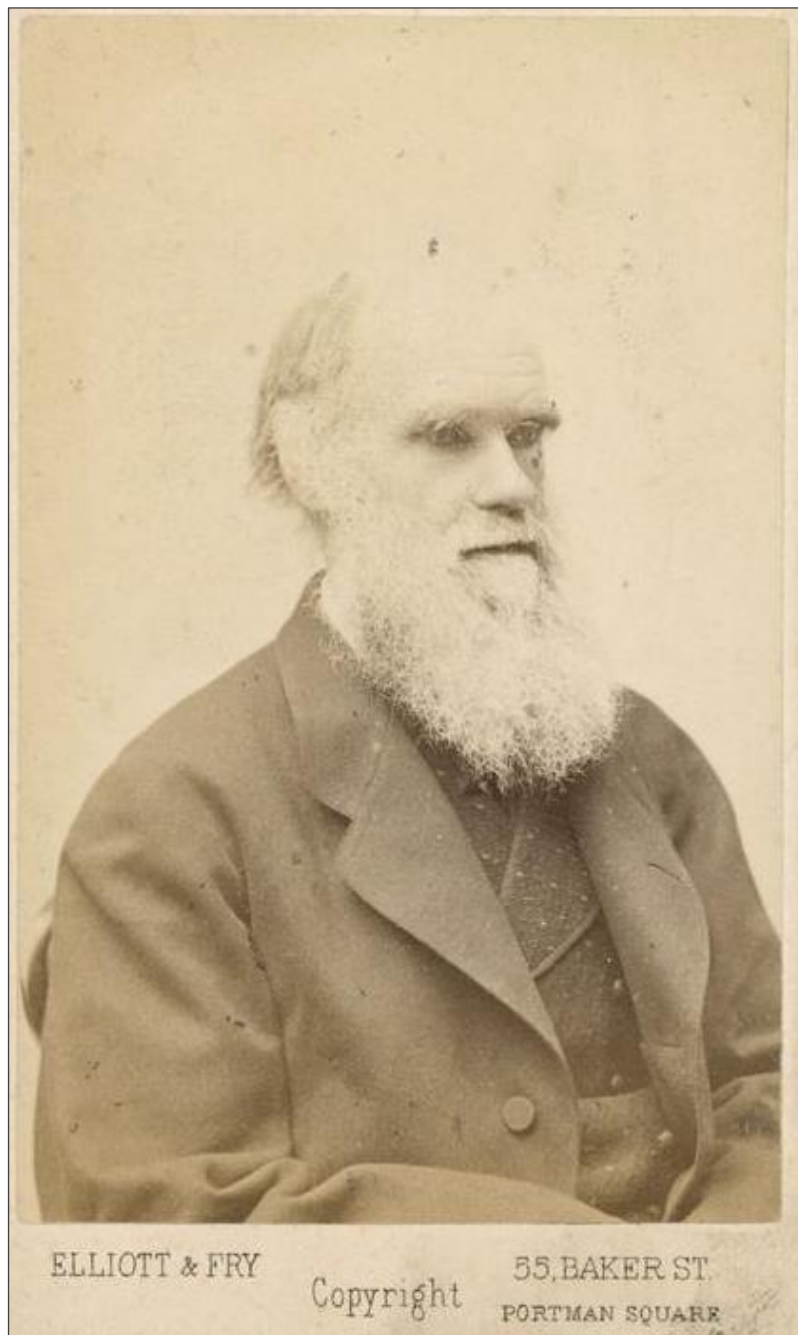


Figure 20 Carte de visite photograph of Charles Darwin. Science Museum.

⁶⁷ <https://collection.sciencemuseum.org.uk/objects/co8018692/charles-darwin-photograph-portrait>

Figure 21 shows how Amazon Rekognition has classified this image.



Figure 21 Amazon Rekognition's classification of the Darwin portrait. K-Int.

In **figure 22**, Amazon Rekognition has done a facial recognition analysis of the Darwin portrait. Interestingly, his beard has been missed, probably due to the image contrast.

```

emotions: [
  {
    confidence: 1.370303,
    value: "SAD"
  },
  {
    confidence: 95.52601,
    value: "CALM"
  },
  {
    confidence: 1.0966095,
    value: "SURPRISED"
  },
  {
    confidence: 0.20327064,
    value: "HAPPY"
  },
  {
    confidence: 0.70936453,
    value: "ANGRY"
  },
  {
    confidence: 0.6247756,
    value: "CONFUSED"
  },
  {
    confidence: 0.46965432,
    value: "DISGUSTED"
  }
],
eyeglasses: {
  confidence: 99.979385,
  value: false
},
eyes: {
  status: {"confidence":95.62993,"value":"open"}
},
eyewear: true,
features: [
  {
    coordinates: {"x":0.5567772,"y":0.33349112},
    value: "eyeLeft"
  },
  {
    coordinates: {"x":0.64928544,"y":0.33295032},
    value: "eyeRight"
  },
  {
    coordinates: {"x":0.5737971,"y":0.41330495},
    value: "mouthLeft"
  },
  {
    coordinates: {"x":0.6485412,"y":0.4125277},
    value: "mouthRight"
  }
]

```

Figure 22 Amazon Rekognition's facial recognition analysis of the Darwin portrait. K-Int.

Finally, **figure 23** below shows what Google’s Cloud Vision web entity plugin ⁶⁸ makes of the same image. Impressively, it has identified the subject as ‘Charles Darwin’ from the image alone, as well as extracting several other entities through optical character recognition of the text beneath the portrait, sometimes inaccurately (eg ‘The Fry Art Gallery’).

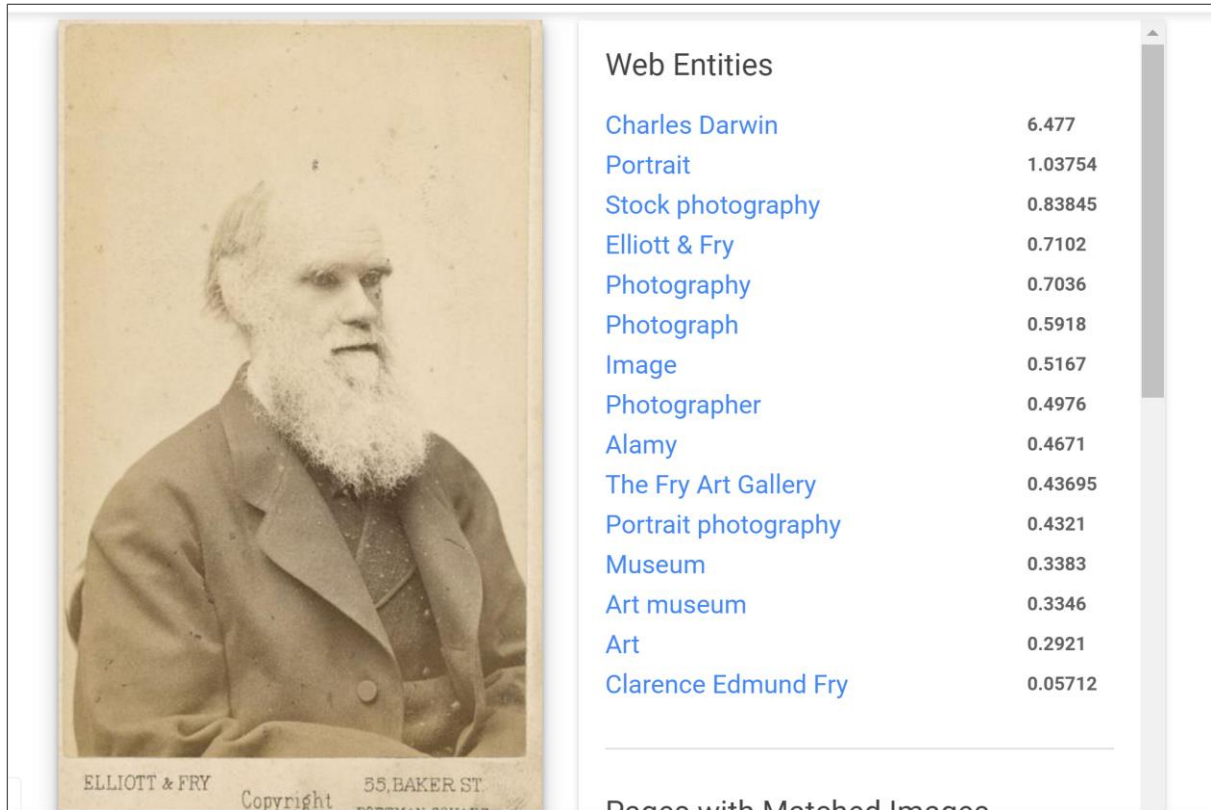


Figure 23 Google Cloud Vision analysis of entities in the Darwin portrait. K-Int.

⁶⁸ <https://cloud.google.com/vision/docs/detecting-web>

5. Use scenarios

This section considers the extent to which the principles demonstrated in the prototype could be of help to users in the five use-case scenarios suggested in the scoping phase. To repeat the caveats noted earlier, the prototype did not have a user interface and only contained item-level content from four sources. Nonetheless, it is possible to extrapolate what a fully-formed service built on the same principles might offer in the following scenarios:

- 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, 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.

5.1 Material relating to Charles Darwin

Collections-level data ingested into the prototype included Cornucopia records revealing that the following collections include material relating to 'Charles Darwin'. The first two were among the sample institutions investigated through desk research during the scoping phase.

- Norwich Castle Museum's Natural History Correspondence Collection ('The collection includes letters to Robert Fitch, a local antiquarian, from Charles Darwin.')
- Bath Royal Literary and Scientific Institution's Archives Collection ('The Rev Leonard Jenyns library includes four volumes of letters sent to him by 'men of science'. These hundreds of letters include many from Charles Darwin, Sir Joseph Hooker and Professor Henslow.')
- University of Cambridge, Herbarium of The Department of Plant Sciences (star objects include 'Darwin's collection of 1,200 plants from the voyage of the Beagle').
- University of Cambridge, Museum of Zoology (eg 'Beetles collected by Darwin whilst an undergraduate at Cambridge', etc).
- University of Cambridge, Fitzwilliam Museum, Manuscripts and Printed Books ('hand-written letters by Charles Darwin').
- Liverpool Museum's Biology Collection ('There is also material collected by many distinguished collectors including Captain James Cook and Charles Darwin.')
- Manchester Museum's Biology Collection (which also has 'Darwin material from the voyage of the Beagle').

Delving deeper than level 2 collection summaries, the level 3 and 4 data ingested from the Science Museum includes examples of specific items relating to Charles Darwin. **Figure 24** shows the ones that had images as well as catalogue record data.

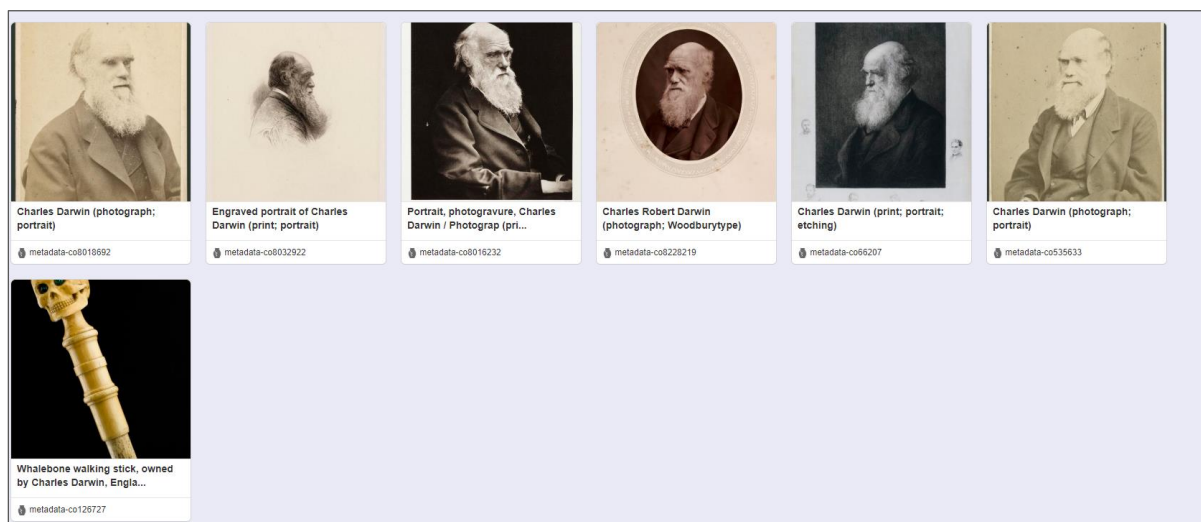


Figure 24 Search results in CIIM for Science Museum records (with images) including ‘Charles Darwin’ in metadata. K-Int.

Figure 25 shows the ingested Science Museum record for the whalebone walking stick owned by Charles Darwin.

Description	
DESCRIPTION	primary Whalebone walking stick with skull form pommel in ivory, once owned by Charles Darwin, probably English, 1839-1881
Metadata	
DATE MADE	1839-1881
IDENTIFIER	A4962
PLACE MADE	England
Title	
VALUE	primary Whalebone walking stick, owned by Charles Darwin, England, 1839-1881 (walking stick)
Type	
BASE	object

Figure 25 CIIM view of Science Museum record for Charles Darwin’s walking stick. K-Int.

Seven English Heritage photographs of Darwin’s home, Down House, are available to view via the ingested Culture Grid data. The Art UK records ingested also included a number of painted portraits of Charles Darwin and some other relevant artworks.⁶⁹

Conclusion

From the relevant data ingested, the curator would have access to item-level information (and many images) of artworks depicting Charles Darwin, and items belonging to him, from

⁶⁹ See https://artuk.org/discover/artworks/view_as/grid/search/keyword:charles-darwin/page/3

institutions as varied as the National Portrait Gallery, Science Museum and English Heritage. Collections summaries in the ingested Cornucopia data would point the curator in the direction of several other institutions with potentially relevant material. The collection-level information ingested could be supplemented, through desk research of the kind carried out in the scoping phase, to indicate whether the curator could seek out further item-level information online at the relevant institution’s website. If not, the institution-level data would at least provide contact details for an enquiry.

5.2 Images of Egyptian ceramics

In this scenario, the academic researcher was specifically looking for digitised images of ancient Egyptian ceramics licensed for non-commercial re-use, and also for information about the collectors responsible for the Egyptian ceramics held in cultural institutions.

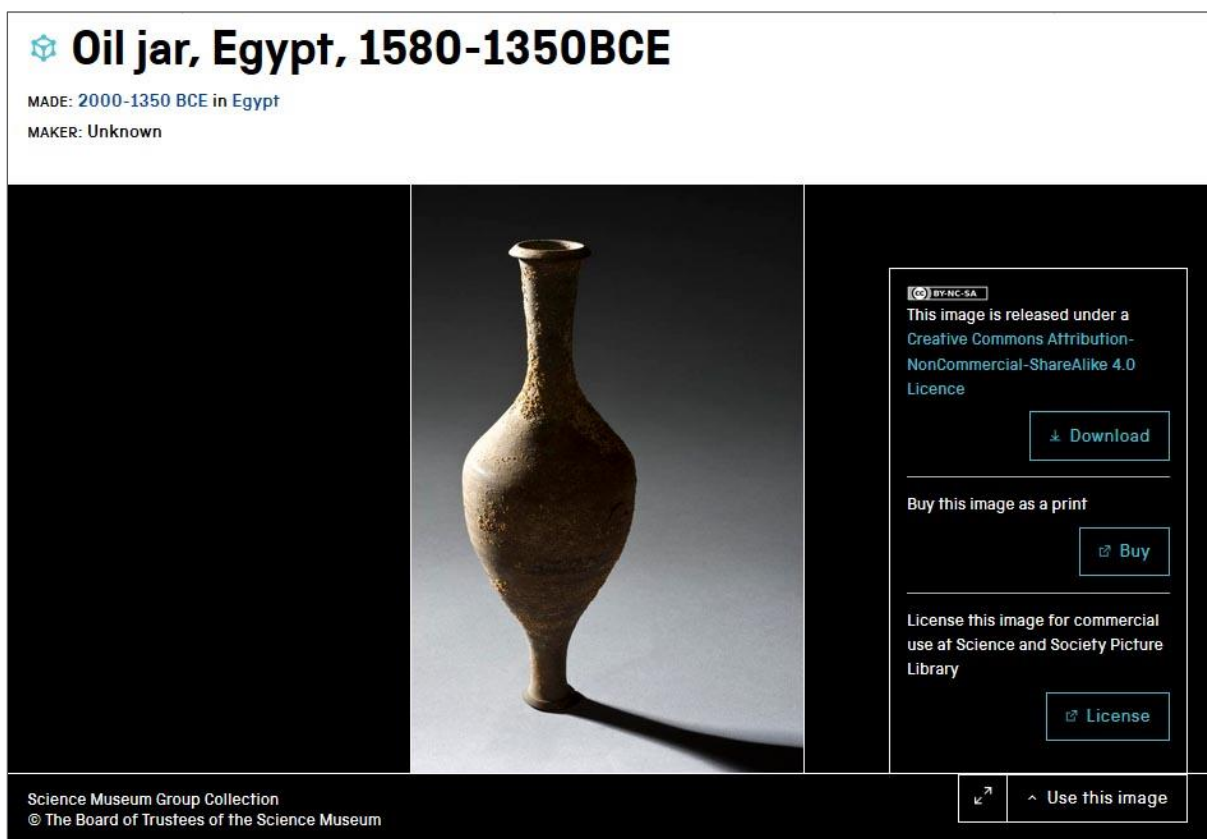


Figure 26 Screenshot from the Science Museum website of an image licensed for non-commercial re-use. Science Museum.

Among the image records ingested from the Science Museum are several ceramic items relating to ancient Egyptian medicine (especially from the later Greek and Roman periods). The published metadata gives no information about the provenance of these items, but the images are licensed for re-use under the Creative Commons BY-NC-SA (Attribution-NonCommercial-ShareAlike) licence, which lets others use and modify the image non-commercially, as long as they credit the Science Museum and license their new creations under the same terms.⁷⁰ Figure 26 above shows one of these.

⁷⁰ <https://creativecommons.org/licenses/>

Part of a record about an Egyptian ceramic jar in Nottingham City Museums has already been mentioned (see **figure 18**). There is an image associated with this record, as can be seen in **figure 27**.⁷¹ This jar is one of 157 images of 'Egyptian ceramics' ingested from Nottingham via Culture Grid. Unfortunately, the source data does not include any information about the copyright status of this image or whether it is licensed for re-use.

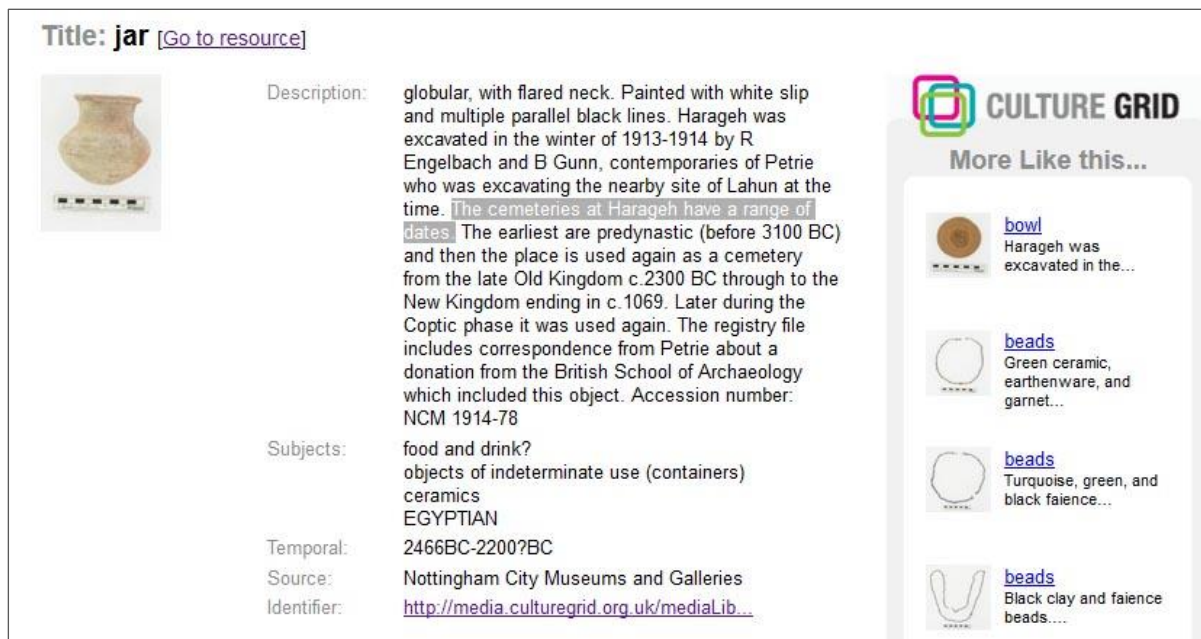


Figure 27 Screenshot of ingested Egyptian jar record as it appears on Culture Grid. Collections Trust.

The ingested Cornucopia data suggests that at least 247 collections have 'Egyptian' material. These summaries often contain the names of noted excavators and collectors. For example, if the researcher in this scenario were interested in how the Egyptologist Sir William Matthew Flinders Petrie distributed his finds around the country's museums, a quick search of the Cornucopia data would suggest no fewer than 85 collections to follow up.

Conclusion

Even from the limited data ingested, the researcher in this scenario would be able to find images of ancient Egyptian ceramics explicitly licensed for non-commercial re-use. Moreover, the associated data, even the collection-level summaries from Cornucopia, generally provide useful provenance information, including named sites and their excavators.

5.3 Anglo-Saxon material from Essex

Searching the ingested Cornucopia and Culture24 venue data reveal that three museums in Essex report having 'Saxon' material.

- Saffron Walden Museum
- Burnham-on-Crouch and District Museum
- Thurrock Museum

⁷¹ <http://www.culturegrid.org.uk/search/3419946.html>

Thurrock Museum does not have its own website, just a few pages on the council website. Burnham does have a website, but it does not include any images of Saxon material. Saffron Walden Museum’s website does not have a searchable database but does have pages featuring highlights of each collection, as shown in **figure 28**.⁷² Five of these images are of Anglo-Saxon objects, and a further two are of Viking objects from the same period (the coins being Viking copies of Anglo-Saxon originals). However, the accompanying information is brief and there is no indication of whether the images are licensed for re-use.

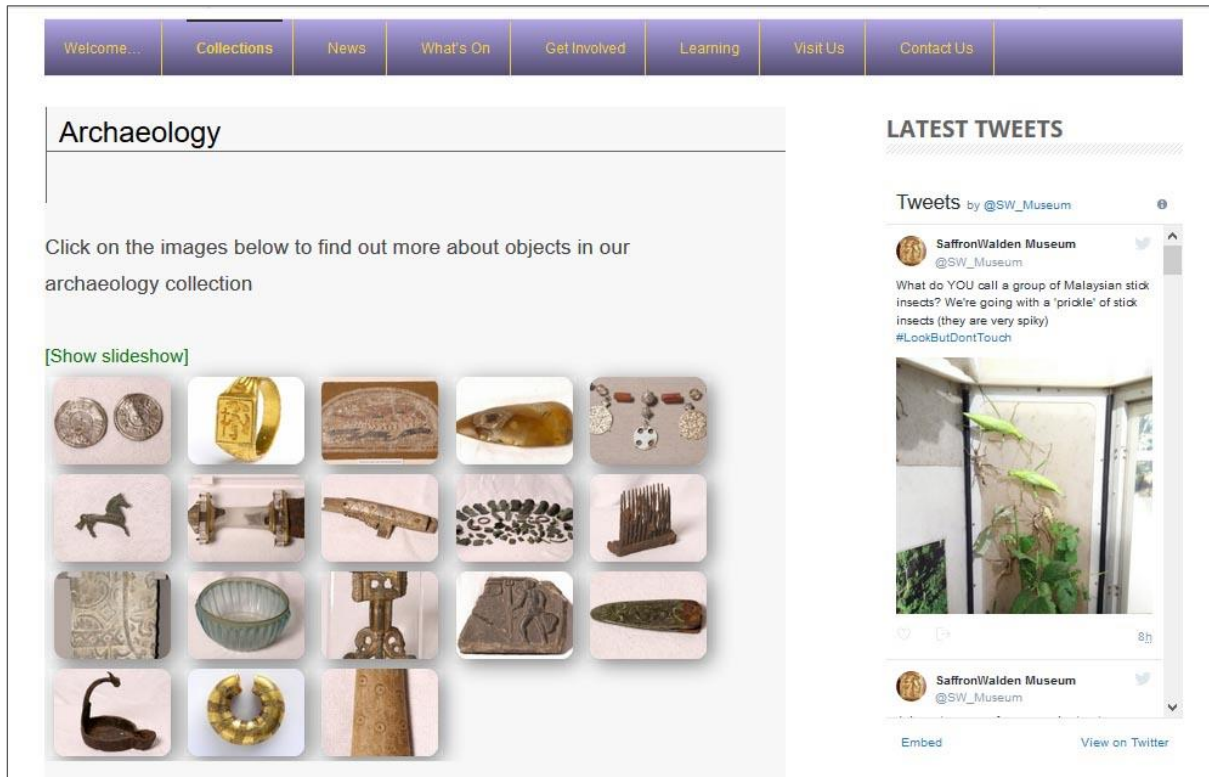


Figure 28 Highlights of Saffron Walden Museum’s archaeology collection. Saffron Walden Museum.

Conclusion

From the Cornucopia data, and a website link from Culture24’s data, the teacher could have been directed to the Saffron Walden Museum website. Following the navigation path ‘collections > archaeology’ the teacher would have found the page shown as **figure 28** with just one click but would then have to click on each image in turn to find out which were of Anglo-Saxon objects.

⁷² <https://saffronwaldenmuseum.swmuseumsoc.org.uk/discover/archaeology/>

5.4 Farming-related material

The collection-level data ingested from Cornucopia includes 160 records that mention the word ‘farming’, rising to 328 if the broader search term ‘agriculture’ is used.⁷³

At item level, the data ingested into the prototype for the test included material relating to ‘farming’ from three sources:

- Art UK
- National Gallery
- Nottingham City Museums (via Culture Grid)

Sample screenshots of content ingested from all three are shown below. **Figure 29** shows the results of a search for ingested Art UK records using the search term ‘farming’. It is easy to see how a work with the title ‘The Home Farm’ has been found through this search. The data ingested for ‘The Evening Meal’ includes the tags ‘farming and fishing’, ‘animals, farm’, ‘hay’, ‘sheep’ and ‘shepherd’.

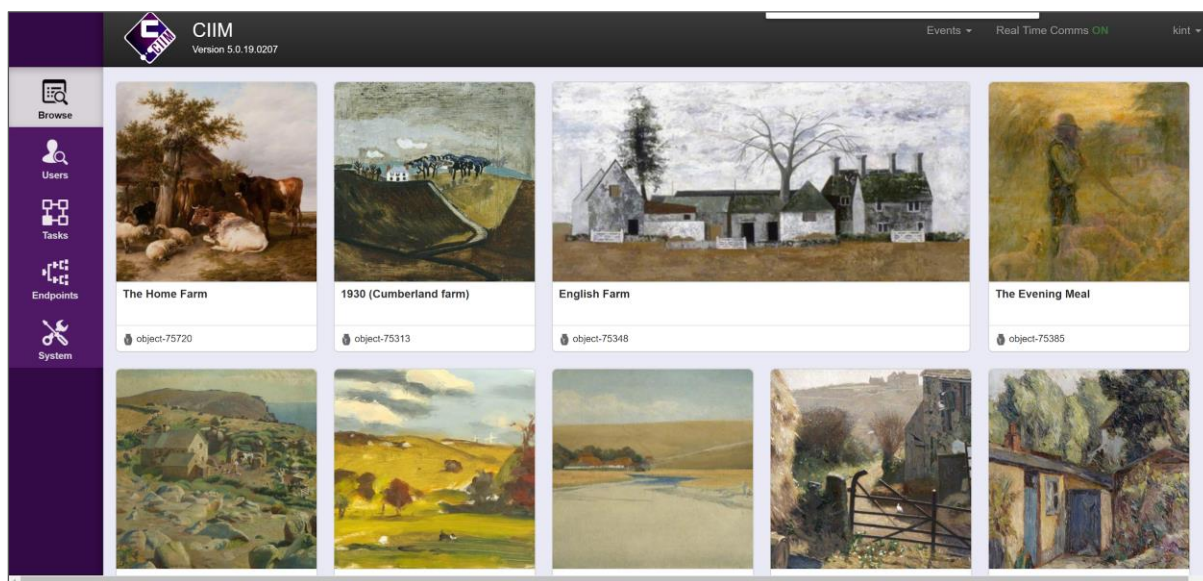


Figure 29 Farming-related records ingested from Art UK into the prototype. K-Int.

Many National Gallery paintings appear on the Art UK site,⁷⁴ and data on these was therefore ingested into the prototype through that route. However, to demonstrate a different technical mechanism, National Gallery object and library records were also ingested directly from the institution’s *Elasticsearch* API (minus images at the current beta stage). **Figure 30** shows one National Gallery record relating to ‘farming’ ingested that way.

⁷³ See, for example, the UK Archival Thesaurus for the relationship between the two terms in a structured thesaurus: <https://ukat.aim25.com/thesaurus/f6/mt635/2520/>

⁷⁴ <https://artuk.org/visit/venues/the-national-gallery-london-2030>

Description #1	
CREATED	2018-12-10 13:50:58.397
FORMATTED	The strong horizontal composition of the sketch suggests that it was made as a study for a panorama painting. The colours are muted, the forms not sharply distinguished from one another, and the paint applied thinly, almost like watercolour. The space seems compressed, with the farm buildings and ruins in the middle distance forming a single horizontal line. (less)
SOURCE	TMS
STATUS	Active
LONG TEXT	The strong horizontal composition of the sketch suggests that it was made as a study for a panorama painting. The colours are muted, the forms not sharply distinguished from one another, and the paint applied thinly, almost like watercolour. The space seems compressed, with the farm buildings and ruins in the middle distance forming a single horizontal line. (less)
Author	
SUMMARY TITLE	The National Gallery (London)
Name	
DISPLAY	The National Gallery (London)
SORT	National Gallery (London)

Figure 30 Record including 'farm' ingested from the National Gallery into the prototype. K-Int.

Nottingham Museums do not currently have an online collection as part of their website, but images and metadata for 9,685 items are available via the legacy aggregator Culture Grid.⁷⁵ The screenshot below shows some of these, ingested into the prototype, as the results of a search using the keyword 'farming'.

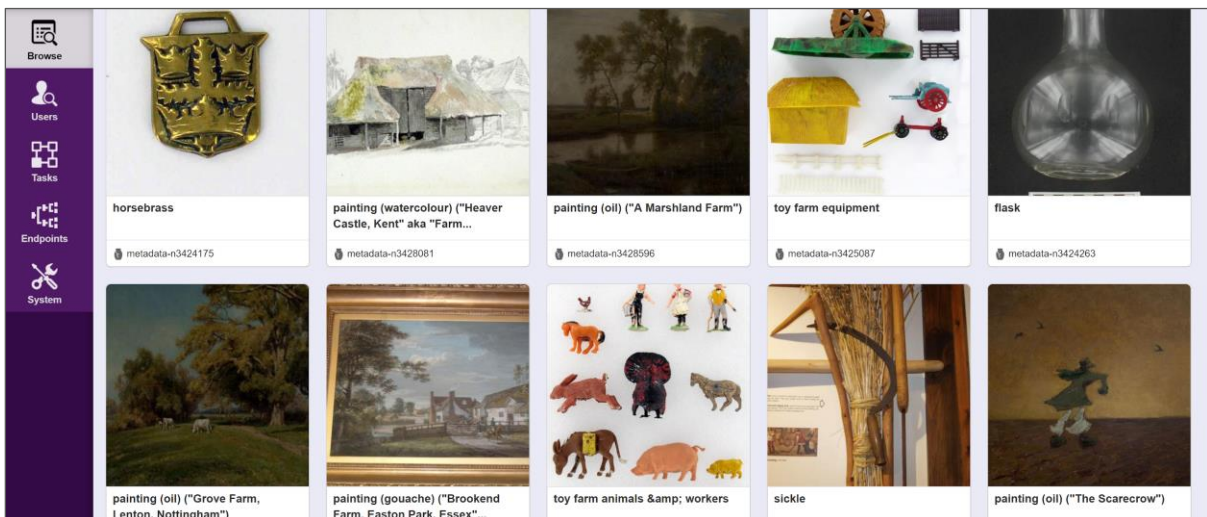


Figure 31 Farming-related records ingested from Nottingham Museums into the prototype. K-Int.

75

http://www.culturegrid.org.uk/search/#!/culturegrid:cqCollection=Nottingham_City_Museums_and_Galleries:cqHasThumbnail=true:query=*

Figure 32 shows part of the ingested record for the painting of ‘Grove Farm, Lenton, Nottinghamshire’.

Name	
VALUE	primary Nottingham and Notts.
SUMMARY TITLE	farming
Name	
VALUE	primary farming
SUMMARY TITLE	animals
Name	
VALUE	primary animals
SUMMARY TITLE	rivers and lakes
Name	
VALUE	primary rivers and lakes
Title	
VALUE	primary painting (oil) - "Grove Farm, Lenton, Nottingham"

Figure 32 Keywords from the ingested record for the painting ‘Grove Farm, Lenton, Nottinghamshire’. K-Int.

Conclusion

The challenge in this scenario is not finding content relating to ‘farming’, but in being to dig down further to find material relating to sub-themes and specific farming activities. Terminology sources commonly used by cultural heritage institutions when cataloguing their collections (such as the Art & Architecture Thesaurus⁷⁶ and Social History and Industrial Classification⁷⁷) could be used by the prototype to help index the ingested content.

5.5 Monk’s Hall Hoard

This scenario illustrates the need for the kind of data enhancement tools available as plug-ins to the prototype, and also some of the challenges they would face.

In fact, the enquirer who wanted information about the ‘Monks Hall Hoard’ is potentially in luck. There are digitised images online of many of the medieval coins discovered near Monks’ Hall in Eccles in 1864. Most were donated to the British Museum by the Duchy of Lancaster the following year. As **figure 33** below shows, the British Museum has put photographs of 325 of the coins online.⁷⁸

As well as presenting its online collection on its own website, the British Museum publishes it as open data via a ‘Sparql endpoint’.⁷⁹ Though it did not form part of the test, the British Museum’s open data could easily be ingested into the prototype via this route.

⁷⁶ <http://www.getty.edu/research/tools/vocabularies/aat/>

⁷⁷ <https://cidoc-dswg.org/wiki/SHIC>

⁷⁸ https://www.britishmuseum.org/research/collection_online/search.aspx?searchText=eccles+hoard

⁷⁹ Eg, the data for the first coin shown above looks like this:

<https://collection.britishmuseum.org/resource/?uri=http%3A%2F%2Fcollection.britishmuseum.org%2Fid%2Fobject%2FCMB11322>

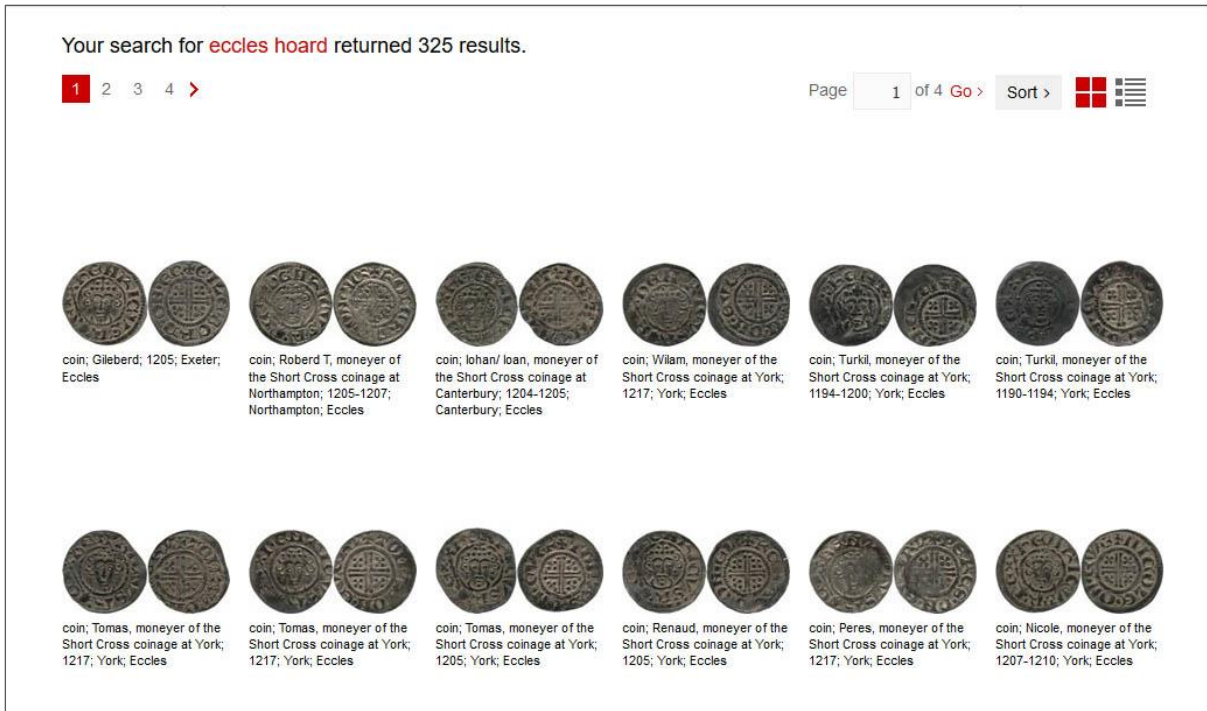


Figure 33 Search result for 'Eccles Hoard' on British Museum's Collection online. British Museum

There is, however, a problem: trying to find the 'Monks Hall Hoard' draws a blank, as the British Museum knows it as the 'Eccles Hoard' and nowhere mentions 'Monks Hall'. Moreover, the record shown in **figure 34** below incorrectly has the find location as the Scottish Border town of Eccles, instead of the Lancashire town of the same name.

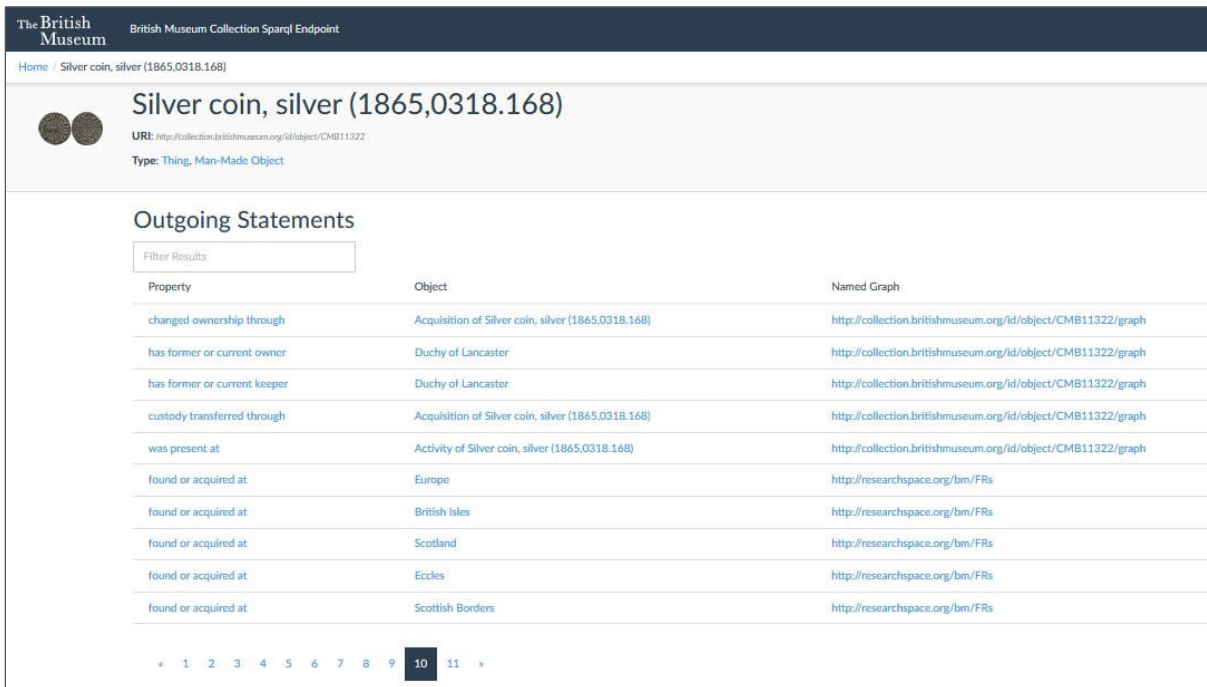


Figure 34 Part of the open data record for one of the Eccles Hoard coins on the British Museum Collection Sparql Endpoint

If the prototype also had access to Historic England’s dataset on Listed buildings, it would know that the address of Monk’s Hall is ‘Wellington Road, Eccles, M30 9RD’⁸⁰ and the kind of AI services demonstrated during the test might be able to join the dots.

⁸⁰ <https://historicengland.org.uk/listing/the-list/list-entry/1162896>

6. Conclusion

This final section of the report evaluates the prototype test against the success factors agreed in the scoping phase, and notes some lessons learned. It goes on to consider the potential benefits and risks if the proposed framework were developed into an aggregator supporting the cultural heritage sector and its audiences.

6.1 Evaluation against success factors

During the scoping phase, it was agreed that the following were useful success factors against which to evaluate the test of 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 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.

6.1.1 Be helpful to users

The test data demonstrated that, with a suitable user interface, the proposed framework would allow end users a single point of access to data at multiple levels from a wide range of institutions. Even collection-level summaries of the kind found in the Cornucopia data would be very helpful to researchers (see **section 5.2**) and occasionally to less-specialist users if links were given to relevant online content elsewhere (see **section 5.4**). Within the very limited test data ingested, it was possible to show how the framework would also allow users to find relevant item-level content where this was available.

In some cases, the users in the test scenarios would have found relevant images but been left unsure whether they were licensed for non-commercial re-use. This problem could be mitigated if institutions contributing to any future aggregator were required to supply metadata about the copyright status and re-use permissions of their content (without any particular licensing model being imposed by the aggregator).

Although only a relatively small number of records were ingested for the test, the need for advanced search strategies in any eventual user interface was clear. To take just one example that illustrates the point, typing 'Charles Darwin' into the search box on Art UK's website yields 54 results.⁸¹ But this includes a fair few portraits of people related to Charles Darwin, and some by his great-grandson, the artist Robin Darwin. The texts accompanying these images all include 'Charles Darwin'. As the quantity of ingested data increases, the

⁸¹ https://artuk.org/discover/artworks/view_as/grid/search/keyword:charles-darwin/page/2#artwork-undefined

number of false positive matches can be unhelpful. The aggregator Europeana, for example, returns 496 results for the search 'Charles Darwin', but the first few are letters by his grandson, Charles Galton Darwin.⁸²

6.1.2 Be helpful to contributing institutions

The test demonstrated that the prototype could ingest data from a range of institutions using various technical means, and without the institutions having to format their contributions to any kind of set template. That by itself lowers a potential barrier to data-sharing by cultural heritage institutions by making it easier for their collections management software providers to offer routine exporting to a future aggregator built along similar lines.

One problem that the consultants have experienced first-hand when working on aggregation projects is that, even if it is technically straightforward for institutions to share their data, the benefits of doing so are not seen as worth the effort. For any eventual aggregator to achieve widespread buy-in from the sector, it will need to add value and help institutions meet needs that are important to them.

As set out in **section 4**, much of the effort within the prototype test was spent exploring the potential of various image recognition and text mining services to enhance the data provided by institutions. In principle, enhancements resulting from the application of these and other AI services could be offered back to the providing institution (clearly labelled to distinguish them from the source data) to be re-ingested (or 'round-tripped'), so improving the quality of their own collections databases.

A further potential benefit is illustrated by the records ingested from Nottingham City Museums via the legacy aggregator Culture Grid. This service does not currently have an online collection as part of its own website. The desk research carried out during the scoping phase reveals that around a third of the institutions sampled, including a number of sizeable local authority services (such as Leicester, Derby, Colchester and Ipswich, Oxfordshire and Warwickshire) do not currently offer a searchable collections database as part of their own websites. An aggregator, with a suitable user interface, would offer a low-cost alternative to each of these services – and more - creating and maintaining such an online service themselves.

6.1.3 Bring together resources from museums, libraries and archives

At the level of collection summaries, the ingested Cornucopia data describes not only museum collections, but also many library and archive collections too. The number of collection-level descriptions of archive collections could have been increased by many thousands if records had been ingested from the aggregator Archives Hub.

At item level, within the test records were ingested from both the National Gallery's main collections database and its library database, as noted in **section 5.4**. The same data feed could also have brought in the same institution's archival database. Similarly, the feed used

⁸² <https://www.europeana.eu/portal/en/search?q=charles+darwin>

to harvest images from the Science Museum could also have been used for its archives data.

Although museums, libraries and archives follow different cataloguing standards (often even within the same institution), the flexible nature of the prototype means there is no technical reason why data from all three could not be harvested by an aggregator built using the same architecture, allowing searches to be made across the different collection types. This can be seen in action at websites such as those of the Royal Armouries⁸³ or Manx National Heritage,⁸⁴ both of which use the same middleware as the prototype to bring together disparate data types so that the user can search across the institution's entire holdings.

6.1.4 Show the potential and pitfalls of agreed emerging technologies

As described in **section 4.2** the various generic AI services applied to the test data had mixed success. Specific entities such as places were often recognised, and sometimes enhanced with additional information such as address details and opening hours. Similarly, useful keywords, and even concepts not actually mentioned in the source data, were successfully analysed and, in some cases, linked with authority references such as DBpedia (see **figure 17**).

However, there were enough mis-identifications and mis-classifications to illustrate the important principle that such AI enhancements should be clearly identified as auto-generated and kept apart from the original source data. There were some curious misses and false positives (eg failing to identify 'Suffolk' as a location, while deciding that 'Egyptology' was one).

Understandably, very specific place names such as archaeological sites caused particular problems to the generic services applied, with Google Places often falling back on business addresses/points of interest when it did not specifically recognise a place. There were also real howlers such as the classification of a record about an ancient Egyptian pot as something to do with 'society/dating', just because the word 'date' appeared (**figure 18**).

Such problems, however, reflect the need to train such AI services with lots of data relevant to cultural heritage collections, as suggested in **figure 19**, rather than any inherent shortcomings of the technologies themselves.

6.1.5 Show how the framework would be scalable

Technical scalability was designed into the architecture of the proposed framework, and built into the prototype, as described in **sections 2 and 3**. In the test, the successful use of different ingest mechanisms and a range of data formats demonstrated that the framework is widely applicable and does not take the 'one size fits all' approach of previous aggregators. In particular, the hierarchy of information means that an institution could initially be represented with just collection-level records that would still be useful, adding item-level records and digitised assets as and when they were ready, in whatever form and through whatever means they were able to provide.

⁸³ <https://collections.royalarmouries.org/#/objects>

⁸⁴ <https://www.imuseum.im/>

6.2 Benefits to the cultural heritage sector

The discussion above considers the extent to which the proposed framework might be helpful to individuals and institutions. Here, some potential sector-wide benefits are considered, based on conversations with the people acknowledged and the published literature cited.

6.2.1 Enabling content curation to reach new audiences

The ambition expressed in the 2016 Culture White Paper, for 'users to enjoy a seamless experience online and have the chance to access particular collections in depth as well as search across all collections,' was noted at the beginning of this report. The subsequent *Culture is Digital* report elaborated:

*'The ambition is with good reason. The richness and variety of our national museum collections are the envy of the world, captivating to audiences and scholars in every corner of the globe.'*⁸⁵

*'... Online curation can unlock access to cultural experiences and audience reach otherwise unimaginable and using social media to showcase digitised material can raise the profile of items, exhibitions and collections.'*⁸⁶

It is important to stress that most of the public benefits likely to flow from the data aggregation proposed in this report would be indirect rather than direct. The aggregator would not be a destination site for the wider public; rather it would be the tool behind limitless end-use scenarios that presented curated content to specific audiences. *Culture is Digital* noted

*'Differences in data standards and openness mean that it is difficult to curate across collections and create new online exhibitions or content for audiences and limiting the educational value of the digitised asset. Unless images are tagged in a certain way, content aggregators will not be able to gather the image when searched. This has implications for modern audiences who expect digital content to be easy to navigate and to be open for them to enjoy, contribute, participate and share.'*⁸⁷

As demonstrated during the test, the proposed aggregator would be able to bring together the raw material for curated content no matter what data standards had been used to create it. The aggregator would also use various tools to enhance the source data, mitigating its tagging inconsistencies to improve discoverability.

Moreover, the aggregator could streamline the current, labour-intensive workflows for publishing curated content, which are often the digital equivalent of hand-crafting illuminated manuscripts. For example, *Culture is Digital* rightly counts Google Arts and Culture among the success stories of digitised collections.⁸⁸ But until recently the only way an institution could contribute content to it was by uploading a spreadsheet or entering information item by

⁸⁵ DCMS, *Culture is Digital*,

https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/687519/TT_v4.pdf, p45.

⁸⁶ Op cit, p47.

⁸⁷ Op cit, p50.

⁸⁸ Op cit, p46.

item. Over a seven-year period the well-resourced Metropolitan Museum of Art only managed to contribute 757 artworks to the platform this way. Then in 2018 the Met launched a new **API** and collaborated with Google Arts and Culture to increase that number to 205,000 artworks.⁸⁹ The national aggregator proposed in this report would have the critical mass to work with Google Arts and Culture in a similar way, providing a less laborious route for any content provider who wanted to scale up their presence on this platform.

6.2.2 Supporting dynamic collections management

One of the priorities identified by the 2017 *Mendoza Review* of museums in England was ‘dynamic collection curation and management’:

‘Dynamic collections curation and management are the fundamental point of museums – to protect and take care of the collections they hold, and to make them accessible to the public, not just physically, but meaningfully as well. This is not without its challenges ... [such as] less available curatorial time and expertise, and the ongoing need for a sensible approach to both growing and rationalising collections. There are good examples of where sharing skills and infrastructure can help to overcome these issues; this is a particular area where a strategic framework for how the national museums’ work with the rest of the sector will benefit museums across the country.’⁹⁰

Addressing this priority would be a lot easier if those working collaboratively across the sector could routinely search across the collections data that is currently siloed within individual museums. For example, a ‘sensible approach to both growing and rationalising collections’ might allow curators to find out what else was in the country’s 1,700 other museums before deciding to acquire or dispose of an item.

The proposed aggregator could also provide a long-term home for the results of important, but short-lived, projects such as regional or subject-specific collections reviews. As well as national-level work to describe museum collections (such as Cornucopia), the past few decades have seen many such projects conducted at regional level (eg in the South West) or by Subject Specialist Networks. An example of the latter is the collections-mapping project carried out by the Islamic Art and Material Culture SSN in 2013-14, with ACE funding.⁹¹ The outputs included a report, downloadable as a PDF, but the promised online database, which might well link the reported insights to specific items, does not seem to have been delivered. A retrospective ‘review of collections reviews’ that aimed to rescue and repurpose valuable data currently languishing in spreadsheets across the country would be very worthwhile and would enhance the second (and, in many cases, the third) level of the proposed framework.

6.2.3 Strategic partnership with higher-education sector

It is more than ten years since the Research Information Network concluded: ‘What researchers need above all is online access to the records in museum and collection

⁸⁹ <https://www.metmuseum.org/blogs/now-at-the-met/2018/met-collection-api>

⁹⁰ DCMS, *The Mendoza Review: an independent review of museums in England*, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/673935/The_Mendoza_Review_an_independent_review_of_museums_in_England.pdf, p10.

⁹¹ <http://krc.orient.ox.ac.uk/iamcssn/index.php/en/map>

databases to be provided as quickly as possible, whatever the perceived imperfections or gaps in the records.’⁹² The Natural History Museum’s Digital Collections Programme Manager adds that researchers need to know what proportion of the total holdings have been catalogued or digitised, even if this is just a rough estimate, so they are aware of the scale of undocumented material that might be of interest, rather than assuming that the digitised items are all there is.⁹³ NHM does this with a simple dashboard on the landing page of its data portal.⁹⁴

The potential for cultural heritage institutions to work in partnership with the higher-education sector is obvious for those that are themselves part of universities or, like some of the nationals, are recognized as independent research organisations (IROs) by UK Research and Innovation (UKRI).⁹⁵ But it is also true of smaller institutions: between 2016-18 the ACE-funded Museum-University Partnership Project (MUPI) ‘demonstrated how the higher education sector can be opened up to smaller and medium sized museums whose unique collections and engagement expertise are often an underutilised resource,’⁹⁶ and also published useful research into the potential of data aggregation within models of digital networking between museums and universities.⁹⁷

6.2.4 Being part of international research and development

Some of the larger and more technically-sophisticated cultural heritage institutions are able to participate in advanced research and development projects through their own efforts. NHM, for example, is aggregating its collections data through the Global Biodiversity Information Facility (GBIF) and has introduced tools to allow the scientific community to cite more accurately its use of the museum’s dynamic datasets.⁹⁸ And the British Library is collaborating with the Turing Institute on the £9.2 million *Living with machines* project, which will digitise millions of pages from newspapers published during and immediately following the industrial revolution, combine this data with other sources (such as geospatial data and census records) and develop new AI tools to unearth patterns in the data.⁹⁹

As the national library, the British Library also collaborates directly with Europeana¹⁰⁰, the continent’s main aggregator of digital cultural heritage and its most extensive research and development ecosystem in this field.¹⁰¹ However, the default model for smaller institutions is that Europeana harvests their data from a network of national and subject-specific aggregators. While its current project, Europeana Common Culture,¹⁰² will continue and build upon an experiment in harvesting linked data begun by the National Museum of the Netherlands and Dutch Digital Heritage Network,¹⁰³ this is likely to be feasible only for large,

⁹² ‘Discovering physical objects: meeting researchers’ needs’, *Research information network*
<http://www.rin.ac.uk/system/files/attachments/Discovering-objects-appendices.pdf>

⁹³ Pers comm

⁹⁴ <http://data.nhm.ac.uk/>

⁹⁵ www.ukri.org/files/funding/tcs/eligible-independent-research-organisations-pdf

⁹⁶ <https://www.publicengagement.ac.uk/nccpe-projects-and-services/completed-projects/museum-university-partnership-initiative>

⁹⁷ Alexandra Reynolds, Sammy Field, Jane Cameron and Lindsay Moreton, ‘Exploring Digital Network Museum-University Partnerships’, *National Co-ordinating Centre for Public Engagement*

https://www.publicengagement.ac.uk/sites/default/files/publication/mupi_digital_networks_report_2018_final_jun.pdf

⁹⁸ <https://naturalhistorymuseum.blog/2019/04/11/our-evolving-data-portal-digital-collections-programme/#more-13211>

⁹⁹ Dr Mia Ridge, pers comm

¹⁰⁰ <https://www.bl.uk/international-engagement/networks>

¹⁰¹ <https://pro.europeana.eu/what-we-do>

¹⁰² Project began 10 January 2018; no public web pages yet

¹⁰³ <https://github.com/nfreire/data-aggregation-lab>

technically-sophisticated institutions. Most English museums wanting to join the Europeana ecosystem in the foreseeable future, therefore, will need to do so through a national aggregator.¹⁰⁴

However, we have no national aggregator. The legacy aggregator Culture Grid¹⁰⁵ has been the UK pipeline to Europeana in recent years but stopped receiving new content in 2015 after its funding was discontinued.

Europeana is a founding partner of the ambitious European Time Machine FET project, a cutting-edge AI collaboration that is currently in the planning stage.¹⁰⁶ While it is open to individual institutions to become interim members of the new Time Machine Organisation, the terms, obligations and costs of participating in the eventual project are still being worked out. Collections data aggregated through Europeana, on the other hand, will be available to the Time Machine project without the contributing institutions having to do anything else.

6.2.5 A strategic, cross-sector approach to gathering audience data

While individual institutions may have analytics systems in place to monitor the traffic to and within their online collections, and even (like NHM) to track citations of individual digital assets, going beyond that will not be easy. Indeed, Europeana has struggled to implement the statistics dashboard it launched in beta version back in 2016.¹⁰⁷ (Interestingly, one of the technical challenges that led Europeana to suspend it was the lack of an authority file for institutions of the kind created for the prototype test described in this report.¹⁰⁸)

The Audience Agency suggests that the question of who uses digitised collections, and how, might usefully be included within the remit of the ambitious cross-sector Culture Finder framework it is currently scoping. That would allow the use of aggregated digital collections to be tracked in a **GDPR**-compliant way on behalf of all participating institutions, and the results interpreted within the overall context of users' online interactions with all forms of culture. Digital fingerprinting technologies could be applied at the aggregator level to allow the onward journeys of downloaded or shared assets to be tracked with greater precision than is currently attempted outside the commercial sector.¹⁰⁹

6.2.6 Maintaining authoritative lists of cultural heritage institutions

The National Lottery Heritage Fund (NLHF) has been working closely with ACE to try to improve the consistency of data about applications to the two funders. Inconsistent data about applicant organisations is proving especially challenging, and NLHF would welcome the kind of authority file proposed for the framework.¹¹⁰

¹⁰⁴ Harry Verwayen and Henning Scholz, pers comm

¹⁰⁵ <http://www.culturegrid.org.uk/>

¹⁰⁶ www.timemachine.eu

¹⁰⁷ <https://pro.europeana.eu/post/introducing-the-europeana-statistics-dashboard>

¹⁰⁸ Harry Verwayen and Henning Scholz, pers comm

¹⁰⁹ Anne Torreggiani and Cimeon Ellerton, pers comm

¹¹⁰ Fiona Talbott, pers comm

6.3 Risks

6.3.1 Confusion about potential audiences for aggregated data

It is worth repeating that the proposed aggregator would not be a destination site aimed at the wider public, nor would most of the collections data it would bring together be the kind of curated content expected by audiences. Rather, the aggregator would allow a wide range of third parties to research, select and re-purpose the raw data. In marketing terms, it would be a business-to-business service, not a business-to-consumer one. In framing the business case for any eventual aggregator built on the proposed framework, it will be important to keep this distinction in mind, and to value the behind-the-scenes use of aggregated data by curators for collections management purposes as much as the more obvious public-facing possibilities.

6.3.2 Duplicate records

If data is drawn from disparate sources, especially a mix of individual institutions and other aggregators, there is a risk of duplicate records. For example, in the test, data was ingested from both the National Gallery and Art UK. The former contributes content to the latter, so duplicate records resulted. In any eventual aggregator, a mechanism would be needed for identifying possible duplicates and deciding whether to prefer one source over another. For example, the National Gallery might want its own records to take precedence; another institution might prefer Art UK's enhanced version of its data.

6.3.3 Mixing up original source data and processed versions

The benefit of allowing institutions to send their content to the aggregator in whatever format they choose or can manage ('lenient ingest') does mean that those receiving and processing it have a certain amount of work to do in order to 'model' the data. This should not cause problems if an original copy of the source data is kept, and the contributing institution is able to review the imported data and the way it has been processed before it goes live.

6.3.4 Broken links

A further risk is the ever-present danger of broken links as contributing institutions rename their online content, move it around or otherwise fail to maintain it. Any framework for mapping digitised collections would need agreed standards for ***persistent identifiers***¹¹¹ and other aspects of good digital preservation practice that are beyond the scope of this study, but essential for any long-term digitisation strategy.

6.3.5 Lack of long-term commitment

Above all, there is the risk that an aggregator service is started without regard to the long-term need to nurture such a fundamental bit of infrastructure. There is the danger that cultural heritage institutions, software providers and developers of third-party applications that re-use its aggregated data would be left high and dry if the service were not used, maintained and supported. As the history of aggregation initiatives in the UK cultural heritage

¹¹¹ <https://dpconline.org/handbook/technical-solutions-and-tools/persistent-identifiers>

sector shows, ¹¹² since the late 1990s there has been a pattern of short-term funding, by a succession of short-lived commissioning bodies, that has held back early advances in this field.

¹¹² See **appendix A** of this study's scoping report

Appendix A: Acknowledgements

Tao-Tao Chang, Research Grants Manager, V&A Research Institute (on secondment to AHRC)

Cimeon Ellerton, Chief Operating Officer, The Audience Agency (TAA)

Scott Furlong, Director, Collections and Cultural Property, Arts Council England (ACE)

Zelina Garland-Rowan, Senior Manager, Collections and Museums Programmes, ACE

Helen Hardy, Digital Collections Programme Manager, Natural History Museum

Sarah Madden, Digital Engagement Officer, South West Museum Development Programme

Chris Michaels, Digital Director, National Gallery

Jon Ray, Programme Manager – Gardens, Libraries and Museums Digital Strategy, University of Oxford

Dr Mia Ridge, Digital Curator for Western Heritage Collections, British Library

Henning Scholz, Partner and Operations Manager, Europeana Foundation

Fiona Talbott, Head of Museums, Libraries and Archives, National Lottery Heritage Fund

Anne Torreggiani, Chief Executive Officer, TAA

Harry Verwayen, Executive Director, Europeana Foundation

Appendix B: 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

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>

GDPR

General Data Protection Regulation. An EU regulation to protect personal data.

See: <https://gdpr-info.eu>

IIIF

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

See: <https://iiif.io/>

HTTP

HyperText Transfer Protocol. HTTP is the underlying protocol of the World Wide Web and defines how messages are formatted and transmitted, and what actions Web servers and browsers should take in response to various commands.

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

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 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 specification for embedding structured metadata into web pages that can be read and used by search engines to give better results.

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

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>

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>

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>