

Research and Innovation Action

Social Sciences & Humanities Open Cloud

Project Number: 823782 Start Date of Project: 01/01/2019 Duration: 40 months

Deliverable 5.16 Report on making heritage science data FAIR (Open data in Heritage Science and Archaeology)

Dissemination Level	PU
Due Date of Deliverable	30/04/22, M40
Actual Submission Date	30/04/22, M40
Work Package	WP5: Innovations in Data Access
Task	Task 5.6 Issues in providing Open Data in Heritage Science and Archaeology
Type	Report
Approval Status	Waiting EC approval
Version	V1.0
Number of Pages	p.1 – p.79

Abstract: This deliverable presents a documentation of the mappings and procedures used to model Heritage Science related data into FAIR, CIDOC CRM based, open data resources. The report gives overview of the work relating to the modelling of two existing collaborative systems from the National Gallery and presents work by CNR on updating current digital documentation software, MOVIDA, to export technical examination results as FAIR data.

The information in this document reflects only the author’s views and the European Community is not liable for any use that may be made of the information contained therein. The information in this document is provided “as is” without guarantee or warranty of any kind, express or implied, including but not limited to the fitness of the information for a particular purpose. The user thereof uses the information at his/ her sole risk and liability. This deliverable is licensed under a Creative Commons Attribution 4.0 International License.



SSHOC, “Social Sciences and Humanities Open Cloud”, has received funding from the EU Horizon 2020 Research and Innovation Programme (2014-2020); H2020-INFRAEOSC-04-2018, under the agreement No. 823782

History

Version	Date	Reason	Revised by
0.1	14/03/2021	Initial draft report of CNR activities	Anna Amat
0.2	30/09/2021	Initial draft report of NG modelling work	Orla Delaney
0.3	30/03/2022	Completion of draft of final work for NG modelling work	Joseph Padfield
0.4	20/04/2022	Draft of completed document – after the completion of the required live digital presentations of the work.	Joseph Padfield
0.5	21/04/2022	Correcting heading numbers and references adding in final descriptions and links	Joseph Padfield
1.0	28/04/2022	Checking content summarise and referencing. Adding in references to the collaborations around DataVerse.	Joseph Padfield

Author List

Organisation	Name	Contact Information
NG	Joseph Padfield	Joseph.padfield@nationalgallery.org.uk https://orcid.org/0000-0002-2572-6428
NG	Orla Delaney	Project Position Ended (as above) or https://orcid.org/0000-0002-7114-7929
CNR	Anna Amat Alberti	anna.amatalb@gmail.com

Executive Summary

This deliverable reports on work carried out within SSHOC Task 5.6 - Issues in providing Open Data in Heritage Science and Archaeology. It focusses specifically on issues related to working with Heritage Science data and examines the accessibility and interoperability of such data. Starting with two distinct, but related, non-standard datasets, covering the documentation and study of old master paintings, the work created new fully semantic, linkable, shareable, machine-readable FAIR¹² datasets mapped to the standard CIDOC-CRM⁵ ontology and other external Linked Open Data resources. The work also examined how such data mapping procedures and FAIR datasets could be incorporated into existing digital documentation software, such as the MOVIDA database software¹. This report begins by providing some background on the notion of FAIR data and the CIDOC CRM before going on to describe the work creating and using FAIR data.

The original, existing datasets are described along with some of the key concepts and technologies used within the mapping and modelling procedure. The report then goes on to describe some of the preliminary work carried out to develop some new tools and procedures that would be exploited during the main mapping process, including the development of automated processes to create web presentations (Simple Site⁴⁴, Simple IIF Discovery⁶¹ and the Dynamic Modeller⁴⁵) and the re-formatting and opening up some existing semantic models⁶⁵ developed in previous EU projects (IPERION-CH).

The mapping procedures for the two main datasets are then described including details of the processes used and the decisions made. Examples diagrams are provided for several of the key semantic models along with sections of the code used. In addition, to these general descriptions of the processed datasets, access details are also provided for live digital presentations where the full final data sets can be explored and visualised, via a new user-friendly website, complete with a range of predefined example queries to demonstrate the data stored in the system and the relationships that have been modelled. Options for direct access via standard machine-readable systems are also described. Additional access details are also provided for all of the software, web tools, datasets and data repositories created within the work.

An introduction to the MOVIDA software is then provided outlining its history and how it can exploit FAIR data. Details are then provided documenting the key metadata terms used to describe objects and examinations within the MOVIDA software and how they can be mapped to CIDOC-CRM based ontologies. Discussion is also included in relation to where the direct mapping of existing terms and relationships are not immediately clear or where the further extension of the existing ontologies could simplify or clarify the description of the included data. A summary of some initial next steps for both the mapping work and the practical application of CIDOC-CRM mapped data with MOVIDA are provided.

¹ The MOVIDA software is described in detail in section 0.

Abbreviations and Acronyms

EU H2020	EU's research and innovation funding programme from 2014-2020 (https://ec.europa.eu/info/research-and-innovation/funding/funding-opportunities/funding-programmes-and-open-calls/horizon-2020_en)
SSHOC	The Social Sciences & Humanities Open Cloud, EU funded H2020 project (https://www.sshopencloud.eu/)
FAIR	Provide guidelines related to improving the F indability, A ccessibility, I nteroperability, and R euse of digital resources (https://www.go-fair.org/fair-principles/)
NG	The National Gallery, London (https://www.nationalgallery.org.uk)
ICOM	The International Council of Museums (https://icom.museum/)
CIDOC	ICOM International Committee for Documentation (https://cidoc.mini.icom.museum/)
CIDOC-CRM	The CIDOC Conceptual Reference Model (CRM) is a standard ontology developed for the Heritage domain. (https://www.cidoc-crm.org/)
IIIF	The International Image Interoperability Framework (https://iiif.io)
CNR	The Italian National Research Council (https://www.cnr.it/)
CNR ISPC	Istituto di Scienze del Patrimonio Culturale (https://www.ispc.cnr.it/)
CNR-SCITEC	Istituto di Scienze e Tecnologie Chimiche "G. Natta" (http://www.scitec.cnr.it/en/ricerca-it/chemistry-for-cultural-heritage)
MOLAB	MO bile LAB oratory – a distributed infrastructure providing coherent access to a set of mobile analytical and examination equipment and related competencies, (https://www.iperionhs.eu/molab)
MOVIDA	MO lab VI sualisation D ata
XML	"Extensible Markup Language (XML) is a markup language and file format for storing, transmitting, and reconstructing arbitrary data." (https://en.wikipedia.org/wiki/XML)

RDF	Resource Description Framework (https://en.wikipedia.org/wiki/Resource_Description_Framework)
RDFS	Resource Description Framework Schema (https://en.wikipedia.org/wiki/RDF_Schema)
CRMsci	CIDOC CRM - Scientific Observation Model (https://cidoc-crm.org/crmsci)
CRMdig	CIDOC CRM - Model for provenance metadata (https://cidoc-crm.org/crmdig)
RRR	The Raphael Research Resource (https://cima.ng-london.org.uk/documentation)
IPERION-CH	Integrated Platform for the European Research Infrastructure ON Cultural Heritage, EU funded H2020 project (http://www.iperionch.eu)
PID(s)	A P ersistent ID entifier(s) (https://tanc-ahrc.github.io/PIDResources/)
GLAM	Galleries, Libraries, Archives, and Museums (https://en.wikipedia.org/wiki/GLAM_(cultural_heritage))
IPERION-HS	Integrating Platforms for the European Research Infrastructure ON Heritage Science, EU funded H2020 project (https://www.iperionhs.eu)
AHRC	The UK Arts and Humanities Research Council (https://www.ukri.org/councils/ahrc/)
CRMinf	CIDOC CRM – Argumentation Model (https://cidoc-crm.org/crminf/)
OWL	Web Ontology Language (https://www.w3.org/TR/owl2-overview/)
SQL	Structured Query Language (https://en.wikipedia.org/wiki/SQL)
MySQL	An open source relational database management system (https://en.wikipedia.org/wiki/MySQL)
API	Application Programming Interface (https://en.wikipedia.org/wiki/API)
HTML	HyperText Markup Language (https://en.wikipedia.org/wiki/HTML)
JSON	JavaScript Object Notation (https://en.wikipedia.org/wiki/JSON)
URL	"...Uniform Resource Locator ... colloquially termed a web address..."

	(https://en.wikipedia.org/wiki/URL)
PHP	PHP: Hypertext Preprocessor, (which is a recursive acronym) (https://en.wikipedia.org/wiki/PHP)
EPSRC	The UK Engineering and Physical Sciences Research Council (https://www.ukri.org/councils/epsrc/)
DOI	Digital object identifier (https://en.wikipedia.org/wiki/Digital_object_identifier)
3M	Mapping Memory Manager (FORTH Centre for Cultural Informatics) (https://www.ics.forth.gr/x3ml-toolkit)
AAT	The Getty's Art and Architecture Thesaurus (http://www.getty.edu/research/tools/vocabularies/aat/)
SPARQL	SPARQL Protocol and RDF Query Language (https://en.wikipedia.org/wiki/SPARQL)
SMK	Statens Museum for Kunst (The National Gallery of Denmark) (https://www.smk.dk)
RDF* or RDF-star	An extension to RDF (https://w3c.github.io/rdf-star/cg-spec/2021-07-01.html)
SPARQL* or SPARQL-star	An extension to SPARQL (https://w3c.github.io/rdf-star/cg-spec/2021-07-01.html)
OCT	Optical coherence tomography (https://en.wikipedia.org/wiki/Optical_coherence_tomography)
NIR	Near Infrared (https://en.wikipedia.org/wiki/Infrared#Regions_within_the_infrared)
XRF	X-ray fluorescence (https://en.wikipedia.org/wiki/X-ray_fluorescence)
SQLite	Is "small, fast, self-contained, high-reliability, full-featured, SQL database engine" (https://www.sqlite.org/)
CATS	CATS - Centre for Art Technological Studies and Conservation

Table of Contents

1	Introduction	8
1.1	FAIR: Data standards in Cultural Heritage	9
1.2	CIDOC-CRM: Data standards in Cultural Heritage.....	10
2	Heritage Science - Mapping existing Data (NG)	12
2.1	The Original Input Datasets	12
2.2	Background and initial technologies used	13
2.3	Supporting activities	19
2.4	The Mapping Work	27
2.5	Summary of accessibility of resources	48
2.6	Increasing the accessibility of Grounds Data.....	49
2.7	Next steps for the RRR and Grounds Datasets.....	51
3	Heritage Science - The MOLAB experience (CNR)	52
3.1	MOVIDA (MOlab Visualisation DAta)	54
3.2	Interoperability between CNR MOVIDA and CIDOC-CRM.....	58
3.3	Schema for: General Information on the artwork (GIA).....	59
3.4	Schema for: General Analytical Campaign (GAI).....	61
3.5	Modelling Issues	66
3.6	Implementation	70
3.7	Next steps with MOVIDA	70
4	Conclusion	72
5	References	73
5.1	List of Figures	74
5.2	List of Tables	75
6	Appendices	76
6.1	Example Simple MOVIDA FAIR XML Dataset.....	76

1 Introduction

Cultural heritage institutions, particularly the departments of those institutions concerned with scientific examination of artworks and artefacts, gather and store vast quantities of data about their collections. This data usually includes detailed provenance information, full documentation records, images of samples taken from the works, and results of any examinations conducted using x-radiography, mass spectrometry and other analytical technologies. This wealth of information is rarely made available to external researchers, and made available in a usable, interoperable format even less often. The work in this task has explored how resources can be made significantly more accessible and interoperable by mapping existing resources to a domain standard ontology.

This report describes two related, but separate, pieces of work, carried out within SSHOC² Task 5.6, which have been exploring issues related to making Heritage Science data more “FAIR”³.

The first, led by the National Gallery (NG)⁴, consisted of mapping two bespoke datasets previously assembled at the Gallery to the linked data ontology CIDOC-CRM⁵. Steps were also taken to make the datasets available both in raw, a searchable format and the possibility of a new curated interface with which art historians and other research professionals could interact was considered. The images referenced in the datasets were made compliant with IIF⁶ and presented using standard compliant IIF viewing software. All code and datasets were published on both Zenodo⁷ and Github⁸, where full code documentation may also be found. This report will describe how the mapping exercise was conducted and to give instruction to future researchers who wish to replicate the approach taken, given that technologies such as CIDOC-CRM are still in the early stages of in-depth adoption by many cultural heritage institutions.

The second, led by CNR ISPC⁹ and CNR-SCITEC¹⁰, examined what level of interoperability was possible between the CIDOC-CRM and an existing piece of cultural heritage digital documentation software,

² The general project website can be found at: <https://www.sshopencloud.eu> [21/04/22]

³ Detailed documentation of the FAIR principles can be found at: <https://www.go-fair.org/fair-principles>. [21/04/22]

⁴ Institutional website: <https://www.nationalgallery.org.uk/>. [21/04/22]

⁵ For more details and documentation see: <https://www.cidoc-crm.org/>. [21/04/22]

⁶ International Image Interoperability Framework (<https://iiif.io/>). [21/04/22]

⁷ Zenodo is an open, free, catch-all data repository supported by CERN (<https://home.cern>), OpenAire (<https://www.openaire.eu>) and the EU Horizon 2020 Programme (<https://ec.europa.eu/programmes/horizon2020>) – For more details see: <https://zenodo.org>. [21/04/22]

⁸ An open, online software repository and development platform. For more details see: <https://github.com>. [21/04/22]

⁹ The CNR Institute of Cultural Heritage Sciences: <https://www.ispc.cnr.it/>. [21/04/22]

¹⁰ The Istituto di Scienze e Tecnologie Chimiche “G. Natta” (<http://www.scitec.cnr.it/en/ricerca-it/chemistry-for-cultural-heritage>)

MOVIDA a standalone java based piece of software that gathers all the generated data from non-invasive analytical spectroscopic investigations and stores it in a single XML file. The software can also be used to analyse the stored data on-the-fly, analyse the multi-technique information recorded and share it with researchers and future users. The work created a pilot database using, when possible, existing schemas and ontologies and suggesting the necessary implementations on the chosen schema when the data from MOVIDA could not be included. The aim of this work was to explore how the data created by MOVIDA could be made more open and FAIR.

1.1 FAIR: Data standards in Cultural Heritage

It has been recommended, by the EU, that all H2020 projects should aim for all the data they produce to be “As Open as Possible, As Closed as Necessary”¹¹. The FAIR principles are a series of guiding ideas to help ensure that current and future users of published data can **F**ind, **A**ccess, work with (**I**nteroperable) and legally **R**euse the data.

The acronym ‘FAIR’ provides an important guiding principle for this research project. First defined in 2016 in *Scientific Data*¹², the FAIR principles dictate that open data should be Findable, Accessible, Interoperable and Reusable and provide specific guidelines for how these conditions might be met. In broad terms, Findability refers to the ability of both humans and computers to locate the dataset and individual pieces of data within; this is achieved by indexing data in an appropriate, searchable resource, providing persistent identifiers (PIDs) for each piece of data, and ensuring that rich metadata is present alongside the dataset. Accessibility is about ensuring that any protocol for querying and manipulating data conforms to a well-documented, widely used standard, and that metadata can persist beyond the deprecation of the rest of the dataset. Interoperability is concerned with the integration of datasets both with other comparable datasets and with applications that may draw on the dataset to present content to users; to fulfil this criterion, data should be described using standardised systems of knowledge representation (of which the CIDOC-CRM is an example) and connected where possible to vocabularies that are themselves FAIR. Finally, Reusability describes procedural boxes that must be ticked so that a user knows how they can use a dataset from both a technical and legal perspective: a license must be chosen and made available, the data must be sufficiently documented as to be replicable, and the dataset’s provenance information must be accessible to users. Taken together, these principles provide not only a sound set of structuring guidelines for a cultural heritage dataset, but also an ontological

¹¹ Quoted from: https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm. [21/04/22]

¹² Mark D. Wilkinson; Michel Dumontier; IJsbrand Jan Aalbersberg; et al. (15 March 2016). "The FAIR Guiding Principles for scientific data management and stewardship". *Scientific Data*. 3 (1): 160018. <https://doi.org/10.1038%2FSDATA.2016.18>. ISSN 2052-4463. PMC 4792175. PMID 26978244. Wikidata Q27942822. [21/04/22]

approach that prioritises persistence and openness; in other words, an approach that better aligns digital technologies and academic ways of working.

- **Findable:** Ensuring that all data is appropriately described with sufficient Meta¹³ and Paradata¹⁴ to ensure future users can understand and put the data in context. The data should also be identified with a unique persistent identifier or PID.
- **Accessible:** Ensuring that data or at least the Meta and Paradata describing the data available online via standard, documented procedures. For example, it has been uploaded to an open, sustainable, online data repository, such as Zenodo.
- **Interoperable:** Ensuring that appropriate common formats and standards are used to organise the data and that terms used to categorise or tag the data come from open standard vocabularies.
- **Re-usable:** Ensuring that data is appropriately documented and licenced for re-use.

1.2 CIDOC-CRM: Data standards in Cultural Heritage

Cultural information comes from a very large domain with multidisciplinary sources that generate different types of information that could be integrated. Among the different types of information gathered there can be, but are not limited to:

- collection descriptions from museums and institutions
- archaeological-cultural heritage data
- heritage science information
- historical information
- industrial processes
- preservation and conservation

All this data is analysed and interpreted generating, in turn, lots of new knowledge and data. Every organisation handles this data using different schemas and methodologies and this heterogeneity hampers interoperability and knowledge sharing. At a certain point it has become clear the need of providing the semantic definitions and to create a common language for domain experts and implementers in order to formulate requirements for information systems.

Created by the International Committee for Documentation (CIDOC) of ICOM, with over 20 years of development it is composed of a set of concepts and relationships that tell us about the possible states that exist in the domain of interest. With the objective to serve as a guide for good practice of conceptual

¹³ Definition: "Metadata is "data that provides information about other data" ..."
(<https://en.wikipedia.org/wiki/Metadata>). [21/04/22]

¹⁴ Definition: "The paradata of a data set or survey are data about the process by which the data were collected ..."
(<https://en.wikipedia.org/wiki/Paradata>). [21/04/22]

modelling, it is organised in an object-oriented approach creating a rich web of information data set that is at the same time human and machine readable. This enables us to exchange data between organisations and professionals. Therefore, it aims to provide the greatest flexibility of systems to become compatible, rather than imposing one particular solution.

CIDOC-CRM provides a common expandable semantic framework to which, in principle, any CH data can be mapped, recognized as an ISO standard under ISO 21127:2014¹⁵. It is developed by a volunteer community dedicated to enable information exchange and integration between a variety of heterogeneous sources of cultural heritage information bringing them together into an integrated environment.

CIDOC-CRM is organized in classes and relations devised for the cultural heritage world. It has been expressed as an object-oriented semantic model, in the hope that this formulation will be comprehensible to both documentation experts and information scientists alike, while at the same time being readily converted to machine-readable formats¹⁶ such as the RDF Schema¹⁷. It can be implemented in any Relational or object-oriented schema. CRM instances can also be encoded in a wide range of other machine-readable formats.

At the time of writing, the latest version of the CIDOC-CRM (7.2.1)¹⁸ has a total of 82 entities and 159 properties. Besides the main branch development, a series of approved and supervised compatible models are being constantly improved. These models are intended to be more specific of a series of areas. Within this task two of these compatible models have been considered. The CRM-sci (Scientific Observation Model)¹⁹ aim to integrate metadata about scientific observation, measurements, and processed data in descriptive and empirical sciences to deal with the heritage science data. Also, the CRM-dig (Model for provenance metadata)²⁰ aim to integrate metadata relating to the production ("provenance") of digitization and digital representations.

¹⁵ The ISO standard for the CIDOC CRM: <https://www.iso.org/standard/57832.html>. [21/04/22]

¹⁶ Definition: https://en.wikipedia.org/wiki/Machine-readable_data. [21/04/22]

¹⁷ The RDF Schema documentation: <https://www.w3.org/TR/rdf-schema/>. [21/04/22]

¹⁸ The Current version of the CIDOC CRM: <https://www.cidoc-crm.org/Version/version-7.2.1>. [21/04/22]

¹⁹ Documentation for the CIDOC CRM extension CRMsci: <https://cidoc-crm.org/crmsci> [21/04/22]

²⁰ Documentation for the CIDOC CRM extension CRMdig: <https://cidoc-crm.org/crmdig> [21/04/22]

2 Heritage Science - Mapping existing Data (NG)

2.1 The Original Input Datasets

Two main datasets, already held by the NG, were used as the basis of this area of work. The first, known as the Raphael Research Resource (RRR) dataset, which was created in 2007, describes paintings by Raphael²¹ held in the collection of the NG, as well as additional entries from the collections of the Metropolitan Museum of Art²², the Musée Condé²³, and others. In total, the dataset describes 27 paintings, and references full-resolution images of each painting as well as images and samples taken during various scientific examination processes. Each painting is richly documented, and the dataset contains full provenance records, to the extent that they were available at the time of its creation; these range from short curatorial descriptions, and scholarly articles, to scans of physical records known as Conservation Dossiers, which describe all the conservation procedures undergone by a given painting. Prior to the start of this research project, the RRR dataset was already presented in a linked data format²⁴; however, the ontology used was bespoke to the NG, using terminology and vocabularies not common to other similar datasets. The dataset could not conform to FAIR standards unless mapped to a more widely used standard ontology.

The second NG dataset used is known as the Grounds database. Developed in 2018 as part of the EU H2020 funded IPERION-CH project²⁵, the Grounds database is a traditional relational database of 44 tables describing the preparation layers used in 16th century Italian paintings, including those in the NG's Raphael collection²⁶. The database went on to incorporate data from the Prado²⁷, Madrid, the Doerner Institut²⁸, Munich, and CATS, Copenhagen²⁹ among other partners, and now contains information for more than 250 paintings across these institutions' collections. The majority of data points in the database

²¹ Raphael - "The Italian painter and architect ..." (<https://www.wikidata.org/wiki/Q5597>) [21/04/22]

²² Institutional website: <https://www.metmuseum.org>. [21/04/22]

²³ Institutional website: <https://www.musee-conde.fr>. [21/04/22]

²⁴ The original RDF presentation of the RRR dataset: <https://rdf.ng-london.org.uk/workshops/lcd>. [21/04/22]

²⁵ H2020 EU funded: Integrated Platform for the European Research Infrastructure ON Cultural Heritage (IPERION-CH) H2020 – for more information see: <http://www.iperionch.eu/> and <https://cordis.europa.eu/project/id/654028> [21/04/22]

²⁶ The development of the database is described within one of the IPERION-CH project deliverables: D.8.6 Two digital research resources available for open access on the web and one working resource (<https://doi.org/10.5281/zenodo.5838339>) [21/04/22]

²⁷ Institutional website: <https://www.museodelprado.es>. [21/04/22]

²⁸ Institutional website: <https://www.doernerinstitut.de>. [21/04/22]

²⁹ CATS - Centre for Art Technological Studies and Conservation, for more details see: <https://www.smk.dk/en/article/the-cats-reference-collection/> [21/04/22]

describe cross-section samples taken from these paintings, describing their source on the painting's surface, as well as observations about their material makeup and descriptions of each layer's colouring and pigmentation. This relational database was presented within a bespoke web resource that enabled users to query the data; however, it was not conceived of as a FAIR resource, rather as one that could be made to conform to FAIR standards at a later date, and as such does not contain any PIDs and is not currently open to users without a password.

Each of these datasets provided distinct challenges that would trouble standard methods of mapping to CIDOC-CRM. The Grounds database's size and complexity made it an unsuitable candidate for the simple use of traditional tools such as the FORTH Centre for Cultural Informatics' Mapping Memory Manager (3M)³⁰ or the open source mapping application Protegé³¹, which would ordinarily be used to map a database containing fewer tables. The RRR came with its own unique challenge, as many of the semantic relationships "covered" by the original schema were inferred rather than specifically stated. It quickly became clear that a bespoke mapping approach would also be needed more fully map the old data to the CIDOC CRM.

This document aims to describe the steps taken to map both datasets to CIDOC-CRM and make them available in an online resource, as well as providing documentation that will enable users to query data and understand how the datasets fit together and interoperate with one another. It also includes an account of the issues encountered during the mapping process and recommendations for future similar mappings, alongside an overview of the kinds of research this data will make possible both within the NG and in the SSHOC community as a whole.

2.2 Background and initial technologies used

This work took place in a context of ongoing digital transformation, both within the NG and throughout the GLAM ecosystem in the UK and beyond. EU-funded projects IPERION-CH and IPERION-HS³² were developing datasets relating to digital collections, and the AHRC's Towards a National Collection³³ had begun to produce outputs that would need to conform to certain data standards. Similarly, the NG's approaching bicentenary celebrations included a mandate to improve the digital collection's functionality for researchers, exemplified by a planned internal Digital Dossier project which, began in earnest in late 2021. Additionally, the upcoming Raphael exhibition at the Gallery could provide an opportunity to showcase a new dataset and accompanying interface to researchers and the general public. Producing

³⁰ 3M software documentation: <https://www.ics.forth.gr/x3ml-toolkit>. [21/04/22]

³¹ Protegé software documentation: <https://protege.stanford.edu/>. [21/04/22]

³² H2020 EU funded: Integrated Platform for the European Research Infrastructure ON Heritage Science (IPERION-HS) H2020 – Project website: <https://www.iperionhs.eu/> and <https://cordis.europa.eu/project/id/871034>. [21/04/22]

³³ Project website: <https://www.nationalcollection.org.uk/> [21/04/22]

an example of FAIR NG data was important for all these efforts, as it would set out steps for achieving the data standards necessary for a functioning interoperable system of GLAM data.

This section introduces the main initial technologies that were used to achieve FAIRness, giving both an overview of their function and a justification for their use. Later sections will give more detail on implementation of the below technologies alongside an account of the technical challenges faced during the project as a whole.

2.2.1 Linked Data and CIDOC-CRM

This project's approach to making heritage science data FAIR entails the transformation of certain datasets from traditional relational formats to linked open data. For this use case, linked open data³⁴ was the most FAIR choice of data structure because of its seamless interoperability: any dataset that is correctly mapped to a linked data ontology can, in theory, be connected to any other similarly mapped dataset. This choice was also made with future data standardisation in mind: projects such as the UK's Towards a National Collection are increasingly exploring and making use of linked data as a means of digitally unifying disparate cultural heritage collections. The CIDOC Conceptual Reference Model (CRM) is the ICOM standard ontology for the linking of heritage data and is overwhelmingly the most widely used ontology to describe this type of data. As well as enabling the interoperability of CRM-mapped collections, it also allows this project to be seamlessly linked with the outputs of projects such as linked.art³⁵, whose bespoke ontology is a derivative of the CIDOC-CRM. The CIDOC-CRM has also been expanded to include a series of compatible models such as CRMdig, which describes digital objects and builds on the CRM's already-existing distinction between a physical object and its digital representation, and CRMinf (Argumentation model)³⁶, which provides functionality that enables the mapping of argumentation. The mappings of the RRR and the Grounds database make use of base CRM, CRMdig and CRMsci, an ontology that describes scientific sampling, and this work includes examples of the usage of each within our output dataset.

2.2.2 Python³⁷

The decision to use Python as the primary technology for the mapping was largely a practical one. From early on it was clear that the complexities inherent to both datasets demanded a custom approach and would not allow us to use existing mapping technologies. Python was chosen both because of existing proficiencies within the mapping team and for its RDFlib package³⁸. RDFlib is an open source library

³⁴ Definition "... is structured data, which is interlinked with other data, so it becomes more useful through semantic queries. ..." (https://en.wikipedia.org/wiki/Linked_data) [21/04/22]

³⁵ Project website: <https://linked.art/>. [21/04/22]

³⁶ Documentation for the CIDOC CRM extension CRMinf: <https://cidoc-crm.org/crminf/>. [21/04/22]

³⁷ The Python programming language (<https://www.python.org/>) [21/04/22]

³⁸ This Python package is documented at: <https://rdflib.readthedocs.io> [21/04/22]

within the Python language whose functionality includes local storage, querying and creation of RDF graphs, as well as OWL-based inferencing and documentation. Crucially for this project, RDFlib's functionality is such that both sets of inputs could be standardised to such an extent that an overarching code structure as well as specific functions could persist across both mappings. This decreased workload and increased the eventual interoperability of the target datasets.

Python's native data science capabilities also made it a particularly suitable language in which to conduct this migration. The open source library Pandas³⁹ is an extremely widely used data science tool, which allows users to transform datasets into 'dataframes' and load them directly into Python, from where various operations can be performed in a few simple lines of code. This became essential to the mapping of the Grounds database in particular, as it enabled SQL views created in the MySQL database where Grounds is stored to be directly imported into Python, drastically reducing processing time by decreasing the number of database calls that were necessary to access the Grounds database.

2.2.3 Persistent identifiers

From this project's inception, it was clear that the use of PIDs would be a crucial part of making any heritage science dataset FAIR. Globally unique, actionable, and persistent, these identifiers provide a citeable reference for any given entity in a dataset, increasing both Findability and Interoperability. The NG is in the process of migrating its published data to a new stable PID system and changing the structure of its PIDs at the same time, so this work had to anticipate these changes before their official implementation. It was also expected that processing the datasets would necessitate the creation of new PIDs: for example, PIDs to refer to ephemeral entities like specific historical events and timespans rather than just the objects in the NG's collection. For this reason, this project's approach to PIDs was designed to be reactive to the NG's system, first querying the NG PID API⁴⁰ for existing records before creating temporary placeholder PIDs for new entities in the datasets. This allowed the project to proceed independently of the work being done on PIDs while enabling it to conform to that work's standards at a later date, as well as creating a stable list of entities that would need new PIDs when the time came. The use of these temporary PIDs and the mapping code has been designed to allow the new "stable" PIDs to be incorporated into new versions of these datasets in the future.

2.2.4 Temporary SSHOC NG PIDs

Although many of the PIDs used in the datasets are considered "temporary", work has been carried out to ensure that they do still function as fully resolvable and citable references, which could still be maintained and incorporated, as alternative IDs, in future versions of the datasets.

³⁹ The Pandas software and documentation can be found at: <https://pandas.pydata.org/>. [21/04/22]

⁴⁰ Details of the NG's current PID system, along with a few examples, can be found at: <https://data.ng-london.org.uk/>. [21/04/22]

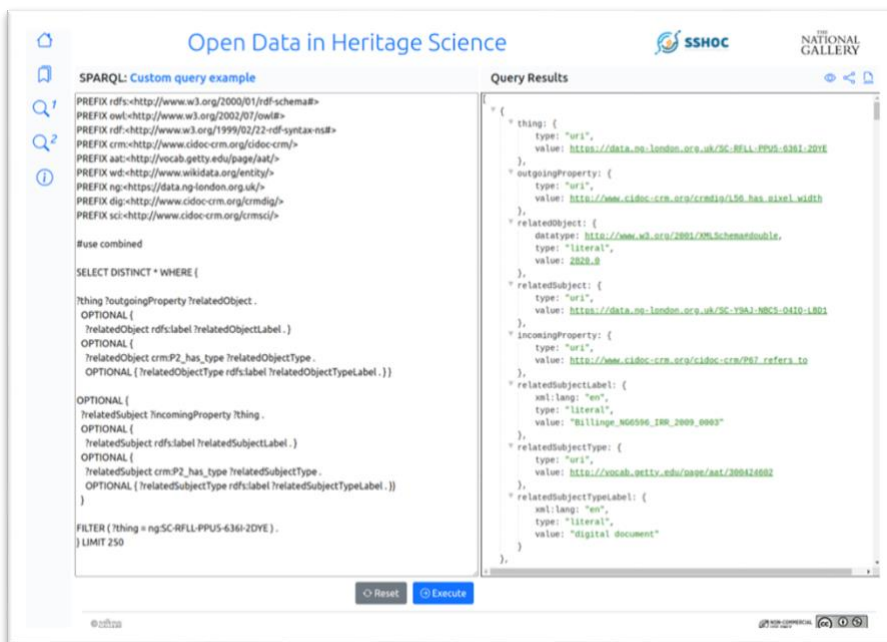
- Entities which already existed within the NG's PID system have resolvable PIDs in the form of: <https://data.ng-london.org.uk/0DOE-0001-0000-0000> (Portrait of Pope Julius II) - data representing these entities comes from the NG's public API⁴⁰.
- New entities, created as part of this project, have been given temporary PIDs, which also resolve to re-useable data, have a similar form but with a "SC-" prefix, such as: <https://data.ng-london.org.uk/SC-RFLL-PPU5-6361-2DYE>⁴¹ (An X-ray image in the RRR)

The temporary PIDs system can return data formatted as "html" or as "JSON" as needed, by adding the appropriate extension to the PID, with "html" being the default option.

- Request JSON: <https://data.ng-london.org.uk/SC-RFLL-PPU5-6361-2DYE.json>
- Request html: <https://data.ng-london.org.uk/SC-RFLL-PPU5-6361-2DYE.html>⁴²

The actual data returned represents the results of a generic default query. see

Figure 1, which is called against both datasets, the RRR and the Grounds database.



The screenshot shows the 'Open Data in Heritage Science' interface. On the left, there is a SPARQL query editor with a 'Custom query example' and a list of prefixes. The query is:

```

PREFIX rdf=<http://www.w3.org/2000/01/rdf-schema#>
PREFIX owl=<http://www.w3.org/2002/07/owl#>
PREFIX rdf=<http://www.w3.org/1999/02/22-rdf-syntax-ns#>
PREFIX crm=<http://www.cidoc-crm.org/cidoc-crm/>
PREFIX aat=<http://vocab.getty.edu/page/aat/>
PREFIX wd=<http://www.wikidata.org/entity/>
PREFIX ng=<https://data.ng-london.org.uk/>
PREFIX dig=<http://www.cidoc-crm.org/crmidig/>
PREFIX sci=<http://www.cidoc-crm.org/crmsci/>

#use combined

SELECT DISTINCT * WHERE {
  ?thing ?outgoingProperty ?relatedObject .
  OPTIONAL {
    ?relatedObject rdfs:label ?relatedObjectLabel .
  }
  OPTIONAL {
    ?relatedObject crm:P2_has_type ?relatedObjectType .
    OPTIONAL { ?relatedObjectType rdfs:label ?relatedObjectTypeLabel . }
  }
  OPTIONAL {
    ?relatedSubject ?incomingProperty ?thing .
    OPTIONAL {
      ?relatedSubject rdfs:label ?relatedSubjectLabel .
    }
  }
  OPTIONAL {
    ?relatedSubject crm:P2_has_type ?relatedSubjectType .
    OPTIONAL { ?relatedSubjectType rdfs:label ?relatedSubjectTypeLabel . }
  }
}
FILTER (?thing = ng:SC-RFLL-PPU5-6361-2DYE) .
LIMIT 250
  
```

On the right, the 'Query Results' section shows a JSON array:

```

[
  {
    "thing": {
      "type": "uri",
      "value": "https://data.ng-london.org.uk/SC-RFLL-PPU5-6361-2DYE"
    },
    "outgoingProperty": {
      "type": "uri",
      "value": "https://www.cidoc-crm.org/crmidig/50_has_nixel_width"
    },
    "relatedObject": {
      "datatype": "http://www.w3.org/2001/XMLSchema#double",
      "type": "literal",
      "value": "2023.8"
    },
    "relatedSubject": {
      "type": "uri",
      "value": "https://data.ng-london.org.uk/SC-Y9A1-NBC3-0410-1801"
    },
    "incomingProperty": {
      "type": "uri",
      "value": "http://www.cidoc-crm.org/cidoc-crm/P87_refers_to"
    },
    "relatedSubjectLabel": {
      "xml:lang": "en",
      "type": "literal",
      "value": "Billings N60596_180_2009_0003"
    },
    "relatedSubjectType": {
      "type": "uri",
      "value": "http://vocab.getty.edu/page/aat/300426002"
    },
    "relatedSubjectTypeLabel": {
      "xml:lang": "en",
      "type": "literal",
      "value": "digital document"
    }
  }
]
  
```

⁴¹ The full IIIF information relation to this images can be found at: https://research.ng-london.org.uk/iiif/pics/tmp/raphael_pyr/N-6596/16.1_X-Ray_images/N-6596-00-000020-PYR.tif/info.json.

⁴² The PID URL with and without the ".html" extension both to a simple human readable landing page which presents the related data, for an example see Figure 2. [21/04/22]

FIGURE 1: A SCREENSHOT SHOWING THE DEFAULT GENERIC SPARQL QUERY USED TO GATHER DATA FOR A GIVEN TEMPORARY NG PID.

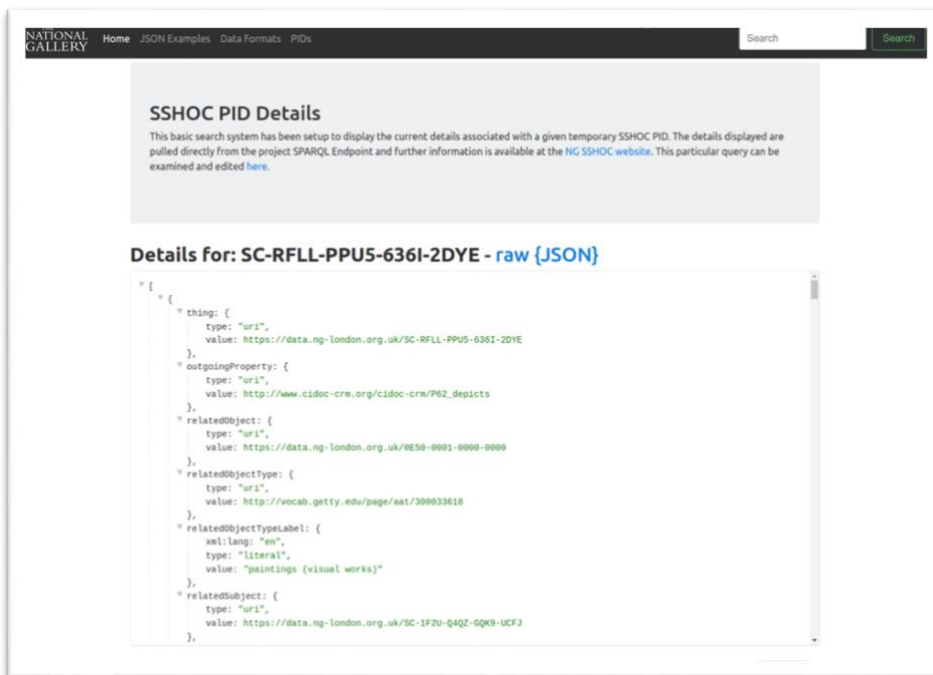


FIGURE 2: A SCREENSHOT OF AN EXAMPLE "HTML" PRESENTATION OF DATA PRESENTED FOR ONE OF THE TEMPORARY SSHOC PIDS. THIS EXAMPLE IS SHOWING THE DATA FOR: [HTTPS://DATA.NG-LONDON.ORG.UK/SC-RFLL-PPU5-636I-2DYE](https://data.ng-london.org.uk/SC-RFLL-PPU5-636I-2DYE).

2.2.5 IIIF (<https://iiif.io>)

All images linked within these datasets already had the potential to be fully compliant with the standard International Image Interoperability Framework (IIIF). IIIF is a protocol that ensures that actual images, as well as the metadata describing them, can also be FAIR, providing detailed metadata in a standardised format and direct access to high quality images that can be displayed and shared without the need for users to download images. The SSHOC project ran alongside an AHRC-funded NG project⁴³ that addressed the practical usage of IIIF and aimed to increase uptake among museum professionals, and this piece of SSHOC research was able to make good use of the IIIF project’s outputs. For IIIF purposes, image metadata is grouped in documents; simple image data in “info” or image information files and complete sets of metadata in documents known as IIIF manifests. These formatted files form the backend datasets for IIIF compliant image viewers. Our datasets and output interface are designed such that image presentations could be created on-the-fly from user-defined inputs: if a user wants to see all the

⁴³ “Practical applications of IIIF as a building block towards a digital National Collection” a Foundation project within the AHRC funded Towards a National Collection Programme (<https://tanc-ahrc.github.io/IIIF-TNC/>). [21/04/22]

paintings ascribed to a particular artist, or all the copies of a particular painting, or all the documentation surrounding a specific cross-section sample, an IIIF-compliant manifest could be generated to serve this data. The use of IIIF also increases the Reusability and Interoperability of our datasets, as they can be referenced and reproduced on external websites and apps simply through the inclusion of the IIIF manifests. As an example, all the infrared images associated to a given painting can be found and presented via a single query, see Figure 3.

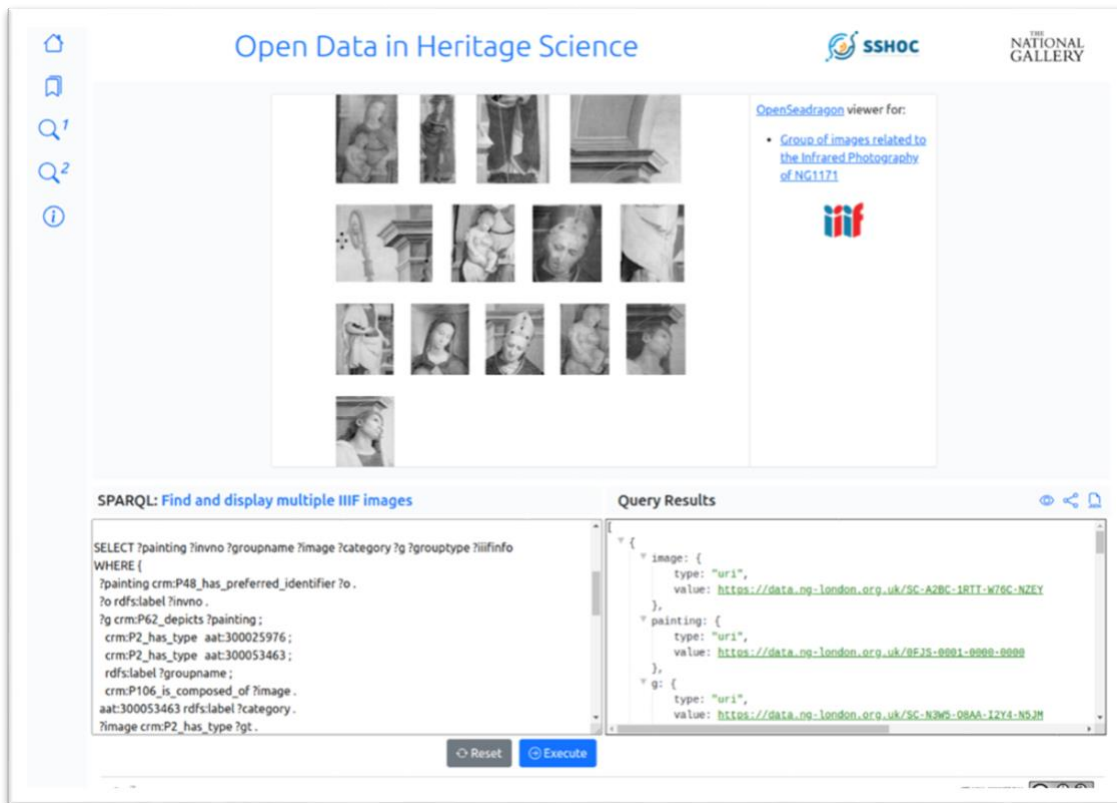


FIGURE 3: A SCREENSHOT OF AN EXAMPLE SPARQL QUERY AND ZOOMING IMAGE PRESENTATION¹ USING THE OPENSEADRAGON¹ IMAGE VIEWER

2.3 Supporting activities

The mapping of the data to CIDOC-CRM, which was expected to represent the bulk of the effort within this part of the task, was carried out by a dedicated Senior Research Fellow at the NG. Due to the impact of COVID-19 the fellowship had to be conducted remotely, so several tools were developed, with support from the SSHOC project, to mitigate the negative impacts of the lack of in-person brainstorming time. These included the Simple Site⁴⁴ system, which enabled a user without programming experience to quickly develop and share a web interface for their work, hosted for free on GitHub, and the Dynamic Modeller⁴⁵, a Mermaid⁴⁶ based tool that simplifies the production of flow diagrams by automatically producing visualisations of relationships and knowledge graphs from simple tab-separated inputs. These tools allowed project staff members to share and compare work with one another and proved extremely useful to the completion of the task quickly and dynamically. These tools and their incorporation into some of the final task outputs have also ensured that some of the process of how the work was carried out could be as FAIR as the final datasets it produced.

2.3.1 Simple Site

Simple Site is a very simple set of processes for creating a standard set of webpages based on a few of json files. It was designed to provide other projects, software or otherwise, an easier way of automatically creating a set of consistent webpages, hosted for free on GitHub, using the existing functionality of GitHub Actions⁴⁷ and GitHub Pages⁴⁸. The idea of developing this Simple site system came from the need to develop an open platform on which to collaboratively present and discuss examples of semantic flow diagrams models, see below. These diagrams had originally been manually created as static word documents but needed to be subsequently exploited within the SSHOC project and developed to create multiple very similar web pages as the models were developed. A custom PHP⁴⁹ script was written to automate the production of the generic aspects of the web-pages such as navigation bars, logos, columns, layout, titles, etc. Additional PHP functions were also created which automatically re-format

⁴⁴ The live example of the Simple Site system can be seen at: <https://jpadfield.github.io/simple-site/> and the code for the system can be found on GitHub at: <https://github.com/jpadfield/simple-site>. [21/04/22]

⁴⁵ The live example of the Dynamic Modeller can be seen at: <https://research.ng-london.org.uk/modelling/> and the code for the system can be found on GitHub at: <https://github.com/jpadfield/dynamic-modelling>. [21/04/22]

⁴⁶ Is a freely available (MIT License) JavaScript based diagramming and charting tool, see: <https://mermaid-js.github.io>. [21/04/22]

⁴⁷ GitHub Actions allow predefined pieces of software to be automatically triggered as a result of users interacting with repositories on GitHub, for examples when they edit a particular file or set of files, for further details see: <https://docs.github.com/en/actions>. [21/04/22]

⁴⁸ GitHub Pages provides a process through which sets of static webpages can be hosted and presented for free on GitHub, for further details see: <https://pages.github.com/>. [21/04/22]

⁴⁹ PHP is a free server side scripting language used to create web pages and run various software processes, for more information see: <https://www.php.net/>. [21/04/22]

simple text-based, model descriptions to create complex interactive flow diagrams. The first set of generic functions were then transferred to GitHub as part of the initial Simple site system.

Subsequent work focussed on more complex data parsing and presentation via specific extensions to the original system⁵⁰. This included the option of creating timeline Gantt charts, simple image galleries, and standard IIIF zooming image presentations, which can all now also be created by editing simple text files.

The use of the free resources, offered by GitHub, along with these new simplified processes, allow groups of researchers to dynamically, collaboratively, explore, discuss and present the results of their work; in an open, reproducible environment, while simultaneously reducing the level of technical know-how required and the need for individual dedicated web-servers. In addition to developing a re-usable collaborative tool for SSHOC, the Simple Site system was also used and directly supported by additional research projects such as the EPSRC funded ARTICT⁵¹ project and 6 AHRC funded foundation projects⁵², which are part of the AHRC Towards a National Collection programme.

The Simple Site system and its documentation have been specifically designed to make the system FAIR – the code for the system has been published and is freely available⁵³ under a defined open license⁵⁴, it is built using open and freely available software libraries⁵⁵, the documentation for the system is created using the system with all of the code and setting available, and in addition to it being published on GitHub the whole system has also been uploaded to the free data repository Zenodo, complete with its own DOI⁵⁶ for citations and referencing.

⁵⁰ For more details about the currently available extensions for Simple Site see: <https://jpadfield.github.io/simple-site/extensions.html>. [21/04/22]

⁵¹ The website for the “Art Through the ICT Lens: Big Data Processing Tools to Support the Technical Study, Preservation and Conservation of Old Master Paintings” or ARTICT project can be found at: <https://art-ict.github.io/artict/>. [21/04/22]

⁵² The GitHub repositories for these projects can be found at: <https://github.com/tanc-ahrc> - further details of the projects themselves can be found at: <https://www.nationalcollection.org.uk/Foundation-Projects>. [21/04/22]

⁵³ <https://github.com/jpadfield/simple-site> [21/04/22]

⁵⁴ <https://github.com/jpadfield/simple-site/blob/master/LICENSE> [21/04/22]

⁵⁵ Such as <https://www.php.net/>, <https://getbootstrap.com/>, and <https://jquery.com/>. [21/04/22]

⁵⁶ The DOI for the Simple Site code dataset on Zenodo -<https://doi.org/10.5281/zenodo.5137663>. [21/04/22]

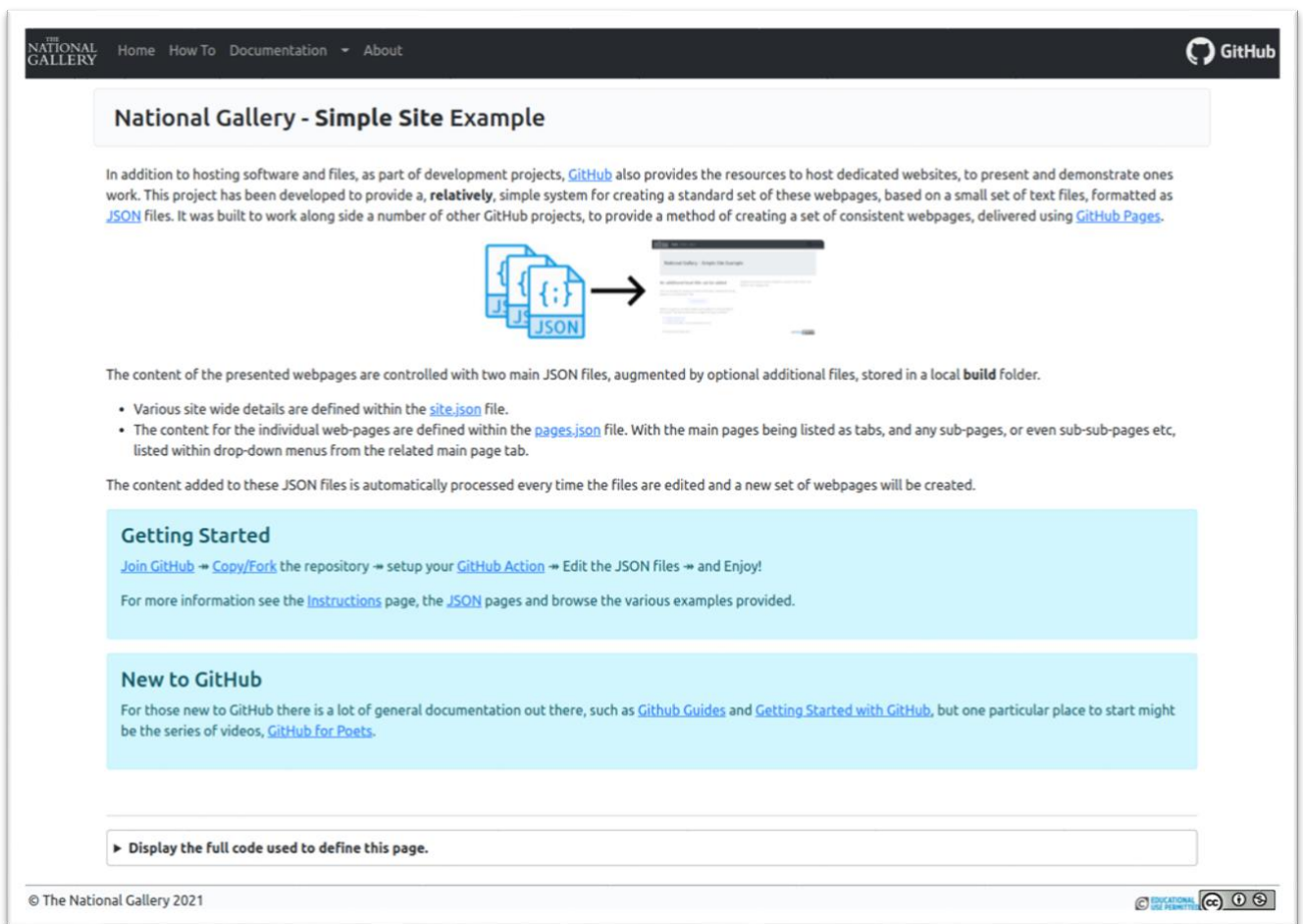


FIGURE 4: A SCREENSHOT OF THE LANDING PAGE FOR THE SIMPLE SITE EXAMPLE AND DOCUMENTATION SITE.

2.3.2 Simple IIIF Discovery

The Simple IIIF Discovery system is built on foundations of several extensions of the Simple Site project and was developed with support from SSHOC along with two other key projects^{32,43}, all of which involved some aspect of making image search and presentation more FAIR. The work was intended to demonstrate how shared international standards, like IIIF, can be used to facilitate simple cross collection interoperability. Basically, making it easier for users to find and explore the rich image content being presented on the web.

The system is based on three main parts: a source, organisation, and then presentation.

- Sources can be any institution or web system that presents IIIF resources via an open, documented API that can process simple keyword-based searches and return related data, including details of the relevant IIIF resources.

- The organisation stage is based on the notion of creating a simplified interface or “end-point”⁵⁷, for each source. These endpoints convert simple keywords into a full API queries and then format the returned results into consistent simple lists of IIIF resources.
- The presentation stage combines a simple search form with the ability to take the simple lists of IIIF resources, provided by one of more endpoints, and automatically presents them with IIIF compliant image viewers.

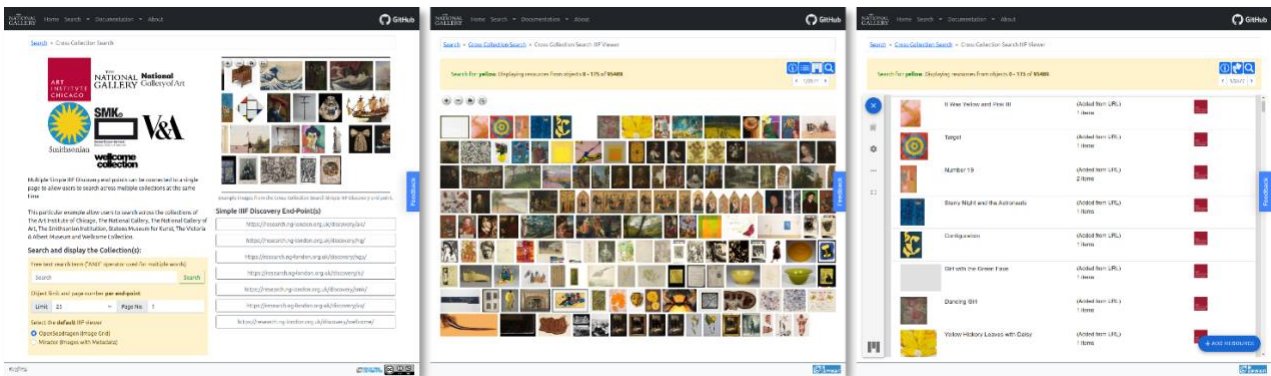


FIGURE 5: SCREENSHOTS FROM THE EXAMPLE SIMPLE IIIF DISCOVERY SITE - SHOWING THE CROSS COLLECTION SEARCH PAGE AND THEN THE RESULTS FOR A SEARCH FOR “YELLOW” PRESENTED IN TWO DIFFERENT IIIF VIEWERS, OPENSEADRAGON⁵⁸ AND MIRADOR⁵⁹.

This combination provides a simple tool that allows researchers to quickly browse potentially huge collections of images without needing to have the time and technical ability to learn how each of the complex source APIs work. This simplified approach, combined across multiple collections, can, for some more generic search terms, result in very large numbers of hits⁶⁰. However, it does provide a very easy way to assess where researcher might want to invest more time.

An example system has been setup to document the system, but also provide users with the option to search across 7 different collections, either individually or as a combined search.⁶¹ Within the example

⁵⁷ The technical details of how these end-points are created and function, including a number of examples, can be found at: <https://research.ng-london.org.uk/ss-iiif/end-points>. [21/04/22]

⁵⁸ Direct access to the OpenSeadragon formatted results can be seen at: <https://research.ng-london.org.uk/ss-iiif/viewer-combined/yellow/25/1/osd>. [21/04/22]

⁵⁹ Direct access to the Mirador formatted results for this search can be seen at: <https://research.ng-london.org.uk/ss-iiif/viewer-combined/yellow/25/1/m>. [21/04/22]

⁶⁰ For example a search for “flowers” in the current cross collection search system returns over 6 million hits, though most of these are from the collection of the Smithsonian Institution: <https://research.ng-london.org.uk/ss-iiif/viewer-combined/flowers>. [21/04/22]

⁶¹ The example system can be found at, <https://research.ng-london.org.uk/ss-iiif>, which currently connect to the collections of The Art Institute of Chicago, The National Gallery, The National Gallery of Art, The Smithsonian Institution, Statens Museum for Kunst, The Victoria & Albert Museum and Wellcome Collection. [21/04/22]

system users can choose to simply browse images within the OpenSeadragon **Error! Bookmark not defined.** viewer or explore a more complex presentation of images, and potentially rich metadata, within the Mirador⁶² viewer. In addition to the documentation for the system, with all the code being published on GitHub⁶³ the whole system has also been uploaded to the free data repository Zenodo, complete with its own DOI⁶⁴ for citations and referencing.

2.3.3 Exploiting previous IPERION-CH CIDOC Models

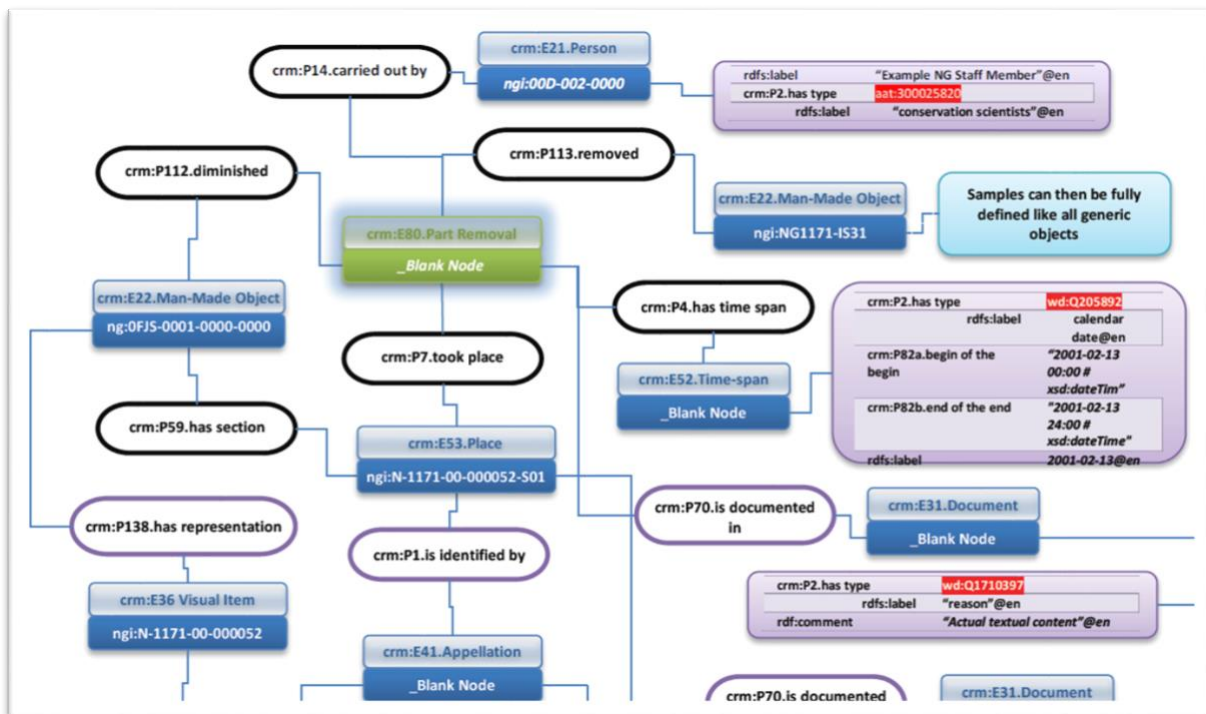


FIGURE 6: A SECTION OF A SCHEMATIC REPRESENTATION OF SOME OF THE MORE COMMON SEMATIC RELATIONSHIPS THAT CAN BE DOCUMENTED DURING THE PROCESS OF TAKING AND DESCRIBING A SAMPLE FROM A GIVEN PAINTING⁶⁵.

Prior to the beginning of the SSHOC task several CIDOC CRM example models had been created, during the IPERION-CH project, describing how the CIDOC-CRM could be used to model aspects of the study of

⁶² Mirador is an "... Open-source, web based, multi-window image viewing platform ..." which is compliant with the IIIF standard: <https://projectmirador.org/>.

⁶³ The code for the Simple IIIF Discovery system can be found at: <https://github.com/jpadfield/iiif-discovery>. [21/04/22]

⁶⁴ The DOI for the Simple IIIF Discovery dataset on Zenodo: <https://doi.org/10.5281/zenodo.5512980>. [21/04/22]

Heritage Science. These models had been documented within one of that project’s deliverables⁶⁵ within Microsoft Word⁶⁶, see

Figure 6, and provided a prototype mapping towards which the larger SSHOC mapping could aspire; they also used Raphael data, from the RRR, as a test case and established much of the logic that would go on to be used during the main mapping work reported here. However, these Word formatted models took quite some time to format and edit, to increase the accessibility and reusability of these models, so that they could be further exploited within SSHOC, they were converted into simple text documents as sets of tab-separated triples within a dedicated GitHub repository⁶⁷. These text files were then combined with an instance of the Simple Site system, described above, to provide an open website which presented each of the models as automatically generated interactive 2D or 3D models⁶⁸ and presented in a dedicated public website⁶⁹, see Figure 7 and Figure 8.

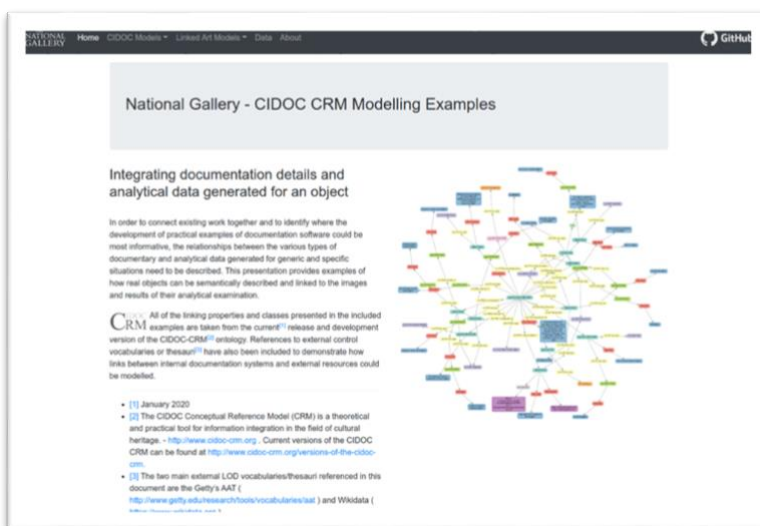


FIGURE 7: A SCREENSHOT OF THE LANDING PAGE OF THE WEBSITE SETUP TO MAKE THE EXISTING CIDOC CRM MODELS FROM THE IPERION-CH PROJECT MORE FAIR.

⁶⁵ Joseph Padfield. (2019). D.8.5 Completed example of prototype designs for integration of various types of documentation and analytical data generated for a single object (1.0). Zenodo. <https://doi.org/10.5281/zenodo.5519016> [21/04/22]

⁶⁶ Commercial software – further details please see: <https://www.microsoft.com/en-gb/microsoft-365/word> [21/04/22]

⁶⁷ All of the prepared model files, describing the IPERION-CH models that were reformatted in SSHOC can be found on GitHub: <https://github.com/jpadfield/cidoc-crm.examples/tree/master/models> [21/04/22]

⁶⁸ The 2D models are presented using the Mermaid Javascript Library (<https://mermaid-js.github.io/>) and the 3D models are presented using updated versions of the D3 Process Map code (<https://github.com/nylen/d3-process-map>). [21/04/22]

⁶⁹ The live presentation website can be found at: <https://jpadfield.github.io/cidoc-crm.examples/> and the related GitHub repository can be found at: <https://github.com/jpadfield/cidoc-crm.examples/>. [21/04/22]

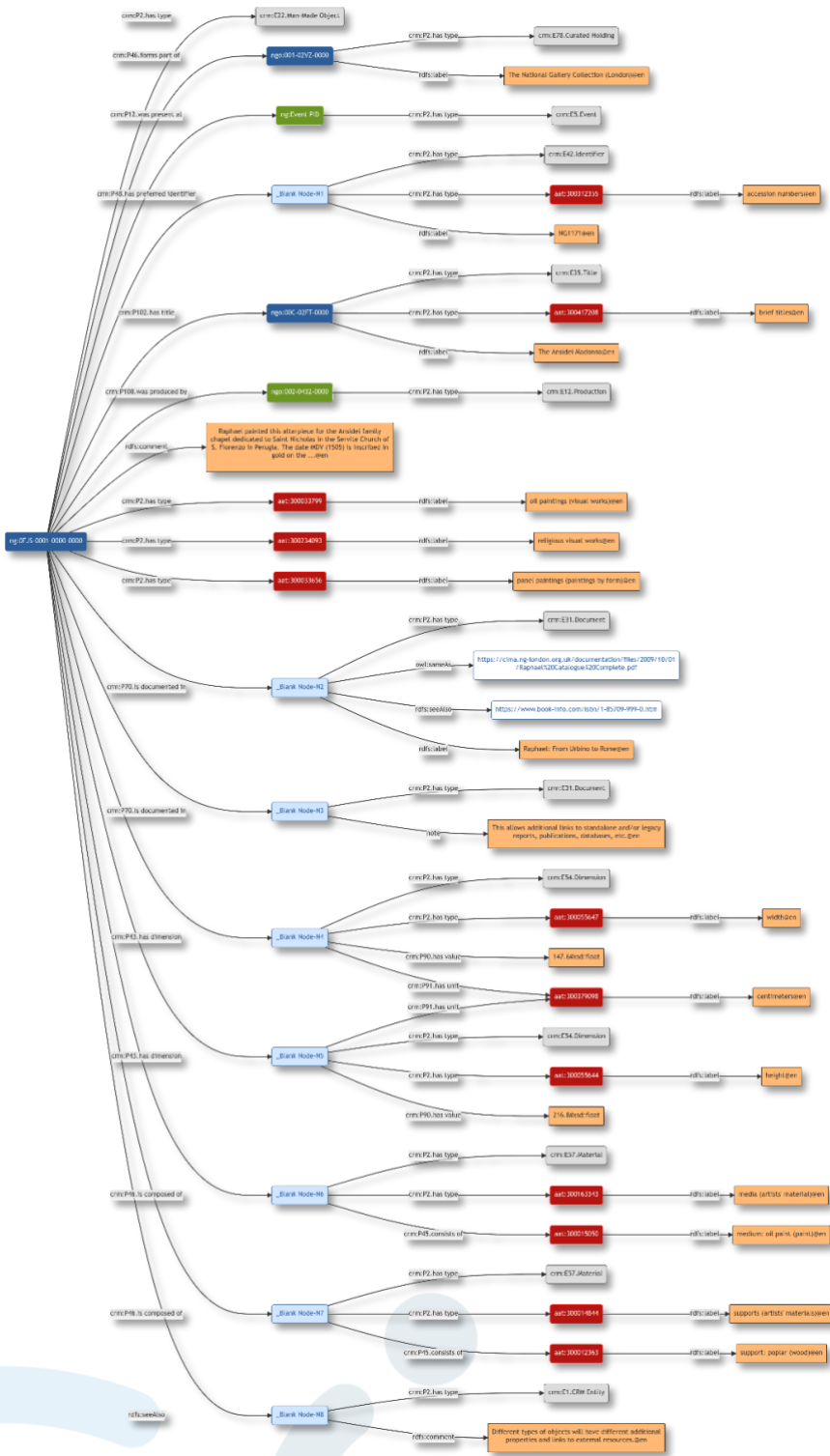


FIGURE 8: AN EXAMPLE SEMANTIC MODEL DESCRIBING A PAINTING IN THE RAPHAEL COLLECTION, CREATED FROM THE PROTOTYPES DEVELOPED IN IPERION-CH

2.3.4 The Dynamic Modeller

To facilitate the live, virtual, collaborative development of semantic maps, required during SSHOC, a more dynamic version of the modelling work done within Simple Site was developed. This system also automatically created flow diagrams directly from simple tab separated sets of data but wrapped it up into a simpler user interface⁷⁰, without the need for a GitHub repository or any other GitHub functionality. The tools simple GUI is split into two sections, see Figure 9; one for the tab separated text and one for the resultant model, with a simple “refresh” button provided to update the model as the text is edited. The system was setup to facilitate modelling discussions during online meetings, with one user sharing the screen and editing the text as required. However, after using the system for a while an even easier, more inclusive mode of use was identified, and the code was updated to allow users to copy and paste text directly from online collaborative spreadsheets.

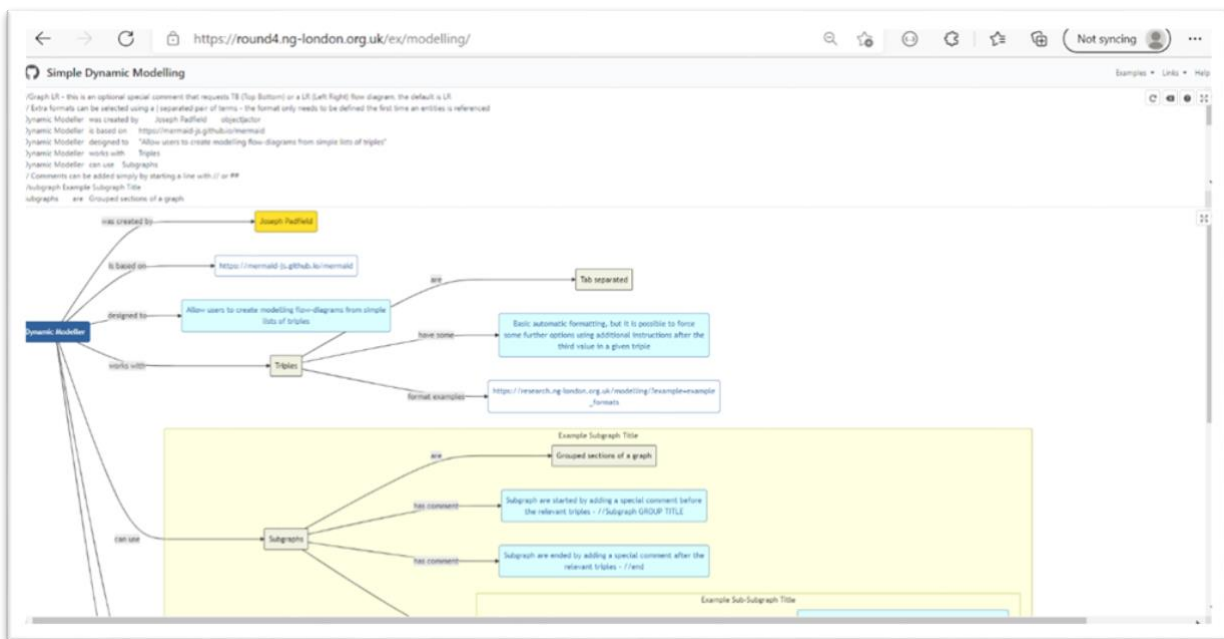


FIGURE 9: THE SIMPLE DYNAMIC MODELLING TOOL

The Dynamic Modelling tool and its documentation have been specifically designed to make the system FAIR – the code for the system has been published and is freely available⁷¹ under a defined open license⁷²,

⁷⁰ The live version of the tool can be found at: <https://research.ng-london.org.uk/modelling/>. [21/04/22]

⁷¹ <https://github.com/jpadfield/dynamic-modelling> [21/04/22]

⁷² <https://github.com/jpadfield/dynamic-modelling/blob/master/LICENSE> [21/04/22]

it is built using open and freely available software libraries⁷³, and in addition to being published on GitHub the whole system has also be upload to the free data repository Zenodo, complete with its own DOI⁷⁴ for citations and referencing and has also be published in the SSHOC Marketplace⁷⁵.

2.3.5 Persistent Identifiers vs Blank Nodes

It was during this stage that the governing principles of the task’s use of PIDs and blank nodes⁷⁶ were established. In general, and throughout the work, PIDs are used as the main identifier for any given datum, and RDF’s blank nodes are used in place of PIDs for data that was identified as less likely to be directly relevant to users beyond its function in this dataset. For example, paintings and other objects are given PIDs, but ephemeral quantities pertaining to them (such as their height and width) were assigned blank nodes. This relates to how semantic relationships are defined; an object would have a “dimension”, which would then have a value and unit etc. Users are expected to search for the “dimension” value of an object rather than the “dimension” semantic node used to connect it to the object.

The assumption made here was that users would query the dataset for objects in their own right and might also search for the events of their production, the collection in which they were held, etc. and other similar quantities, but would be less likely to search directly for more generic quantities such as dimensions, preferred identifier, or specific materials without them being directly connected to specific objects. This logic persisted throughout the work and went on to underpin both output datasets.

2.4 The Mapping Work

2.4.1 Preparatory steps

Traditionally, a data mapping can be divided into two broad phases: a phase of ideation, in which the theoretical structure of the mapping is laid out, and an implementation phase. However, partially due to the constraints imposed by the COVID-19 pandemic, which necessitated that work be done entirely remotely, these phases were compounded, and ideas were prototyped and coded simultaneously. This enabled outputs to be created on the fly that could then be passed into the Dynamic Modeller and

⁷³ Such as <https://www.php.net/>, <https://mermaid-js.github.io/>, <https://getbootstrap.com/>, and <https://jquery.com/>. [21/04/22]

⁷⁴ The DOI for the Dynamic Modelling code dataset on Zenodo - <https://doi.org/10.5281/zenodo.4724103>. [21/04/22]

⁷⁵ SSHOC Market place record for the Dynamic Modeller - <https://marketplace.sshopencloud.eu/tool-or-service/ASSQXO>. [21/04/22]

⁷⁶ Blank nodes are anonymous resources used to establish semantics connections between defined entities within an RDF knowledge graph – further information describing how they work can be found at: https://en.wikipedia.org/wiki/Blank_node. [21/04/22]

presented for discussion over video call; this was intended as a substitute for the post-its and whiteboard scribbling that would ordinarily make up the ideation phase. Doing the mapping this way required a technical system to be put in place from the outset, so the initial phases of the project focussed on establishing a methodology and a broad code structure. From there, different iterations of the mapping could be tested seamlessly by changing small amounts of code.

Several different technologies were explored in the early phases of the mapping. Previous Gallery mappings had been conducted in PHP, which allowed their code to be integrated into webpages and showcased without having to switch languages, but the specific code was outdated relative to modern web development technologies and lacked the specific data science capabilities that would be crucial to handling a complex database such as the Grounds. Open source tools such as Protegé³¹ and 3M³⁰ were considered: they each share the advantage of outputting an XML file describing the logic used in the mapping alongside the mapping outputs themselves, but ultimately were unable to handle the level of specific detail this task required. Python was selected early on based on team capabilities and turned out to be a suitable and flexible technology that brought the mappings up to open source data science standards – arguably a more FAIR choice owing to its extremely broad user base.

Once the technology was chosen, the basic architecture of the mapping code needed to be set out. Within Python's RDFlib package it is possible to load data into a construct known as a Graph, which is an unsorted container that forms the basic structure for any set of RDF code. Graphs hold data in the form of triples, or in Python terms, tuples that resemble RDF triples. Data loaded into a Graph can be queried locally using SPARQL or a Pythonic SPARQL-like set of commands, and manipulated before being passed into a new, output Graph. An output Graph can also be built from data in any other format if it is passed into the new Graph in triplicate form. This functionality provided the basis for this project's code structure: data is loaded into Python, a query isolates the relevant data for a particular part of the mapping (by pulling e.g., all data of RDF type 'Painting'), and new CIDOC-CRM triples are created to describe that part of the mapping and loaded into an RDFlib Graph. From here the data can be exported in common formats such as XML and Turtle and loaded into triple stores to continue development.

Essential to this process is the project's use of PIDs. Because the mapping would involve looping through the same data multiple times, it was crucial that the ability to identify a specific datum using a PID was present from the outset to prevent duplication. For this reason, the previously described system of introducing placeholders where no Gallery PID was available was implemented early in the project. Initially, PIDs were to be stored in an ElasticSearch database, but after some testing it was determined that the fuzzy matching of ElasticSearch would only complicate the process of locating a PID, and a MySQL database would be the simpler (and less computationally expensive) choice. The primary MySQL database of temporary PIDs is stored currently on a local server, but as indicated before they are still fully resolvable, with the relevant relationships formatted as html or JSON.

2.4.2 Mapping Raphael

The first of the two datasets to be mapped was that of the RRR. The input for this mapping was an XML file⁷⁷ containing the RRR dataset expressed in linked data format but using an ontology that had been developed specifically for that dataset. Full ontology documentation was also available in XML format⁷⁷. The basic relations described in the RRR ontology are similar in nature to those described in the CIDOC-CRM: at the centre of the ontology is the class 'RC12.Painting', linked using the property 'RP24.has owner' to an 'RC41.Institution', as well as to an 'RC40.Person' by way of e.g., an 'RC32.Production' event. This use of events as a structuring principle is akin to their use in the CIDOC-CRM, in which no entity is ever presumed to be permanent, rather is temporally bordered in the same way as it might be geographically bordered.

A painting, or object in CRM terms, was taken as the starting point for the mapping. Initially it was difficult to isolate the triples in the input dataset that could be described as referring to any given painting, as the nature of linked data means that the web of relations originating from one object can in theory be infinite. For this reason, somewhat arbitrary differentiations had to be made between entities so that the code could remain relatively simple: for example, a painting's mapping would end with a reference to the PID of an event, which would itself be mapped at a later point. Eventually, the code was divided into a total of eight distinct but overlapping mapping functions, which taken together described the totality of the data; this was not however fully formalized until later in the process of mapping the Raphael data. For this reason, mapping the first test painting (NG1171 "The Ansidei Madonna"⁷⁸) was a protracted process, necessitating the development not only of the mapping function but also of various functions that performed basic tasks that would go on to be key to the rest of the mapping.

One such task was integrating external vocabularies into the mapping. The previous RRR ontology and implementation had suffered from its lack of interoperability: concepts such as 'oil paint' were referenced but not linked to any universal resource that would identify them as the same 'oil paint' referred to in another museum's database, for example. Two vocabularies were used to provide universal identifiers for certain artistic concepts, as well as well-known entities like artists and cultural heritage institutions: these were Wikidata⁷⁹ and the Getty's Art and Architecture Thesaurus (AAT)⁸⁰. Initially, it was hoped that the mapping of concepts could be automated such that a reference to 'oil paint' in the NG's database would be inserted into a query against the AAT, and the resulting ID pulled into the new mapping's Graph.

⁷⁷ The XML files describing the original bespoke ontology and the data instances can be sourced directly (<https://rdf.ng-london.org.uk/2019/09/>) but they have also been included, as inputs within the published modelling code on GitHub (https://github.com/jpadfield/sshoc_raphael_modelling) and on Zenodo (<https://doi.org/10.5281/zenodo.6461653>). [21/04/22]

⁷⁸ The original data presentation for this painting can be seen at https://cima.ng-london.org.uk/documentation/index.php?object_code=NG1171 or the current NG public website can be seen at <https://www.nationalgallery.org.uk/paintings/raphael-the-ansidei-madonna>. [21/04/22]

⁷⁹ The public website for Wikidata can be found at: <https://www.wikidata.org/>. [21/04/22]

⁸⁰ The public website for Getty's Art and Architecture Thesaurus (AAT) can be found at <http://www.getty.edu/research/tools/vocabularies/aat/index.html>. [21/04/22]

However, this would have required either an advanced system of fuzzily matching Gallery inputs to related but differently worded AAT concepts, or the use of natural language processing (NLP) algorithms to determine similarity between entries in both databases. Neither approach was within the scope of this project, so much of the AAT and Wikidata integration work was done manually, by pulling terms from the RRR into a spreadsheet⁸¹ and choosing corresponding terms from the two vocabularies that could replace the original terms in an output dataset. This introduced the possibility of human error in interpreting terms and choosing the most suitable synonym: whether this error is greater than that which would be introduced by NLP remains unknown at this time.

Over the course of the test mapping of NG1171 a code structure was put in place that would form the basis of the rest of the data migration. First, a function called 'map object' is called, see Figure 10. This function queries the RRR database to locate entities of RDF type 'RC12.Painting', then loops through the results and queries the database again to find all triples of which they are the subject, thereby creating a subset of the data that pertains only to those paintings. A PID is generated or selected for each painting, and the subset looped through again to isolate triples that hold specific information about the painting: its title, for example, or its measurements. Information from these subsets-of-subsets provides the inputs for functions such as 'create title triples', which uses these inputs to build CIDOC-CRM triples that are stored in a new Graph. Each characteristic that requires mapping has a corresponding function like 'create title triples', and these triple-creation functions are written to be relevant to various parts of the mapping; for example, the same 'create location triples' function might be called by all of 'map object', 'map institution' and 'map sample', creating outputs that follow the same pattern for each. This construct has an economising effect, reducing duplication and making the mapping code more readable as well as more replicable.

```
def map_object(new_graph, old_graph):
    for painting_id, _, _ in old_graph.triples((None, RDF.type, getattr(RRO, 'RC12.Painting'))):
        for subj, pred, obj in old_graph.triples((painting_id, None, None)):
            subject_PID = generate_placeholder_PID(subj)
            new_graph = create_title_triples(new_graph, subject_PID, subj, pred, obj)
            new_graph = create_medium_triples(new_graph, subject_PID, subj, pred, obj)
            new_graph = create_collection_triples(new_graph, subject_PID, subj, pred, obj)
            new_graph = create_dimension_triples(new_graph, subject_PID, subj, pred, obj)
            new_graph = create_identifier_triples(new_graph, subject_PID, pred, obj)
            new_graph = create_type_triples(new_graph, subject_PID, pred, obj)
            new_graph = create_location_triples(new_graph, subject_PID, subj, pred, obj)

    return new_graph
```

FIGURE 10: THE 'MAP OBJECT' FUNCTION IN THE RAPHAEL MAPPING CODE

⁸¹ The spreadsheet detailing the external terms used in this work is included here as an input within the published modelling code on GitHub (https://github.com/jpadfield/sshoc_raphael_modelling) and on Zenodo (<https://doi.org/10.5281/zenodo.6461653>) [21/04/22]

For straightforward-to-describe entities such as a painting, the mappings to CIDOC-CRM followed the broad contours outlined in the earlier IPERION-CH example, see Figure 8. As well as using terms from CIDOC-CRM, the outputs were also made compatible with RDF-based technologies: each entity within the Graph had an RDF type and an RDFS label, as well as a CRM type. This duplication was intentionally done so that anyone querying the dataset could take either a purely CRM-based approach or opt to use RDF terms. RDFS comments were also used to incorporate notes – either those that were already in the input data, or new notes that were inserted to describe, for example, the process of taking a sample from a painting. Sometimes RDFS comments were used to describe the reasoning behind an action; in this case they were also attached via a blank node and RDF type to a Wikidata ID for the concept of a ‘reason’ and treated as the text of documents. The full planned semantic model, used in SSHOC, for a given an example painting (object) can be seen in Figure 12.

Some creative decisions were made to translate the spirit of the original dataset into a CIDOC-CRM compatible format. In particular, the RRR input dataset had made liberal use of a construct it described as a ‘project category’, effectively a tagging system under which disparate entries were grouped. These tags could describe categories as broad as ‘art history’ or as specific as ‘photomicrographs’ or ‘x-ray images’.

CIDOC-CRM lacks a direct means of dealing with loose semantic tagging, likely owing to the fact that it disrupts the ‘linked-ness’ of a dataset, parcelling pieces of data out into broadly-defined groups rather than mapping the connections between them; however, in the RRR dataset the categories provided a neat way of navigating a user interface that sought to display and compare images in groups (or IIF manifests), so it was important to retain some of the logic of tagging in our new mapping. This was achieved in two ways, initially, where possible, the specific relationships indicated by the categories were more fully modelled. For the more specific tags this was straightforward: an image tagged ‘photomicrographs’ simply became an object measurement whose technique was photomicrography, and which produced the image in question. Broader image categories, such as ‘art history’, were harder to replace with discreetly defined relationships and in these cases the approach was either done away with or the relationship was turned into broader overarching ideas, as was done in the case of the ‘provenance’ category. To describe provenance a ‘history’ event was created for each image in the collection. Any event that would have been tagged ‘provenance’ in the old dataset became a sub-entity of the history event, and documents that described a painting’s provenance were linked to the history event using the property ‘P70 documents’. This approach transformed the tagging system of the old dataset into a useful means of creating specificity in the new one: using IDs from the AAT, the techniques previously described using tags could easily be subsumed into the dataset itself and made into linkable pieces of information.

The second approach to mirror the project categories from the RRR, specifically used for images and digital texts, which provided a more complete representation of the categories, was achieved through the creation of “groups”, of items that depicted (crm:P62) each painting, and giving the group an additional type based on the required category, see Figure 11.

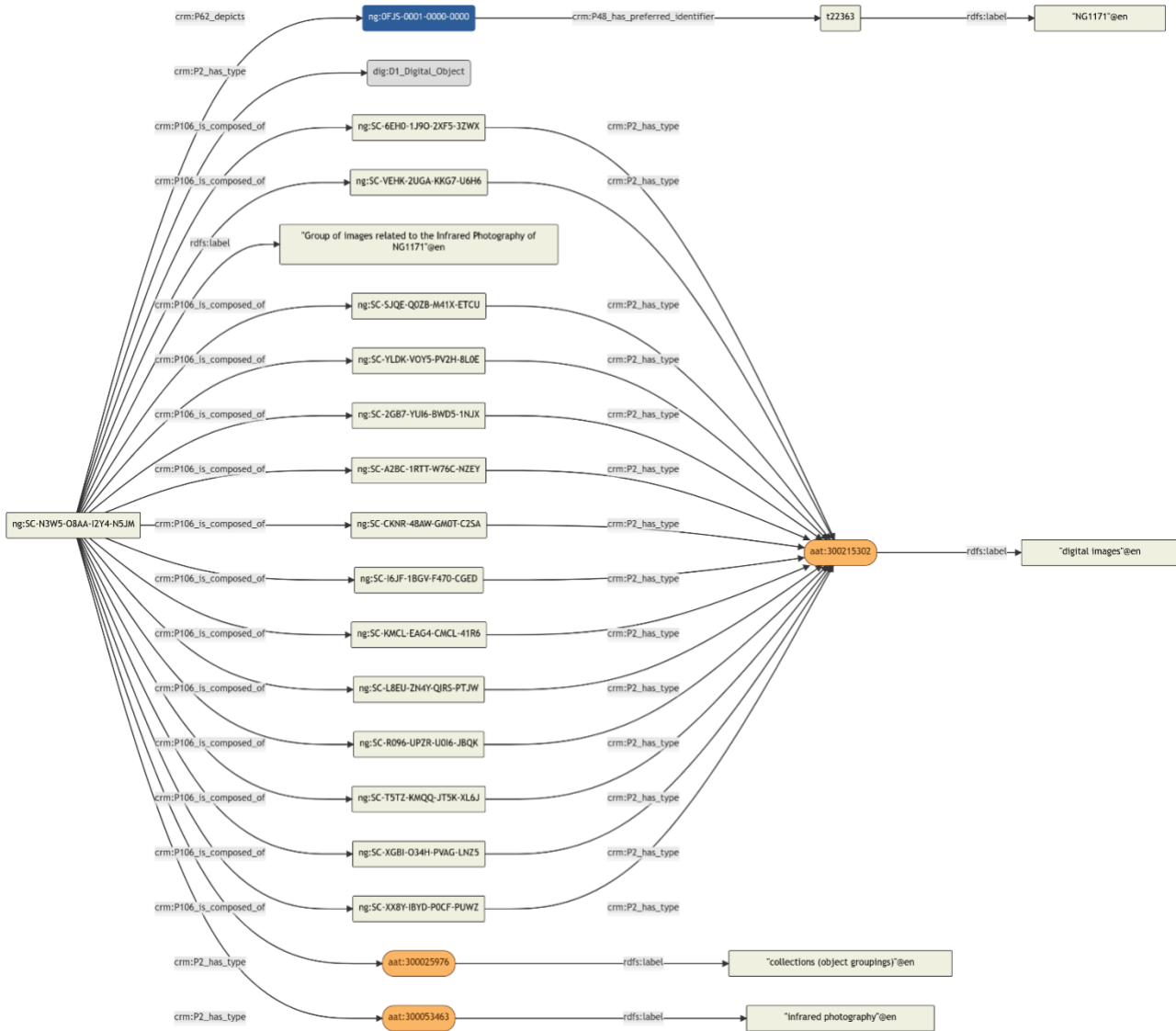


FIGURE 11: AN EXAMPLE DIAGRAM DEMONSTRATING THE RELATIONSHIPS CREATED IN RELATION TO THE FORMATION OF CATEGORY BASED GROUPS OF IMAGE OR DIGITAL TEXTS.

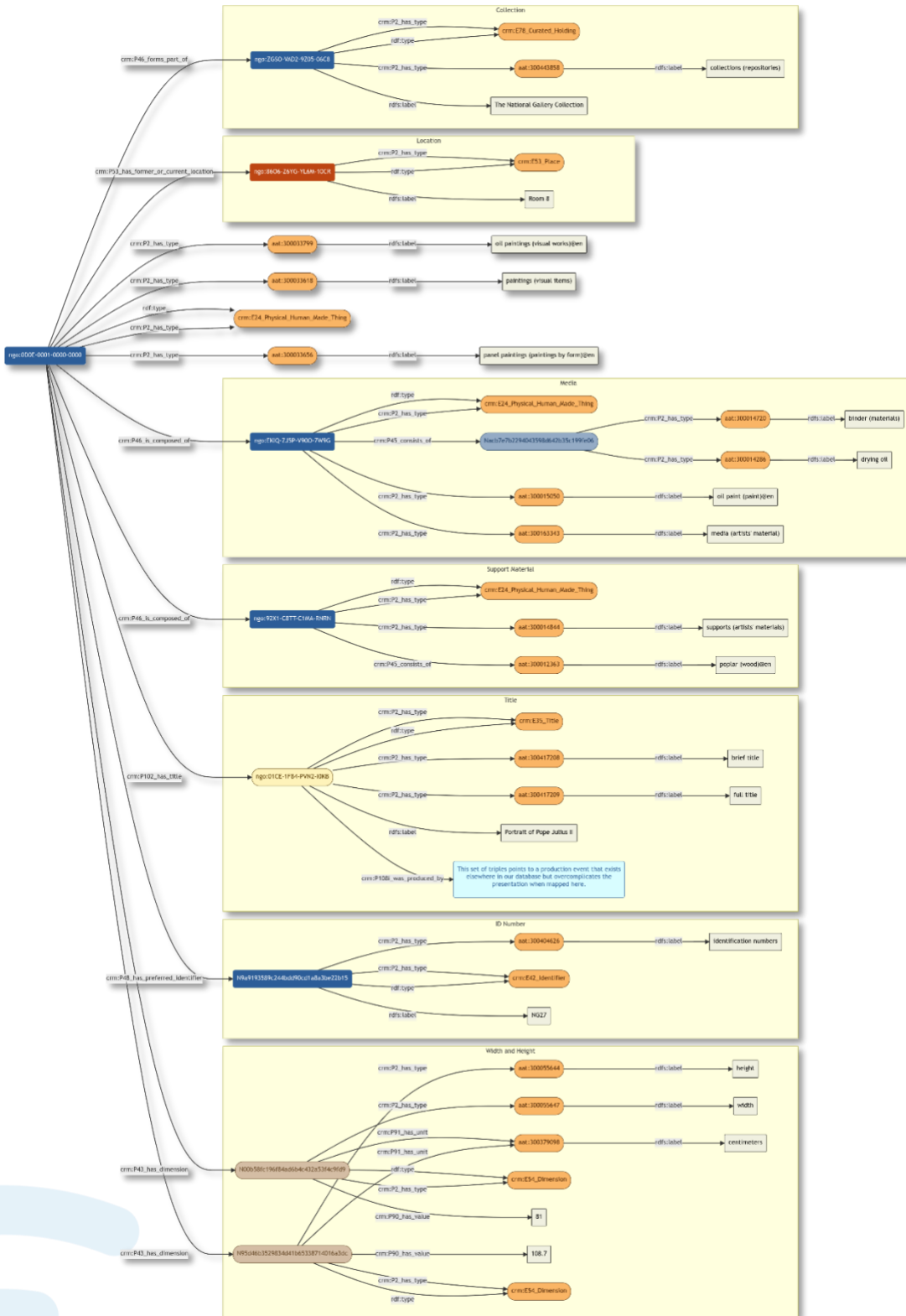


FIGURE 12: A VISUAL REPRESENTATION OF THE PLANNED SEMANTIC RELATIONSHIPS CREATED TO DEFINE AN EXAMPLE PAINTING.

The greatest challenge confronted in the mapping was the question of how to describe an image file in CIDOC-CRM. In the old RRR dataset, images were treated purely as files, with metadata, see Figure 13, only connected to their subject using the property ‘RP40.is related to’. The CRM, however, demands a more holistic view of the relationship between image, file and subject, a view that is expanded through the incorporation of the compatible model CRMdig. CRMdig’s remit is to describe in detail the production of digital derivatives or representations of a work, which was especially useful for this project’s use case of describing files as both semi-ephemeral versions of an image and actual computational structures that could be described using URIs and manifests. A file, then, was mapped both as an ‘E24 Physical Human-Made Thing’ (basic CRM), and as a ‘D1 Digital Object’ (CRMdig). The CRMdig concept ‘D3 Formal Derivation’ described the relationship between files as nonspecific digital entities and their representations as pyramids or thumbnails. Blending basic CRM and CRMdig in this way allowed files to be treated as a hybrid format: they could have pixel-specific dimensions as well as being derived from and representative of an image in a symbolic sense, without entirely separating the file-as-digital-object from the visual image that appeared when the file was opened, see Figure 14.

```
<rdf:Description rdf:about="DSC_0480">
  <rdf:type rdf:resource="https://rdf.ng-london.org.uk/raphael/ontology/RC25_Image"/>
  <rdfs:label xml:lang="en">DSC_0480</rdfs:label>
  <rro:RP17.has_identifier xml:lang="en">DSC_0480</rro:RP17.has_identifier>
  <rro:RP95.has_file_name xml:lang="en">DSC_0480.jpg</rro:RP95.has_file_name>
  <rro:RP233.has_caption xml:lang="en">The Virgin and Child (The Bridgewater Madonna), with frame, NGL065.46. The Duke of Sutherland Collection, on loan to The National Gallery of Scotland @ National Galleries of Scotland.</rro:RP233.has_caption>
  <rro:RP98.is_in_project_category rdf:resource="RCL195.Framing"/>
  <rro:RP40.is_related_to rdf:resource="NGL065.46"/>
  <rro:RP14.has_file_size xml:lang="en">3115182</rro:RP14.has_file_size>
  <rro:RP225.has_width_in_pixels xml:lang="en">2848</rro:RP225.has_width_in_pixels>
  <rro:RP227.has_height_in_pixels xml:lang="en">4288</rro:RP227.has_height_in_pixels>
  <rro:RP30.has_pyramid xml:lang="en">pics/tmp/raphael_pyr/2009/07/14/DSC_0480-PYR.tif</rro:RP30.has_pyramid>
  <rro:RP86.has_no_of_pyramidal_levels xml:lang="en">6</rro:RP86.has_no_of_pyramidal_levels>
  <rro:RP5.has_bit_depth rdf:resource="RCL243.8-bit"/>
  <rro:RP15.has_format rdf:resource="RCL90.jpeg"/>
  <rro:RP96.has_file_path rdf:resource="pics/tmp/raphael_files_2009_07_14"/>
  <rro:RP235.has_order_code xml:lang="en">DSC_0480</rro:RP235.has_order_code>
  <rro:RP243.has_pyramid_server rdf:resource="cima.ng-london.org.uk/fcgi-bin/lipsrv.fcgi"/>
  <rro:RP259.has_thumbnail xml:lang="en">http://cima.ng-london.org.uk/fcgi-bin/lipsrv.fcgi?FIF=pics/tmp/raphael_pyr/2009/07/14/DSC_0480-PYR.tif&QL=75&cnt=1&mid=256&CVT=JPEG</rro:RP259.has_thumbnail>
</rdf:Description>
```

FIGURE 13: TRIPLES DESCRIBING AN IMAGE IN THE OLD RAPHAEL DATA (XML FORMAT)

```
ngo:RMWM-DEIX-ICO0-33P0 a crm:E24_Physical_Human_Made_Thing,
  dig:D1_Digital_Object ;
crm:P108i_was_produced_by ngo:2X9X-0051-ULBW-RARB ;
crm:P149_is_identified_by [ a crm:E42_Identifier ;
  rdfs:label "DSC_0480.jpg"@en ;
  crm:P2_has_type crm:E42_Identifier,
  wd:Q1144928 ] ;
crm:P2_has_type aat:300215302,
  aat:300240903,
  aat:300266224,
  crm:E24_Physical_Human_Made_Thing,
  dig:D1_Digital_Object ;
crm:P43_has_dimension [ a crm:E54_Dimension ;
  crm:P2_has_type aat:300410392,
  crm:E54_Dimension,
  wd:Q270159 ;
  crm:P91_has_unit aat:300265866 ],
[ a crm:E54_Dimension ;
  crm:P2_has_type aat:300265863,
  crm:E54_Dimension ;
  crm:P90_has_value 3.115182e+06 ;
  crm:P91_has_unit aat:300265869 ] ;
crm:P62_depicts ngo:8UR9-IMT7-8G95-6X9A ;
crm:P70_is_documented_in [ a crm:E31_Document ;
  rdfs:label "The Virgin and Child (The Bridgewater Madonna), with frame, NGL065.46. The Duke of Sutherland Collection, on loan to The National Gallery of Scotland @ National Galleries of Scotland."@en ;
  crm:P2_has_type crm:E31_Document ] ;
crm:P70i_is_documented_in ngo:4CBT-XTBS-GIHR-D730 ;
dig:L56_has_pixel_width 2.848e+03 ;
dig:L57_has_pixel_height 4.288e+03 .
```

FIGURE 14: TRIPLES PARTIALLY DESCRIBING THAT SAME IMAGE IN THE NEW RAPHAEL DATA (TURTLE FORMAT)

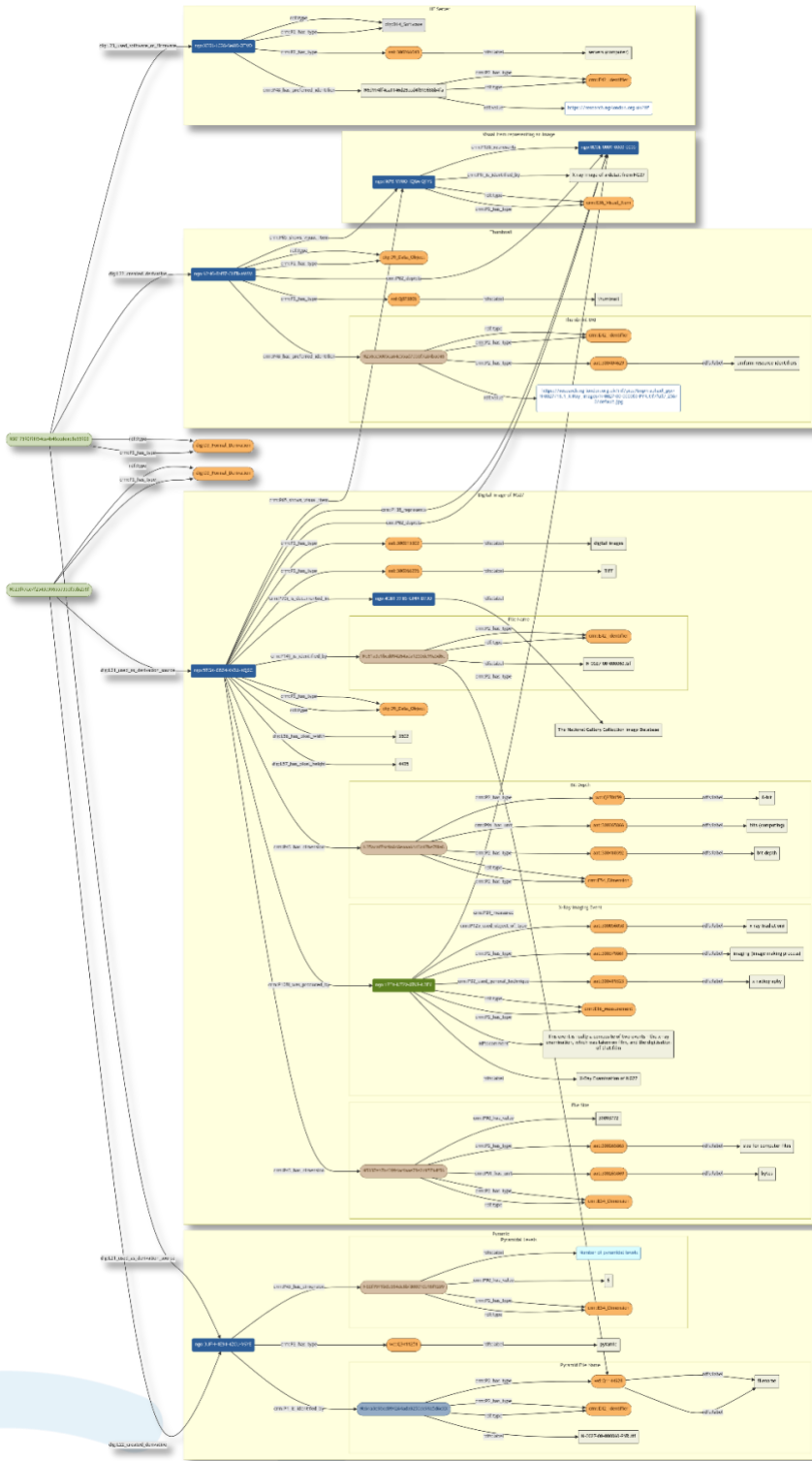


FIGURE 15: A VISUAL REPRESENTATION OF THE PLANNED SEMANTIC RELATIONSHIPS CREATED TO DEFINE A GIVEN IMAGE AND THE VARIOUS DERIVATIVES CONSIDERED IN THE SYSTEM.

The Raphael Research Resource introduced a new and unusual use case for CIDOC-CRM, in the sense that the CRM was used here to update an already existing linked dataset. Beyond the technical challenges associated with building a mapping software from scratch, the RRR also brought an unprecedentedly subjective point of view on its data that had to be either preserved or overhauled. The CRM proved a suitable tool to contend with this, partially because the temporal philosophies of the two ontologies had already been in alignment, and partially due to the ability of the CRM and its extensions to contend with multiple class instantiations for the same object.

2.4.3 Mapping the Grounds Database

The Grounds database was the second of the two main datasets to be mapped to the CIDOC-CRM. During its building phase, which was part of the IPERION-CH project²⁵ undertaken in 2018, the plan to transform it into a linked dataset by migrating it to the CRM was already in existence, so the database was constructed with this in mind, see **Error! Reference source not found..** A number of tables include reference to the AAT vocabulary, and it shares the same event-based structure as the RRR's ontology and the CRM itself. Efforts were also made to standardise what could otherwise have been open-input fields: characteristics such as colour are limited by a drop-down list in the database's input interface. The Grounds database is largely made up of data held by the NG, though there is information from other institutions, and it is still being circulated and accepting contributions from other holders of data on 16th century Italian preparation layers and beyond. This fact means that its standardisation can never be guaranteed: data belonging to other institutions is often patchier than NG data, and won't have been held in the same structure, so the mapping needed to account for this disparity in input data by being flexible and reactive to its inputs in its design.

It was initially planned that much of the code that had been written for the RRR would be reused in the Grounds database mapping. This was possible to an extent, but the vastly different structures of both datasets necessitated an entirely new overarching code structure. One major issue came from the size of the Grounds database and the need to store its tables temporarily whilst running the mapping code. Specifically, in order to assemble input values for quantities that would be connected together in triplicate within CIDOC-CRM, a number of tables would have to be joined: for example, the notion that 'The Ansidei Madonna' was created in a production event by Raphael in 1505 can be expressed using four CRM triples, but uses information from a minimum of four tables in the Grounds database, as well as the tables that act as bridges interconnecting those source tables. Initially, the database, which was held in MySQL on a remote server, would be loaded in its entirety into Python's Pandas library, and the table joins necessary to create the output triples conducted locally within Python. However, this was quickly found to be too computationally expensive to be sustainable given the complexity of the joins, so a solution needed to be found within the MySQL database itself. This solution came in the form of SQL views, temporary tables that could be created within SQL and loaded into Pandas in the same manner as ordinary SQL tables. The creation of SQL views was done in tandem with the mapping itself, as it was difficult to plan, ahead of time, precisely what views would be needed.

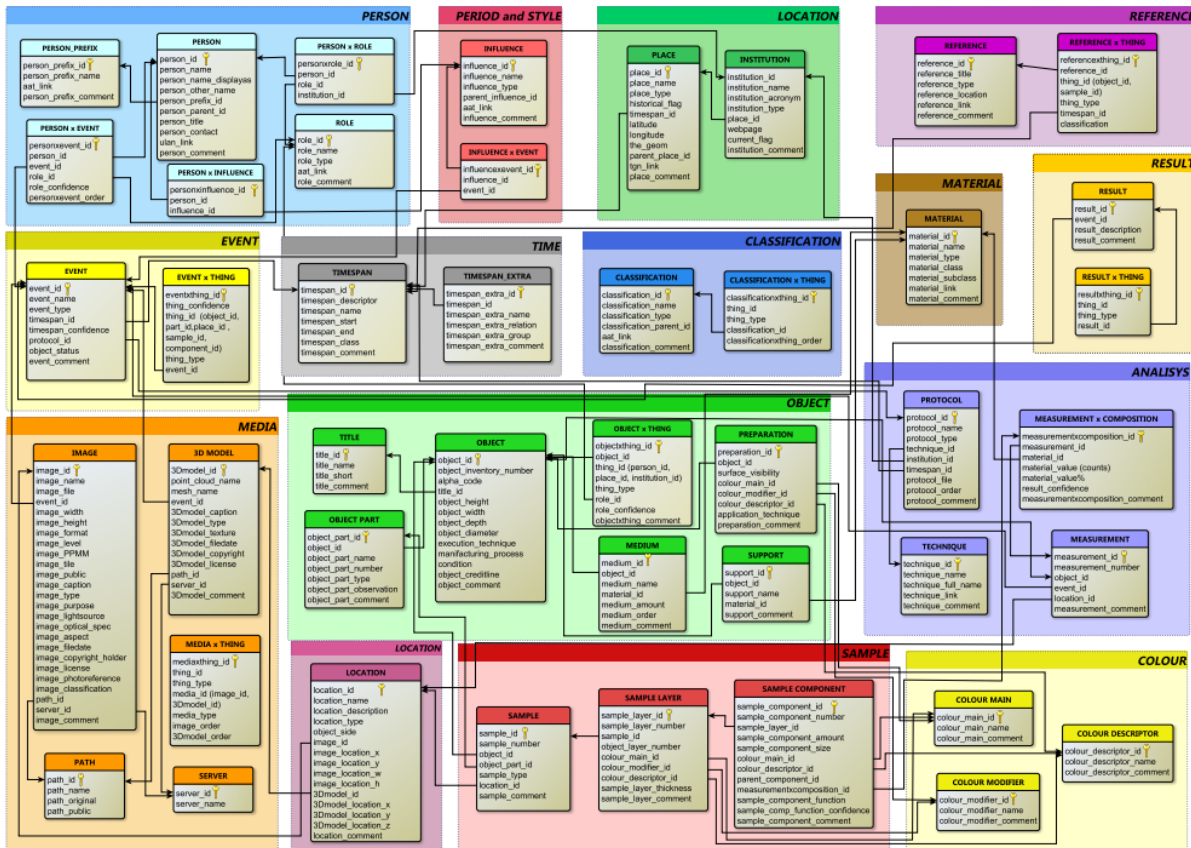


FIGURE 16: A DIAGRAM OF THE TABLES IN THE GROUNDS MYSQL DATABASE, SHOWING HOW THE TABLES ARE GROUPED AND CONNECT TOGETHER.

Conceptually, the Grounds data is similar to the RRR in the sense that it is structured around the idea of an object, which represented a painting in the database. Objects are linked to events, which are linked to people, timespans, samples and so on. One key process that makes up much of the detail in the Grounds database is sampling: as can be seen in the table diagram, see Figure 16, three tables are devoted in their entirety to the examination and sub-examination of cross section samples. It was important for the mapping to deal with hierarchical subjects, both in the context of the samples and in other tables where the concept of a ‘parent’ appears. Additionally, the prominence of samples in the data would necessitate that attention be paid to how objects and their parts were represented in the CRM; the extension CRMsci became a useful tool in expressing this.

The broad code structure put in place to map the RRR was transferable to Grounds work with only minor changes. As before, the mapping was broken up into overlapping functions that, taken together, produced triples to describe the data as a whole. Seven basic categories make up the Grounds mapping, each with a corresponding mapping function (such as the ‘map object’ construct described in the previous section). Once all the SQL views are ingested into Pandas data-frames, these data-frames are passed in groups as the inputs to the mapping functions: because of the functions’ complexity, a single one might require up to six SQL views as its inputs. Within the mapping function, each data-frame is looped through, and its values passed to various triple-creation functions, many of which are lifted wholesale from the

RRR's mapping code. PIDs are created for the values that need to persist, and newly created triples are loaded into a new RDFlib Graph, which is slowly built as the mapping progresses and eventually is output in XML or Turtle format.

```

CREATE VIEW object_event_influence AS
SELECT
  e.*,
  i.influence_name,
  i.influence_id
FROM
  (
    SELECT
      eo.*,
      t.timespan_name,
      t.timespan_start,
      t.timespan_end,
      t.timespan_descriptor,
      t.timespan_comment
    FROM
      timespan t
    INNER JOIN (
      SELECT
        e.*,
        o.object_inventory_number,
        o.object_creditline,
        o.manufacturing_process
      FROM
        object o
      RIGHT JOIN (
        SELECT
          e.event_name,
          e.event_id,
          e.event_type,
          e.timespan_id,
          e.timespan_confidence,
          et.thing_id
        FROM
          event e
          INNER JOIN (
            SELECT
              *
            FROM
              eventXthing
            WHERE
              thing_type = 'object'
          ) et ON e.event_id = et.event_id
        ) e ON o.object_id = e.thing_id
      ) eo ON t.timespan_id = eo.timespan_id
    ) e
  LEFT JOIN (
    SELECT
      i.influence_name,
      i.influence_id,
      ie.event_id
    FROM
      influence i
      INNER JOIN influenceXevent ie ON i.influence_id = ie.influence_id
    ) i ON e.event_id = i.event_id
  
```

FIGURE 17: SQL CODE TO CREATE A VIEW JOINING OBJECT, EVENT, AND INFLUENCE TABLES

The major distinction between the two datasets was the attention paid to samples within the Grounds database, see Figure 18. This built on the image mapping already developed for the RRR: samples were digitally imaged as well as mounted in resin and stored in a physical store, and the mapping needed to account for their imaging as well as the specific processes involved in their acquisition and storage. Splitting each sample into two ideas, joined with a unifying PID, enabled this. Along with the RRR's image logic, CRMsci was used to build a new set of logic to describe the sampling and mounting process. CRMsci differentiates a sample from its sample site and expresses the change in state of an object as a result of having been sampled as an E13 Attribute Assignment, a construct familiar from base CRM. This attention to both the sample and the object from which it was taken became crucial to the mapping of a database that gave each equal importance and weight. Though CRMsci simplified the description of sampling in the mapping, it was still necessary to develop a logic that would express samples in terms of their constituent parts as well as handle observations about those parts, such as their colour or their chemical

The strict hierarchy of the Grounds database's sample tables is representative of a more subtle notion of hierarchy evident throughout the database. Unusually, the database contains a table detailing artists' influences, which suggest a sense of lineage that is both temporal and ideological. CRMInf was investigated as a means of translating this to CIDOC-CRM but was more complex than this use case required; instead, base CRM's 'P15 was influenced by' was used in conjunction with an overarching event describing the creation of an artist's body of work, to satisfy P15's domain condition ('E7 Activity'). What is lacking in this representative choice is any acknowledgement of the subjectivity of the idea of influence: it would possibly be preferable to express a statement of influence with reference to the person or entity making that statement, but this information was not available in the input dataset. The 'influence' table also introduced the idea of a 'parent' entity that would recur throughout the database. A table containing a 'parent' column indicated a hierarchy (familial, ideological, geographical, or otherwise) that could be viewed by joining the table to itself. Usually, the parental relationship could be expressed simply: in the case of actual parenthood, CRM properties existed to describe it, and in the case of samples, influences or places the appropriate property was often the same as that which had pointed to the child entity in the first place.

Being much larger than the RRR dataset, the Grounds database needed to be built remotely. To make this easier, its code was split such that each broad section of the mapping became its own Graph, with these Graphs combined into one after all were built. This structure proved useful and was reverse engineered into the Raphael code later. Each mapping took several hours to run in total on a small remote server, but once an initial set of outputs had been produced there was little need to run the code again save for a major update.

2.4.4 Making datasets FAIR

In order to develop presentations for linked data on the web, it must first be made available within an accessible "triple store". These are repositories that hold named graphs, which operate similarly to the Graph concept already encountered in RDFlib. They act as databases for linked data and can be queried using the proprietary query language SPARQL, as well as through structured web APIs. Choosing a triple store for this data was trivial: the NG had already been using Blazegraph⁸² to store the old RRR data, and Blazegraph was also the triple store of choice for the British Museum's application ResearchSpace⁸³, where it is hoped the new mapped data could be displayed in the future. A Blazegraph instance was set up on an NG public server, and the created datasets were ingested into two named graphs that could be queried individually or as a combined source.

⁸² For more information and to download the actual software see <https://blazegraph.com/>. [21/04/22]

⁸³ ResearchSpace has been specifically designed to work with semantically modelled Heritage data, similar to the datasets produced here, for more information see <https://researchspace.org/>. [21/04/22]

2.4.4.1 DIRECT LIVE ACCESS TO THE SSHOC DATASETS

Direct access to the data can be achieved using a correctly formatted URL following the pattern:

https://END-POINT-URL?format=json&query=FULLESCAPEDSPARQLQUERY

For the RRR data:

- Base URL: <https://rdf.ng-london.org.uk/bg/>
- RRR suffix: namespace/sshoc-raphael/sparql
- Example Query: **SELECT * WHERE { ?s ?p ?o . } LIMIT 10** (which simply select the first 10 triples)
- Escaped Query: `SELECT+%2A+WHERE+%7B%0A++%3Fs+%3Fp+%3Fo+.%0A%7D+LIMIT+10`
- <https://rdf.ng-london.org.uk/bg/namespace/sshoc-raphael/sparql?format=json&query=SELECT+%2A+WHERE+%7B%0A++%3Fs+%3Fp+%3Fo+.%0A%7D+LIMIT+10>

For the Grounds data:

- Base URL: <https://rdf.ng-london.org.uk/bg/>
- RRR suffix: namespace/sshoc-grounds/sparql
- Example Query: **SELECT * WHERE { ?s ?p ?o . } LIMIT 10** (which simply select the first 10 triples)
- Escaped Query: `SELECT+%2A+WHERE+%7B%0A++%3Fs+%3Fp+%3Fo+.%0A%7D+LIMIT+10`
- <https://rdf.ng-london.org.uk/bg/namespace/sshoc-grounds/sparql?format=json&query=SELECT+%2A+WHERE+%7B%0A++%3Fs+%3Fp+%3Fo+.%0A%7D+LIMIT+10>

For the Combined dataset:

- Base URL: <https://rdf.ng-london.org.uk/bg/>
- RRR suffix: namespace/sshoc-combined/sparql
- Example Query: **SELECT * WHERE { ?s ?p ?o . } LIMIT 10** (which simply select the first 10 triples)
- Escaped Query: `SELECT+%2A+WHERE+%7B%0A++%3Fs+%3Fp+%3Fo+.%0A%7D+LIMIT+10`
- <https://rdf.ng-london.org.uk/bg/namespace/sshoc-combined/sparql?format=json&query=SELECT+%2A+WHERE+%7B%0A++%3Fs+%3Fp+%3Fo+.%0A%7D+LIMIT+10>

2.4.4.2 FORMATTED LIVE ACCESS TO THE SSHOC DATASETS

Although direct access to semantic data via a simple SPARQL end-point can be useful for developers it can be difficult for less technical users to engage with and also it is difficult for all levels of users to exploit data via this type of interface without further documentation and ideally a series of good worked examples.

Initial tests were made to set up an instance of ResearchSpace that would store and display this project's output data. ResearchSpace is a powerful tool for the development of narratives around linked data in the cultural heritage sector: it enables institutions to turn semantic connections into stories told by the paths between nodes. Users can create their own narratives and explore the collection thoroughly in a

self-directed way or turn to curated narratives provided by domain experts. Unfortunately, this work was constrained by the length of the dedicated research fellowship position and general project time limits and did not progress beyond the initial testing phase. It is hoped that it will be possible to continue the approach of using ResearchSpace or other compatible software systems beyond the scope of the SSHOC project.

A dedicated website has been setup to provide guided access to the new SSHOC datasets⁸⁴, see Figure 19.

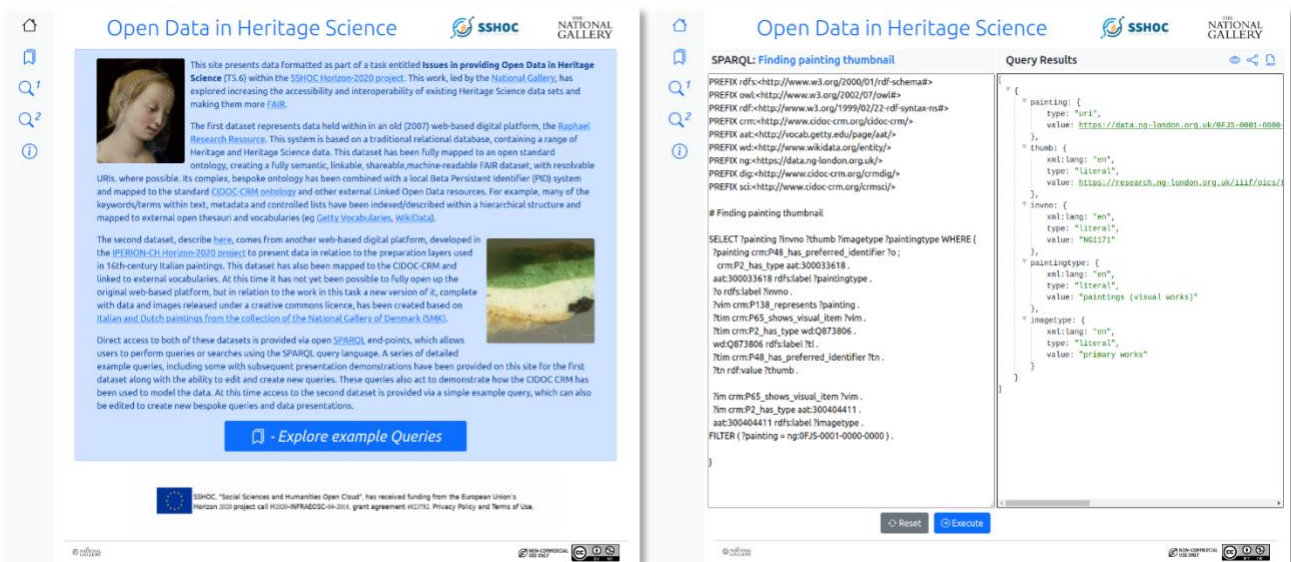


FIGURE 19: SCREENSHOTS OF THE WEBSITE⁸⁴ DEDICATED TO THE PRESENTATION OF THE NEW SSHOC DATASETS – WITH THE LANDING PAGE ON THE LEFT AND AN EXAMPLE QUERY ON THE RIGHT.

The website, (<https://rdf.ng-london.org.uk/sshoc/>), introduces the details of the two main datasets providing both context and explanation for the work. Access to both SPARQL end-points has been provided with a simple user interface, shown in Figure 19, with the current SPARQL query shown on the left and the results of the query on the right. The current query can also be edited as needed, to create custom queries, that can be executed within the same user-interface.

A series of detailed example queries are provided for the RRR datasets, allowing the user to see how the data can be explored and to document how the CIDOC-CRM has been used to model the data. The examples range from requesting simple lists of entities to requesting the more complex sets of data required to create new interactive data presentations, particularly demonstrating how the images within

⁸⁴ The website is live and can be accessed at: <https://rdf.ng-london.org.uk/sshoc/> - it is anticipated that the exploitation of these data sets will continue beyond the end of SSHOC so further development of the site, and its content is planned. [21/04/22]

the system can be surfaced and re-used via the IIIF standard, see Figure 3 and Figure 19. Further development, beyond the end of the SSHOC project is planned to fully validate and provide similar worked examples for the Grounds dataset.

In addition to the initial user-interface users are also provided with sharable links, see Figure 21, either to the current page or to the raw data returned by the current query. This makes it possible for users to bookmark what they are working on, cite queries or reuse the resultant data in another system, increasing the FAIR nature of the resource. A third tool option provides links back to the work previous described before, see section 2.3.4, and automatically reformats the results of the current query so that it can be visualised in the Dynamic Modeller system.

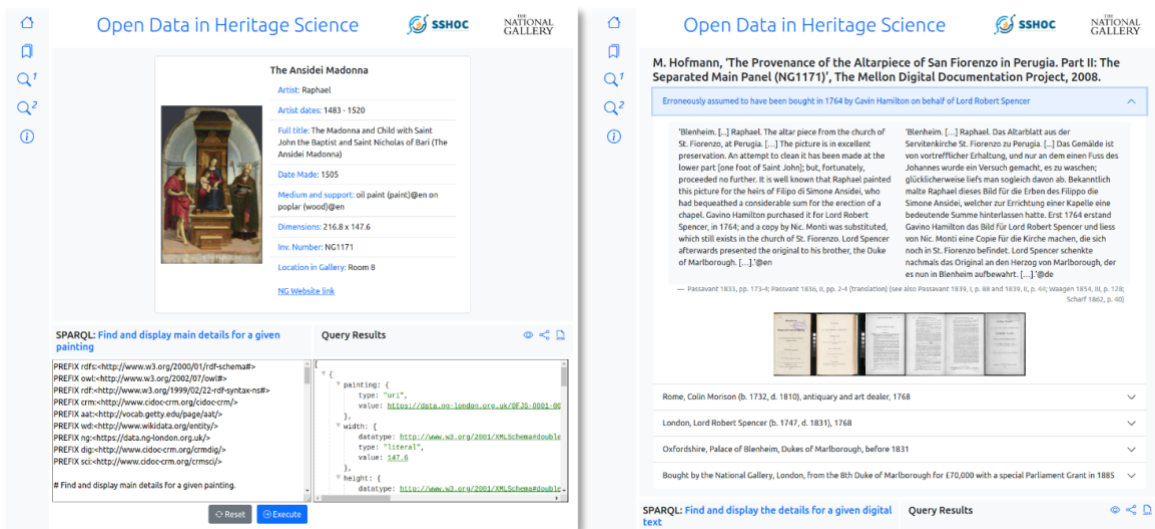


FIGURE 20: SCREENSHOTS OF THE DEDICATED WEBSITE⁸⁴ – THE EXAMPLE ON THE LEFT DEMONSTRATES HOW A QUERY CAN BE USED TO GATHER ALL OF THE KEY DATA FOR A GIVEN PAINTING INCLUDING A IIIF BASED THUMBNAIL IMAGE AND THE EXAMPLE ON THE RIGHT TAKEN FROM THE EXAMPLE QUERY THAT SHOWS HOW ALL OF THE DETAILS FOR A COMPLEX, NESTED SET OF PROVENANCE DIGITAL TEXTS CAN BE REQUESTED AND THEN PRESENTED INLINE WITH THE RELEVANT IIIF THUMBNAILS.

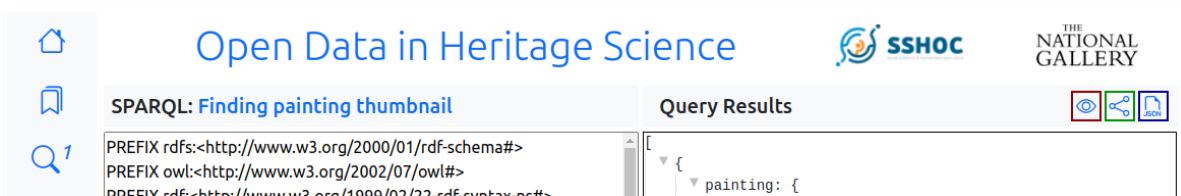


FIGURE 21: DETAIL OF WEBSITE SCREENSHOT HIGHLIGHTING THE ADDITIONAL TOOLS PROVIDED TO INTERACT WITH QUERY RESULTS: THE EYE ICON (HIGHLIGHTED IN RED) ALLOWS THE USER TO VISUALISE THE RESULTS OF THE QUERY AS A FLOWDIAGRAM IN THE DYNAMIC MODELLER SYSTEM. THE SHARE ICON (GREEN) PROVIDES A SHARABLE LINK TO THE SITE INCLUDING THE CURRENT QUERY AND THE JSON ICON (BLUE) PROVIDES AN ALTERNATIVE LINK TO JUST THE RAW RESULTS OF THE CURRENT QUERY FORMATTED AS A JSON DOCUMENT.

The reformatting process uses the relationships indicated in the query and the returned data to create a full set of triples, modelling all the relevant relationships, these triples are then passed to the Dynamic modeller as tab separated text allowing the system to automatically create a semantic flow or

relationship diagram. This functionality was added to improve the clarity of the relationships modelled in the data by providing this visual representation. However, during the later stages of the modelling process this functionality also proved very helpful in testing that all the expected relationships has been correctly modelled and that all the returned data did all connect up as intended.

Once the system was full functional and the modelling had been checked with the full set of example queries an additional set of “Mapping Model” examples were added to the system. This set of final example queries were more complex than the original examples as they were designed to represent the full set of relationships modelled in the data for all the major entities in the RRR; objects, artists, images (in two parts), samples, institutions, and documents. To improve the speed of the system, the actual displayed queries are set to only return a few variables of data⁸⁵ and the full model of relationships can then be visualised by pushing the formatted results to the Dynamic modeller. This process allowed the relationships in the live data to be directly compared with the original theoretical SSHOC models, such as those shown in Figure 12, Figure 15 and Figure 18, with live data models, such as the one shown in Figure 22.

Queries to generate the following full Mapping Models have been included in the system:

- Find Full Object Model: <https://rdf.ng-london.org.uk/sshoc/?example=object-full-model>
- Find Full Artist Event Model: <https://rdf.ng-london.org.uk/sshoc/?example=artist-full-model>
- Find Full Image Model: <https://rdf.ng-london.org.uk/sshoc/?example=image-full-model>
- Find Pyramid Image Model: <https://rdf.ng-london.org.uk/sshoc/?example=image-pyramid-model>
- Find Full Sample Model: <https://rdf.ng-london.org.uk/sshoc/?example=sample-full-model>
- Find Full Institution Model: <https://rdf.ng-london.org.uk/sshoc/?example=institution-full-model>
- Find Full Document Model: <https://rdf.ng-london.org.uk/sshoc/?example=document-full-model>

These full mapping models provide a clear and accessible documentation of how the relationships in the data set have been modelled. A further static presentation has also been added to the presentation website⁸⁶ to show how all of these “Model” diagrams connect, see Figure 23.

2.4.4.3 SUSTAINED ACCESS TO PUBLISHED VERSIONS OF THE DATASETS

In addition to the live access to the datasets created in this work, described in sections 2.4.4.1 and 2.4.4.2,

⁸⁵ As these queries have not been fully optimised it was found that several of took too long to process in the provided GUI, this was generally due to the high level of results required to capture all the possible variations required. The “image” model was particularly complicated, so it was separated into two examples. This issued will be resolved as the system is optimised during future development.

⁸⁶ The full connected example model relating to the example queries, for the RRR dataset, can be explore at: <https://rdf.ng-london.org.uk/sshoc/?diagram=1> [25/04/22]

to ensure the sustained accessibility of the datasets, mapping code and presentations code, described in this report have also be published within open public repositories, see Table 1.

The generic free, open data repository Zenodo⁷ was used as the default repository for all the prepared datasets and they have been directly linked to SSHOC Zenodo community⁸⁷.

To explore alternative data repository opportunities, in collaboration with the Dutch national centre of expertise and repository for research data (DANS)⁸⁸ within SSHOC, it was also possible to published both of the final datasets directly within a prepared SSHOC WP5⁸⁹ instance of DataVerse⁹⁰, hosted by DANs, with direct links included in Table 1.

⁸⁷ The SSHOC community on Zenodo can be found at: <https://zenodo.org/communities/sshoc/>.

⁸⁸ For more details see the institutional website at: <https://dans.knaw.nl/>.

⁸⁹ The SSHOC WP5 instance of Dataverse can be accessed at: https://dataverse.nl/dataverse/SSHOC_WP5.

⁹⁰ Dataverse is an example of “Open source research data repository software”, detailed documentation and the software can be accessed via the project website: <https://dataverse.org/>.

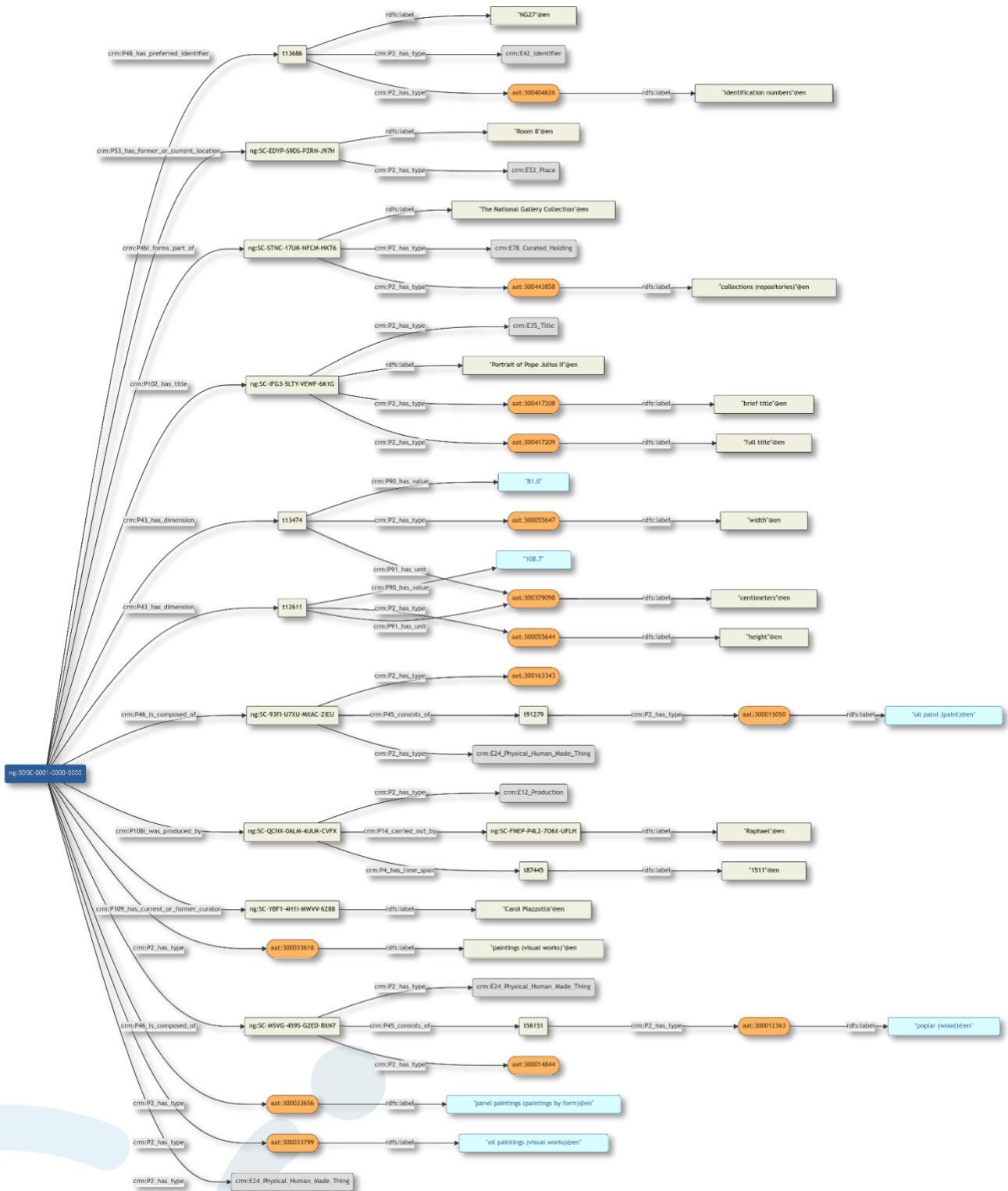


FIGURE 22: AN IMAGE OF THE SEMANTIC MODEL FOR AN OBJECT - AUTOMATICALLY GENERATED FROM THE LIVE DATA, WHICH CAN BE COMPARED WITH THE ORIGINAL THEORETICAL MODEL IN FIGURE 12.

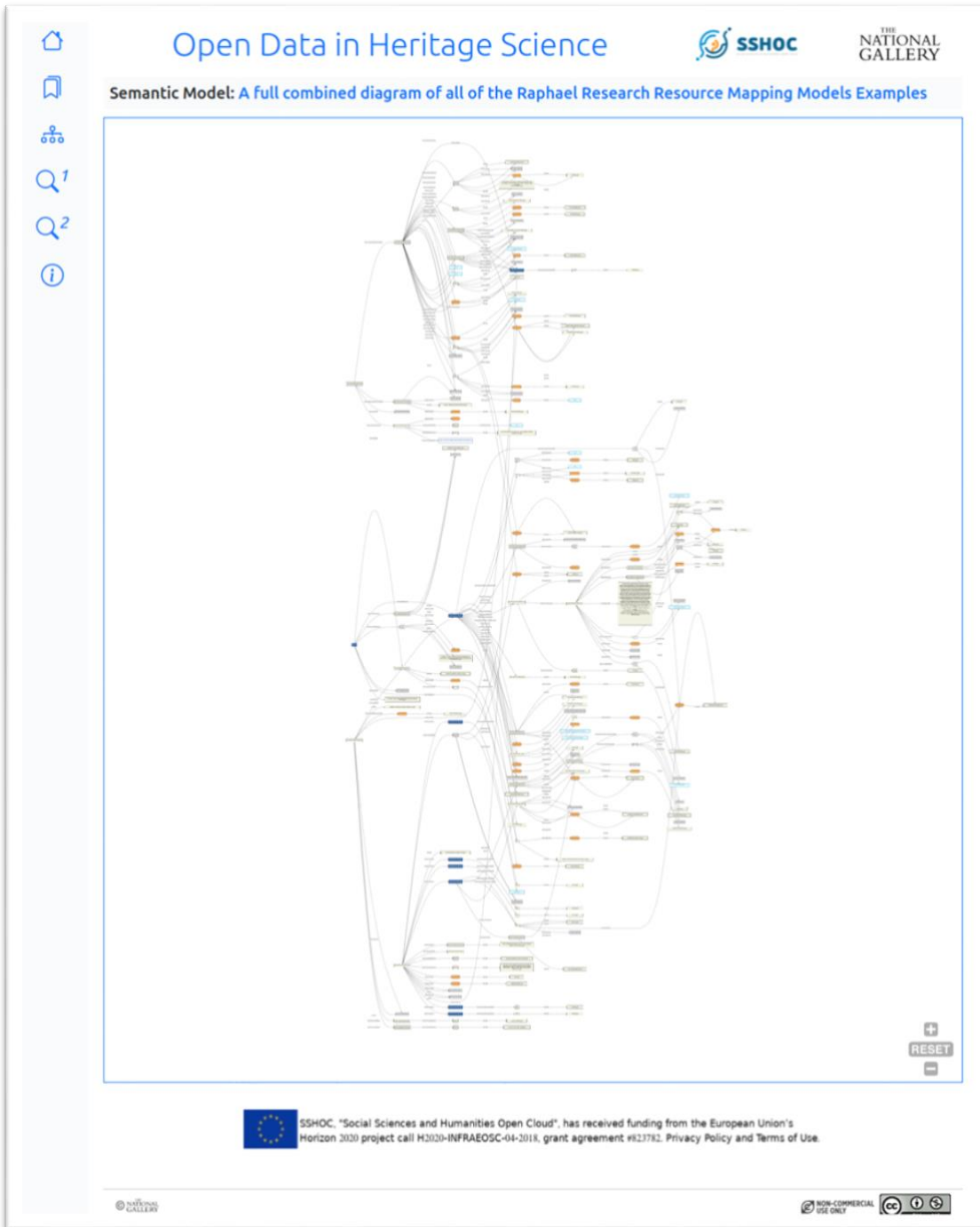


FIGURE 23: A SCREENSHOT OF THE SEMANTIC CONNECTIONS GENERATED BY COMBINING THE OUTPUT OF ALL OF THE FULL MAPPING MODELS QUERIES FOR THE RRR⁸⁶.

2.5 Summary of accessibility of resources

TABLE 1: MAPPING EXISTING DATA - DIRECT ACCESS DETAILS FOR THE RESULTS

Name	Format	Licence	Link
Simple Site	Doc/Presentation	CC BY-NC 4.0 ⁹¹	https://jpadfield.github.io/simple-site/
	Code	GPL-3.0	https://github.com/jpadfield/simple-site
	Dataset	GPL-3.0	https://doi.org/10.5281/zenodo.4504844
Simple Discovery	Doc/Presentation	CC BY-NC 4.0 ⁹¹	https://research.ng-london.org.uk/ss-iiif/
	Code	GPL-3.0	https://github.com/jpadfield/iiif-discovery
	Dataset	GPL-3.0	https://doi.org/10.5281/zenodo.5512980
Dynamic Modeller	Doc/Presentation	CC BY-NC 4.0 ⁹¹	https://research.ng-london.org.uk/modelling/
	Code	GPL-3.0	https://github.com/jpadfield/dynamic-modelling
	Dataset	GPL-3.0	https://doi.org/10.5281/zenodo.4724103
SSHOC Data Presentation	Doc/Presentation	CC BY-NC 4.0	https://rdf.ng-london.org.uk/sshoc/
	Code	GPL-3.0	https://github.com/jpadfield/sshoc-data-presentation
	Dataset	CC BY-NC 4.0	https://doi.org/10.5281/zenodo.6462987
SSHOC RRR Dataset	Source	Custom	https://cima.ng-london.org.uk/documentation/
	Doc/Presentation	CC BY-NC 4.0	https://rdf.ng-london.org.uk/sshoc/
	Dataset (Zenodo)	CC BY-NC 4.0	https://doi.org/10.5281/zenodo.6476541
	Dataset (DataVerse)	CC BY-NC 4.0	https://doi.org/10.34894/FO2VHB
SSHOC Grounds Dataset	Source	Custom	https://research.ng-london.org.uk/iperion/
	Doc/Presentation	CC BY-NC 4.0	https://rdf.ng-london.org.uk/sshoc/
	Dataset (Zenodo)	CC BY-NC 4.0	https://doi.org/10.5281/zenodo.6478779
	Dataset (DataVerse)	CC BY-NC 4.0	https://doi.org/10.34894/GZTK50
SSHOC RRR Modelling Code	Code	CC BY-NC 4.0	https://github.com/jpadfield/sshoc_raphael_modelling
	Dataset	CC BY-NC 4.0	https://doi.org/10.5281/zenodo.6461653
SSHOC Grounds	Code	CC BY-NC 4.0	https://github.com/jpadfield/sshoc_grounds_modelling
	Dataset	CC BY-NC 4.0	https://doi.org/10.5281/zenodo.6461469

⁹¹ All images and metadata sourced from external APIs/Sources are subject to the IPR set by the content owners.

Modelling Code			
SMK Grounds	Doc/Presentation	CC BY-NC 4.0 ⁹²	https://research.ng-london.org.uk/iperion-smk
Data	Doc/Presentation	CC BY-NC 4.0	https://research.ng-london.org.uk/ss-smk/

2.6 Increasing the accessibility of Grounds Data

The work carried out within T5.6 of SSHOC also provided technical support for the production of a new open version of the Grounds Database, with all of its content made re-usable under a defined creative-commons licences - <https://research.ng-london.org.uk/iperion-smk>, see Figure 24. A generous grant from The Samuel H. Kress Foundation⁹³ for the project "The digitization of cross-sections from Italian and Dutch paintings" at The National Gallery of Denmark (SMK)⁹⁴ enabled the digitization and analyses of cross-sections from a total of 158 Italian 14th to 17th C. and 17th C. Dutch paintings from the SMK collection to be made available in an open access art and technology research database on ground layers. With samples from the collections of Nationalmuseum Stockholm⁹⁵ and Museum of National History, Frederiksborg Castle in Hillerød⁹⁶ the database includes an additional 11 paintings. Effort within the SSHOC project re-formatted the provided data and enabled it to be presented within the original IPERION-CH Grounds Database GUI. Further work was also carried out to open this data to non-specialists, allowing access to the same images via a simple keyword search option, though a dedicated Simple IIIF Discovery site⁹⁷, see Figure 25. the development of which was supported by the AHRC funded Practical IIIF project⁴³, the SSHOC project and the H2020 IPERION HS project³².

⁹² Some of the content presented has been released under more open CC licences, further details are provided within the source IIIF manifests presented by the system.

⁹³ Institutional website: <https://www.kressfoundation.org/>. [21/04/22]

⁹⁴ Institutional website: <https://www.smk.dk/>. [21/04/22]

⁹⁵ Institutional website: <https://www.nationalmuseum.se/en/om-nationalmuseum>. [21/04/22]

⁹⁶ Institutional website: <https://dnm.dk/en/>. [21/04/22]

⁹⁷ The live Grounds Data Simple IIIF discovery site can be found at: <https://research.ng-london.org.uk/ss-smk/>. [21/04/22]

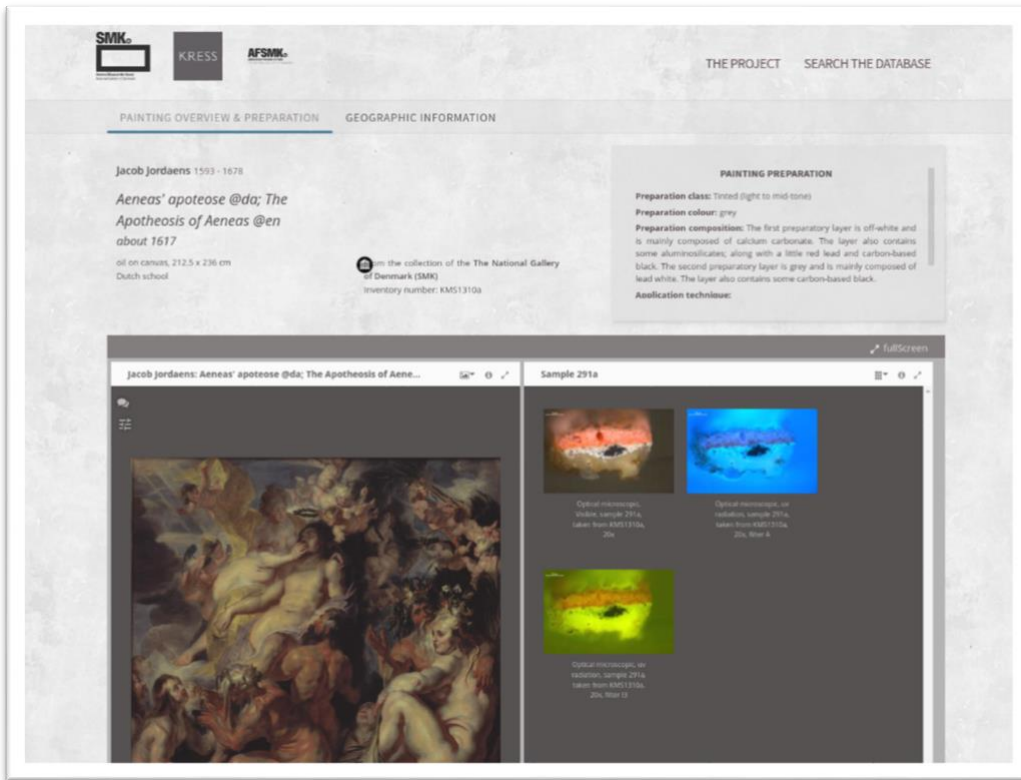


FIGURE 24: A SCREENSHOT OF THE NEW OPEN VERSION OF THE GROUNDS DATABASE BASED ON DATA FROM THE SMK.

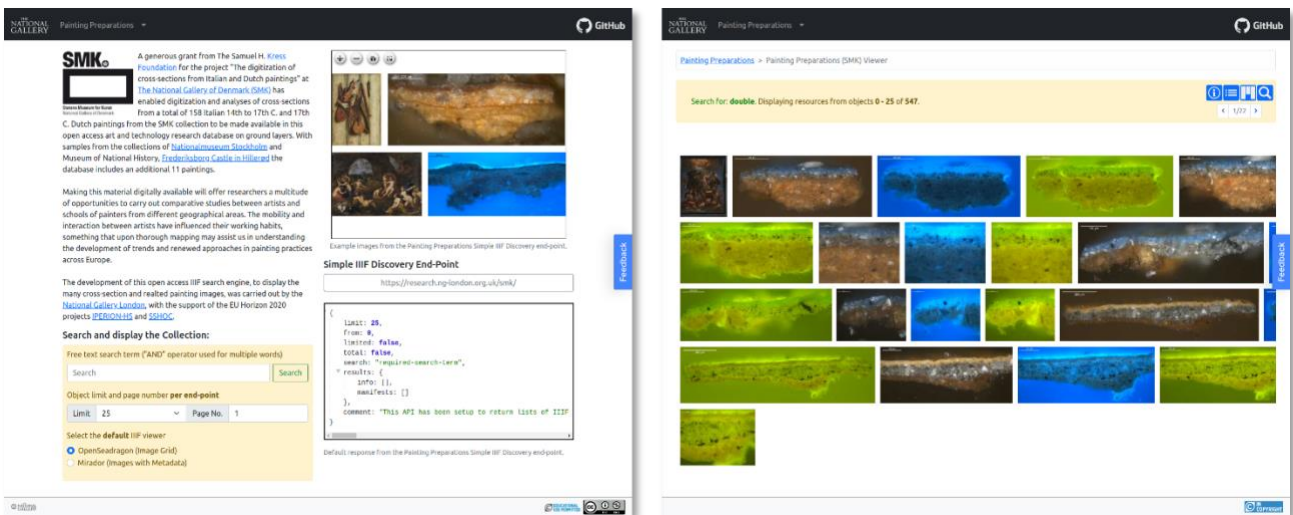


FIGURE 25: SCREENSHOTS FROM THE DEDICATED GROUNDS DATA SIMPLE IIIF DISCOVERY WEBSITE.

2.7 Next steps for the RRR and Grounds Datasets

Taken together, this project's output datasets make up a unique example of how heritage science data might be made FAIR using the CIDOC-CRM. Through their availability as datasets and metadata in Zenodo they partially fulfil the Findable criterion; further augmented by the registration and full documentation of the NG PIDs used within the datasets. The Blazegraph SPARQL endpoint, IIF integration, and painstaking provenance documentation within the dataset make it Accessible, as does, in part, the Zenodo availability. Interoperability is provided through the use of the CIDOC-CRM, RDF, and OWL ontologies, as well as the IIF image standard and the Wikidata and AAT vocabularies. It is worth pointing out that the two datasets are also internally interoperable: through their mutual PID database they have been made to point to a single identifier where the same entity is included in both. This means that for certain Raphael paintings it is possible to view specialist provenance data from the RRR directly alongside preparation layer data from the Grounds database: an authentically new perspective on Raphael that heretofore had not been possible for external researchers to achieve before. Finally, the project's data is Reusable owing to its documentation, part of which is contained in this text and much of which is included within the dedicated Github repositories. Github also houses the project's mapping code, which any user can fork and further develop. Finally, as part of their presentation in Zenodo, a license agreement has been attached to both datasets, and it has been made clear from within the data where there are absences due to copyright issues. Better display facilities for the data would surely improve each of the FAIR categories, but the current state of the dataset is sufficient to meet all the criteria.

This project's work not only connects two important Heritage datasets together and provides a model for the further interlinking of future datasets, but also opens NG data to collaboration with other cultural heritage institutions that hold CIDOC-CRM assets, this is particularly true with the Grounds database, which is already being extended to include new data as shown. As well as this, modelling work provides an example of how scientific analytical work can become institutional knowledge: as assets are developed atop a CRM-based dataset, its content can become integral to future presentation of the paintings within, and subsets of the results of scientific research can be made available in simplified formats for the public. This also suggests the possibility of custom interfaces for researchers: by encoding detailed research data, this project creates numerous levels for users to engage with Heritage content, which ought to be especially useful in a research context. These outcomes can have real impact on engagement and on the NG's ongoing digital development and demonstrate that making data FAIR has potentially positive effects on wider research communities well beyond the initial narrower target user base of the original datasets.

3 Heritage Science - The MOLAB experience (CNR)

Heritage Science is an interdisciplinary scientific field for the study of Cultural Heritage. In the recent decades this field has gained great attention. Several EU financed projects (CHARISMA-FP7⁹⁸, Eu-ARTECH-FP⁹⁹ and IPERION-CH²⁵) have led to the creation of a stable research infrastructure in Heritage Science (E-RIHS)¹⁰⁰. This infrastructure offers access to instruments, methodologies, expertise, and data for advancing knowledge and innovation in the conservation and restoration of cultural heritage by integrating national world-class facilities at research centres, universities, and museums.

Within this framework a trans-national access is offered through the complementary platforms ARCHLAB, FIXLAB, MOLAB and DIGILAB¹⁰¹.

ARCHLAB offers access to specialised knowledge and organized scientific information. This includes different types of data coming mainly from published and unpublished data sets and archives from European Museums, galleries, and research institutions. Technical images, analytical data and conservation documentation can also be consulted through the ARCHLAB access.

FIXLAB offers access to large-scale and medium-scale facilities. Scientists can examine samples or artworks through particle accelerators and synchrotrons, neutron sources; non-transportable analytical instruments to mention some. These sophisticated scientific investigations help to reveal their microstructure and chemical composition and give insights into materials, alteration and degradation phenomena as well as revealing information on historical technologies or artwork authenticity.

MOLAB¹⁰² has the goal of bringing science to artwork without the need to move the object or take samples. Is a unique set of non-invasive portable instrumentation for the study of Cultural Heritage. MOLAB access users can thus perform complex in-situ multi-disciplinary investigations of unmovable

⁹⁸ EU funded: Cultural heritage Advanced Research Infrastructures: Synergy for a Multidisciplinary Approach to Conservation/Restoration (CHARISMA-FP7) – for more information see: <https://cordis.europa.eu/project/id/228330>. [28/04/22]

⁹⁹ EU Funded: Access, research and technology for the conservation of the European cultural heritage (EU-ARTECH) – for more information see: <https://cordis.europa.eu/project/id/506171>. [28/04/22]

¹⁰⁰ EU Funded: European Research Infrastructure for Heritage Science (E-RIHS) - for more information see: <http://www.e-rihs.eu/>. [28/04/22]

¹⁰¹ For summary description of each access see: <http://www.e-rihs.eu/access/> [28/04/22]

¹⁰² MOLAB is a distributed infrastructure providing coherent access to a set of mobile analytical and examination equipment and related competencies, (<https://www.iperionhs.eu/molab>) [28/04/22]

objects such as monuments or archaeological sites or smaller artworks as paintings whose fragility and/or extraordinary value make them hardly movable.



FIGURE 26: MOLAB ACCESS AT THE ESTORICK COLLECTION OF MODERN ITALIAN ART, LONDON, UK.

DIGILAB is the new proposed digital platform for E-RIHS and will facilitate access to digital services and scientific data concerning tangible heritage. A key objective is to provide guidelines to harmonize the myriad of data and formats available in an interdisciplinary approach, making them FAIR. It will include searchable registries of multidimensional images, analytical data, and documentation from large academic as well as research and heritage institutions.

In the framework of DIGILAB and MOLAB the CNR and Università degli Studi di Perugia has been using and developing, for almost 10 years now, MOVIDA, a standalone java-based system that gathers all the generated data from non-invasive analytical spectroscopic investigation. The data is stored in a single XML file and the software can be used to analyse the store data on-the-fly, analyse the multi-technique information recorded and share it with the MOLAB access users.

This work explored **the interoperability between MOVIDA and a FAIR data set**. Creating a pilot database using, when possible, existing schemas and ontologies and suggesting the necessary implementations on the chosen schema when the data from MOVIDA cannot be included.

This section of the report describes MOVIDA, and the data held by the software, including a short introduction to the reference schema that has been chosen (CIDOC-CRM). The interoperability /compatibility between MOVIDA and CIDOC-CRM, in particular the data that needs to be shared is defined, and two different schemas are proposed. The first schema regards the general data of the artwork (Author, period, conservation state...). For this schema, CIDOC-CRM has all the necessary entities and relationships and therefore a plugin for MOVIDA has been developed. The plugin and the pilot data generated are presented. The second schema regards the data of the analytical campaign (Actors, dates, techniques and instruments, types of data...). For this data set, it was not possible to connect all the information to the proper entities and relationships of CIDOC-CRM, within the work in this task. Therefore, existing entities were used when possible and new classes and properties have been suggested when the CIDOC-CRM did not seem to fulfil the required needs.

3.1 MOVIDA (MOlab Visualisation DAta)

The main advantage of a multi-technique approach to complex cultural heritage objects is the amount of interconnected information that can be obtained, providing a greater insight into the composition and alteration phases of the artwork. The amount of data obtained can be huge and heterogeneous and hence the need to have a suitable tool to store data on the fly and to share and analyse it in a second step.

With this aim MOVIDA was developed 10 years ago. MOVIDA is a standalone piece of, java based, software that deals with data and information of a spectroscopic non-invasive investigation. The software is mainly used for punctual spectroscopic techniques and the data is stored in an xml file. The software links raw data to the specific area or point of the artwork where it has been taken. Different comments on the measurement and the results inferred from data, as well as the experimental conditions, can also be stored and linked to the investigated spot and to the raw files that can be plotted with the software.

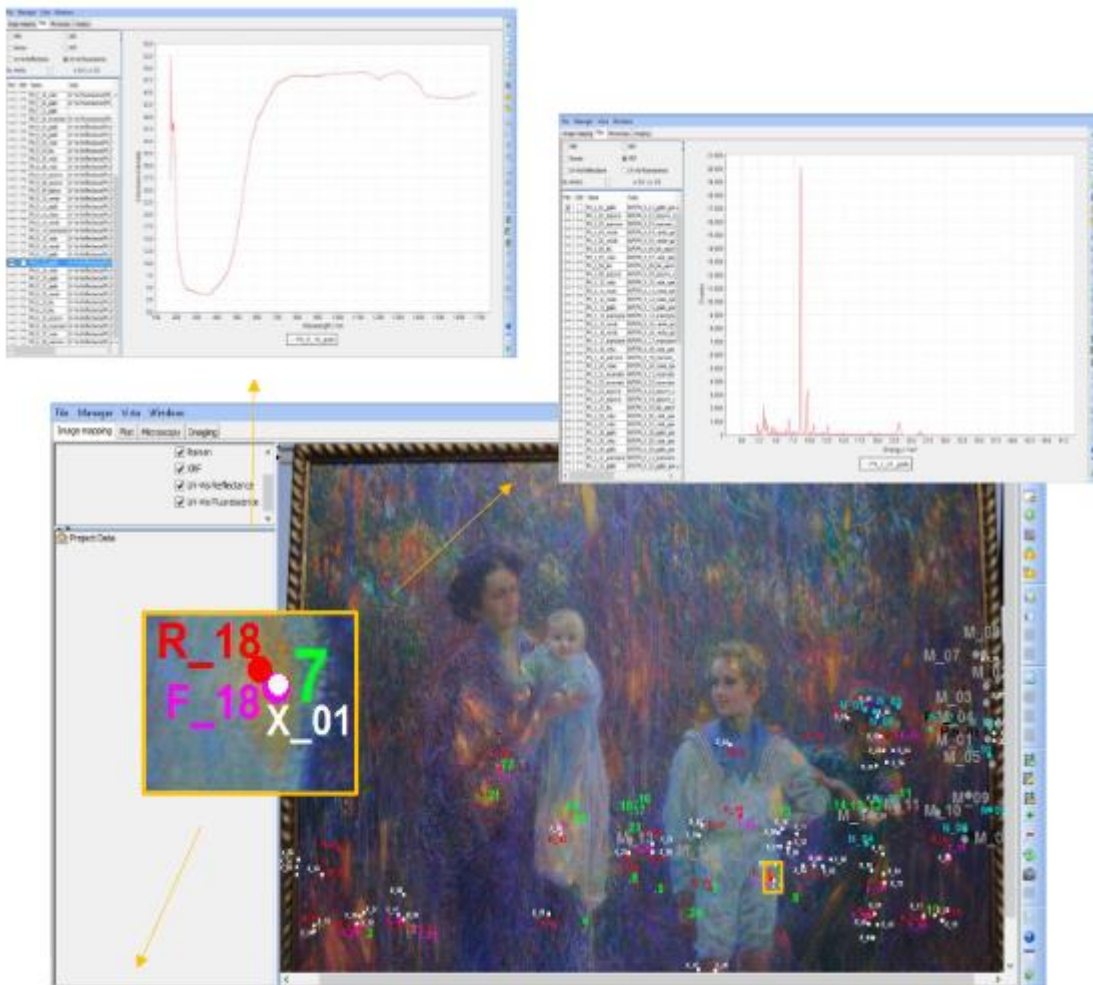


FIGURE 27: SCREENSHOTS OF MOVIDA FROM THE MOLAB ACCESS AT THE GALLERIA NAZIONALE DI ARTE MODERNA, ROMA, IT.

As stated, all the data is stored in a single xml file, for an example see Figure 28.

```
<?xml version="1.0" encoding="UTF8"?>
<data>
  <tecnica name="MIR" added="YES">
    <Punt namePagina="V1" nameTecnica="M" nameNumero="01"
      nameDesc="paper">
      <ResPunt text="Cellulose"></ResPunt>
      <Imatge name="V.2014.56.1Louis_XIV" indImg="0"
nameCoord="0.024461839348077774;0.22936449944972992"
labCoord="0.05859375;0.21379980444908142"></Imatge>
      <DataPlot namePlot="" pathPlot="MIR/V1_M_01_paper.dpt">
        <expCond name="Spec. res." value=""></expCond>
        <expCond name="Scans" value=""></expCond>
        <expCond name="Check" value=""></expCond>
        <expCond name="Backg" value=""></expCond>
      </DataPlot>
    </Punt>
  </tecnica>
  <tecnica name="NIR" added="NO">
    <expCondTec name="Spec. res." value=""></expCondTec>
    <expCondTec name="Scans" value=""></expCondTec>
    <expCondTec name="Check" value=""></expCondTec>
    <expCondTec name="Background" value=""></expCondTec>
  </tecnica>
  <tecnica name="XRF" added="YES">
    <expCondTec name="Voltage" value="40 kV"></expCondTec>
    <expCondTec name="Current" value="100 uA"></expCondTec>
    <expCondTec name="Integr. Time" value="40 s"></expCondTec>
    <expCondTec name="A (Fitting)" value="-0.30717"></expCondTec>
    <expCondTec name="B (Fitting)" value="0.02256"></expCondTec>
    <expCondTec name="C (Fitting)" value="0"></expCondTec>
    <expCondTec name="Fitting Needed" value="YES"></expCondTec>
    <Punt namePagina="V1" nameTecnica="X" nameNumero="01"
      nameDesc="blue">
      <DescPunt text="Possible other blue pigments (lapis) check by MIR and
        UV-Vis."></DescPunt>
      <DescPunt text="Ca, Fe, Pb, Zn, Cu from the support in all the spectra."></DescPunt>
      <ResPunt text="Ca, Pb, Fe, Cu, Zn, K, Si, Ti, (Mn, Sr, Al)"></ResPunt>
      <Imatge name="V.2014.56.1Louis_XIV" indImg="0"
nameCoord="0.5981387495994568;0.7713202834129333"
labCoord="0.6099796295166016;0.7560790181159973"></Imatge>
      <DataPlot namePlot=""
        pathPlot="XRF/V1_X_01_blue_spectrum.spt">
        <expCond name="Voltage" value="40 kV"></expCond>
        <expCond name="Current" value="100 uA"></expCond>
        <expCond name="Integr. Time" value="40 s"></expCond>
        <expCond name="A (Fitting)" value="-0.30717"></expCond>
        <expCond name="B (Fitting)" value="0.02256"></expCond>
        <expCond name="C (Fitting)" value="0"></expCond>
        <expCond name="Fitting Needed" value="YES"></expCond>
      </DataPlot>
    </Punt>
  <imatge name="V.2014.56.1Louis_XIV"
    path="Images/V.2014.56.1_Louis_XIV_W.Vaillant_1660_red.jpg">
    <imgTec name="MIR" color="255;51;51" size="0.8" sizeLabel="2.8"></imgTec>
    <imgTec name="NIR" color="0;0;0" size="1.5" sizeLabel="10.0"></imgTec>
    <imgTec name="XRF" color="0;0;0" size="0.5" sizeLabel="3.3"></imgTec>
  </imatge>
</data>
```

FIGURE 28: MINIMAL EXAMPLE OF THE MOVIDA XML FILE GENERATED

Data is organised in relation to images and techniques. For each technique, a series of points where measurements have been made are linked to the image of the artwork through spatial coordinates starting from the left top corner. For each point there is a series of experimental conditions, the path to the raw file and two text areas for comments and results. Each measurement point has a unique name that relates it to the image and the technique and sequentially orders all the measures coming from the same instrument. The first version of the software has been successfully used since 2010 by the CNR and Università degli Studi di Perugia in the MOLAB accesses performed through the CHARISMA-FP7 and IPERION-CH and is used nowadays within ERIHS-IT. Though it has been demonstrated to be a valuable tool, the constant and rapid evolution of the field and the need of extending the software horizons to the latest techniques such as hyperspectral imaging techniques, or OCT has brought the need of developing a new version of the software. The new version of the software is now in its final development stages, see Figure 29. Some of the differences with the previous versions are:

- The capability to graphically compare punctual and imaging and hyperspectral data.
- New types of data that can be uploaded (xml, multi and hyperspectral, video, 3d surface data...)
- The data can now be linked to areas of different shapes (not only point as in previous version) through a polygon or a freehand tool.

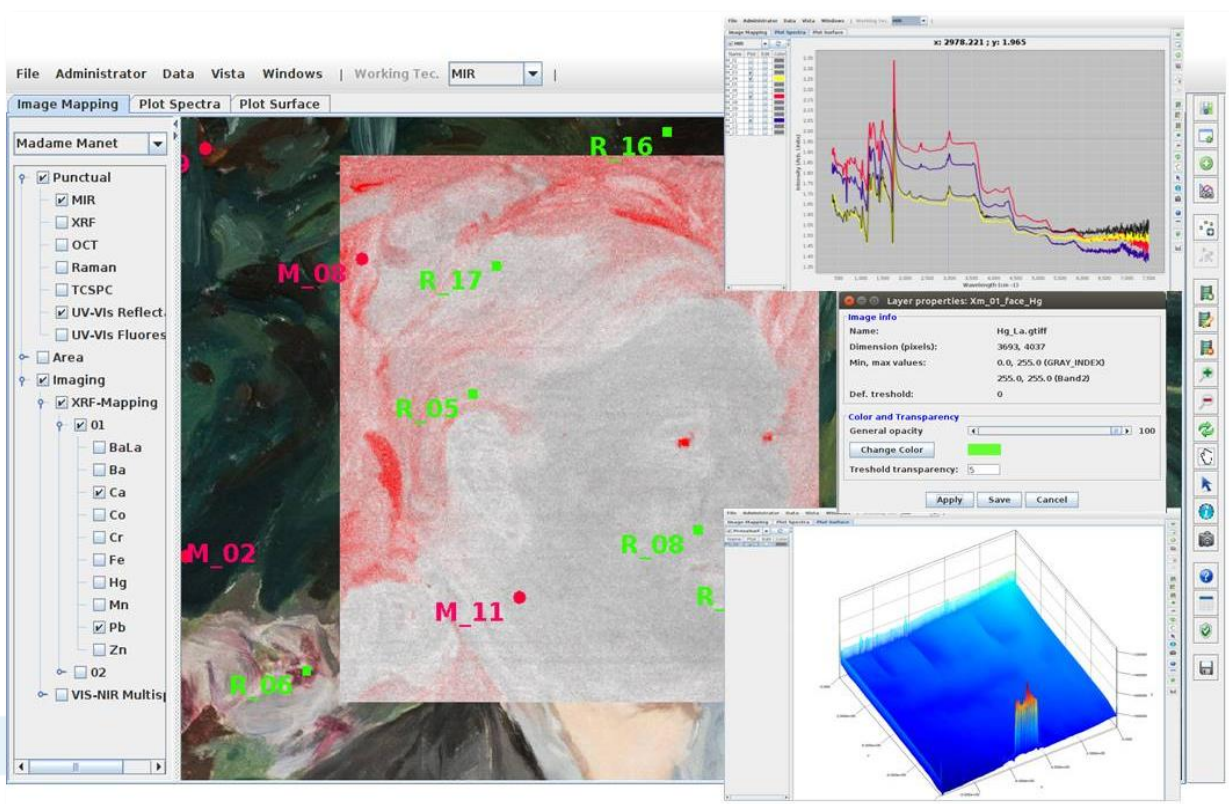


FIGURE 29: SCREENSHOT OF THE USE OF MOVIDA 2.0. VISUALISATION OF A DETAIL OF MADAME MANET IN THE CONSERVATORY WITH THE SPOTS WHERE DATA FROM MIR AND RAMAN HAS BEEN REGISTERED. PRESENCE OF CALCIUM AND LEAD FROM XRF HYPERSPECTRAL ANALYSIS ON THE FACE OF MADAME MANET OCT DATA

Given the increased complexity of the information handled, the data is stored in a sqlite DB¹⁰³ with the schema reported in Figure 30.

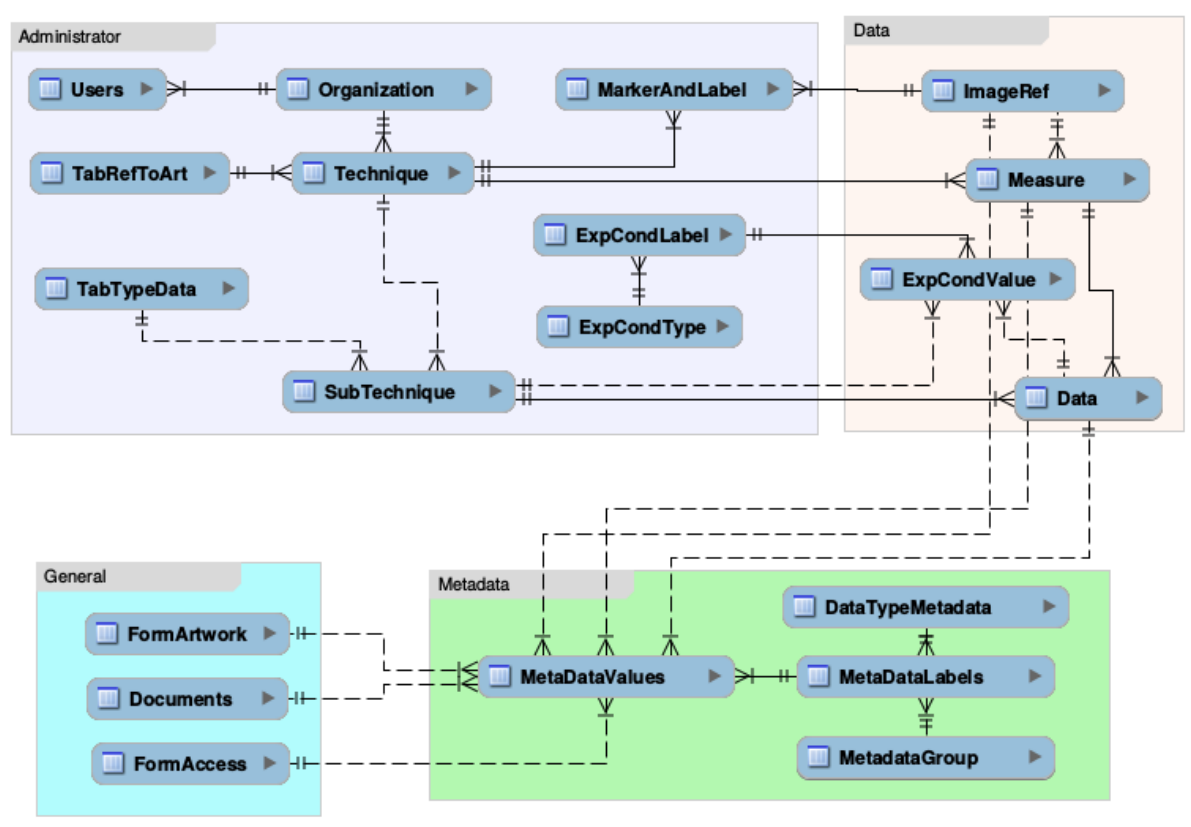


FIGURE 30: SQLITE SCHEMA FOR MOVIDA 2.0

Regarding the issues related to the SSHOC task 5.6, two general forms have been included with the intention of creating a FAIR set of data that can be easily exported to existing registries and/or databases.

- General information on the artwork:
 - Legal owner and location (contact points, ids...)
 - Type of artwork, author, and period
 - Size and conservation state.
- General information on the investigation:
 - Institutes involved (responsibilities, operators, and dates)
 - Instruments used

¹⁰³ For more details and access to the software see: <https://www.sqlite.org/index.html> [28/04/22]

- Type of data generated

The amount of information to be included to these two general areas has been limited. Considering the present stage of development of FAIR standards in Heritage Science at this stage it would not be useful to include specific information on each data recorded (spectra, images...). These assessments will be further developed in the next chapters.

3.2 Interoperability between CNR MOVIDA and CIDOC-CRM

The analysis developed here as part of the SSHOC task 5.6 had the objective to check the interoperability between MOVIDA and a FAIR data set. Data handled by MOVIDA is very complex and goes very deep: from very general information to the single measurement data (experimental condition, area of the measurement or position of the spectral peaks of a single measure, etc.).

Within this task, due to the allocated time, the scope of the work was limited to an initial examination of the upper or general level of semantic descriptions rather than exploring a full semantic description of every data point. The objective was to analyse the feasibility of generating a minimal FAIR data set from MOVIDA that could allow users to find datasets answering general questions such as, “Find me ...”:

- Analytical campaigns on XVIII century paintings using XRF
- Analytical campaigns on Picasso’s paintings
- Analytical campaigns on oil paintings using OCT, XRF and MIR
- Analytical campaigns on paintings of the XIX century having condition state: discoloration of yellow
- Analytical campaign performed by MOLAB-CNR using MIR
- Analytical campaign performed on Sculptures since 2019

After which a user could then choose to ask for the complete MOVIDA dataset to analyse for themselves and to access the “second level” or full set of data.

It was a choice to include data from the artwork and from the analytical campaign performed. Cultural Heritage semantics have a long history and are highly developed and, therefore, a creation of schema based on CIDOC-crm including all the data recorded regarding the general information of the artwork has been enabled. On the contrary, Heritage Science field is much newer, and not fully developed in CIDOC-CRM. Recently the extension (CRMsci) has been proposed but it's only in its first steps development and we have not found the proper structure to store all the data regarding an analytical investigation campaign. Therefore, it has been necessary to separate data into two main blocks. One concerning the general information on the artwork and the other concerning the proper heritage science data and this task has been organised with a two-fold purpose. The first area of work was to incorporate into **MOVIDA 1.0** a tailored module for the automatic gathering of contextual information of the artwork compliant with the CIDOC-CRM scheme. By employing common metadata standards to describe artwork details, the information will be easily gathered in a registry and could be also uploaded to an online

resource like the SSH Open Cloud in the future. The second area of work was to formulate a proposal to include some additional Entities and Properties and/or to change the domain or range of the existing ones to allow the data of the analytical campaign to be included in a FAIR dataset. These two areas of work can be summarised as follows:

General information on the artwork (GIA):

- Legal owner and location (contact points, ids...)
- Type of artwork, author, and period
- Size and conservation state.

General information on the analytical campaign (GAI):

- Institutes involved (responsible actors, operators, and dates)
- Instruments used
- Type of data generated

3.3 Schema for: General Information on the artwork (GIA)

CIDOC-CRM has demonstrated to be a good framework to store general data describing an artwork and the data has been implemented using the following schema, see Figure 31.

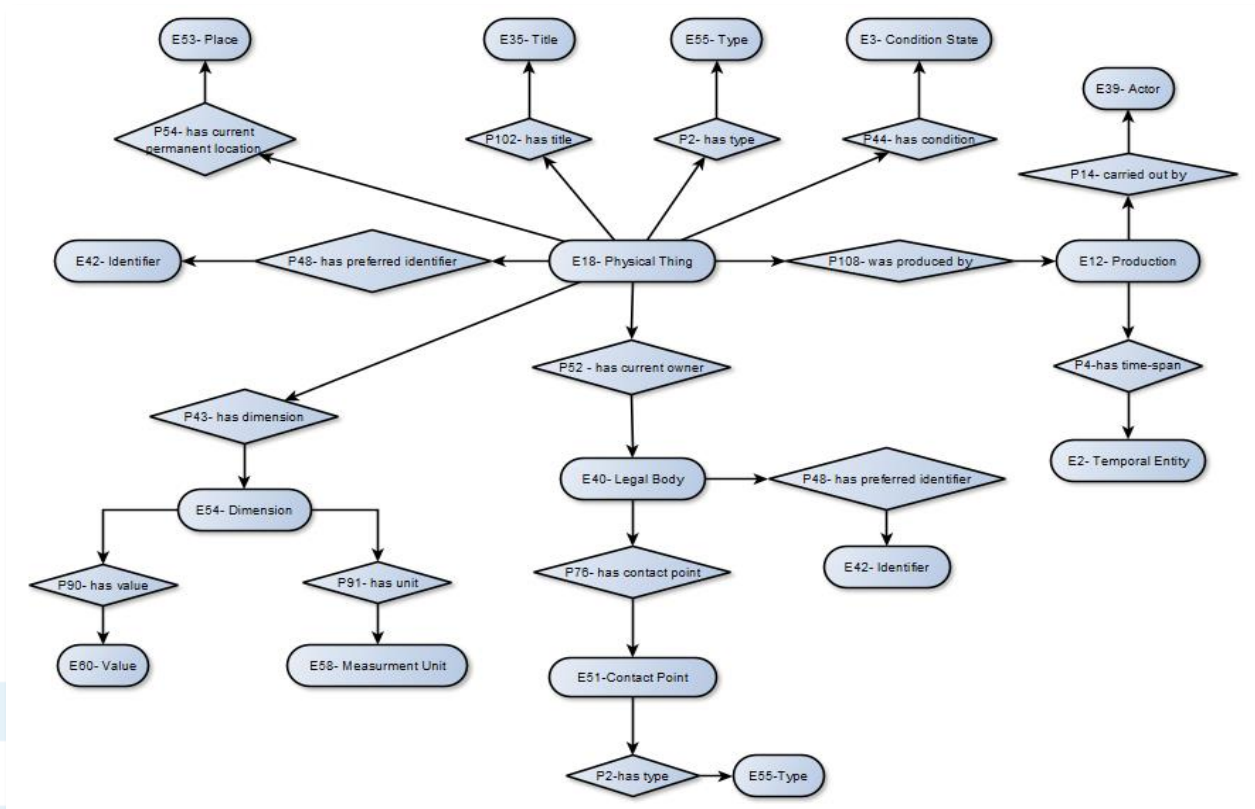


FIGURE 31: SCHEMA OF THE GENERAL INFORMATION OF THE ARTWORK USING CIDOC-CRM.

The information is centred on the artwork (E18- Physical Thing) and from this entity different blocks of data are reported. In the following there is in the detail the schema dividing entities and properties and with an example of its use.

3.3.1 Identifier of the artwork, type, title and location and condition state

E42: Identifier	P48: has preferred identifier
E3: Condition State	P44: has condition
E35: Title	P102: has title
E55: Type	P2: has type

Example: Madame Manet in the conservatory(E35) has id (P48) Id-xxx-xxx-xxx (E42) presents (P44: 1→n) craquelure (E3) is (P2) oil painting (E55)

3.3.2 Owner, contact data and location

E40: Legal Body	P52: has current owner
E42: Identifier	P48: has preferred identifier
E51: Contact point	P76: has contact point
E55: Type	P2: has type

Example: Madame Manet in the conservatory has current owner (P52) “Nasjonalmuseet for kunst, arkitektur og design” (E40) with identifier (P48) lxxx-xxx-xxx. The museum (P52) has contact data (P76):

- + 47 22 20 04 04 (E51) type of (P2) phone number (E55)
- info@Nasjonalmuseet.org (E51) type of (P2) email (E55)

3.3.3 Author, type of artwork and date

This data has been linked using the class production since no direct relationship was available.

E12: Production	P108: was produced by
E39: Actor	P14: carried out by

E2: Temporal Entity	P4: has time span
---------------------	-------------------

Example: Madame Manet in the conservatory was produced with technique (P108) oil painting (E12):

- by (P14) Edouard Manet (E39)
- in (P4) 1879 (E2)

3.3.4 Dimension of the artwork

E54: Dimension	P43: has dimension
E60: Value	P90: has value
E58: Measurement Unit	P91: has unit

Example: Madame Manet in the conservatory has (p43):

- height (E54) of 60 (E60) cm (P91, E58)
- width (E54) of 100 (E60) cm (P91, E58)

This set of data has been implemented in MOVIDA. Its implementation and an example file will be described in section 3.6.

3.4 Schema for: General Analytical Campaign (GAI)

After having analysed all the features (entities and properties) present in CIDOC-CRM and the CRMsci extension it was determined that the existing options did not seem to fully meet the requirements of the MOVIDA data set, and some additional entities and properties could be required. The following section will highlight the relationship that still need to be described by proposing changes that would allow MOVIDA to generate a file fully compliant with CIDCO-CRM, complete with the data from a multi-technique scientific investigation of a cultural heritage object.

The information intended for delivery as a FAIR file regarding an analytical campaign is:

- Artwork general information (see previous section)
- Person responsible for the campaign
- Period of the campaign
- Techniques used
 - Participant, Institution
 - Dates
 - Instruments

- Type of data generated

Figure 32 represents the proposed schema. For the sake of readability green rectangles group information that is used more than once (such as persons involved, range of time or other defined IDs). These groups of data are described in Figure 33. Entities and procedures that represent relationships that may require extensions to the existing CIDOC schema are highlighted in red.

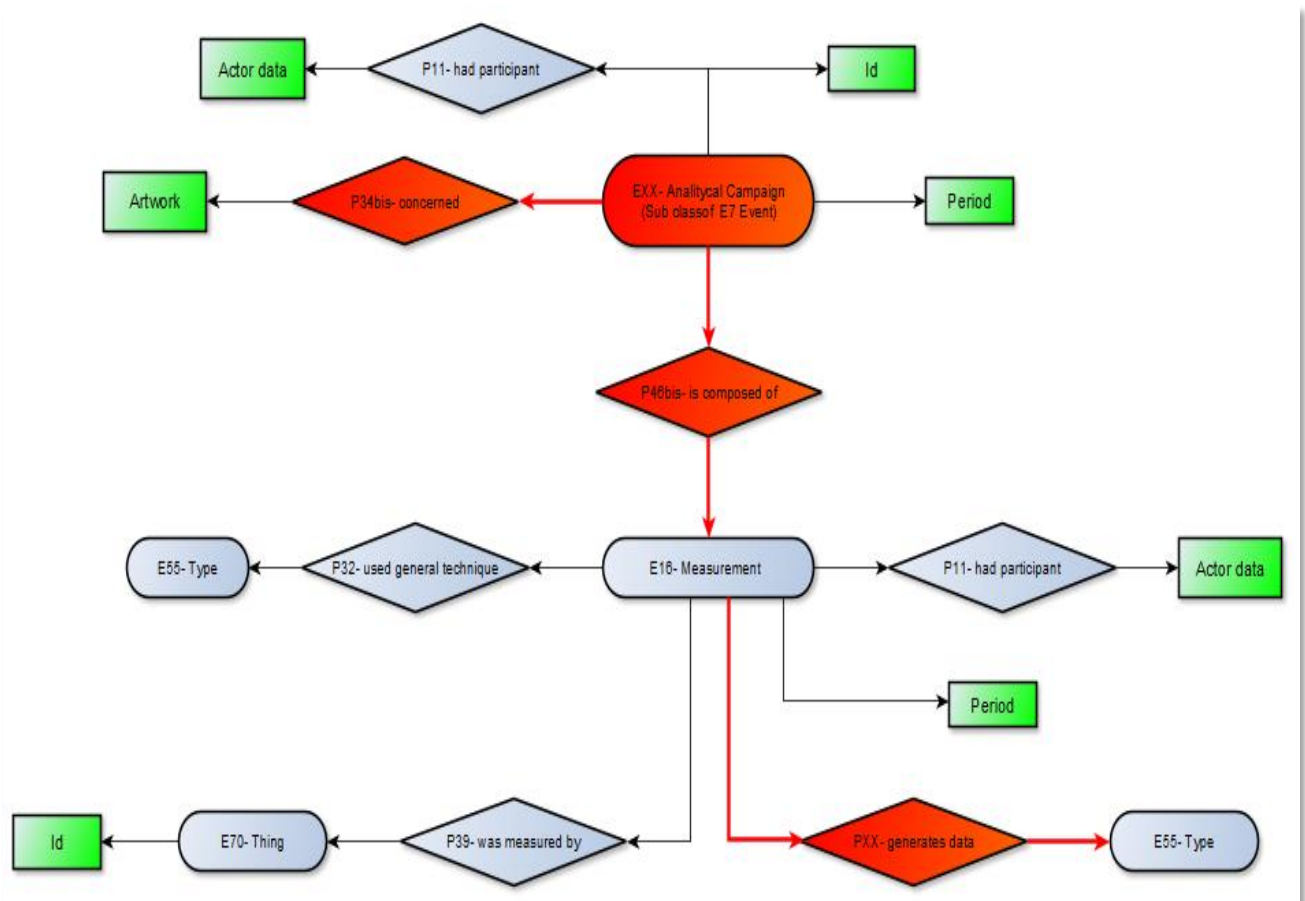


FIGURE 32: MAIN SCHEMA PROPOSED FOR THE ANALYTICAL CAMPAIGN DATA

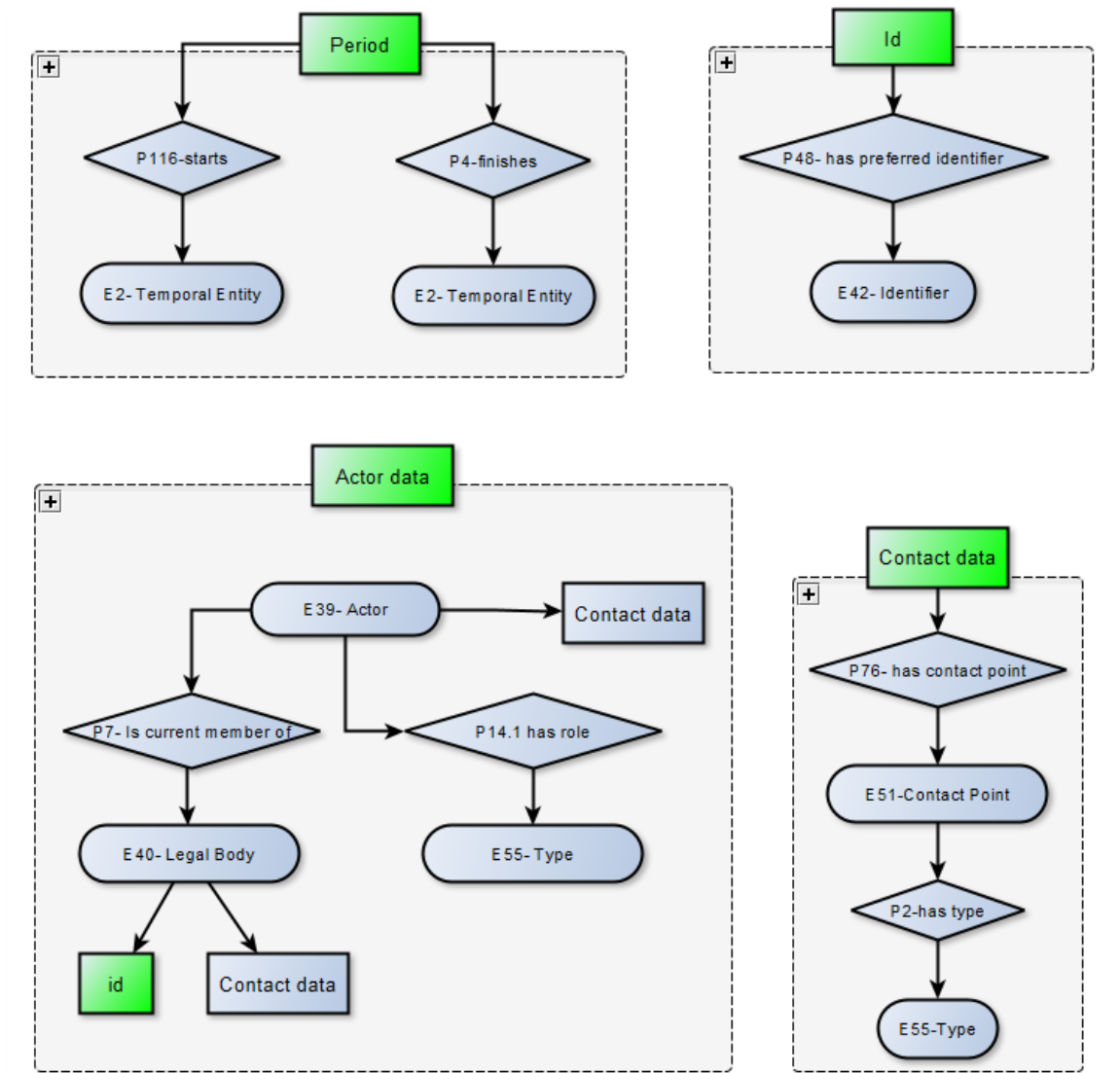


FIGURE 33: BLOCKS OF INFORMATION THAT ARE REPEATED ALONG THE MAIN SCHEMA

The information is centred on the concept of an analytical campaign. This entity does not exist now but could be a subclass of E7-Activity (see below for specifications). From this entity (EXX-Analytical campaign) different blocks of data are reported below.

3.4.1 Identifier of the campaign, responsible and dates, techniques

E42: Identifier	P48: has preferred identifier
E16: Measurement	P11: had participant

Actor data block	P46bis: is composed of
Period block	P34bis: concerned
General information of artwork block	

Example: Analytical campaign (EXX)

- was performed on the painting Madame Manet in the Conservatory (links to artwork block)
- headed by (P11¹⁰⁴) Mario Rosi (links to Actor data block)
- within the MOLAB access with id (P48) ID-xxx-xxx
- started the 13/05/2020 and finished the 25/06/2020 (Period block)
- used techniques (P46bis)
 - XRF(E16)
 - OCT(E16)
 - NIR (E16)

3.4.2 Techniques, Instruments, type of data, operators, dates

E16: Measurement	P32: used general technique
E55: Type	P11: had participant
E70: Thing	P39: was measured by
Period block	PXX: generates data
Actor data block	

Example: OCT measurements (E16)

- were performed by (P11)

¹⁰⁴ The idea of a person being in charge is not specifically indicated with crm:P11-had participant, this specific relationship could be achieved with the creation of a new sub-property that specifically indicates a person is in charge (responsible) of an activity alternatively in the future it should be possible to define additional properties of a triple rather than of a given entity using RDF*, (once it is completed), thus allowing one to indicate that an Activity - had a participant - in the role of - leader - further details relating to the development of RDF* can be found at: <https://w3c.github.io/rdf-star/cg-spec/2021-07-01.html> [28/04/22]

- Piotr Targowski
- Mario Rossi
- started the 13/05/2020 and finished the 25/06/2020 (Period block)
- used general technique (P32) Optical Coherence Tomography (E55)
- were measured using (P39) Phoenix MICRON™ Image-Guided OCT (E70) with id XXXXX (ID block)
- generating (PXX)
 - Images (E55)
 - xml File (E55)
 - 3D surface cube (E55)

3.4.3 Period

E2: Temporal Entity	P144: Starts P4: finishes
---------------------	------------------------------

Example: Analytical campaign (EXX)

- started (P144) 22/10/2020 (E2)
- finished (P4) 05/12/2020 (E2)

3.4.4 Actor and Contact data

E39: Actor E51: Contact Point E55: Type E40: Legal Body E42: Identifier	P76: has contact point P2: has type P14.1 has role P7: is current member of P48: has preferred identifier
---	---

Example: Mario Rossi (E39)

- has contact data (P76):
 - + 39 22 20 04 04 (E51) type of (P2) phone number (E55)
 - info@mariorossi.org (E51) type of (P2) email (E55)
- has role (P14.1) researcher (E55)

- is current member of (P7) CRN (E40) has contact data (P76):
 - + 39 075 44 44 44 (E51) type of (P2) phone number (E55)
 - info@cnr.it (E51) type of (P2) email (E55)

3.5 Modelling Issues

The CIDOC CRM provides a wide range of options for describing data captured with Heritage Science activities, and there can often be more than one way to model a relationship. During the development of this area of work, as documented in above section 3.4, it was determined that the existing CRM classes and properties did not provide sufficient options for fully describing the required relationships and a few modifications were proposed; the creation or modification of 1 Entity and 3 Properties. Developed descriptions of these changes are included below along with additional alternative options identified during the collaborative examination of this work.

One key area, which relates to this work, but also to several practical applications of the CIDOC CRM, is the notion of attaching properties to relationships rather than just to entities. For example, an event can be described as having P11-had_participant specific person, but this relationship does not define the role of that person during the event. They might have been a leader, a participant, a spectator, a subject, etc. Properties can be ascribed to the event or the person but not to the specific relationship. This is a limitation with the underlying RDF technology. At this time this extra information can be captured through the addition of a range of new properties or by constructing more complex semantic connections between the event and the person, making use of additional nodes in the knowledge graph on which the P2_has_type or similar properties could be used to define roles, positions etc. These options can both work, but are not very practical, the first can quickly result in a very large ontology and the second option can result in much more complex and non-standard data structures which will diminish interoperability.

However, there is a proposed solution to this problem, but it is still in development, so was not exploitable during this work. The underlying RDF and SPARQL technologies are being developed, to allow users to “make statements about other statements”, to form what is being called RDF-Star (RDF*) and SPARQL-star (SPARQL*). This development¹⁰⁵ will become very useful for the modelling of Heritage Science data and is recommended for future work once it is formalised and incorporated into the various pieces of software, like MOVIDA or Blazegraph, that are used to create or interact with the semantic data.

¹⁰⁵ A draft report on this work has been published by W3C: <https://w3c.github.io/rdf-star/cg-spec/2021-07-01.html> [28/04/22]

3.5.1 Defining an Analytical Campaign

a) EXX- Analytical Campaign¹⁰⁶

Introduce a new subclass of E7- Activity specific for a scientific analytical campaign, on an artwork.

Subclass of: E7 Activity

Superclass of: No superclasses

Scope Note: To compromise the activities that result in the scientific study of classes E18 Physical Thing

It specializes the notion of activity into the investigation of an artwork with the respective techniques to be recorded.

Examples: The MOLAB access campaign concerned Madame Manet in the Conservatory – which included multiple different examination of the work using different techniques.

In First Order Logic: $EXX \supset E7$

Properties: P34bis concerned: E18 Physical Thing

Inherited properties: Inherited from E7

Inherited references: Inherited from E7

Update: Now an analytical campaign could simply be defined as an Event that has type “campaign”, but the notions of further nested examination events within the campaign might be lost. Another option might be to use the E4 Period class which could then be used to group the examination events together in time, as this meets the scope description of an E4 Period. Further worked examples of this type of entity will be explored as the work on MOVIDA develops to check if all the other related relationships might be solved by either of these alternatives.

3.5.2 Defining the focus of an Examination

b) P34bis- concerned

Introduce a new property based on P34 that links the activity of performing an analytical campaign to the physical thing (artwork) that is being studied

The present structure of P34 concerned is as follows

¹⁰⁶ In this case the XX in EXX is just used to indicate that there no defined entity number for this proposed class in the CRM.

Domain: E14 Condition Assessment

Range: E18 Physical Thing

Subproperty of: E13 Attribute Assignment.

P140 assigned attribute to (was attributed by): E1 CRM Entity

Superproperty of: --

Quantification: many to many, necessary (1,n:

Scope Note: This property identifies the E18 Physical Thing that was investigated an EXX Analytical campaign activity.

Examples: The MOLAB access campaign (EXX) concerned (P34bis) Madame Manet in the Conservatory (E18)

Update: This would probably need an update to the definition of an E14 Condition Assessment, as the scope note indicates that it is targeted at the preservation state of an object rather than also including the broader notions of an objects physical characteristics or even its physical composition. An alternative approach is to consider an examination as measuring a dimension of and object or identifying a component part of the object, but this would depend on the type of examination. Another option would be to consider the properties P12.occurred_in_the_presence_of or more specifically P16.used_specific_object, with the development of RDF* these could be further described with the use roles or P16.1 mode of use.

3.5.3 Linking related examination activities

c) P46bis- is composed of

The present property P46-is composed of links physical things to physical things. It is necessary to introduce a similar property that can link activities to a major and more generic activity. Example: Analytical campaign composed of a series of subactivities related to each technique. The structure proposed is therefore:

P46bis- is composed of

Domain: EXX Analytical campaign

Range: E16- Measurements

Subproperty of: to be determined

Superproperty of: to be determined

Quantification: many to many (0,n;0,n)

Scope Note: This property allows instances of EXX- Analytical Campaign to be subdivided into minor actions as E16-Measurements

Examples: The MOLAB access (EXX) included (P46bis) XRF, MIR and Raman measurements (E16)

Update: If Campaigns can be considered as Events or Periods then this relationship could be managed using P9.consists_of(forms_part_of) or P10.falls_within(contains).

3.5.4 Defining generated Data

d) Pxx- generates data - “of type”

No property able to link a measurement with the type of data has been found. Therefore, suggestion is to create a new property with the following structure:

Domain: E16- Measurements

Range: E55- Type

Subproperty of: To be determined

Superproperty of: To be determined

Quantification: many to many (0, n:0,n)

Scope Note: This property allows instances of E55- Type to be associated with the action that created them. In particular in this case with the measurement (E16)

Examples: Optical Coherence Tomography measurements (E16) generated (Pxx):

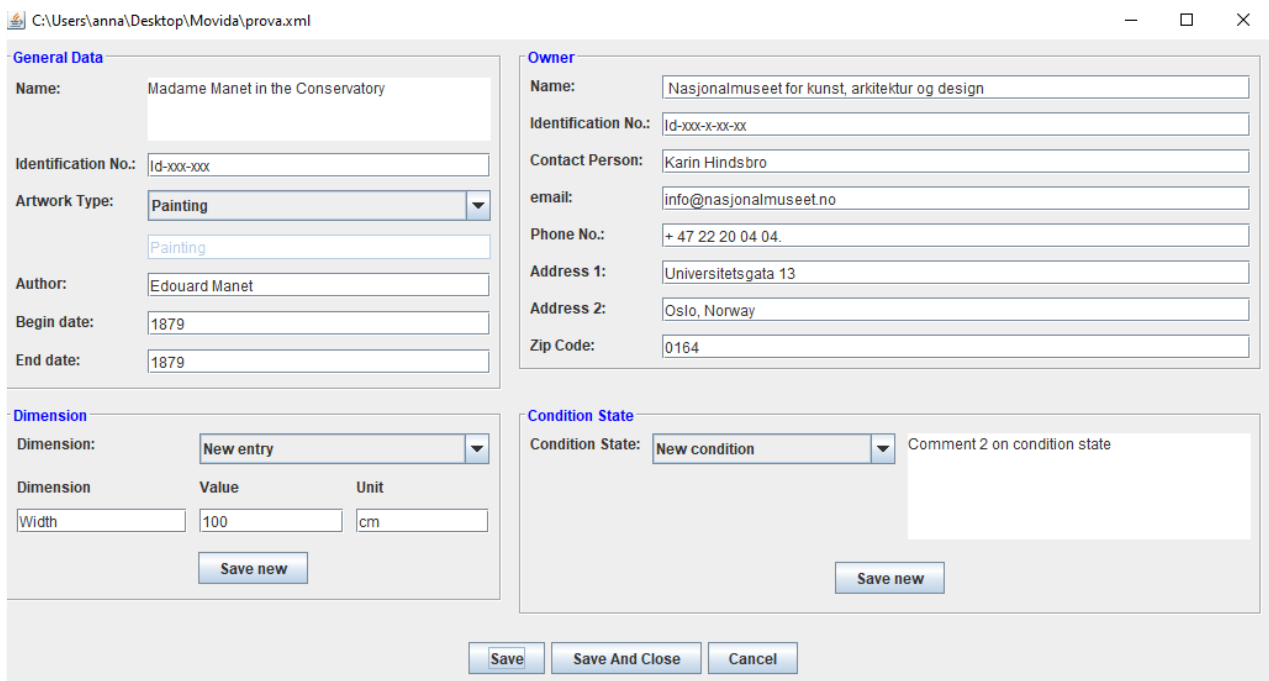
- xml file (E55)
- video (E55)
- 3D surface data (E55)

Update: Simply defining the type of an actual data file that has been created is straightforward, for example the E16-Measurement - P108.has produced – Data (or a E24 physical representation of the data) – P2.has_type – XML file. Alternatively, the newer CRMdig provides the option of L11.had_output – Data (D1 Digital Object) – P2.has_type – XML file. However, both options fail to define the notion of “a type of examination” - “will create” - “a type of data”, which is the purpose of this proposed property. One option could be to say a given E16-Measurement – P2.has_type – “XML generating event”, but this would need to be further complicated to link back to the specific xml file (E55). From a CRM point of view “PXXcreates data” would be a short cut of a more complex query – lists the files types of all of the data files created during events in which this type of E16-Measurement was used. There are also issues to consider that not all examination of a given type might create the same files, particularly as a given technique develops, so this notion of ascribing a type to the data that is usually generated might need to be linked to types of equipment or even specific instances of equipment. This issue will need to be explored further as the more and more Heritage Science data is semantically described and further use cases of how it will be exploited are defined.

3.6 Implementation

As stated before only the general information related to the artwork is being implemented at this stage. The module has been included in MOVIDA 1.0 since it is the version in current use but it's been developed as a java extension so it can be automatically included in a future MOVIDA 2.0.

The data is inserted through a general data form, see Figure 34.



The screenshot shows a web-based form with the following sections and fields:

- General Data:**
 - Name: Madame Manet in the Conservatory
 - Identification No.: Id-xxx-xxx
 - Artwork Type: Painting (dropdown menu)
 - Author: Edouard Manet
 - Begin date: 1879
 - End date: 1879
- Owner:**
 - Name: Nasjonalmuseet for kunst, arkitektur og design
 - Identification No.: Id-xxx-x-xx-xx
 - Contact Person: Karin Hindsbro
 - email: info@nasjonalmuseet.no
 - Phone No.: + 47 22 20 04 04.
 - Address 1: Universitetsgata 13
 - Address 2: Oslo, Norway
 - Zip Code: 0164
- Dimension:**
 - Dimension: New entry (dropdown menu)
 - Table with columns: Dimension, Value, Unit. Row: Width, 100, cm.
 - Save new button
- Condition State:**
 - Condition State: New condition (dropdown menu)
 - Comment 2 on condition state (text area)
 - Save new button

At the bottom of the form are three buttons: Save, Save And Close, and Cancel.

FIGURE 34: GENERAL ARTWORK INFORMATION UTILITY ON MOVIDA 1.0

Once the data is inserted it is saved in a separate xml file. The file has the same name as the base project with _GI (General Information) after the main project name. For example, if the MOVIDA project is stored in a file named MadameManetConservatory.xml a file named MadameManetConservatory_GI.xml will be created. The file, compliant with the CIDOC-CRM standards will be ready to be uploaded to the desired registry. When the user opens a project with MOVIDA the software searched for the _GI.xml file. If the file exists, the information is uploaded, and the user can modify it using the form in Figure 34. If the file has not been yet created (old project) the systems automatically create a stub file with no information, which can be edited later as the work progresses. A simple example of the structure of a created file is included in section 6.1.

3.7 Next steps with MOVIDA

The MOVIDA software is a key component of the process of documenting the continuing work of MOLAB¹⁰² within their national and international research activities. The development and exploitation

of this work carried out in SSHOC and this tool in general, is continuing within current EU funded projects, such as IPERION-HS³². The addition of standardised formats and descriptions, for the results produced in the examination of Heritage works, is key to ensuring interoperability across institutions and nations. The software is currently freely provided to users during MOLAB examinations, but a full open version has not yet been published. As part of the ongoing development of MOVIDA, work is being carried out to assess the complex IPR issues relating to the history of the software, over the past 10 years, to facilitate its future open release¹⁰⁷.

¹⁰⁷ Once it has been completed details and the code for the open release should be published at: <https://github.com/E-RIHS/movida>.

4 Conclusion

Modelling all aspects of Heritage Science data is a complex and expanding endeavour, but even as this complexity grows, with the introduction of new techniques, methods of data processing, collaborations, etc. the benefits of FAIR data and FAIR work practices become more evident. Heritage Science research is a very interdisciplinary field exploiting expertise from nuclear physics to art history. Only by being guided by the principles of FAIR can one ensure that the results of this research can be further developed and linked together, allowing it to be fully exploited in the years to come.

The CIDOC CRM is described as “... a theoretical and practical tool for information integration in the field of cultural heritage.”¹⁰⁸ It and its related compatible models¹⁰⁸ provide an improving generic guide towards how the complex relationships required within the study of Heritage Science can be captured. Worked examples, like those reported here, can be invaluable in understanding how the theory can be used in practice and as more of them are developed a stronger consensus can be reached across the field. This work demonstrates how complex worked examples, of semantic models, can be documented and shared in open re-usable formats, to help facilitate their future development and support the establishment of such a consensus. Work with the MOVIDA system also showed how such standardised mappings can be incorporated in documentation systems, allowing researcher to explore the benefits of FAIR data created automatically during the documentation process.

Mapping the existing NG datasets considered within this task was a complex process and although the planned mapping process has been successful, further work, particularly for the Grounds Database, will be needed to fully integrate these new datasets together. The connection with Dynamic Modeller system proved invaluable to the process of checking the relationships and connections modelled in the new datasets, providing a rapid way of visualising the results and further development of the Dynamic Modeller is already underway.

The examination of FAIR Heritage Science data and widening the use of these types of interoperable models and exploring how they can be used within software systems like MOVIDA, is already continuing in the partner institutions and related projects, particularly within the IPERION-HS³². The data resources developed in this task will continue to be the focus of active research ensuring that they can be integrated into new research resources and institutional digital repositories. The documentation of the work, developing the tools and the datasets, within GitHub, Zenodo and this Deliverable, along with the presentation and accessibility of the results within open live public websites has also highlighted the importance of ensuring that one’s work can also be as open and FAIR and the data it produces.

¹⁰⁸ A current list of the available compatible models can be found at: <https://www.cidoc-crm.org/collaborations>.
[27/04/22]

5 References

- The Go-Fair website - <https://www.go-fair.org/fair-principles>
- The CIDOC CRM website - <https://www.cidoc-crm.org/>
- The current version of the CIDOC CRM - <https://www.cidoc-crm.org/Version/version-7.2.1>
- The IIF website - <https://iif.io>
- Documentation for the IIF standard - <https://iif.io/api/>
- The Zenodo data repository - <https://zenodo.org>
- The CNR Institute of Cultural Heritage Sciences - <https://www.ispc.cnr.it/>.
- Data management documentation from the EU - https://ec.europa.eu/research/participants/docs/h2020-funding-guide/cross-cutting-issues/open-access-data-management/data-management_en.htm
- Mark D. Wilkinson; Michel Dumontier; IJsbrand Jan Aalbersberg; et al. (15 March 2016). "The FAIR Guiding Principles for scientific data management and stewardship". Scientific Data. 3 (1): 160018. <https://doi.org/10.1038%2FSDATA.2016.18>. ISSN 2052-4463. PMC 4792175. PMID 26978244. Wikidata Q27942822
- The ISO standard for the CIDOC CRM: <https://www.iso.org/standard/57832.html>.
- The RDF Schema documentation: <https://www.w3.org/TR/rdf-schema/>.
- Documentation for the CIDOC CRM extension CRMsci - <https://cidoc-crm.org/crmsci>
- Documentation for the CIDOC CRM extension CRMdig - <https://cidoc-crm.org/crmdig>
- Digital presentation of the original RDF presentation of the RRR dataset - <https://rdf.ng-london.org.uk/workshops/lcd>
- EU H2020 IPERION-CH Project Deliverable - D.8.6 Two digital research resources available for open access on the web and one working resource (<https://doi.org/10.5281/zenodo.5838339>)
- 3M software documentation - <https://www.ics.forth.gr/x3ml-toolkit>
- Protegé software documentation - <https://protege.stanford.edu/>.
- H2020 EU funded: Integrated Platform for the European Research Infrastructure ON Cultural Heritage (IPERION-CH) H2020 – Project website: <https://www.iperionhs.eu/> and <https://cordis.europa.eu/project/id/871034>
- Documentation for the CIDOC CRM extension CRMinf - <https://cidoc-crm.org/crminf>
- Documentation and access to the NG's PID system - <https://data.ng-london.org.uk>
- Collaborating project, Practical IIF, website and doc. - <https://tanc-ahrc.github.io/IIF-TNC/>
- Documentation and code for the "Simple Site" system - <https://github.com/jpadfield/simple-site>.
- Documentation and code for the "Dynamic Modeller" system - <https://github.com/jpadfield/dynamic-modelling>.
- Documentation for the Mermaid JavaScript library - <https://mermaid-js.github.io> .
- Documentation for the GitHub Actions process - <https://docs.github.com/en/actions>.
- Documentation for the GitHub Pages system - <https://pages.github.com>.
- Documentation for the Extensions to the "simple Site" system - <https://jpadfield.github.io/simple-site/extensions.html>.
- Doc. and code for the "Simple IIF Discovery" system - <https://github.com/jpadfield/iif-discovery>.
- Joseph Padfield. (2019). D.8.5 Completed example of prototype designs for integration of various types of documentation and analytical data generated for a single object (1.0). Zenodo. <https://doi.org/10.5281/zenodo.5519016>.

- Orla Delaney, & Joseph Padfield. (2022). SSHOC - Raphael Resource CIDOC Mapping (v1.0). Zenodo. <https://doi.org/10.5281/zenodo.6461654>.
- Documentation for Getty's Art and Architecture Thesaurus (AAT) - <http://www.getty.edu/research/tools/vocabularies/aat/index.html>.
- Documentation and the Blazegraph software - <https://blazegraph.com>.
- Documentation for the ResearchSpace software - <https://researchspace.org>.
- Task live data presentation website - <https://rdf.ng-london.org.uk/sshoc/>.
- Public access to the Grounds Database focussed Simple IIF Discovery website - <https://research.ng-london.org.uk/ss-smk/>.
- Draft documentation for the RDF* and SPARQL* technologies - <https://w3c.github.io/rdf-star/cg-spec/2021-07-01.html>.

5.1 List of Figures

Figure 1: A screenshot showing the default generic sparql query used to gather data for a given temporary NG pid.	16
Figure 2: A screenshot of an example "html" presentation of data presented for one of the temporary sshoc pids. This example is showing the data for: https://data.ng-london.org.uk/SC-RFLL-PPU5-636I-2DYE	17
Figure 3: A screenshot of an example sparql query and zooming image presentation using the openseadragon image viewer	18
Figure 4: A screenshot of the landing page for the simple site example and documentation site.	21
Figure 5: Screenshots from the example Simple IIF Discovery site - showing the cross collection search page and then the results for a search for "yellow" presented in two different IIF viewers, OpenSeadragon and Mirador.	22
Figure 6: A section of a Schematic representation of some of the more common semantic relationships that can be documented during the process of taking and describing a saMple from a given painting.	23
Figure 7: A screenshot of the landing page of the website setup to make the existing CIDOC CRM models from the IPERION-ch project more fair.	24
Figure 8: An example semantic model describing a painting in the Raphael collection, created from the prototypes developed in IPERION-CH	25
Figure 9: The Simple Dynamic Modelling Tool	26
Figure 10: The 'map object' function in the Raphael mapping code	30
Figure 11: An example diagram demonstrating the relationships created in relation to the formation of category based groups of image or digital texts.	32
Figure 12: A visual representation of the planned semantic relationships created to define an example painting.	33
Figure 13: Triples describing an image in the old Raphael data (XML format)	34
Figure 14: Triples partially describing that same image in the new Raphael data (Turtle format)	34
Figure 15: A visual representation of the planned semantic relationships created to define a given image and the various derivatives considered in the system.	35

Figure 16: A diagram of the tables in the Grounds MySQL database, showing how the tables are grouped and connect together.....	37
Figure 17: SQL code to create a view joining object, event, and influence tables	38
Figure 18: A visual representation of the planned semantic relationships created to define a given sample and how it is connect to the painting it was taken from.	39
Figure 19: Screenshots of the website dedicated to the presentation of the new SSHOC datasets – with the landing page on the left and an example query on the right.....	42
Figure 20: Screenshots of the dedicated website – The example on the left demonstrates how a query can be used to gather all of the key data for a given painting including a IIIF based thumbnail image and the example on the right tken form the example query that shows how all of the details for a complex, nested set of provenance digital texts can be requested and then presented inline with the relevant IIIF thumbnails.....	43
Figure 21: Detail of website screenshot highlighting the additional tools provided to interact with query results: the eye icon (highlighted in red) allows the user to visualise the results of the query as a flowdiagram in the Dynamic modeller system. the share icon (green) provides a sharable link to the site including the current query and the JSON icon (blue) provides and alternative link to just the raw results of the current query formatted as a JSON document.	43
Figure 22: An image of the semantic model for an object - automatically generated from the live data, which can be compared with the original theoretical model in Figure 11.....	46
Figure 23:A screenshot of the semantic connections generated by combining the output of all of the full Mapping Models queries for the RRR.	47
Figure 24: A screenshot of the new open version of the Grounds database based on data from the SMK.	50
Figure 25: Screenshots from the dediacted Grounds Data Simple IIIF discovery website.	50
Figure 26: MOLAB access at the Estorick Collection of Modern Italian Art, London, UK.	53
Figure 27: Screenshots of MOVIDA from the MOLAB access at the Galleria Nazionale di Arte Moderna, Roma, IT.	54
Figure 28: Minimal example of the MOVIDA xml file generated.....	55
Figure 29: Screenshot of the use of MOVIDA 2.0. Visualisation of a detail of Madame Manet in the Conservatory with the spots where data from Mir and Raman has been registered. Presence of Calcium and Lead from XRF hyperspectral analysis on the face of Madame ManeT OCT Data.....	56
Figure 30: Sqlite schema for MOVIDA 2.0.....	57
Figure 31: Schema of the general information of the artwork using CIDOC-CRM.	59
Figure 32: Main schema proposed for the Analytical Campaign data.....	62
Figure 33: Blocks of information that are repeated along the main schema	63
Figure 34: General artwork information utility on MOVIDA 1.0	70

5.2 List of Tables

Table 1: Mapping existing data - Direct access details for the results	48
--	----

6 Appendices

6.1 Example Simple MOVIDA FAIR XML Dataset

Example of the FAIR xml file (compliant with CIDOC-crm) created with the new MOVIDA utility

```
<?xml version="1.0" encoding="UTF-8" standalone="no"?>
<rdf:RDF
  xmlns:crm="http://www.cidoc-crm.org/cidoc-crm/"
  xmlns:rdf="http://www.w3.org/1999/02/22-rdf-syntax-ns#"
  xmlns:rdfs="http://www.w3.org/2000/01/rdf-schema#">
<rdf:Description rdf:about="Sample MOVIDA">
  <crm:E18_Physical_Thing>
    <crm:P103_has_title>
      <crm:E35_Title>
        <rdfs:label>Madame Manet in the Conservatory</rdfs:label>
      </crm:E35_Title>
    </crm:P103_has_title>
    <crm:P2_has_type>
      <crm:E55_Type>
        <rdfs:label>Painting</rdfs:label>
      </crm:E55_Type>
    </crm:P2_has_type>
    <crm:P48_has_preferred_identifier>
      <crm:E42_Identifier>
        <rdfs:label>Painting</rdfs:label>
      </crm:E42_Identifier>
    </crm:P48_has_preferred_identifier>
    <crm:P44_has_condition>
      <crm:E3_Condition_State>
        <rdfs:label>Comment 1</rdfs:label>
      </crm:E3_Condition_State>
    </crm:P44_has_condition>
    <crm:P44_has_condition>
```

```
<crm:E3_Condition_State>
  <rdfs:label>Comment 2 on condition state</rdfs:label>
</crm:E3_Condition_State>
</crm:P44_has_condition>
<crm:P43_has_dimension>
  <crm:E54_Dimension>
    Height
    <crm:P91_has_unit>
      cm
      <crm:E58_Measurment_Unit>
        <rdfs:label />
      </crm:E58_Measurment_Unit>
    </crm:P91_has_unit>
    <crm:P90_has_value>
      81.0
      <crm:E60_Valye>
        <rdfs:label />
      </crm:E60_Valye>
    </crm:P90_has_value>
  </crm:E54_Dimension>
</crm:P43_has_dimension>
<crm:P43_has_dimension>
  <crm:E54_Dimension>
    Width
    <crm:P91_has_unit>
      <crm:E58_Measurment_Unit>
        <rdfs:label />
      </crm:E58_Measurment_Unit>
    </crm:P91_has_unit>
    <crm:P90_has_value>
      100.0
      <crm:E60_Valye>
        <rdfs:label />
      </crm:E60_Valye>
  </crm:E54_Dimension>
</crm:P43_has_dimension>
```

```
</crm:P90_has_value>
</crm:E54_Dimension>
</crm:P43_has_dimension>
<crm:P108_was_produced_by>
  <crm:E12_Production>
    <rdfs:label />
    <crm:P14_carried_out_by>
      <crm:E39_Author>
        <rdfs:label>Edouard Manet</rdfs:label>
      </crm:E39_Author>
    </crm:P14_carried_out_by>
  </crm:E12_Production>
  <crm:P4_has_time_span>
    <crm:E32_Temporal_Entity>
      <rdfs:label>1879</rdfs:label>
    </crm:E32_Temporal_Entity>
  </crm:P4_has_time_span>
</crm:E12_Production>
</crm:P108_was_produced_by>
<crm:P52_has_current_owner>
  <crm:E40_Legal_Body>
    <rdfs:label> Nasjonalmuseet for kunst, arkitektur og design</rdfs:label>
    <crm:P48_has_preferred_identifiier>
      <crm:E40_Identifier>
        <rdfs:label>Id-xxx-x-xx-xx</rdfs:label>
      </crm:E40_Identifier>
    </crm:P48_has_preferred_identifiier>
    <crm:P76_has_contact_point>
      <crm:E51_Contact_Point>
        <rdfs:label>info@nasjonalmuseet.no</rdfs:label>
      </crm:E51_Contact_Point>
    </crm:P76_has_contact_point>
    <crm:P76_has_contact_point>
      <crm:E51_Contact_Point>
        <rdfs:label>+ 47 22 20 04 04.</rdfs:label>
      </crm:E51_Contact_Point>
    </crm:P76_has_contact_point>
```

```
</crm:E51_Contact_Point>
</crm:P76_has_contact_point>
<crm:P76_has_contact_point>
  <crm:E51_Contact_Point>
    <rdfs:label>Universitetsgata 13</rdfs:label>
  </crm:E51_Contact_Point>
</crm:P76_has_contact_point>
<crm:P76_has_contact_point>
  <crm:E51_Contact_Point>
    <rdfs:label>Oslo, Norway</rdfs:label>
  </crm:E51_Contact_Point>
</crm:P76_has_contact_point>
<crm:P76_has_contact_point>
  <crm:E51_Contact_Point>
    <rdfs:label>0164</rdfs:label>
  </crm:E51_Contact_Point>
</crm:P76_has_contact_point>
</crm:E40_Legal_Body>
</crm:P52_has_current_owner>
</crm:E18_Physical_Thing>
</rdf:Description>
</rdf:RDF>
```