**Title: Is your Cultural Heritage Collection AI Ready? A Methodology for Semantic Enrichment of Cultural Images and Photographs.**

Yalemisew Abgaz, Dublin City University, Dublin, Ireland

The recent advancement in Artificial Intelligence (AI) has paved the way for the wide adoption of new tools and techniques in numerous disciplines. Galleries, Libraries, Archives, and Museums (GLAMs) are adopting AI-based solutions to efficiently organise, analyse, and utilise their digital collections. The application of AI-based solutions in GLAMs is mainly based on the foundational work librarians, archivists and museologists did in digitising their collections in a machine-readable format. Following the digitisation effort, the organisation of the digitised resources by integrating metadata that provides useful information to properly utilise the resources paved the way for the application of AI solutions. Nowadays, GLAMs have started exploiting the technology in digital image processing, semantic enrichment, and interlinking of historical and cultural collections including images, photographs, drawings, sketches and other archival collections. To efficiently utilise these AI solutions and assist non-technical experts who are working in GLAMs, a methodology that works not only for AI experts but also for all stakeholders is a necessary condition. In this paper, I discuss a methodology that has been used in projects that are dedicated to the organisation of cultural heritage collections using AI-based solutions. The methodology has three phases: the preparation phase focuses on domain understanding, acquisition of target collection, and ontology selection; the analysis phase focuses on semantic enrichment (annotation) and knowledge graph generation; the deployment and exploration phase focuses on focuses on the implementation of the solutions and exploitation of the semantically enriched AI-Ready resources using the AI-based solutions. This paper will further present two case studies where the methodology is applied and presents the lessons learned from the two projects.

## Introduction

In the past decade, the application of Artificial intelligence (AI) in disciplines other than computer science has become widespread. The technological advancement to support high-end computing, availability of very large volume, variety, and velocity of data, and the emergence of several algorithms that improved the state of the art paved the way for the adoption of AI-based applications. AI-based applications support numerous tasks that, otherwise would have been done by humans in areas such as customer support, business intelligence, financial predictions, gene analysis and others. The application of AI is further extended to the social sciences and digital humanities domain. In the digital humanities domain, GLAMs have adopted AI-based applications into their workflow to provide better support and improved user experience in accessing, analysing, and presenting resources of cultural and historical importance.

GLAMs hold a large number of cultural and historical resources. These resources can be viewed from different perspectives. Timewise, these collections cover artefacts from ancient times to modern-day collections. GLAMs that are specialised in supporting specific eras (such as WWII collections) or those focused on providing services independent of time, or GLAMs that focus on the specific types of collections such as books, historical sites, etc. Based on their objectives, GLAMs respond differently to the need for adopting AI systems.

The adoption of AI systems in GLAMs did not happen overnight but rather took several slow but concrete steps including collecting and preservation, digitisation, semantic enrichment, and exploitation. In the early ages of GLAMs, the focus was on acquiring a piece of work of huge importance of preserving a resource that is at risk. The digitisation process enables the conversion of such resources into digital formats. Following the success of digitising resources and making them available to the public digitally, the focus shifts towards enriching such digital resources using generic and domain-specific metadata. Now, with the availability of a large collection of resources and the availability of several metadata standards, it has become possible to further exploit the collections using different types of AI systems.

What the current technology offers to GLAMs is the opportunity to exploit very powerful AI systems including computer vision, natural language processing, question answering, information search and retrieval and user support using conversational applications. The use of AI algorithms in supporting one or more of these applications mainly depends on the availability of digital resources, sufficient metadata based on widely accepted standards, and the availability of highly interconnected, rich semantics that goes beyond the descriptive metadata. In simple terms, there should be a large volume of usable datasets.

However, there are challenges in exploiting the full capacity of AI systems when it comes to historical and cultural collections. The challenge in GLAMs mainly revolves around the availability of digital representations, metadata, and rich semantics for the content of the target digital resource. The first challenge, to some extent, has been addressed by digitisation, whereas the second challenge was addressed by generic and descriptive or technical metadata which was created before or during the digitisation of the resources.

AI become part of the solution in the effort to semantically enrich digital resources. AI-based applications that are capable of learning from large volumes of training data showed a significant improvement in the past ten years. These systems are currently exploited to incorporate additional semantics by analysing the content of images, texts, and voices based on the training data provided to them. Once the training phase saturates, they can be used to annotate new, previously unseen content.

In 2021, we published a journal paper on the topic of “A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies” (Abgaz, et al., 2021) that outlines the methodology adopted to build AI-based applications for European image collection in the context of the ChIA[1] project. In 2018, we also published another journal paper on the topic of “Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers” (Abgaz, et al., 2018) which focuses on semantic modelling, enrichment, and exploitation of a large collection of Bavarian Dialects in the context of the ExploreAt![2] project. In this paper, we presented our experience in transforming a digital collection into something that can be exploited by AI-based systems – making it AI-ready. We will present the three phases of our methodology along with the two case studies (ExploreAt! and ChIA) by exploring how the methodology is applied to solve the problems in the respective projects. Finally, we will present the lessons that we learn from the ExploreAt! and ChIA projects.

[1] Add the chia project link here.
[2] Add the ExploreAt project here

## Background

In the past, the focus of GLAMs was to acquire a wide range of collections that have historical, cultural, political, educational, or monetary value. Back then, GLAMs focused on acquiring as many original collections. However, in many circumstances, these collections were not effectively exploited but kept behind the shelves behind the counter. A huge catalogue of the collections in big cabinets served as an interface between the users and the collections. This approach has sustained much of the collections by minimising direct access to the valuable collections but resulted in a very low level of usage due to the need for passing through multiple gatekeepers. The major problem at this stage was that the required information is only available in one place, and it was not easily accessible to the wider community of users. From the collection's point of view, the valuable information is mostly kept in the minds of the experts and is not explicitly linked to the objects. Such tacit knowledge remained untapped until the further adaptation of new technologies happened.

### Digitisation

The invention of computers and digital technologies had revolutionised the way data is stored, organised, and accessed in GLAMs. The digitisation of a considerable amount of historical and cultural collections (books, images, photographs, sketches. paintings, drawings, etc.) improved the visibility of the resources and facilitated easier access to the information by replicating the digital copies of the collection. Digitisation plays a major role in addressing the problem of the physical barrier to accessing such resources. However, as the technology grows, and multiple collections were digitised, it became difficult to process and exploit the collections which led to the use of descriptive information about the collections as a form of metadata.

### The Adoption of Metadata Standards

Metadata in a form of catalogues has been around. GLAMs organise their resources using classification and cataloguing methods that enabled them to represent the subjects of the contents of documents, determine the appropriate classification and provide significant metadata about the documents including the authors, title, publisher, editor, and the physical state of the document including the dimension, number of pages, material and keywords that provide additional information about the content. Although this is mainly done in a physical drawer (See Figure 1), such a scheme has laid the foundations for grouping resources that have the same or similar topics together although the materials are stored far away from each other on shelves. The catalogues are usually produced in several copies enabling searching by various facets of the documents including Author, Subject, Title, and Year entries

(See Figure 2) among the most widely used methods. Subject headings that organise collections using broader and narrower terms by capturing one or more of the essence of the content that has been covered in the books or any other materials.

*Figure 1. Paper-based catalogue card cabinets.*

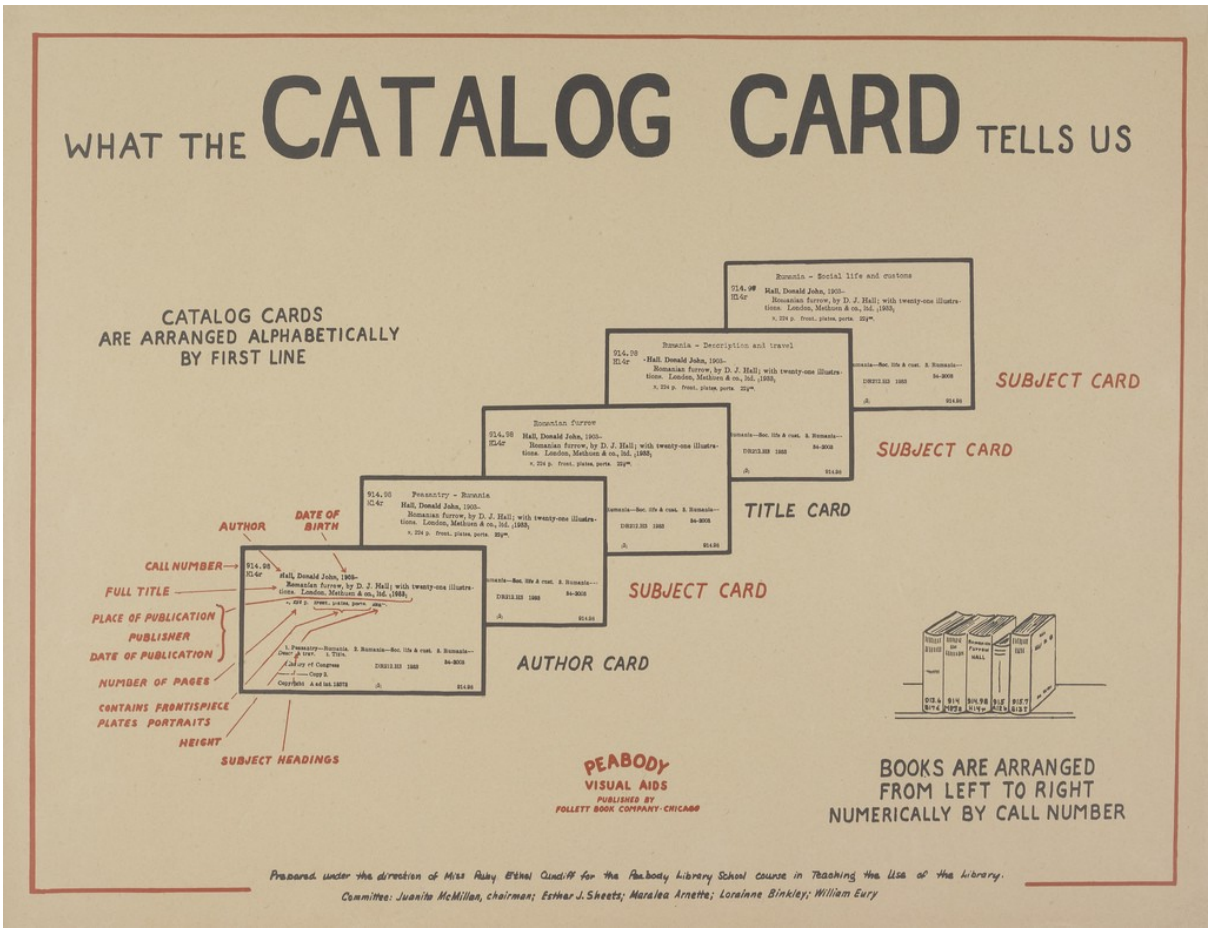


*Figure 2. The detailed content of a catalogue.*

Fast forwarding, the introduction of semantic web technologies and the adoption of metadata standards and taxonomies as a form of ontologies and knowledge graphs paved a new way for the exploitation of the contents of historical and cultural collections (Jones, 2012). Again, text-based cultural collections benefited the most from this. This is because, the focus of the research was on text-based data until a significant shift in processing multimedia content (image, audio, video) occurred. In recent years, the significant advancement of technology in image, audio, and video processing enabled GLAMs to enrich and make available their widely untapped image collection (Alexiev, 2018).

## Data in GLAMs

**Past Data**

It is common for GLAMs to deal with data which was collected before the invention of modern-day digital technologies. Historical collections in published media are usually stored in GLAM. However, there exists a large volume of historical and cultural collections in the hands of individuals and private collectors. These collections included handwritten communications, sketches and paintings of buildings, public figures, and utilities, and publications including books, periodicals, and newspapers. Among the oldest collections, periodicals including the oldest newspapers such as *Gazzetta di Mantova*, *Avisa Relation oder Zeitung* and *La Gazeta* from the seventeenth century are available in different archival centres. With the emergence of the internet in the 1990s, GLAMs started the digitisation process to maintain their online presence. These collections became available in digital formats. GLAMs that deal with archival collections mostly deal with multiple versions and formats of such collections. Due to this, there are major challenges GLAMs face in their effort to utilise such resources by AI systems.

Some of the problems and challenges include the digitisation of the resources, extraction of meaningful information from the sources, and the transformation of the data into some usable digital format which is suitable for AI-based applications.

**Challenge 1:** The mass digitisation of such resources happened mostly in the 1980s and 1990s in many countries as the use of computers became a common practice. However, there are still several GLAMs that preserve collections that are available only physically. Digitised collections that exist in a form of images also pose challenges to the modern-day GLAMs. The proposed solution to convert these digitised scans into their textual equivalents, mostly in the case of books, newspapers, periodicals, journals, etc has been done manually and using optical character recognition (OCR) technologies. Although sketches, drawings and paintings are digitised, mostly they remain in the form of images. However, the

textual description of the contents of the images that provide very useful metadata required the extraction of the information from the images.

**Challenge 2**: Another challenge faced is the extraction of meaningful information from the resources. The digitised text and images as they appear in the original collection have limited relevance without additional information provided about the content. For example, the digitised image of the oldest newspaper does make little sense for modern-day users unless some additional information to help the users understand and interpret the information contained in the newspaper is provided. To achieve this, it is necessary to collate information from different sources. For example, the availability of the meaning of the words used in the newspaper, their translation, the setting, and the history during the time the newspaper was published make these resources more usable. The task of making this information available fell on the shoulders of museologists, archivists, and librarians who are subject experts. Thus, the second challenge is to semantically enrich the resources with rich semantics.

**Challenge 3**: The last but not the least challenge in making such collections available and usable is the transformation of the information into a format that is suitable for modern-day AI applications. Although some GLAMs still preserve huge collections that are not yet digitised, most digitised collections exist in a format that is not suitable for AI-based applications. The digital collections mostly exist in a form of unstructured, or semi-structured that require significant pre-processing of the data such as data cleaning. Since semantic interlinking and enrichment are recent phenomena in computer science, most collections yet must pass through this step to be efficiently exploited.

While more technology-advanced GLAMs are addressing challenge 3, several others are struggling to resolve challenge 2 or even challenge 1. The GLAMs that are facing challenges 1 and 2 are in a better position as there are many tools and technologies that they could exploit without going through the challenges others had been facing. However, addressing these challenges is not a trivial task of picking an off-the-shelf application, but it requires a deep understanding of the domain of the collection and the meaningful transformation of the resources.

**Present Data**

The present-day GLAMs are mostly focused on resources that are already available digitally since their creation (born-digital resources). These digitally available resources usually come with very rich descriptions associated with them, mostly prepared by the original creators of the artefacts. However, an existing challenge that is still rolling down from the past is for those GLAMs that have a huge collection of historical and cultural resources to make them available online in a format that can be used efficiently by existing AI technologies. The major objective of GLAMs

currently is making the existing historical and cultural resources not only digitally available but also exploiting them efficiently to support their maximum usability by users.

Although this appears to be a noble objective, there are significant challenges associated with it. The first is providing standardised, machine-readable, and rich metadata about the resources to make them suitable for Artificial intelligence applications. A significant effort has been made in this direction by PREMIS[3] (PREservation Metadata: Implementation Strategies). PREMIS is the international standard for metadata to support the preservation of digital objects and ensure their long-term usability. This is a significant contribution to transforming analogue collections into their equivalent digital format.

## Methodology

In this section, a general methodology used to transform historical and cultural collections into AI-ready data is presented. The methodology is organised into preparation, analysis, deployment and exploration phases. These three phases are empirically derived from multiple case studies from projects that aim at transforming existing datasets into suitable formats for AI applications.

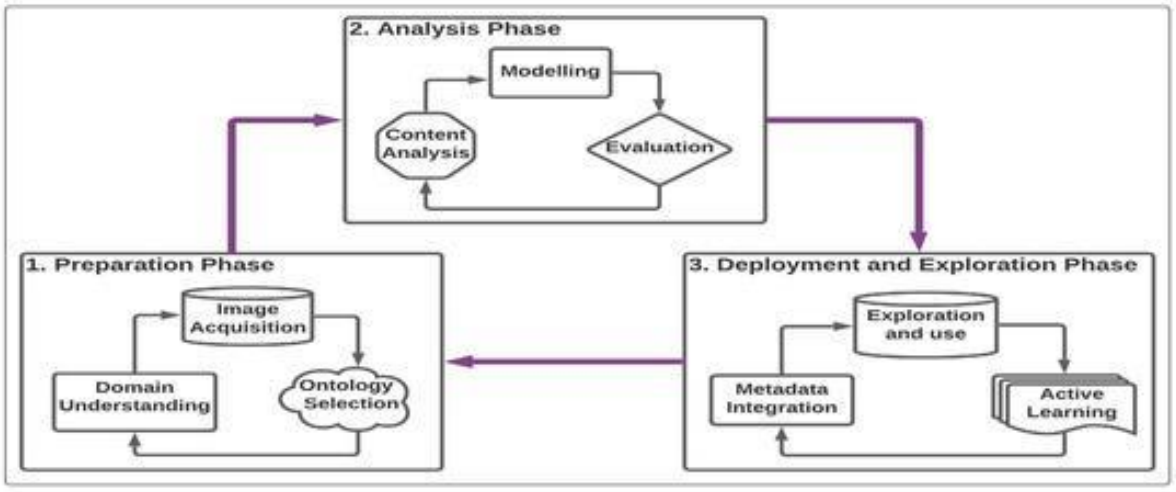


*Figure 3. A three-phase methodology for semantic enrichment of historical and cultural collections.*

***The Preparation Phase.***

[3] https://www.dpconline.org/handbook/glossary#P

It is important to remember that cultural and historical collections are complex and diversified in their content and format. Although this diversity is an indicator of the quality of the collection, it is also the source of many challenges. In this phase, the main tasks include domain understanding which focuses on the exploration of the subject of the collection, determining the content and deciding the boundaries of the domain. Another crucial step in this phase is the acquisition of the collection. This is usually achieved through digitisation. The collection then becomes the subject of the domain understanding stage as it is carefully investigated to determine what is covered in the collection to understand the domain and then to select the ontologies that can represent the content. The ontology selection goes together with the acquisition and domain understanding steps. For example, if we take a collection focusing on cultural food images, the domain understanding will focus on the subject of food images and define what can be considered a food image. The domain understanding phase makes a clear distinction between the food images and non-food images. Based on this understanding, the collection will be filtered and all the images that qualify will become the target of the study. By iteratively investigating the contents of the food images, we may further expand our domain understanding to further explore the collection. The level of detail we would like to incorporate to enrich the collection determines the type and nature of the ontology we select. A fine-grained semantics can be achieved by selecting domain-specific ontologies, whereas generic semantics can be embedded using more general metadata and subject headings.

***Analysis Phase.***

The analysis phase focuses on the extraction of the contents of the collection. This phase includes content analysis, modelling, and evaluation. Content analysis is an extension of domain understanding in that it also focuses on the content of the collection but goes deep into the actual content of the individuals in the collection. The content analysis is usually subjective and may require two or more annotators to reach a consensus. For a very large collection, it is not possible to manually analyse all the contents of the collection. An alternative approach is to exploit existing AI applications to perform the task automatically. For that purpose, building models using training datasets and testing the performance of the model using test and validation sets is a common approach in existing AI-based solutions. The overall objective of the analysis phase is to extract as much detailed and accurate information about the content of the individual collection and semantically interlink it with the selected generic and specific ontologies or taxonomies.

***Integration and Exploitation Phase***

The integration of the semantics generated in the analysis phase with already existing semantics about the content is the focus of this stage. This integration will link the content to standard vocabularies, and existing rich datasets such as DBpedia. The integration creates subject-predicate-object triples that be organised

into a semantically rich knowledge graph. This knowledge graph will be used to support efficient semantic exploration.

A major lesson learned from applying this project is that the methodology should be implemented iteratively and there is no single solution that fits all the problems.

## Case Studies

In this section, I present two case studies that are very relevant to anyone who considers making their collection AI-ready. The first case study focuses on a century-old collection of Bavarian Dialect data by the Austrian Academy of Sciences. The second case study focuses on the ChIA project data collection that represents a collection which is hosted by Europeana, one of the largest aggregators of historical and cultural images. The third case study is more general and focuses on any organisation that deals with routine document collection. Now let's look at each of these case studies in detail.

**Case Study 1: ExploreAt: A Collection of Bavarian Dialect Data**

The Bavarian dialect data collection is a historical non-standard language resource that is collected from 1913 to 1998. The data is collected by experts using questionnaires distributed around present-day Austria, Czech Republic, Slovakia, Hungary and northern Italy. Most of the original data were collected using handwritten paper slips distributed around the Habsburg monarchy (Wandl-Vogt, 2008, Wandle-Vogt, nd)

The aim of the project was to document the Bavarian Dialect and ruler life in Austria. This objective was achieved through a continuous collection of data for a long period which resulted in a huge collection of data in diversified mediums and formats including handwritten text on a paper card, typewritten text on a card and even some digitised data in the later stage of the project.

The Bavarian dialect data collection includes original documents that contain information about the questionnaires (See Fig 4.c), questions and data collectors involved in the project, and answers to the questions (See Fig 4. a and Fig 4.b), from numerous sources. The data further contains location and time-related information. Around 109 main questionnaires, nine additional questionnaires (Ergänzungsfragebögen) and two Mundartgeographischer Fragebogen der Münchner und Wiener Wörterbuchkommissionen questionnaires and other additional freestyle questionnaires were used to collect the data. 24,382 individual questions (See figure corresponding to the available questionnaires had been used resulting in around 3.6 million hand- or type-written answers. 11,157 individuals who had been assigned to various functions including collectors, editors, and coordinators have been recorded.

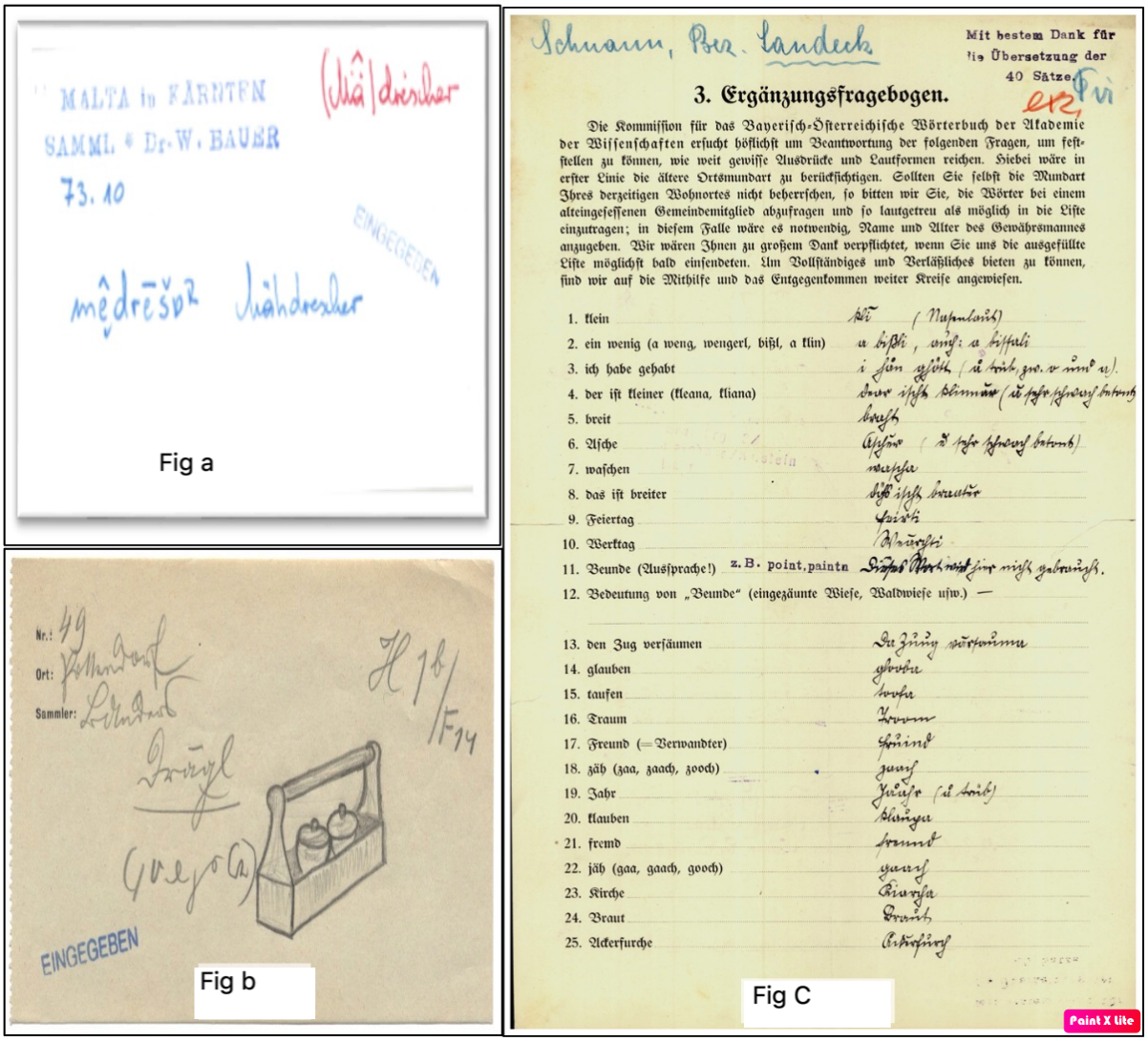


*Figure 4. Questionnaires and answers represent text and sketches.*

The data format of this collection was initially a handwritten paper-based format. The current version of the data has passed through several stages of digitisation. The main objective of this project is to make this resource useful for both humans and AI agents with the aim of supporting the analysis of complex patterns over time and exploring interesting links between the resources. The primary target users here are lexicographers, linguists, historians, biologists, and in general citizen scientists.

***Preparation Phase***

For the Bavarian dialects, the collection mainly focuses on textual and multi-media data. The text represents all the questionnaires used to collect data and the answers arrive in different formats mainly written on paper slips and sometimes with drawings and sketches (See Fig 4. b).

This collection has already been transformed using several digitisation steps. Firstly, the collected data has been scanned and a digital representation of most of the paper slips exists. This resulted in a huge collection of digital scans of the paper slips that became accessible to users online. Secondly, the data is transformed into TEI/XML where the text from the scanned copies of the data had been extracted and encoded. Thirdly, part of the extracted text is converted into a database for easy search and retrieval. Finally, all the data from the database, TEI-XML, and scanned copies are combined to create a five-star linked dataset with additional ontologies providing rich semantics to the data. The important lesson we learned to create AI-ready data is summarised below.

During the domain understanding phase, the role of subject-matter experts has been very significant. The domain experts know details about the numerous questionnaires, the descriptive metadata associated with them, and further semantics that cannot be extracted from existing documents. When dealing with such historical and cultural datasets, it is important to participate in the subject matter experts at the earlier stage of the project.

During the data preparation phase of the project, the DBO dataset is the most challenging project for various reasons. The two most outstanding ones are, firstly there are several issues with the quality of the data. Some of the issues include inconsistency in recording data fields such as place names, dates, measurements etc. Secondly, the length of time the data collection took (>80 years) made the data cleaning phase difficult.

When you have a historical and cultural dataset that spans such a long period, it is important to note that place names, administrative provinces and even borders of countries change over time. Further, cultural aspects also change significantly. For example, the word “Donkey” was mentioned several times along with “Wedding” compared to insignificant co-occurrence of donkey and wedding.

Achieving a higher accuracy in such datasets is not possible. Thus, anyone working in the preparation of such historical and cultural datasets should only deal with the problem incrementally. Cleaning the data to a certain level and should not anticipate a complete accuracy.

Another important factor in the preparation stage is to find a suitable ontology/metadata that is capable of accurately and eloquently describing the content of the collected data. This step is usually ignored or overlooked. However, to semantically represent such historical datasets, a proper analysis of the domain the collection represents (say in DBO, colour is one important aspect) is required. For this step, although there exist so many metadata and ontologies, the provision of rich semantics is usually restricted. The effort to precisely express the semantics of the dataset results in long and complex subject analysis, exploration of existing

metadata and ontologies and development of such metadata to represent the content. In the effort to semantically enrich the dataset, several ongoing projects ended up doing all the steps for several years. And yet, there has not been comprehensive semantics to describe the collection.

Two ways can be followed here to achieve the goal of producing an AI-ready dataset. The first one is focusing on the aspects that are generic to the whole dataset without going into the details and providing rich semantics. The second option is to narrow down the focus to a specific part of the collection and provide rich and deep semantics, but applicable to a subset of the dataset. For example, the ExploreAt! project focus was on the main questionnaires.

***Analysis Phase***

In the analysis phase, the contents of each questionnaire are extracted and analysed. This includes the extraction of the questionnaire number along with the formats, title, authors, year of publication etc. Meaningfully connecting the questionnaires to the questions and their corresponding answers and developing a new semantic model to represent such interlinking were some of the outputs of this phase. Most of the modelling work focused on extracting patterns of dates, place names and other text-based datasets. This resulted in the creation of a new ontology for representing a Lexical Data Collection and ANalysis (OLDCAN) Ontology.

***Integration and Exploitation Phase***

The extracted data were semantically linked to existing metadata such as Dublin core and FOAF. Other domain-specific subject headings such as the DBpedia ontology have been used to interlink the questionnaires to the more semantically rich collections. The final data set includes 720 questionnaires and 24,382 questions. There are also 11,157 individuals semantically annotated along with 65,839 paper slips, 98,272 answers, 8,218 multimedia files and 16,839 sources. 2.8 million triples organised into eight named graphs were generated.

The dataset has been used as an input for several follow on applications. Some of them include supporting exploration of the questionnaires using visualisations (Doren, et al, 2018, Rodríguez Díaz, et al., 2019) and others use the dataset for the analysis of changes in the meaning of words over time and uncertainties in the dataset (Rocha Souza).

The major lesson learned from this project can be summarised as follows. Firstly, we should expect a reasonable level of complexity when we deal with historical and cultural collections. The complexity may come from the data, format, time, and many more factors. Second, we should take subject experts on board as early as possible. Third, we should focus on the content represented in each item in the collection that

contains the piece of information that makes a difference in AI applications in addition to generic metadata. It should be noted that extracting such content and semantically enriching it with new or existing ontologies is not a trivial task.

**Case Study 2: ChIA Data Collection**

The ChIA project deals with data aggregated from different GLAMs in Europe and across the world. The Chia dataset focuses on digitised image collections representing paintings, drawings and sketches of food, farming, construction, war, machinery and a lot more. This collection has standardised metadata based on the European Data Model (EDM) providing descriptive metadata about the image collection. However, most of the time, the metadata does not provide full information to understand the details included in the contents of a particular image. For example, the description of a piece of image could contain a Dublin Core Description (dc: a description of “water painting by Picasso”), without providing further information about the content of the painting.

The dataset considered in the ChIA project contains a unique identifier of the resources and additional metadata which is provided by either the aggregator (Europeana) or by the contributor. The identifiers make the ChIA dataset in a better position in terms of structure and semantics. This is because some of the datasets included in the collection have passed through a checkpoint which requires the inclusion of minimum metadata when the images are made available to the public.

We learned lessons from the different phases of the ChIA project. To benefit others, we organised the discussion using the original phases. During this whole process of enriching the data, a multidisciplinary collaboration has been exploited and benefited the project significantly. Perhaps, that is the first lesson that we learned from this project. In the process of making such collections AI-ready, it is important to build a multidisciplinary team of subject experts, metadata specialists, Artificial intelligence specialists, and librarians, museologists or archivists who know the collection in detail. Furthermore, depending on the collection and the purpose, you may need to include AI experts who understand AI applications and algorithms. This enables building a comprehensive understanding of the data, composing a descriptive dataset, and ensuring the future applicability of the solution to other problems.

***Preparation Phase***

During the preparation phase of the ChIA project, we focused on domain understanding, image acquisition and ontology selection. During the domain understanding phase, the domain experts and mostly the Europeana local Austria experts play a crucial role in eliciting the knowledge contained in the collection and its interpretation. All the relevant information about the collection should be extracted, defined and made available centrally. A shared document suffices for this

purpose, however, as the project continues, it is important to keep the different versions of the document. Later in the second and third phases of the project, this document will serve as a dictionary for all members and future users to refer to.

For example, exploring the collection, the source of the dataset and the specific categories that are considered in the project helped the rest of the team to gain a useful understanding of the content. Even, understanding small things such as the difference between a painting, drawing, and sketch makes a significant impact because, when these categories are applied to the data, the explanation component of the AI system uses them as an input to describe the content.

Another lesson learned here is related to the acquisition of the input dataset. It is important to fully understand the available collection in addition to domain understanding. In the ChIA project, we learned that different metadata of the same image is submitted to Europeana from different sources with different URLs. We further noticed that the metadata of the images varies. It is important to aggregate such information and if possible unify the metadata. What we propose here is either to merge all the metadata into one by voiding the duplicate entries, or select the most authoritative one. The process to reach at such a decision is challenging by itself.

The next step to consider is to think about the completeness of the data. It is important to understand the features that are crucial to describe any resource in the collection and determine the required and optional features of the dataset. further, it is crucial to clearly define or link them with existing definitions. One of the required fields for the images in the ChIA project, for example, is the URI. An image should contain a unique identifier. An image that comes from different sources may have several unique identifiers. An agreement has to be reached about what to do if we have collections that do not have unique identifiers such as excluding the resources from the dataset or assigning a unique identifier. When the resources are available, assigning unique identifiers is preferable as most GLAMs still handle collections that are not yet assigned unique identifiers. Additional descriptions of the resources such as authors, creation date, titles, labels etc also need some sort of decision as to what to do when these descriptions are missing or unknown. This step mostly focuses on the creation of complete, clean and usable data. It is worth noting that any AI system will not learn much from missing data. In this step, make sure all the necessary data is collected, uniquely identified and sufficiently described with minimum attributes.

Domain understanding in combination with data preparation led to the identification of useful metadata to enrich the collection. From our experience in the ChIA project, it is crucial to look into existing metadata and domain-specific ontologies before deciding to create a new one. Most often one or more of the existing metadata provides the level of detail we would like to capture about the collection. However, when dealing with relatively less explored collections, it becomes crucial to create a

new ontology from scratch. In our case, we deliberately avoided creating new ontologies as the project timeline was very short to accommodate this task. However, we aggregated several concepts from other ontologies into one ontology to centrally access the semantics.

*Figure 5. A sample food ima depicting several fruits, animals, and birds.*

*Figure 6. Another sample food image.*

***Analysis Phase***

The analysis phase of the ChiA project focused on extracting or collecting data that can be used to provide rich semantics about a particular collection. For example, what information does a particular image or painting contain not only in its metadata but also in the content? An image of a banquet may have numerous fruits, vegetables and poultry products. Such levels of detail may not be provided in the dataset and may become a requirement for the training and testing of the AI application.

In general, the analysis phase ensures the existence of sufficient and useful information that can be directly consumed by the AI system. It is important to decide in the earlier stage whether the AI system will be used to classify and predict data. This means that when you have a collection of millions of resources, it is difficult to go through all the resources and manually annotate all the information contained in the collection. In such cases, as in ChIA, we need to build an AI system that would learn patterns from sample datasets (training sets) that are manually annotated. Once the system demonstrates satisfactory accuracy, the system could be used to predict the different categories of the image content.

Most of the existing classification systems are built to predict labels or categories of collections. When predictive AI applications are used, best-performing models are selected based on predefined criteria. From our experience with the ChIA project, it is crucial to distinguish between annotations provided by humans and that of AI systems. We also recommend the inclusion of the accuracy as an indicator of the level of confidence in the accuracy of the predicted category.

***Integration and Exploitation Phase***

The integration and exploration phase in the ChIA project involves the application of an AI model to predict the different classes of the appealingness of food images. A pilot model is developed in phase 2 to generate new annotations of images. The new annotation includes the confidence level of the prediction and could be used to inform the users of the source of the annotation which is from an AI model that predicted the appealingness of the image. We consider such annotations as semantic enrichments which can be used to group similar images based on their categories or indirectly by linking the category labels with other ontologies. For example, an appealing image is linked to aesthetic pleasure[4]. This allows us to aggregate all the appealing images. In addition, the aesthetic pleasure class in ontology is linked to other classes. This will allow us to create indirect links between images within the collection or outside of the collection.

Further to semantic interlinking, we focused on the exploitation of the semantic link. As semantic interlinking and the creation of the so-called “Knowledge graph” of the

[4] https://ontobee.org/ontology/MFOEM?iri=http://purl.obolibrary.org/obo/MFOEM_000039

images opens a way for building several AI-based applications including semantic question answering, chatbot systems, interactive visualisation and exploration. For example, in the ChIA project, the semantic annotations were tried in a novel way of aggregating images that were not as such related to each other and were used to develop a chatbot system that is trained based on the knowledge graph generated from the annotation including the EDM and other metadata standards.

Once your data is assigned a unique identifier (URI) and all the necessary annotations are supplied, there will be numerous ways of exploiting the data (Dorn, 2020). One important thing that should be emphasised here is the need for high-quality annotation in a form of correctness, possibly completeness, and richness of annotations.

## Conclusion

In this paper, we discussed the methodology we adopted to transform historical and cultural collections in general and cultural images in particular. The main question is how to transform an existing collection into a dataset that can be efficiently utilised by existing and future AI-based systems. One of the core areas of the transformation process is the preparation phase where the domain of the collection is analysed, necessary data is acquired and relevant metadata, taxonomies, or ontologies are selected. Following this, the analysis of the actual content of the collection provides all the necessary information that needs to be known about the collection. This goes beyond collecting generic metadata such as the author, title, or year of publication of a piece of work. This phase semantically enriches the collection either by manually annotating the collection or by training an AI system to accurately extract the contents of the collection. In both cases, the aim of the analysis phase is to extract as much information and semantically enrich the collection. Once detailed information about the collection is extracted, the dataset will be generated. This dataset could be used in numerous other AI-based applications.

However, the major challenges of converting historical and cultural collections come from understanding the domain of the collection and correctly and accurately representing the information contained in the collection. The second major challenge is identifying the appropriate ontology or taxonomy to organise the actual content of the collection. A semantically rich domain-specific ontology serves better in this context. However, finding such an ontology becomes difficult and may lead to the development of a new semantic model representing the domain content.

The proposed methodology touches only the tip of the iceberg. When dealing with the actual collection, there will be several other factors that will require detailed and time-consuming research. However, in the effort to apply any AI solution, we should always ask whether we provide a semantically rich dataset. One way of ensuring that

is by creating a rich and semantically interlinked dataset with an accurate and detailed annotation that transcends beyond the likes of descriptive metadata and the use of generic taxonomies. AI-ready data set would be a dataset possibly with a five-star linked open data, rich with content and with complete coverage.

## Bibliographies


Abgaz Y, Rocha Souza R, Methuku J, Koch G, Dorn A. A Methodology for Semantic Enrichment of Cultural Heritage Images Using Artificial Intelligence Technologies. J Imaging. 2021 Jul 22;7(8):121. doi: 10.3390/jimaging7080121. PMID: 34460757; PMCID: PMC8404920.

Abgaz, Yalemisew, Amelie Dorn, Barbara Piringer, Eveline Wandl-Vogt, and Andy Way. 2018. "Semantic Modelling and Publishing of Traditional Data Collection Questionnaires and Answers" *Information* 9, no. 12: 297. https://doi.org/10.3390/info9120297

Wandl-Vogt, E.; Kieslinger, B.; O´Connor, A.; Theron, R. exploreAT! Perspektiven einer Transformation am Beispiel eines lexikographischen Jahrhundertprojekts. Available online: http://docplayer.org/16597238-Exploreat-perspektiven-einer-transformation-am-beispiel-eines-lexikographischen-jahrhundertprojekts.html (accessed on 22 November 2018).

Wandl-Vogt, E. Wie man ein Jahrhundertprojekt zeitgemäß hält: Datenbankgestützte Dialektlexikografie am Institut für Österreichische Dialekt- und Namenlexika (I Dinamlex) (mit 10 Abbildungen). In Bausteine zur Wissenschaftsgeschichte von Dialektologie / Germanistischer Sprachwissenschaft im 19. und 20. Jahrhundert. Beiträge zum 2. Kongress der Internationalen Gesellschaft für Dialektologie des Deutschen; Praesens: Vienna, Austria, 2008; pp. 93–112.

Abgaz Y., Dorn A., Piringer B., Wandl-Vogt E., Way A. A semantic model for traditional data collection questionnaires enabling cultural analysis. In: McCrae J.P., Chiarcos C., Declerck T., Gracia J., Klimek B., editors. Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018) European Language Resources Association (ELRA); Paris, France: 2018.

Dorn A., Wandl-Vogt E., Abgaz Y., Benito Santos A., Therón R. Unlocking cultural conceptualisation in indigenous language resources: Collaborative computing methodologies; Proceedings of the LREC 2018 Workshop CCURL 2018; Miyazaki, Japan. 7–12 May 2018.

Abgaz Y., Dorn A., Preza Diaz J.L., Koch G. Towards a Comprehensive Assessment of the Quality and Richness of the Europeana Metadata of food-related Images; Proceedings of the 1st International Workshop on Artificial Intelligence for Historical Image Enrichment and Access; Marseille, France. 11–16 May 2020; Marseille, France: European Language Resources Association (ELRA); 2020. pp. 29–33.

Preza Diaz J.L., Dorn A., Koch G., Abgaz Y. A comparative approach between different Computer Vision tools, including commercial and open-source, for improving cultural image access and analysis; Proceedings of the 10th International Conference on Advanced

Computer Information Technologies (ACIT'2020); Deggendorf, Germany. 16–18 September 2020; - DOI

Rodríguez Díaz, A.; Benito-Santos, A.; Dorn, A.; Abgaz, Y.; Wandl-Vogt, E.; Therón, R. Intuitive Ontology-Based SPARQL Queries for RDF Data Exploration. *IEEE Access* **2019**, *7*, 156272–156286. [CrossRef]

Dorn A., Abgaz Y., Koch G., Díaz J.L.P. Harvesting Knowledge from Cultural Images with Assorted Technologies: The Example of the ChIA Project. In: Lykke M., Svarre T., Skov M., Martínez-Ávila D., editors. Knowledge Organization at the Interface: Proceedings of the Sixteenth International ISKO Conference, 2020 Aalborg, Denmark. 1st ed. Ergon-Verlag; Baden, Germany: 2020. pp. 470–473. International Society for Knowledge Organziation, (ISKO) - DOI

Jones D., O'Connor A., Abgaz Y.M., Lewis D. Proceedings of the 2nd International Conference on Multilingual Semantic Web. Volume 775. DEU; Aachen, Germany: 2011. A Semantic Model for Integrated Content Management, Localisation and Language Technology Processing; pp. 38–49.

Alexiev V. Museum linked open data: Ontologies, datasets, projects. Digit. Present. Preserv. Cult. Sci. 2018;VIII:19–50.

Rocha Souza R., Dorn A., Piringer B., Wandl-Vogt E. Towards A Taxonomy of Uncertainties: Analysing Sources of Spatio-Temporal Uncertainty on the Example of Non-Standard German Corpora. Informatics. 2019;6:34. doi: 10.3390/informatics6030034. - DOI.