Candela, G, Pereda, J, Saez, D, Escobar, P, Sanchez, A, Villa Tores, A, Palacios, AA, McDonough, K and Murrieta-Flores, P

An ontological approach for unlocking the Colonial Archive

http://researchonline.ljmu.ac.uk/id/eprint/19227/

Article

An ontological approach for unlocking the Colonial Archive

GUSTAVO CANDELA, Universidad de Alicante, Spain
JAVIER PEREDA, Liverpool John Moores University, United Kingdom
DOLORES SÁEZ, Universidad de Alicante, Spain
PILAR ESCOBAR, Universidad de Alicante, Spain
ALEXANDER SÁNCHEZ, Universidad de Alicante, Spain
ANDRÉS VILLA TORRES, evulpo AG, Switzerland
ALBERT A. PALACIOS, University of Texas at Austin, USA
KELLY MCDONOUGH, University of Texas at Austin, USA
PATRICIA MURRIETA-FLORES, Lancaster University, United Kingdom

Cultural Heritage institutions have been exploring new ways of making available their catalogues in digital format. Recently, new approaches have emerged as methods to reuse and make available the contents for computational purposes. This work introduces a methodology to transform digital collections into Linked Open Data following best practices. The framework has been applied to Indigenous and Spanish colonial archives based on the collection *Relaciones Geográficas of Mexico and Guatemala* provided by the LLILAS Benson Latin American Studies and Collections. The results of this work are publicly available. This work aims at encouraging Cultural Heritage institutions to publish and reuse their digital collections using advanced methods and techniques.

CCS Concepts: • **Information systems → Semantic web description languages**; • **Computing methodologies → Artificial intelligence**; **Computer vision**; **Machine learning**.

Additional Key Words and Phrases: linked open data, metadata, collections as data, knowledge graph

ACM Reference Format:
Gustavo Candela, Javier Pereda, Dolores Sáez, Pilar Escobar, Alexander Sánchez, Andrés Villa Torres, Albert A. Palacios, Kelly McDonough, and Patricia Murrieta-Flores. 2022. An ontological approach for unlocking the Colonial Archive. *Proc. ACM Meas. Anal. Comput. Syst.* 37, 4, Article 111 (August 2022), 19 pages. https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

During the last decade, Cultural Heritage (CH) organisations have been publishing and providing access to their catalogues in digital format. Recently, new approaches have emerged as methods to reuse and make available the contents for computational purposes [43]. Many organisations, in particular libraries, have adopted the concept of *Lab* to explore the benefits of applying advanced research methods to their collections [34].

Authors' addresses: Gustavo Candela, gcandela@ua.es, Universidad de Alicante, Spain; Javier Pereda, Liverpool John Moores University, United Kingdom, J.Pereda@ljmu.ac.uk; Dolores Sáez, Universidad de Alicante, Spain, md.saez@ua.es; Pilar Escobar, Universidad de Alicante, Spain, mpilar.escobar@ua.es; Alexander Sánchez, Universidad de Alicante, Spain, alexander.sanchez@ua.es; Andrés Villa Torres, evulpo AG, Switzerland, andres.villatorres@evulpo.com; Albert A. Palacios, University of Texas at Austin, USA, aapalacios@austin.utexas.edu; Kelly McDonough, University of Texas at Austin, USA, kelly.mcdonough@utexas.edu; Patricia Murrieta-Flores, Lancaster University, United Kingdom, p.murrieta@lancaster.ac.uk.

2476-1249/2022/8-ART111 $15.00
https://doi.org/XXXXXXX.XXXXXXX

Organizations and research initiatives have explored the benefits of applying the Semantic Web principles to their catalogues and materials [4, 11, 29]. Recent advances in technology based on Artificial Intelligence and Machine Learning have paved the way for a major scientific initiative involving CH institutions [42, 52]. Novel approaches have been proposed to use Semantic Web in combination with AI for Digital Humanities research [25].

The purpose of this work is the transformation of Indigenous and Spanish colonial archives into readable and accessible data using AI technologies, including transcribed texts, linked information, and automated search and analysis of pictorial elements.[1]

The main contributions of this article are as follows: (a) a method to transform a digital collection into Linked Open Data; (b) a Linked Open Data dataset based on the collection *Relaciones Geográficas of Mexico and Guatemala* provided by LLILAS Benson Latin American Studies and Collection from the University of Texas at Austin; c) a compilation of reproducible Jupyter Notebooks to analyse and reuse the dataset; and (d) the results obtained after the application of the method proposed. Furthermore, these contributions are intended to encourage CH organisations to make available their collections as LOD applying Semantic Web technologies, and to provide wider access to researchers, interested public, and indigenous communities.

The paper is organised as follows: after a brief description of the state of the art in Section 2, Section 3 describes the method to transform a digital collection into LOD and its application. Examples of use of the final dataset and a discussion are included in Section 4. The paper concludes with an outline of the adopted methodology, and general guidelines on how to use the results and future work.

2 RELATED WORK

Cultural Heritage institutions have explored the benefits of providing computational access to their collections. In this context, Labs have emerged as a crucial element in organisations to approach external researchers by means of the inclusion of open datasets, the use and publication of APIs and the creation of collaboration research programs. Some examples are the Data Foundry at the National Library of Scotland, the LC Labs at the Library of Congress, and the British Library Labs.

The Semantic Web was introduced by Tim Berners-Lee as an extension of the traditional Web based on standards provided by the W3C such as Resource Description Framework (RDF), Ontology Web Language (OWL) an SPARQL Protocol, and RDF Query Language (SPARQL) to provide machine-readable information [32]. In addition, the author suggested a 5-star deployment schema to make available open data [56] that was extended with two additional stars [26]. Linked Open Data refers to the publication of information under permissive and open licenses, providing URIs to identify resources and including links to external repositories [2].

CH institutions have adopted the Semantic Web principles by making available their contents as LOD using known standard vocabularies such as CIDOC-CRM[2] and linking to external repositories [1, 11, 13]. Libraries have adopted a leading role in this sense providing their catalogues as LOD using different vocabularies. For instance, the Library of Congress has published a Linked Data Service to provide their contents using Bibliographic Framework (BIBFRAME) as main vocabulary.[3] Other approaches such as the National Library of France and Spain have published its catalogues as LOD using a vocabulary based on the Functional Requirements for Bibliographic Records (FRBR) promoted by IFLA [15, 53]. FRBR has been used in other domains such as Functional Requirements for Information Resources for datasets [10]. The British Library published its contents as LOD based on the Bibliographic Ontology (BIBO).[4] Other initiatives such as Europeana and the Digital Public Library of America (DPLA) have adopted the Europeana Data Model (EDM) as the main vocabulary to describe their contents.[5]

[1] https://unlockingarchives.com/

[2] https://www.cidoc-crm.org/

[3] https://id.loc.gov/

[4] https://www.bl.uk/collection-metadata/downloads

[5] https://pro.europeana.eu/page/edm-documentation

Table 1. Overview of CH projects published as LOD.

Institution	Vocabulary	Description
Amsterdam Museum	EDM	Artists and works
Biblioteca Nacional de España	FRBR	Spanish literature
Biblioteca Virtual Miguel de Cervantes	RDA	Spanish literature
Bibliothèque Nationale de France	FRBR	Metadata about authors, works and topics
British Library	BIBO	British National Bibliography
Digital Public Library of America	EDM	American works
Europeana	EDM	European works
Library of Congress	BIBFRAME	Authors and works
Museo del Prado	CIDOC-CRM	Artists and works
RijksMuseum	EDM	Artists and works
Smithsonian American Art Museum	CIDOC-CRM	Artists and works
Spanish Civil War research project	EDM, CIDOC-CRM	Photographic archives
WarSampo knowledge graph	CIDOC-CRM	Data about World War 2

Furthermore, reference models such as the Open Archive Information System (OAIS) was developed to identify how these systems that hold the information will interact to manage the data objects. Within the OAIS protocol, it is paramount to implement the correct knowledge standardisation through the use of such controlled vocabularies to help maintaining the long-term preservation of these digital collections [31].

The Smithsonian American Art Museum has made available information about artists and works as LOD using CIDOC-CRM to foster new opportunities for discovery, research, and collaboration [50]. The Museo del Prado in Spain has adopted CIDOC-CRM and FRBR as the metadata model to describe its collection.[6] The Amsterdam Museum Linked Open Data set is a five-star Linked Data representation and comprises the entire collection of the Amsterdam Museum [12]. In addition, Linked Art is a community of museum and CH professionals collaborating to define a data model to describe Art.[7]

National initiatives have explored the benefits to use LOD in Digital Humanities such as the Linked Open Data Infrastructure for Digital Humanities (LODI4DH) in Finland [24] that has made available the WarSampo knowledge graph, a LOD service based on CIDOC-CRM for publishing data about World War 2, with a focus on Finnish military history [29]. Another approach is based on the Spanish Civil War including a photographic archive as LOD [48].

The Linked Open Data Cloud presents datasets with different domains (geography, media, linguistics, etc.) that have been published using the Linked Data format.[8] In addition, Wikidata provides a section that includes SPARQL endpoints of a list of LOD projects that have been linked from the platform.[9] Table 1 shows an overview of the projects and vocabularies used to make available digital collections and materials as LOD using the Semantic Web principles.

The enrichment of the datasets with external repositories has become a challenge for the research community. The use of advanced techniques such as Named Entity Recognition (NER) and Entity Linking enables the recognition of entities in the text [5, 30]. In this sense, repositories such as DBpedia and Wikidata have been

[6] https://www.museodelprado.es/en/grafo-de-conocimiento/modelo-ontologico

[7] https://linked.art

[8] https://lod-cloud.net

[9] https://www.wikidata.org/wiki/Wikidata:Lists/SPARQL_endpoints

extensively used by CH organisations to enrich their catalogues and to create links to resources such as authors and works.[10] Wikibase is a proven and maintained open software for generic knowledge base maintenance that has been previously tested in CH organisations [21, 28].

Advances in technology have provided a wide variety of tools to transform the original sources into LOD, but also to enrich them with external repositories, reproduce the transformation process and assess their quality. For example, OpenRefine[11] is a tool for cleaning, transforming and enriching data with external services. Many modules have been developed as Python packages such as RDFLib[12] to work with RDF, and spaCy,[13] NLTK and Gensim [46] for natural language processing purposes. With regard to reproducibility, Jupyter[14] has emerged as a powerful tool to enable researchers to reproduce the results [7]. In terms of quality, several approaches have proposed data quality criteria for LOD as well as methods based on the use of SPARQL and Shape Expressions (ShEx) to describe and assess LOD quality [6, 16]. Regarding the storage of RDF, many organisations have used Virtuoso providing a public SPARQL endpoint. Other initiatives are based on dump files and additional storage systems such as RDF4J[15] and Jena.[16] Although the recommendation by the W3C is to reuse standard vocabularies instead of creating new ones, Protégé[17] enables researchers to create their own ontologies, as well as importing previously generated ones and modify them.

Previous works have focused on the challenges for the next generation of Web technologies based on the Semantic Web and AI [22, 25, 40]. The combination of Machine Learning and Semantic Web in order to facilitate the work of curators in CH organisations have been explored in the past [3]. In addition, previous works have measured the impact of the Optical Character Recognition (OCR) errors produced by the digitisation systems on natural language processing (NLP) tasks [51].

Despite the current technological innovations, the vast majority of the technologies used to describe knowledge remain rooted in Western understandings of the world and dominated by a white, patriarchal perspective [14, 47]. These challenges have been flagged up by CIDOC-CRM through the creation of Issue 530: Bias in Information [9]. This points to the inability of models to describe the information of different world views. The absence of diverse resources and non-western ways of thinking, describing and classifying results in the effect of *scarcity bias*. While this issue is increasingly recognised, it is still assumed that technologies such as AI and Machine Learning will be able to fill knowledge gaps in some way, whilst the research is carried out in isolated (often Anglo-centric) sylos. This is known to result in workflows limited by the research objectives of creators that are usually embedded in western systems [14], and commonly limited to the cosmovision of the Global North. While the above-mentioned efforts provide an extensive demonstration of how to transform, publish, assess and reuse LOD, to our best knowledge, none of the previous approaches have considered the use of Indigenous and Spanish colonial archives to apply innovative research methods. Working with indigenous and colonial datasets from the region once known as Mesoamerica expands efforts to a vast territory with shared cosmovision and biopolitical relationships, interconnecting 364 indigenous languages and variants and at least 68 different cultures. This covers a large geopolitical territory from North America, to parts of Central America and the Caribbean. This research uses as an example the *Relaciones Geograficas de Nueva España* (or the Geographic Reports of New Spain). This collection is the result of a survey with thousands of pages describing the situation of indigenous peoples and Europeans at the time. Spanish and Indigenous nobles, rulers, administrators, and

[10] See, for example, the property British Library system number at Wikidata https://www.wikidata.org/wiki/Property:P5199.

[11] https://openrefine.org/

[12] https://rdflib.readthedocs.io/en/stable/

[13] https://spacy.io/

[14] https://jupyter.org/

[15] https://rdf4j.org/

[16] https://jena.apache.org/

[17] https://protege.stanford.edu

scholars participated in this effort which included a rich view of different ethnic groups and diverse cosmovisions held in this territory during the sixteenth century.

This work will be useful in the efforts to diversify views in technological development, and for the CH community to identify best practices and guidelines regarding the publication and reuse of LOD datasets. This research might also open the door to the use of other advanced computational methods, such as Computer Vision (CV) approaches to integrate a body of commonly non-accessible knowledge into the LOD and AI ecosystems, also helping in the creation of new understandings about this historical period.

3 A LLILAS BENSON CASE STUDY: THE SIXTEENTH-CENTURY GEOGRAPHIC REPORTS OF NEW SPAIN

This section presents the method used to transform a Cultural Heritage digital collection into RDF and the datasets provided by the LLILAS Benson as part of the project *Unlocking The Colonial Archive: Harnessing Artificial Intelligence for Indigenous and Spanish American Collections* funded by the AHRC, UK Research and Innovation (UKRI) and the US National Endowment for the Humanities (NEH).

The project includes several datasets from relevant organisations such as the LLILAS Benson library (US) and the General Archive of the Nation (Mexico). In particular, this example is based on the collection of *Relaciones Geográficas de Nueva España* belonging to Mexico and Guatemala held by the LLILAS Benson. As said, this is a corpus of historical documents and paintings (maps) regarded as one of the most important datasets for the history of Early Colonial Mexico and Guatemala. Compiled from 1577 to 1585, the dataset is the result of a questionnaire ordered by Philip II, the king of Spain at the time. In 50 questions, the mandate ordered to collect all sorts of relevant information about people, ways of life, health, economic and cultural resources, languages, climate, social, economic, and military organisations, as well as available foods, plants, and animals, among many others. As such, the collection tends to be one of the major sources used by historians, archaeologists, and anthropologists interested in the history of Mexico and Latin America. Therefore, its importance cannot be understated. Furthermore, written in a combination of Spanish with around 69 different indigenous languages, and containing geographic and historical information for thousands of towns and villages, the Relaciones have always been of interest to indigenous communities, that until recently, have usually only been available to specialists. The LLILAS Benson dataset contains 81 items including the textual reports and its maps, and is available online using a CC0 license.[18] The metadata of each record is provided as JSON, including all the information related to the item, such as title, source collection, publication date, description and list of authors.

3.1 A method to transform a Cultural Heritage digital collection into RDF

Following previous approaches, the method proposed to transform a Cultural Heritage digital collection into RDF works in 6 steps: i) identification of the original sources; ii) extraction of metadata; iii) data mapping to the main vocabulary; iv) generation of RDF; v) data enrichment; and vi) quality assessment.

The identification of the original sources includes several aspects, such as how the data has been made available for the public and the license. In general, datasets are available as dump files (e.g., zipped files), but organisations are starting to adopt APIs to expose data enabling developers to access their contents and perform a wide range of *scientific examination* and analyses. Some examples of API include International Image Interoperability Framework (IIIF), Open Archive Initiative-Protocol for Metadata Harvesting (OAI-PMH) and advanced initiatives such as SPARQL. With regard to the licenses, datasets may be available using permissive and open licenses such as CC0 and CC-BY, but in other cases, the license may not be clear or based on national regulations. For example, the Data Foundry at the National Library of Scotland provides datasets under a CC0 license. Additional aspects to consider to select a dataset include coverage to provide information about missing items or periods, data quality,

[18] https://collections.lib.utexas.edu/.

trustworthiness (e.g., manual or automated metadata curation) and particular features such as the type of contents (e.g., text, image and map), and the metadata format used to describe the contents (ALTO, XML, Dublin Core, etc.). Most importantly, enabling access to the data through the APIs can facilitate the implementation of interactive tools for engagement, citizen science and crowdsourcing processes, that can help promote its findability and reusability.

The following step corresponds to the extraction of metadata from the original sources. In general, and depending on the content, CH organisations provide the datasets using different formats such as plain text, comma-separated values (CSV), JavaScript Object Notation (JSON) and Extensible Markup Language (XML). Some examples are A Medical History of British India provided by the Data Foundry,[19] Chronicling America[20] and the Museum of Modern Art (MoMA) [41]. More advanced approaches are based on RDF, a standard model for data interchange on the Web promoted by the W3C. RDF enables the definition of triples subject-predicate-object to add information to resources using predicates (e.g., Shakespeare is_the_author_of Hamlet). The datasets may be published as dump files or using a SPARQL endpoint. For instance, the Library of Congress provides bulk download files for their LOD datasets.[21] Some examples of datasets providing a SPARQL endpoint include the Smithsonian American Art Museum and the World War I as Linked Open Data from the Linked Data Finland platform.[22]

Each format can be read and analysed using a particular open-source Python or Java module. For instance, and regarding the different formats, many packages have been published for Python. In this sense, tabular data published as CSV can be analysed using Pandas. XML can be read and analysed with a wide range of libraries, such as Pymarc for MARCXML. RDFLib is a pure module for working with RDF including parsers, serializers and a SPARQL implementation.

The data mapping step requires the knowledge of several technologies (e.g. RDF and OWL) to understand the selected vocabulary to define the resources correctly. Each vocabulary may include different classes and properties to define the final dataset. For example, Figure 1 shows the main classes and properties defined in the EDM vocabulary. Other initiatives present additional approaches that differentiate in terms of the level of granularity to describe the works providing a hierarchy based on different classes. For instance, FRBR and RDA provide four classes (Work, Expression, Manifestation and Item) to define the resources, while BIBFRAME provides two classes (Work and Instance). In both cases, the top-level class Work reflects a conceptual entity, and the rest, the materialization of the individuals that can be physical or digital in nature. However, in some cases, the separation using different classes can be inconvenient or even impossible to achieve. In case a new vocabulary is required, tools such as Protégé provides an editor to define the classes and properties.

The generation of RDF is based on the original dataset and the data mapping. Tools such as OpenRefine enable the mapping between an original dataset provided as a CSV or JSON file and the selected vocabulary by using its interface. Each metadata column is mapped to a property defined by the vocabulary. For instance, a title provided as a column in a CSV file is mapped to a Dublin Core property `dc:title`. OpenRefine supports GREL[23] to manipulate strings and enables the previsualisation of the transformation to preview the final result. Other approaches to create RDF are based on Python and Java modules such as RDFLib and Jena.

The data enrichment is performed using external repositories. In terms of how the links are created, OWL provides the property `owl:sameAs` to define the external links. Other approaches are based on the creation of particular properties such as the case of Wikidata to link the resources provided by a dataset (e.g., authors and works). Additional datasets can be used for different domains such as Virtual International Authority File (VIAF)

[19] https://doi.org/10.34812/2w0t-3f08

[20] https://chroniclingamerica.loc.gov/help/

[21] https://id.loc.gov/authorities/subjects.html

[22] https://www.ldf.fi/dataset/ww1lod/index.html

[23] https://docs.openrefine.org/manual/grelfunctions

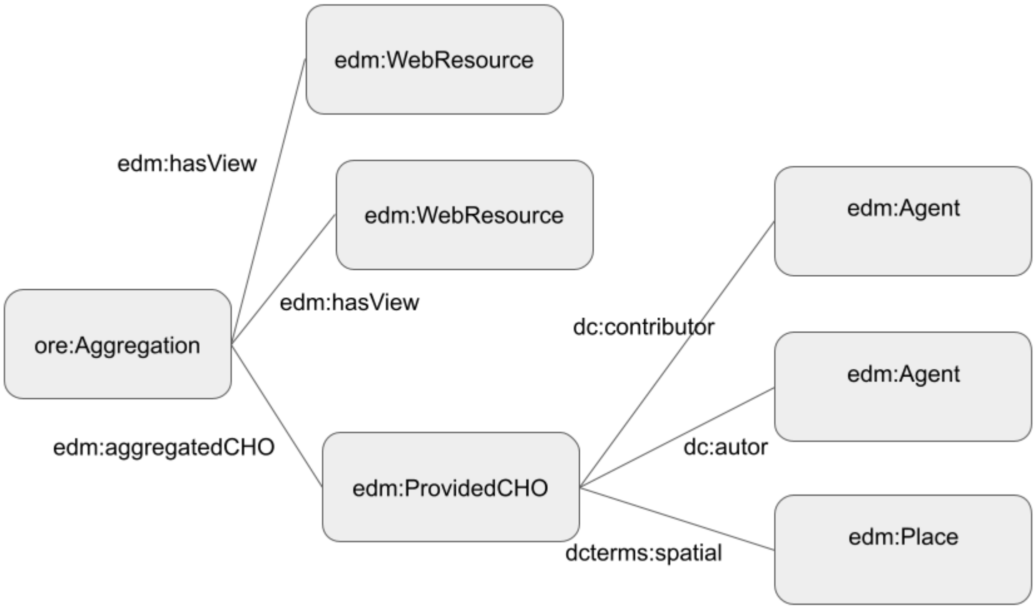


Fig. 1. Representation of the main classes for resources in the EDM vocabulary EDM made available by Europeana.

for authors and GeoNames for geographic locations. While OpenRefine includes reconciliation services for knowledge bases such as Wikidata, there are additional tools based on Entity Linking methods to assign a unique identity to entities mentioned in the text. Previous works are based on DBpedia and Wikidata knowledge bases [30, 35].

Previous approaches have proposed a criteria to assess the quality of LOD including several aspects. In that sense, SPARQL can be used to query the data and test the quality (e.g., number of resources typed as Person or the list of properties used by a particular class) [16]. For instance, Listing 1 shows an example of SPARQL query to retrieve the number of resources typed as `edm:Place`. Other approaches are based on advanced assessment methods such as ShEx that provides a human-readable text to assess and describe RDF [6]. A ShEx schema provides node and triples constraints defining the structure of a resource as RDF data. Listing 2 shows an example of ShEx schema to assess a resource using EDM as vocabulary and typed as `edm:Place`.

```
SELECT (COUNT(DISTINCT ?s) as ?total)
WHERE {
  ?s a edm:Place
}
```

Listing 1. A SPARQL query to retrieve the number of resources using EDM as vocabulary and typed as `edm:Place`.

```
edm-shex:Plac
e{
    rdf:type  [edm:Place]  ;
    geo:long   xsd:string   +;
    geo:lat   xsd:string   +;
    skos:prefLabel   rdf:langString   +
}
```

Listing 2. A ShEx to validate a place described using EDM as vocabulary.

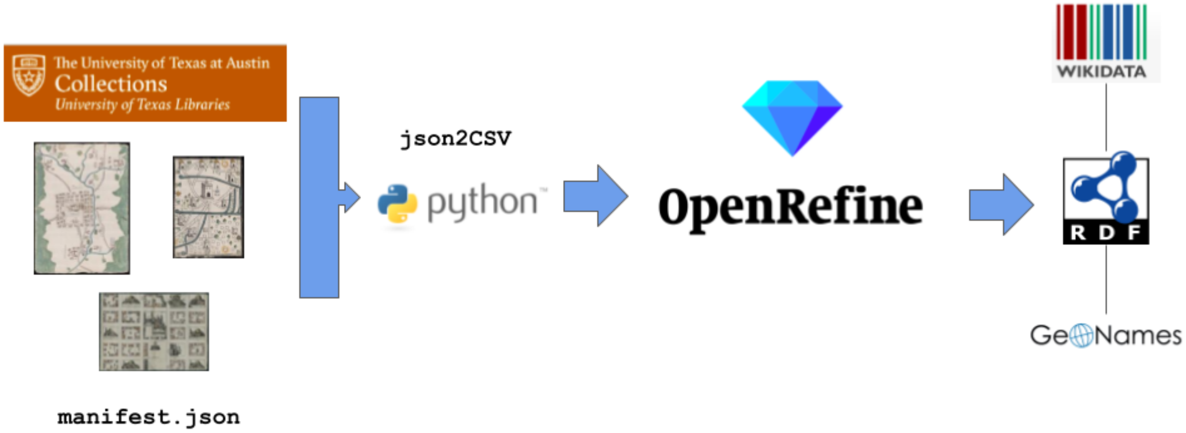


Fig. 2. Methodology to transform the original sources into RDF using open source tools.

Compared to other approaches, and as an example of application of the method described in Figure 2, the transformation into RDF of the collection *Relaciones Geográficas de Nueva España* is based on the following features:

- the metadata has been extracted and mapped to EDM as main vocabulary by means of OpenRefine. EDM has been selected since it is an international initiative used by relevant institutions and the original sources are based on bibliographic information. Table 2 shows the URL patterns used to create the RDF resources according to the EDM vocabulary and Figure 3 shows a representation of a record as EDM vocabulary.
- the dataset has been enriched with external repositories such as Getty Vocabularies, Wikidata and GeoNames. For instance, Getty has been used for the properties `dc:coverage` and `edm:hasType` to enrich the final dataset with entities related to centuries and type of content.[24] In addition, geographic locations have been linked to GeoNames and Wikidata using the property `owl:sameAs`.

[24] See, for example, the resources sixteenth century http://vocab.getty.edu/aat/300404510 and historical maps http://vocab.getty.edu/aat/300028233 available at Getty Vocabularies.

- a data quality assessment has been performed using ShEx since it provides a human-readable text description of the datasets while is a recent method to assess RDF. This step includes two tasks: i) automatic extraction of the ShEx schemas using SheXer [17]; and ii) using the Python module PyShEx to test the ShEx schemas against a random sample of the RDF data.
- the dataset has been described using the Vocabulary of Interlinked Datasets (VoID) concerned with metadata about RDF datasets [55]. Table 3 shows an overview of the final dataset.
- a code repository is available at GitHub including all the scripts developed and the final dataset.[25] Following guidelines and best practices, a reproducible and runnable on the cloud collection of Jupyter Notebooks is provided as part of the project to show how the final dataset can be explored.

Table 2. URL patterns for the final dataset.

Class	URL pattern
edm:Place	../place/id
ore:Aggregation	../aggregation/id
edm:ProvidedCHO	../cho/id
edm:Agent	../agent/id

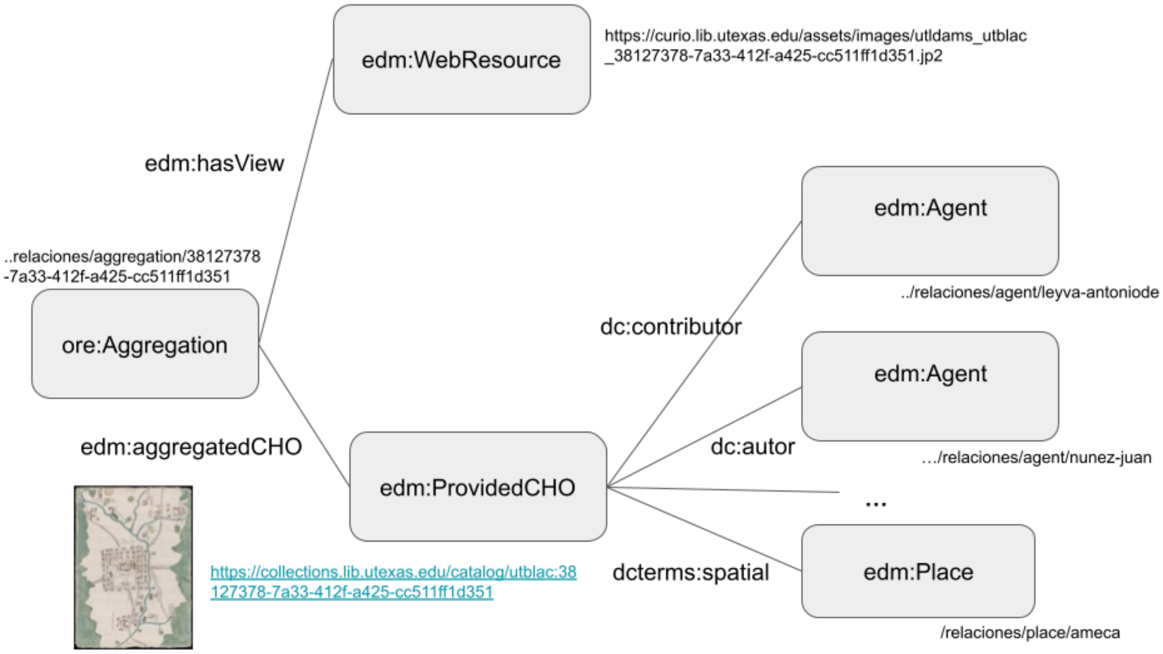

edm:WebResource
https://curio.lib.utexas.edu/assets/images/utldams_utblac
_38127378-7a33-412f-a425-cc511ff1d351.jp2
edm:hasView
..relaciones/aggregation/38127378
-7a33-412f-a425-cc511ff1d351
ore:Aggregation
edm:Agent
../relaciones/agent/leyva-antoniode
dc:contributor
edm:Agent
dc:autor
.../relaciones/agent/nunez-juan
edm:aggregatedCHO
edm:ProvidedCHO
...
https://collections.lib.utexas.edu/catalog/utblac:38
127378-7a33-412f-a425-cc511ff1d351
dcterms:spatial
edm:Place
/relaciones/place/ameca

Fig. 3. Representation of a record using EDM as main vocabulary.

[25] https://github.com/hibernator11/UCA-relacionesgeograficas

Table 3. Overview of the final dataset.

Description	N. of items
N. of triples	3767
N. of properties	34
N. of classes	7
O. of external links	125

4 EXPLORING AND REUSING THE DATASET AS LOD

This section introduces examples of reuse based on the LOD repository created in this work. In addition, a detailed description of how the data can be used to generate new knowledge is presented.

4.1 Geographic locations: GeoNames and Wikidata

Previous works and projects have explored the use of a historical gazetteer for historical places such as the DECM Historical Gazetteer including the Relaciones Geográficas [38]. Other approaches are based on the use of reproducible Jupyter Notebooks to provide a zommable an interactive map [7].

This example reuses the data provided as RDF to query the dataset in order to retrieve the latitude and longitude about the resources typed as `edm:Place`. Listing 3 shows the SPARQL query to retrieve the geographic information. Then, the data is represented as a map using the Python module folium[26] as is shown in Figure 4. Each point in the map provides a link to Wikidata. The representation has been included as a Jupyter Notebook.

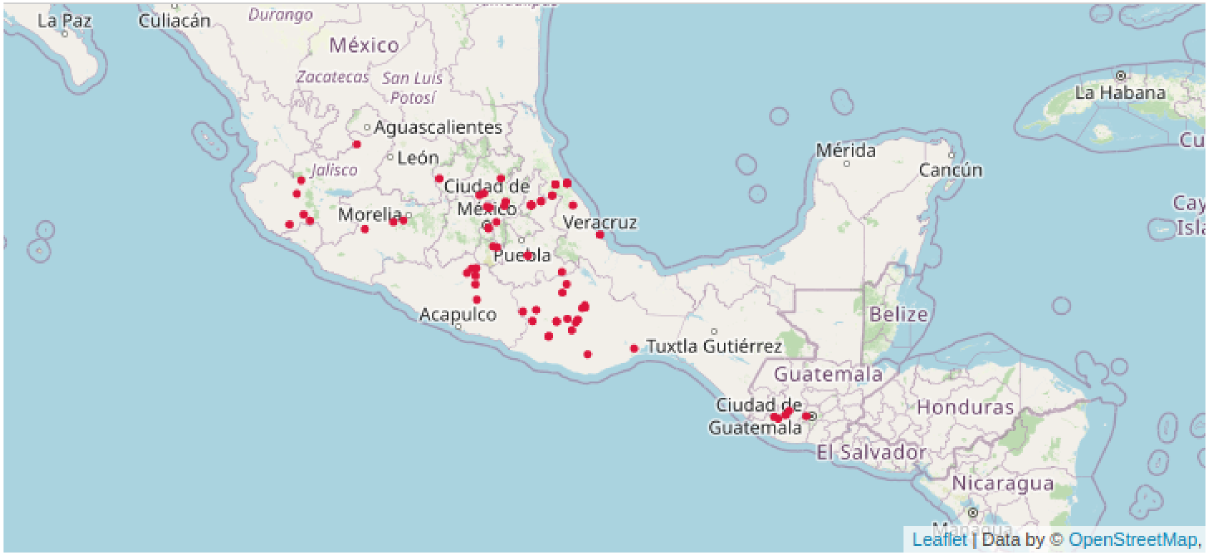


Fig. 4. Map representing the geographic locations provided by the original sources. This map has been generated using the Python module folium.

[26] http://python-visualization.github.io/folium/

```
PREFIX skos: <http://www.w3.org/2004/02/skos/core#>
PREFIX edm: <http://www.europeana.eu/schemas/edm/>
PREFIX wgs: <http://www.w3.org/2003/01/geo/wgs84_pos#>

SELECT DISTINCT ?p ?lat ?long ?lbl ?wikidata ?geonames
WHERE {
    ?p rdf:type edm:Place .
    ?p skos:prefLabel ?lbl .
    ?p wgs:long ?long .
    ?p wgs:lat ?lat .
    ?p owl:sameAs ?wikidata .
    FILTER ( strstarts(str(?wikidata), "https://www.wikidata.org/wiki/") ).
    ?p owl:sameAs ?geonames .
    FILTER ( strstarts(str(?geonames), "https://www.geonames.org/") )
}
```

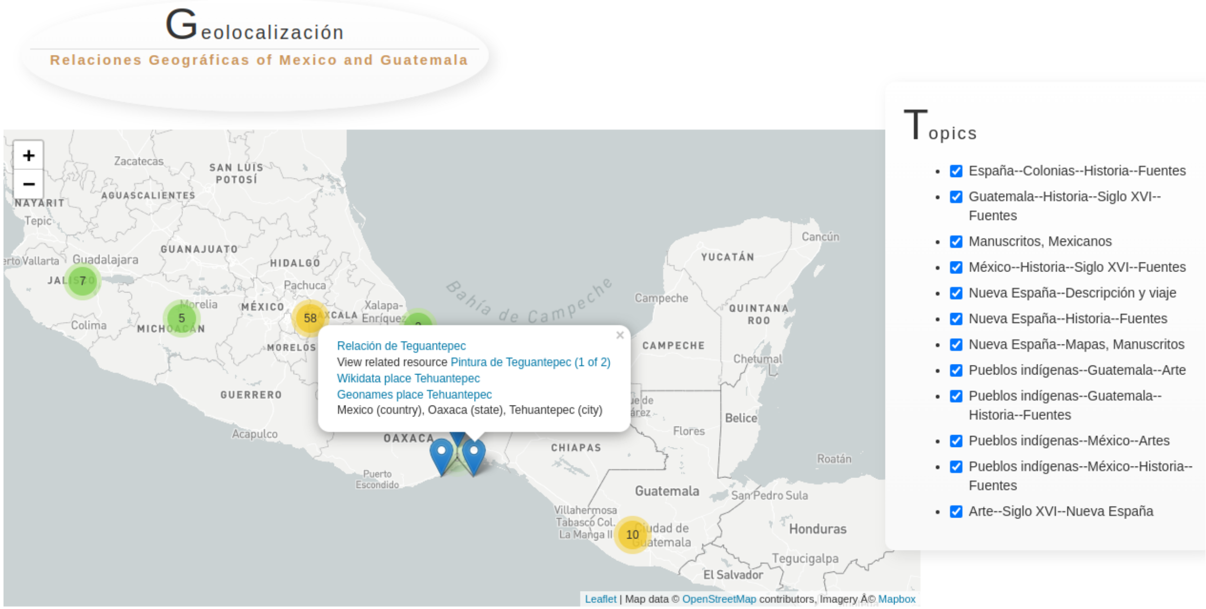
Listing 3. SPARQL query to retrieve the geographic information in the dataset.

A refined version of the map enables the user to select the subjects and is shown in Figure 5. A user-friendly interface has been developed that shows the works in the dataset according to their location (`edm:Place`) and subject (`dc:subject`) based on an interactive map. Depending on the selected topics, each point on the map presents the number of records. By zooming in we can see information about the work as well as the links to Wikidata and GeoNames. The following steps have been followed: i) a CSV file has been generated with the data for visualization through an Extract Transform and Load (ETL) process; ii) the different topics used for faceting the data have been retrieved; and iii) the data and topics have been loaded into a map using LeafLet[27] –an open source JavaScript library–, OpenStreetMap[28] –a collaborative project to create free and editable maps– and Mapbox[29] to customize the map.

4.2 Generating new knowledge from old sources - Reutilisation and Third Party Use

While the method presented here is capable of transforming available metadata into LOD, the research community is working to achieve the same results for information contained within these collections that, to a certain extent remain "hidden", only available to highly skilled specialists, or that are difficult to process through traditional means of research due to the sheer volume of information. Current approaches using AI techniques in the Digital Humanities have proven to be successful in this regard, automatically identifying, cross-referencing and analysing very large volumes of data of historical interest from these collections using Natural Language Processing techniques, and are achieving an acceleration in the generation of knowledge [39]. However, although the nature of historical information is to be linked (for example, historical figures such as the conquistador Hernando Cortés tend to be mentioned in many different documents, or many documents might relate to one place), researchers can only make these connections through painstakingly reading thousands of pages. These connections might then be recorded in a book or article, and they remain again "locked" until someone else discovers them either in such publications or the historical documents again. In this sense, linking data from information contained in digital collections beyond metadata has the potential to unleash new discoveries and

[27] https://leafletjs.com/
[28] https://www.openstreetmap.org/
[29] https://www.mapbox.com/



Relación de Teguantepec
View related resource Pintura de Teguantepec (1 of 2)
Wikidata place Tehuantepec
Geonames place Tehuantepec
Mexico (country), Oaxaca (state), Tehuantepec (city)

Topics

- ☑ España--Colonias--Historia--Fuentes
- ☑ Guatemala--Historia--Siglo XVI--Fuentes
- ☑ Manuscritos, Mexicanos
- ☑ México--Historia--Siglo XVI--Fuentes
- ☑ Nueva España--Descripción y viaje
- ☑ Nueva España--Historia--Fuentes
- ☑ Nueva España--Mapas, Manuscritos
- ☑ Pueblos indígenas--Guatemala--Arte
- ☑ Pueblos indígenas--Guatemala--Historia--Fuentes
- ☑ Pueblos indígenas--México--Artes
- ☑ Pueblos indígenas--México--Historia--Fuentes
- ☑ Arte--Siglo XVI--Nueva España

Fig. 5. Interactive map representing the geographic locations in which the user is able to select the subjects.

achieve the identification of previously unseen patterns. In the case of pictorial documents, further challenges might also emerge. Before the arrival of the Europeans to America, Mesoamerican cultures used an expression and communication system based on orality and paintings. On the one hand, stories or events, mythical and theological understandings, genealogies, and history were preserved through the *tlamatinime* or wise men and women that used orality, music and dance as mnemonic devices, preserving the community's memory. On the other, this knowledge was also depicted through the creation of images where pictography, ideographic symbolism, and phonetic mediation were combined with the use of composition, colour, and the spatiality of the elements in the paintings to tell a story [27]. This pictorial discourse existed in parallel with the oral one, making one think in images rather than words, although the codices could be also used as part of oral performances. Therefore, the Nahua (Aztec) codices, for instance, express very complex sets of information. For example, research by Garduño-Monroy [20], has showcased this by identifying the complexity in the depiction of natural processes and events in the Telleriano-Remensis Codex. The document recorded earthquakes through the amalgamation of the glyphs of movement (*ollin*) and earth (*tlalli*). Furthermore, the glyph *tlalli* is also used as a quantitative value measuring seismic intensity (see Figure 6). The fact that the Aztecs had a form of Richter Scale, is widely unknown, but even if technologies such as Geographic Information Systems can be used to record this kind of seismic data, as many other current technologies, they are not ready to ingest or make use of indigenous knowledge and cosmovisions (views and ways of understanding the world) such as the ones expressed by the Mesoamerican codices.

Another example is the maps of the *Relaciones Geográficas*. These paintings were the result of the Relaciones questionnaire asking for a map of each of the towns it reached. While the Spanish crown expected maps in the cartographic tradition of the time, what it got in response was a combination of the Mesoamerican spatial tradition of painting codices with the new European conceptions of space (see Figure 7). These maps contain important information that, beyond historical interest, can have legal implications and are still used today by

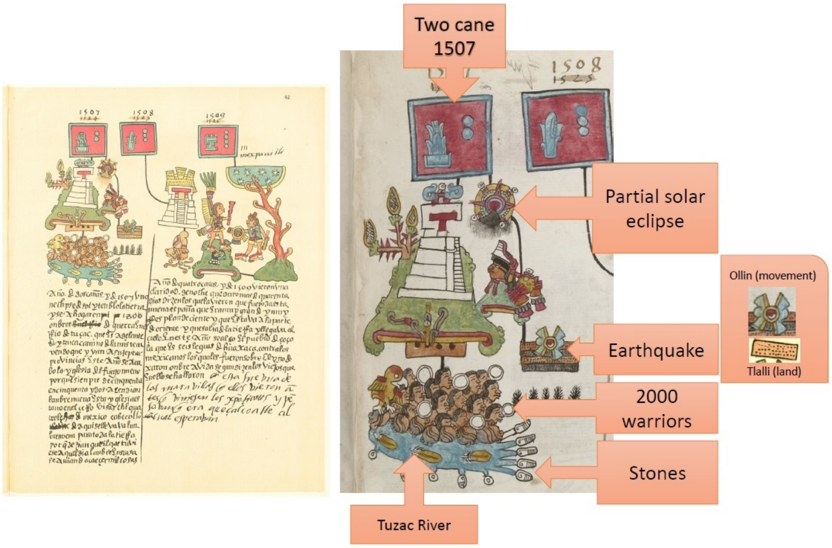


Fig. 6. Codex Telleriano-Remensis (Paris, Bibliothèque nationale de France, MS Mexicain 385 (Codex Telleriano-Remensis), ca. 1563, fol. 42r. Courtesy of gallica.bnf.fr / BnF. Interpretation of codex by Garduño Monroy [20].

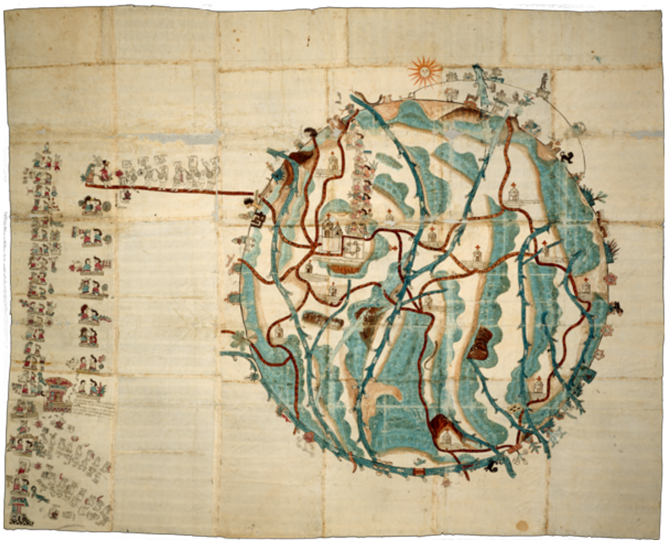


Fig. 7. The Relación Geográfica de Teozacoalco map. Courtesy of the Benson Latin American Collection, University of Texas at Austin (JGI XXV 03).

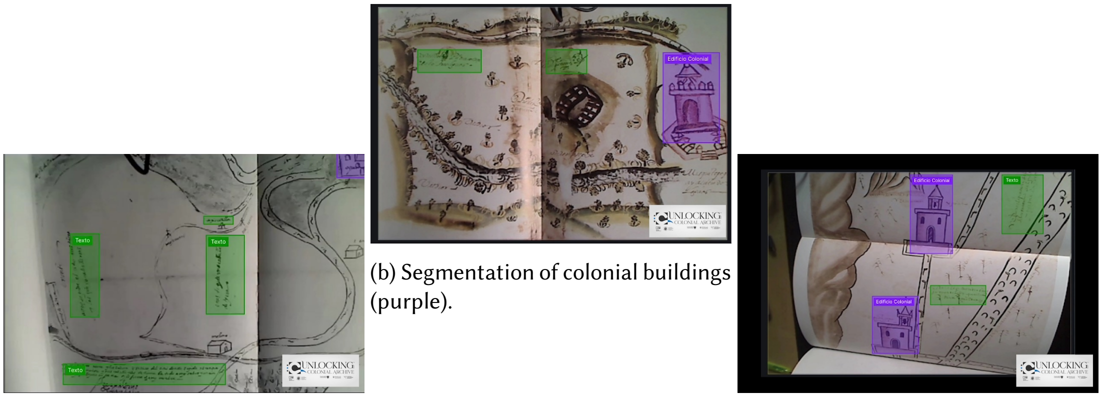
indigenous communities for religious or social purposes, as well as to defend their rights and lands [49]. The maps contain a wide range of features such as toponymic glyphs painted in the Mesoamerican style, place names and annotations already in Latin scripture, and another iconography that is shared by many of these. As such, the historical and social information contained in these can be of great importance. However, this knowledge remains to a certain extent, invisible, either accessible just to a few specialists.

Regarding textual sources, one of the main benefits of automating entity recognition through machine learning processes is that it can help expand access to information in archives, whilst also diversifying it [33]. In this realm, our work has previously contributed to the creation of novel approaches combining methods from Corpus Linguistics, Natural Language Processing and Geographic Information Sciences, where the method called Geographical Text Analysis enables the automated identification and analysis of historical information at a large scale (i.e. the identification of entities within in thousands of documents, that otherwise would take a lifetime to explore) [37]. The consideration of information classification in this research has opened the door for an ontological change, both in a philosophical and computational form. In the same way, research in the Unlocking the Colonial Archive project is looking to extend this kind of work to *Mesoamerican* visual language and cosmovisions. For this, there is previous evidence of the benefits of the combination of image captioning metrics [23, 45], where CV approaches can help generate descriptions of the visual sources. However, CV methodologies are primarily successful with natural images, which are pre-iconologic. Previous studies in iconology by Panofsky [44], have highlighted three core layers of analysis when approached through CV. First, there is a Pre-Iconographical Description, which deals with natural subject matters such as factual or expressional elements. On a second layer, an analysis takes place concerning the allegories and stories of the subject. This is the case with artistic paintings, which display representations of the world. Finally, on a third layer, there is an iconographical interpretation, that aims to flesh out the intrinsic meaning and the values of symbols. CV approaches that engage with Mesoamerican paintings and writings have to engage with this third layer of interpretation, where as in the case of the historical texts, this knowledge is invisibilized and accessible to just a few specialists.

The application of multimodal CV in order to integrate vision and language is an ongoing research challenge [8, 19, 54]. Following previous approaches, we have begun experimenting with CV techniques to detect key iconography that describes cultural elements and visual landmarks, as well as the *glosas* or small textual inscriptions that sometimes accompany these paintings and codices. In this case, we first carried out an experiment to create artistic representations using map paintings from *Mapilu*, a collection of maps from the National Archive of Mexico and the *Relaciones* collection from LLILAS Benson. This experiment aimed to detect and replicate key iconography elements such as water, roads and a type of settlement called *estancias*. We used the Pix2Pix[30] model to translate diverse locations from current Google Maps satellite images into the representation of the maps from our collections (see Figure 9). A second experiment used the YOLOV4[31] model to automatically detect *estancias*, architectural features such as churches and houses, and calligraphic elements (see Figure 8). By identifying these, we will have the ability to produce data objects that can be made reusable for machine and human use. For instance, key settlements and historical place names can be linked into historical gazetteers and become part of the scientific examination process, as well as data that can be used by indigenous communities. Furthermore, by being able to identify and compare similar iconography and representations from hundreds of maps, historians and archaeologists might not only be able to locate specific features and archaeological sites in the landscape, but also understand how this tradition of "writing by painting" changed when it intertwined with the European cosmogony. In this way, connecting these kinds of information in LOD opens a variety of possibilities. This approach will allow us to identify features of interest that would either take a lifetime, or that were impossible before due to the scale of the collections. This will not only have a substantial impact on research in fields such as historical and human geography, archaeology, history, and anthropology, but it will also help computer science diversify knowledge while expanding the accessibility of these datasets for indigenous communities.

[30] https://phillipi.github.io/pix2pix/

[31] https://github.com/maudzung/YOLO3D-YOLOv4-PyTorch



(a) Segmentation of calligraphic text (green).

(b) Segmentation of colonial buildings (purple).

(c) Live recognition segmentation on a physical codex via webcam.

Fig. 8. Examples of object recognition using YOLOv4.

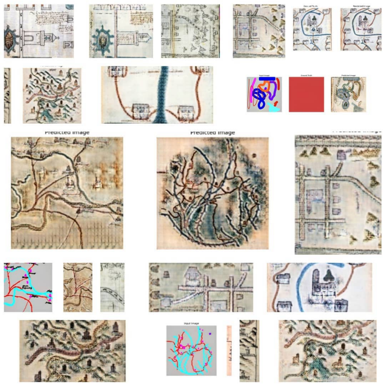


Fig. 9. Annotation and experiments of map data sample for the Pix2Pix model.

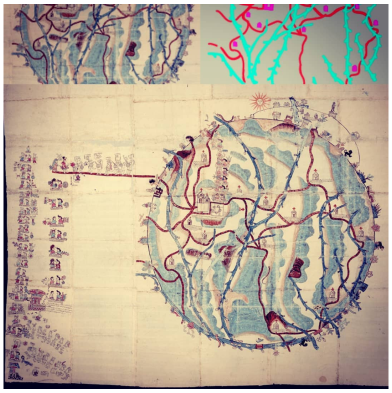


Fig. 10. Image annotation for translation of map of Tezoacalco for Pix2Pix model.

Fig. 11. Image translation of Google Maps using Pix2Pix.

4.3 Discussion

Computational technologies, including, the Web, the Semantic Web, and Artificial Intelligence, have benefited a wide range of communities by enabling access, accelerating knowledge production, and helping preserve the provenance and sustainability of information. However, a wide range of this world's knowledge remains invisible to current technologies. Decoloniality has aimed to make such knowledge visible to displace Western rationality as the only "framework and possibility of existence, analysis and thought" [36, p. 17]. Such invisibility of data is related not only to the technologies and representations of knowledge but also to the tools that make use of these sets of information. Therefore, it is important to begin including counter and alternative narratives to frameworks within AI and other widely spread computational tools. This includes technologies such as Geographic Information Systems, as well as Ontologies and controlled vocabularies to diversify cosmovisions in a way that becomes applicable to our current socio-technical systems. For example, EDM is a vocabulary used by relevant institutions that provides a data model to group resources using views, and to describe authors, places and subjects, among other classes. However, this approach has limitations, and the data mapping could be improved in several ways. For instance, the roles provided as text in the original sources (e.g., signer, artist, interpreter, collector, scribe, contributor, witness, etc.) can be mapped to external vocabularies such as the BnF Roles to enable a better representation and description of the data. In addition, subjects provided as a string in the original sources have been described using the property `dc:subject` and they could be improved by using the vocabulary Simple Knowledge Organization System (SKOS). Similarly, locations are provided as strings describing the country, state and the city (e.g., Mexico (country)| Oaxaca (state)| Santa María Peñoles (city)), and they could be separated to provide individual resources.

With regard to the enrichment, and due to the limited number of resources, locations and authors have been manually identified in external datasets including Wikidata and GeoNames. In addition, centuries provided as text in the original sources have been linked to the appropriate resource at Getty vocabularies using the property `dc:coverage`. This step could be improved by using advanced techniques to apply methods such as Entity Linking.

With regard to the data quality assessment, the ShEx schemas automatically generated were slightly updated in terms of cardinalities. The tool SheXer enables the use of several parameters, such as the number of resources to analyse to generate the ShEx schemas. A low number of resources may result in an incomplete or incorrect ShEx schema.

Finally, this work is just the start, and these are only a few observations, but if scientists are to provide methods to challenge technical and cultural Western hegemony, there is a requirement to collect and analyse new forms of knowledge and provide new imaginative ways of how these data objects can be engaged [14, p. 53]. While this is our aim, larger conversations and work with Global South scholars and communities need to be carried out in order to imagine and create technologies and datasets that are more inclusive of different worldviews. As mentioned before, we have started to work in this area, particularly focusing on the inclusion of more diverse data in the AI and LOD ecosystems, including the creation of decolonial datasets and scarcity biases within ontologies. Our work presents novel means to 'unlocking' non-Western conceptual knowledge and ways of understanding space that have remained inaccessible to the current state-of-the-art computational methods and helps bridging them to further research pipelines such as crowd-sourcing and citizen science methods, geospatial analysis, and the generation and expansion of data modelling in LOD.

5 CONCLUSIONS

In the last decade, there has been a growing interest in making available digital collections published by CH organisations as LOD. Based on previous work, we defined a method described in Section 3 for making available these collections as LOD. The method was applied to an important historical collection published by LLILAS Benson. The discussion explains why this research is important and our evaluation showed several examples of reuse that can be useful to encourage other institutions, particularly those with colonial data, to publish their collections as LOD. Their reuse, enrichment and assessment are becoming increasingly relevant to make the content visible and accessible to both humans and computers, whilst opening the door to the diversification of data and a wide range of knowledge from the Global South, as well as to the creation decolonial computational methods.

Future work to be explored includes the application of the method to additional datasets, the inclusion of the full text and the use of additional vocabularies to describe the contents. In this sense, we are aiming to further develop the ontology for annotation based on previous work carried out by *El Taller*, a group of scholars and *nahuatlatos*[32] formed in 1978 [18]. Their research engaged with multimodal understandings in the depiction of

[32] *Nahuatlatos* are interpreters of Nahuatl and Spanish who help contextualise and translate indigenous knowledge.

toponyms, people names, architectural features, artistic, and phonetic-pictorial elements, among many others in these kinds of documents. Finally, we believe that our research and workflow will help in the integration of knowledge from the Global South in computational research, as well as in the AI and LOD ecosystems.

ACKNOWLEDGMENTS

This research has been supported by UKRI AHRC/NEH Project AH/V009559/1, "Unlocking Colonial Archive: Harnessing Artificial Intelligence for Spanish American Historical Collections".

REFERENCES

[1] Samer Abdallah, Emmanouil Benetos, Nicolas Gold, Steven Hargreaves, Tillman Weyde, and Daniel Wolff. 2017. The Digital Music Lab: A Big Data Infrastructure for Digital Musicology. *J. Comput. Cult. Herit.* 10, 1, Article 2 (jan 2017), 21 pages. https://doi.org/10.1145/2983918

[2] Chris Bizer, Tom Heath, and Tim Berners-Lee. 2008. Linked Data: Principles and State of the Art. http://www.w3.org/2008/Talks/WWW2008-W3CTrack-LOD.pdf

[3] Anna Bobasheva, Fabien Gandon, and Frédéric Precioso. 2022. Learning and Reasoning for Cultural Metadata Quality: Coupling Symbolic AI and Machine Learning over a Semantic Web Knowledge Graph to Support Museum Curators in Improving the Quality of Cultural Metadata and Information Retrieval. *ACM Journal on Computing and Cultural Heritage* 15, 3 (2022), 40:1–40:23. https://doi.org/10.1145/3485844

[4] Gustavo Candela, Pilar Escobar, Rafael C. Carrasco, and Manuel Marco-Such. 2018. Migration of a library catalogue into RDA linked open data. *Semantic Web* 9, 4 (2018), 481–491. https://doi.org/10.3233/SW-170274

[5] Gustavo Candela, Pilar Escobar, Rafael C. Carrasco, and Manuel Marco-Such. 2019. A linked open data framework to enhance the discoverability and impact of culture heritage. *J. Inf. Sci.* 45, 6 (2019), 756–766. https://doi.org/10.1177/0165551518812658

[6] Gustavo Candela, Pilar Escobar, María Dolores Sáez, and Manuel Marco-Such. 2023. A Shape Expression approach for assessing the quality of Linked Open Data in libraries. *Semantic Web* 14, 2 (2023), 159–179. https://doi.org/10.3233/SW-210441

[7] Gustavo Candela, María Dolores Sáez, Pilar Escobar, and Manuel Marco-Such. 2022. Reusing digital collections from GLAM institutions. *J. Inf. Sci.* 48, 2 (2022), 251–267. https://doi.org/10.1177/0165551520950246

[8] Eva Cetinic. 2021. Towards Generating and Evaluating Iconographic Image Captions of Artworks. *J. Imaging* 7, 8 (2021), 123. https://doi.org/10.3390/jimaging7080123

[9] CIDOC-CRM. 2021. Issue 530: Bias in data structure. https://cidoc-crm.org/Issue/ID-530-bias-in-data-structure

[10] Karen Coyle. 2022. Works, Expressions, Manifestations, Items: An Ontology. *Code4Lib Journal* 53 (2022). https://journal.code4lib.org/articles/16491

[11] Edie Davis and Bahareh Rahmanzadeh Heravi. 2021. Linked Data and Cultural Heritage: A Systematic Review of Participation, Collaboration, and Motivation. *ACM Journal on Computing and Cultural Heritage* 14, 2 (2021), 21:1–21:18. https://doi.org/10.1145/3429458

[12] Victor de Boer, Jan Wielemaker, Judith van Gent, Marijke Oosterbroek, Michiel Hildebrand, Antoine Isaac, Jacco van Ossenbruggen, and Guus Schreiber. 2013. Amsterdam Museum Linked Open Data. *Semantic Web* 4, 3 (2013), 237–243. https://doi.org/10.3233/SW-2012-0074

[13] Christophe Debruyne, Gary Munnelly, Lynn Kilgallon, Declan O'Sullivan, and Peter Crooks. 2022. Creating a Knowledge Graph for Ireland's Lost History: Knowledge Engineering and Curation in the Beyond 2022 Project. *J. Comput. Cult. Herit.* 15, 2, Article 25 (apr 2022), 25 pages. https://doi.org/10.1145/3474829

[14] Catherine D'ignazio and Lauren F Klein. 2020. *Data feminism.* MIT press.

[15] Bermes Emmanuelle, Boulet Vincent, and Leclaire Céline. 2016. Améliorer l'accès aux données des bibliothèques sur le web : l'exemple de data.bnf.fr. http://library.ifla.org/1447/1/081-bermes-fr.pdf

[16] Michael Färber, Frederic Bartscherer, Carsten Menne, and Achim Rettinger. 2018. Linked data quality of DBpedia, Freebase, OpenCyc, Wikidata, and YAGO. *Semantic Web* 9, 1 (2018), 77–129. https://doi.org/10.3233/SW-170275

[17] Daniel Fernandez-Álvarez, Jose Emilio Labra-Gayo, and Daniel Gayo-Avello. 2022. Automatic extraction of shapes using sheXer. *Knowledge-Based Systems* 238 (2022), 107975. https://doi.org/10.1016/j.knosys.2021.107975

[18] Joaquín Galarza and Keiko Yoneda. 1982. *Mapa de Cuauhtinchan no. 3:(glifos: catalogo-diccionario).* Vol. 3. Archivo General de la Nación.

[19] Noa Garcia and George Vogiatzis. 2018. How to Read Paintings: Semantic Art Understanding with Multi-modal Retrieval. In *Computer Vision - ECCV 2018 Workshops - Munich, Germany, September 8-14, 2018, Proceedings, Part II.* 676–691. https://doi.org/10.1007/978-3-030-11012-3_52

[20] Víctor Hugo Garduño-Monroy. 2016. Una propuesta de escala de intensidad sísmica obtenida del códice náhuatl Telleriano Remensis. *Arqueologia Iberoamericana* 31 (July 2016), 9–19. https://doi.org/10.5281/zenodo.1318345

[21] Jean Godby, Karen Smith-Yoshimura, Bruce Washburn, Kalan Davis, Karen Detling, Christine Fernsebner Eslao, Steven Folsom, Xiaoli Li, Marc McGee, Karen Miller, Honor Moody, Holly Tomren, and Craig Thomas. 2019. Creating Library Linked Data with Wikibase: Lessons Learned from Project Passage. https://doi.org/10.25333/faq3-ax08

[22] Jim Hendler and Tim Berners-Lee. 2010. From the Semantic Web to social machines: A research challenge for AI on the World Wide Web. *Artif. Intell.* 174, 2 (2010), 156–161. https://doi.org/10.1016/j.artint.2009.11.010

[23] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2021. Clipscore: A reference-free evaluation metric for image captioning. *arXiv preprint arXiv:2104.08718* (2021).

[24] Eero Hyvönen. 2020. Linked Open Data Infrastructure for Digital Humanities in Finland. In *Proceedings of the Digital Humanities in the Nordic Countries 5th Conference, Riga, Latvia, October 21-23, 2020*. 254–259. http://ceur-ws.org/Vol-2612/short10.pdf

[25] Eero Hyvönen. 2020. Using the Semantic Web in digital humanities: Shift from data publishing to data-analysis and serendipitous knowledge discovery. *Semantic Web* 11, 1 (2020), 187–193. https://doi.org/10.3233/SW-190386

[26] Eero Hyvönen, Jouni Tuominen, Miika Alonen, and Eetu Mäkelä. 2014. Linked Data Finland: A 7-star Model and Platform for Publishing and Re-using Linked Datasets. In *The Semantic Web: ESWC 2014 Satellite Events - ESWC 2014 Satellite Events, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*. 226–230. https://doi.org/10.1007/978-3-319-11955-7_24

[27] PATRICK JOHANSSON K. 2001. La imagen en los códices nahuas: consideraciones semiológicas. *Estudios de Cultura Náhuatl* 32 (oct. 2001), 629. https://dialnet.unirioja.es/servlet/articulo?codigo=2264870

[28] Joonas Kesäniemi, Mikko Koho, and Eero Hyvönen. 2022. Using Wikibase for Managing Cultural Heritage Linked Open Data Based on CIDOC CRM. In *New Trends in Database and Information Systems - ADBIS 2022 Short Papers, Doctoral Consortium and Workshops: DOING, K-GALS, MADEISD, MegaData, SWODCH, Turin, Italy, September 5-8, 2022, Proceedings*. 542–549. https://doi.org/10.1007/978-3-031-15743-1_49

[29] Mikko Koho, Esko Ikkala, Petri Leskinen, Minna Tamper, Jouni Tuominen, and Eero Hyvönen. 2021. WarSampo knowledge graph: Finland in the Second World War as Linked Open Data. *Semantic Web* 12, 2 (2021), 265–278. https://doi.org/10.3233/SW-200392

[30] Kai Labusch and Clemens Neudecker. 2022. Entity Linking in Multilingual Newspapers and Classical Commentaries with BERT. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*. 1079–1089. http://ceur-ws.org/Vol-3180/paper-85.pdf

[31] Brian Lavoie. 2000. Meeting the challenges of digital preservation: The OAIS reference model. *OClC Newsletter* 243 (2000), 26–30.

[32] Tim Berners Lee. 1998. Semantic Web roadmap. https://www.w3.org/DesignIssues/Semantic

[33] Mrinalini Luthra, Konstantin Todorov, Charles Jeurgens, and Giovanni Colavizza. 2023. Unsilencing colonial archives via automated entity recognition. *Journal of Documentation* (2023). https://doi.org/10.1108/JD-02-2022-0038

[34] Mahendra Mahey, Aisha Al-Abdulla, Sarah Ames, Paula Bray, Gustavo Candela, Caleb Derven, Milena Dobreva-McPherson, Katrine Gasser, Sally Chambers, Stefan Karner, Kristy Kokegei, Ditte Laursen, Abigail Potter, Armin Straube, Sophie-Carolin Wagner, and Lotte Wilms. 2019. *Open a GLAM lab*. International GLAM Labs Community, Book Sprint, Doha, Qatar. 164 pages. https://doi.org/10.21428/16ac48ec.f54af6ae

[35] Pablo N. Mendes, Max Jakob, Andrés García-Silva, and Christian Bizer. 2011. DBpedia spotlight: shedding light on the web of documents. In *Proceedings the 7th International Conference on Semantic Systems, I-SEMANTICS 2011, Graz, Austria, September 7-9, 2011 (ACM International Conference Proceeding Series)*, Chiara Ghidini, Axel-Cyrille Ngonga Ngomo, Stefanie N. Lindstaedt, and Tassilo Pellegrini (Eds.). ACM, 1–8. https://doi.org/10.1145/2063518.2063519

[36] Walter D Mignolo and Catherine E Walsh. 2018. *On decoloniality: Concepts, analytics, praxis*. Duke University Press.

[37] Patricia Murrieta-Flores, Diego Jiménez-Badillo, and Bruno Martins. 2022. Digital Resources: Artificial Intelligence, Computational Approaches, and Geographical Text Analysis to Investigate Early Colonial Mexico. In *Oxford Research Encyclopedia of Latin American History*.

[38] Patricia Murrieta-Flores, Diego Jiménez-Badillo, Bruno Emanuel da Graça Martins, Mariana Favila-Vázquez, Raquel Liceras-Garrido, and Katherine Bellamy. 2020. DECM Gazetteer. https://doi.org/10.6084/m9.figshare.12367385.v2

[39] Patricia Murrieta-Flores, Diego Jiménez-Badillo, and Bruno Martins. 2022. Digital Resources: Artificial Intelligence, Computational Approaches, and Geographical Text Analysis to Investigate Early Colonial Mexico. https://doi.org/10.1093/acrefore/9780199366439.013.977

[40] Clemens Neudecker. 2022. Cultural Heritage as Data: Digital Curation and Artificial Intelligence in Libraries. In *Proceedings of the Third Conference on Digital Curation Technologies (Qurator 2022), Berlin, Germany, Sept. 19th-23rd, 2022 (CEUR Workshop Proceedings, Vol. 3234)*, Adrian Paschke, Georg Rehm, Clemens Neudecker, and Lydia Pintscher (Eds.). CEUR-WS.org. http://ceur-ws.org/Vol-3234/paper2.pdf

[41] Museum of Modern Art. 2016. *The Museum of Modern Art (MoMA) Collection*. https://doi.org/10.5281/zenodo.45581

[42] Thomas Padilla. 2019. Responsible Operations: Data Science, Machine Learning, and AI in Libraries. https://doi.org/10.25333/xk7z-9g97

[43] Thomas Padilla, Laurie Allen, Hannah Frost, Sarah Potvin, Elizabeth Russey Roke, and Stewart Varner. 2019. Final Report — Always Already Computational: Collections as Data. https://doi.org/10.5281/zenodo.3152935

[44] Erwin Panofsky. 1939. *Studies in iconology: humanistic themes in the art of the Renaissance*. Routledge, New York. 399 pages. https://doi.org/10.4324/9780429497063

[45] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event (Proceedings of Machine*

Learning Research, Vol. 139), Marina Meila and Tong Zhang (Eds.). PMLR, 8748–8763. http://proceedings.mlr.press/v139/radford21a.html

[46] Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*. ELRA, Valletta, Malta, 45–50. http://is.muni.cz/publication/884893/en.

[47] Roopika Risam. 2019. *New Digital Worlds: Postcolonial Digital Humanities in Theory, Praxis, and Pedagogy.* Northwestern University Press, Evanston, Illinois.

[48] Jesús Robledano-Arillo, Diego Navarro Bonilla, and Julio Cerdá-Díaz. 2020. Application of Linked Open Data to the coding and dissemination of Spanish Civil War photographic archives. *J. Documentation* 76, 1 (2020), 67–95. https://doi.org/10.1108/JD-06-2019-0112

[49] Ethelia Ruiz Medrano. 2010. *Mexico's Indigenous Communities. Their Lands and Histories, 1500-2010.* University Press of Colorado.

[50] Pedro A. Szekely, Craig A. Knoblock, Fengyu Yang, Eleanor E. Fink, Shubham Gupta, Rachel Allen, and Georgina Goodlander. 2014. Publishing the Data of the Smithsonian American Art Museum to the Linked Data Cloud. *Int. J. Humanit. Arts Comput.* 8, supplement (2014), 152–166. https://doi.org/10.3366/ijhac.2014.0104

[51] Daniel van Strien, Kaspar Beelen, Mariona Coll Ardanuy, Kasra Hosseini, Barbara McGillivray, and Giovanni Colavizza. 2020. Assessing the Impact of OCR Quality on Downstream NLP Tasks. In *Proceedings of the 12th International Conference on Agents and Artificial Intelligence, ICAART 2020, Volume 1, Valletta, Malta, February 22-24, 2020*. 484–496. https://doi.org/10.5220/0009169004840496

[52] Daniel van Strien, Mark Bell, Nora Rose McGregor, and Michael Trizna. 2021. An Introduction to AI for GLAM. In *Proceedings of the Second Teaching Machine Learning and Artificial Intelligence Workshop, September 8+13, 2021, Virtual Conference*. 20–24. https://proceedings.mlr.press/v170/strien22a.html

[53] Daniel Vila-Suero, Boris Villazón-Terrazas, and Asunción Gómez-Pérez. 2013. datos.bne.es: A library linked dataset. *Semantic Web* 4, 3 (2013), 307–313. https://doi.org/10.3233/SW-120094

[54] Easton Wollney and Miglena Sternadori. 2019. Feminine, Competent, Submissive: A Multimodal Analysis of Depictions of Women in U.S. Wartime Persuasive Messages During World War I and World War II. *Visual Communication Quarterly* 26, 1 (2019), 3–21. https://doi.org/10.1080/15551393.2018.1530600

[55] World Wide Web Consortium. 2011. Describing Linked Datasets with the VoID Vocabulary. https://www.w3.org/TR/void/

[56] World Wide Web Consortium. 2013. Linked Data Glossary. https://www.w3.org/TR/2013/NOTE-ld-glossary-20130627/