Research and Innovation Action

Social Sciences & Humanities Open Cloud

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

Report on **Milestone 49**

Heritage Science and Humanities Pilot alpha release

Dissemination Level	PU
Due Date of Milestone	30/06/2020 (M18)
Actual Achievement Date	**31/03/2021**
Lead Beneficiary/LTP	16. CNR
Work Package	WP9 - Data Communities
Task	Task 9.4 Heritage Science and Humanities
Version	V1.1
Number of Pages	p.1 – p.11

Abstract: Milestone 49 of the SSHOC project concerns the release of a "data pilot" project, based on a modular approach, based on a complete and documented workflow - to be published in the SSHOC Open Marketplace - composed of open access tools and of best practices supporting digital humanities and heritage science data integration. The resulting platform (RESTORE), will foster the integration and interoperability of datasets provided by different GLAM institutions, using different (domain driven) data structures and information standards used in the DH and HS domains, such as: EAD/EAC-CFP, ICCD, TEI, etc. This report provides an overall description of the alpha version of the platform.

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.

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

Authors List

Organisation	Name	Contact Information
CNR	Emiliano Degl'Innocenti	emiliano.deglinnocenti@cnr.it
CNR	Carmen Di Meo	dimeo@ovi.cnr.it
CNR	Francesco Coradeschi	coradeschi@ovi.cnr.it
CNR	Maurizio Sanesi	maurizio.sanesi.eng@gmail.com
DARIAH-IT	Elisa Brunoni	elisa.brunoni@beniculturali.it
FORTH	Athina Kritsotaki Eleni Tsoulouha	athinak@ics.forth.gr tsoulouha@ics.forth.gr

Introduction

The disciplines connected to the fields of Digital Humanities and Heritage Science use different standards, formats, software and tools for the production and the management of datasets which can result in very heterogeneous and non-interoperable resources. This current situation makes it difficult to share the said resources (i.e.: data and related metadata) deemed significant for each domain (i.e., Material history and culture, History, History of Art, History of Literature, Archaeology, Philology, applied Physic and Chemistry, etc.), or interdisciplinary workflows in research, and it is therefore almost impossible to achieve a full data integration on relevant and similar topics.

Among the SSHOC MS49 goals there is the support for heterogeneous data management for the fields of DH and HS, to make them interoperable, following the FAIR approach, with a specific focus on machine readability and semantization (i.e.: using a *PERSON - EVENT - OBJECT*- related logic). To contribute to the implementation of such a framework we are proposing a sustainable model for data management - covering ingestion, normalization, mapping and modeling - as well as a set of tools supporting a number of domain standards and procedures in use in different research fields. The proposed model is based on the CIDOC-crm conceptual framework, thus is fully compatible with the SSHOCro (i.e., the SSHOC reference ontology).

Description of the Milestone

To achieve the goals envisioned for MS49, listed above, the CNR-OVI and DARIAH.it teams collected different sets of data resulting from the digitisation of archival, catalographic, and textual resources. The GLAM partners involved in the data pilot - Archives, Libraries, Museums and Research Centres - provided the required resources to test the platform, also developed by OVI. Such material consists of:

- XML files based on the EAD (Encoded Archival Description) and EAC-CPF (Encoded Archival Context-Corporate Bodies, Persons and Families) standards, representing digital descriptions of archival documents and controlled vocabularies of specific items (individuals, organizations, etc)
- Descriptive records of artworks stored in museums' collections, exported in the XML format, together with images (jpgs) and a nucleus of historical catalogues only partly available in digital format (as pdf);
- texts encoded in the XML-TEI (Text Encoding Initiative) format, with data on various lexical items (categories) including anthroponyms and toponyms, lemma and hyperlemma indexes;

The final release of the platform will include a backend infrastructure (i.e. all the modules, scripts, and tools needed to gather, store, align and edit the data provided by the partners) and a set of web interfaces (frontend) to allow the end users to explore (i.e.: search and browse) the data, providing both a traditional search form, with options and filters to refine results, as well as advanced tools for semantic data querying and semantic browsing (i.e.: graph navigation exploiting the relationships between the different resources) based on RDF triples stored in a triplestore.

5.1 Role of the Milestone

The work on MS49, coordinated by CNR-OVI with the collaboration of DARIAH.it, focussed on the development and the release of the alpha version of a data integration platform (now available for testing) aimed at Digital Humanities and Heritage Science researchers and professionals, despite their background (i.e.: GLAM institutions, Academia, or even out of the expertise side – bringing an interest into the Heritage Culture and the Heritage Science). The MS49 marks the delivery of a toolkit for heterogeneous data modeling and integration, as well as establishing a set of good practices for research data management and storage.

5.1.1 Description of integrated standards

Cultural objects	Description	Standard
1.1 archival data authoring tools	tools generally used by national/state archives to describe archival resources in digital format	ISAD-G, XML-EAD ISAAR-CPF, XML-EAC
1.2 cultural heritage data authoring tools	tools used by museums to describe cultural heritage artefacts in digital format	relational tables and XML formats AAT, TGN, ULAN (Getty RI) ICCD (Italy)
1.3 lexicographical/textual data authoring tools	tools used by research institutions to encode digital lexicographical resources	TEI, custom tags

5.1.2 Technical description of pilot prototype

The provided infrastructure improves access to the cultural heritage digital resources preserved by historical sites, state archives, museums, research centres and other memory and cultural institutions, through the elaboration of customized digital environments and innovative methodologies which improve their understanding of the data and encourage their reuse.

The proposed workflow addresses technical issues and problems of different nature (including semantics), along the whole digital resources lifecycle: from the acquisition of the data to its storage, dissemination and reuse.

Data acquisition from GLAM partners (CKAN)

The relevant datasets are produced in different GLAM contexts and have been created and managed using a vast number of digital tools. Given the relatively high level of specialization, these resources not only often stand isolated from other scientific domains but are also disconnected from very similar types of sources. Starting with data integration, the available datasets were stored in a Knowledge Management System or Datastore, which makes all of it available online[1], together with the related metadata.

Parsing of data on the Entity Relationship model

As a first step of the workflow, a series of custom Python scripts were created in order to convert the initial XML files into simplified CSV data, preventing information loss, and tagging them with human readable elements. This process, carried out in cooperation with the data producers, allows detailed semantic identification of the information objects and deeper understanding of the parameters that will be used in the following process (mapping).

In case of data expressed in a E/R (Entity/Relation) model, the procedure also allows the data to be easily ingested and exported from and to traditional database management systems (RDBMS) such as “MariaDB[2]” or “MySQL[3]”.

Data mapping in CIDOC - CRM

After their conversion, in order to be transferable and interoperable, the data undergo a process of semantic modeling, following the CIDOC-crm model (although other ontologies were taken into consideration).

CIDOC-CRM is recognized as an ISO standard, and it is the most widely used ontology in the domain of cultural heritage. It represents a shared model that provides a conceptualization of cultural heritage.

[1] “CKAN - RESTORE: http://ckan.restore.ovi.cnr.it/” [date of access: 09.07.2021].

[2] “MariaDB Server: https://mariadb.org/” [date of access: 12.07.2021].

[3] “MySQL Database Service: https://www.mysql.com/it/” [date of access: 12.07.2021].

CIDOC-CRM proposes a method of official cataloguing; the idea is that the different metadata models used by organizations for different types of cultural items can be transformed into a single compliant card returned to the semantic web.

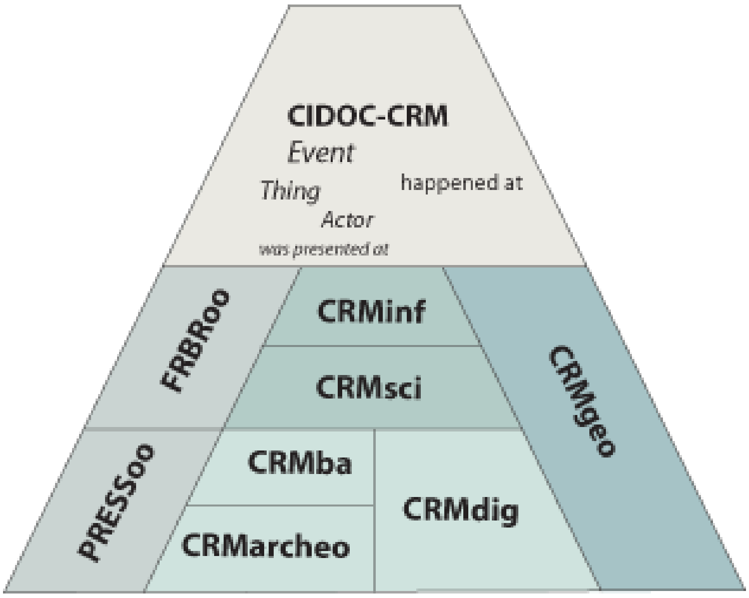


Fig. 1: the CIDOC - crm model showing its derived ontologies

Therefore, a necessary cataloguing of the information following RDF statements, compatible with the CIDOC-CRM standard, has been carried out for the purpose of creating semantic data.

Data Transformation in RDF (Parser in Python, 3M)

Data transformation refers to the process of converting data from the source format (XML, SQL, RAW, etc.) to a destination format (RDF, CIDOC). This process can be managed in different ways, using already available tools - such as Karma[4] or X3ML - or developing customized scripts (as it was done for data parsing).

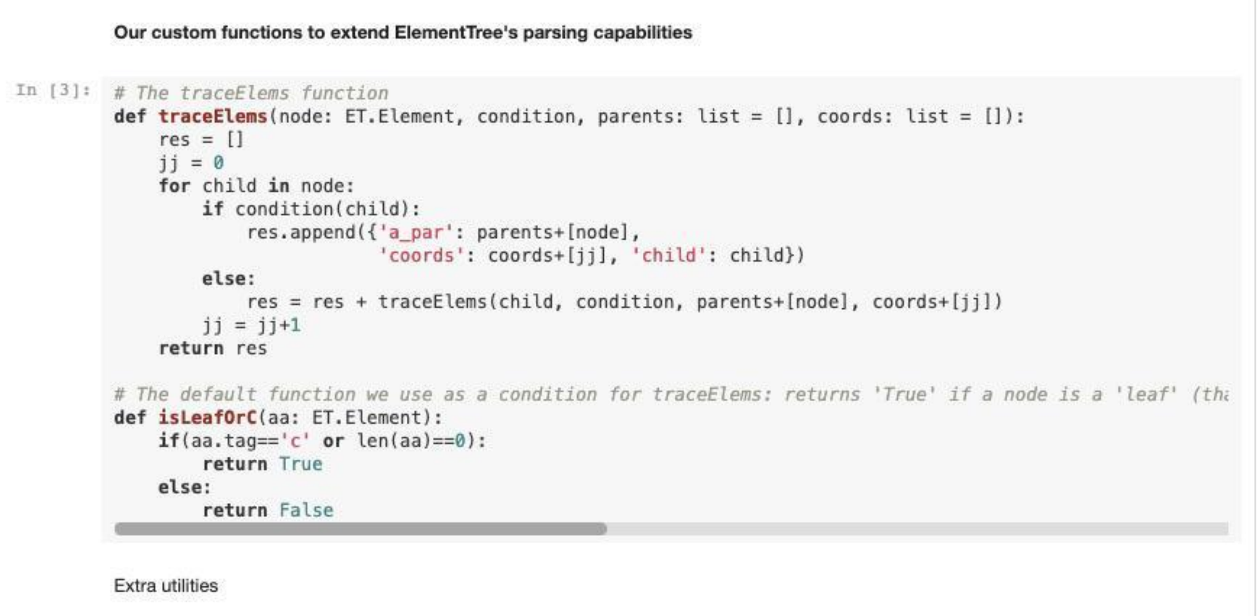
Code storage and documentation (JupyterLab)

All the procedures, documentation, notes and code used to develop the platform are stored in an open-source web application that allows creating and sharing documents that contain live coding: JupyterLab[5],

[4] "KARMA Data Integration Tool: https://usc-isi-i2.github.io/karma/" [date of access: 12.07.2021].

[5] "The Jupyter Notebook: https://jupyter.org/" [date of access: 12.07.2021].

which is language agnostic and supports execution environments in different languages (Python, Julia, R, Haskell, Ruby[6]).



Our custom functions to extend ElementTree's parsing capabilities

```
In [3]: # The traceElems function
        def traceElems(node: ET.Element, condition, parents: list = [], coords: list = []):
            res = []
            jj = 0
            for child in node:
                if condition(child):
                    res.append({'a_par': parents+[node],
                                'coords': coords+[jj], 'child': child})
                else:
                    res = res + traceElems(child, condition, parents+[node], coords+[jj])
                jj = jj+1
            return res

        # The default function we use as a condition for traceElems: returns 'True' if a node is a 'leaf' (tha
        def isLeafOrC(aa: ET.Element):
            if(aa.tag=='c' or len(aa)==0):
                return True
            else:
                return False
```

Extra utilities

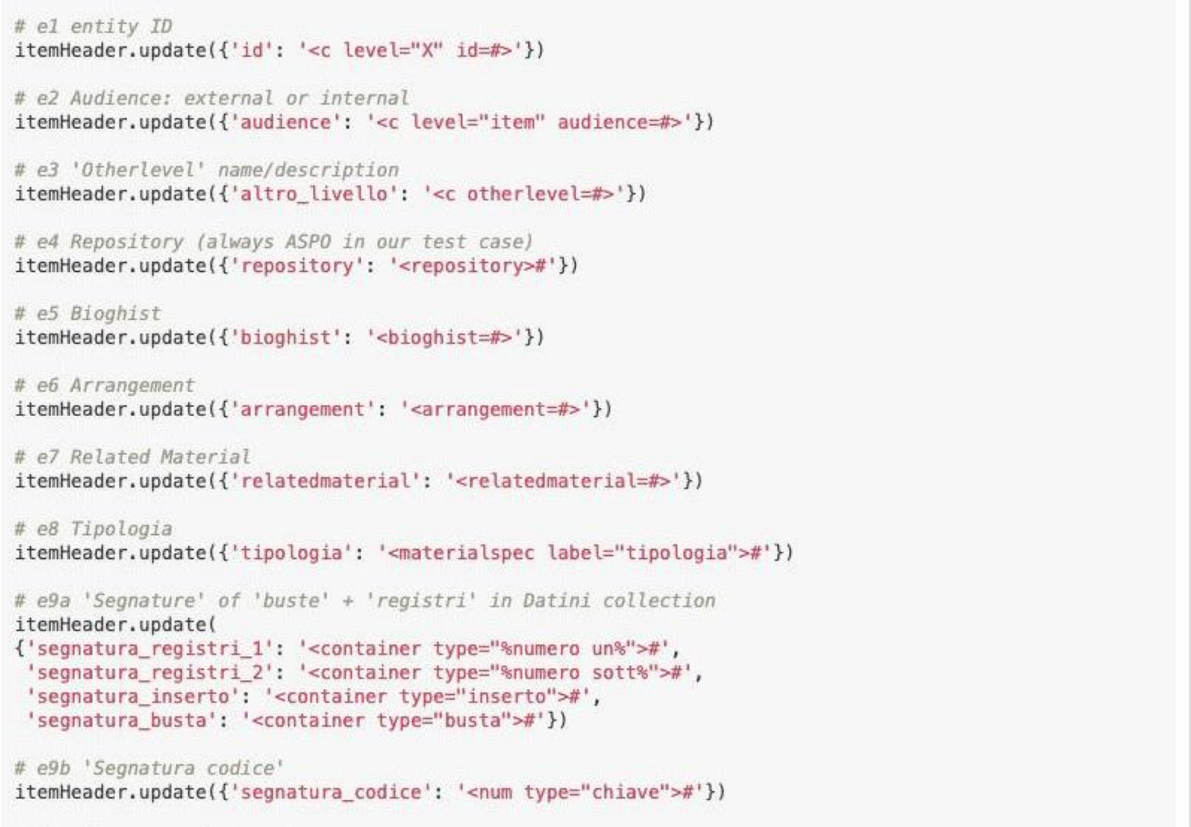
Fig. 2: Custom code documentation - EAD_to_CSV_datini

Control, versioning e documentation (GOGS)

In addition to JupyterLab, a versioning tool has also been used in order to keep track of the various changes made over time within the several documents produced in relation to the in-house generated code. The work to accomplish archival data and standard (XML-EAD) integration, has required a full code documentation, produced by the CNR-OVI team, which is published using a repository named GOGS, open to online consultation[7].

[6] All names refer to programming languages: https://www.python.org/; https://julialang.org/; https://cran.r-project.org/; https://www.haskell.org/; https://www.ruby-lang.org/it/ [date of accesses: 09.07.2021]

[7] “GOGS - RESTORE code documentation: http://dev.restore.ovi.cnr.it:3000/explore/repos” [date of access: 12.07.2021]



```
# e1 entity ID
itemHeader.update({'id': '<c level="X" id=#>'})

# e2 Audience: external or internal
itemHeader.update({'audience': '<c level="item" audience=#>'})

# e3 'Otherlevel' name/description
itemHeader.update({'altro_livello': '<c otherlevel=#>'})

# e4 Repository (always ASPO in our test case)
itemHeader.update({'repository': '<repository>#'})

# e5 Bioghist
itemHeader.update({'bioghist': '<bioghist=#>'})

# e6 Arrangement
itemHeader.update({'arrangement': '<arrangement=#>'})

# e7 Related Material
itemHeader.update({'relatedmaterial': '<relatedmaterial=#>'})

# e8 Tipologia
itemHeader.update({'tipologia': '<materialspec label="tipologia">#'})

# e9a 'Segnature' of 'buste' + 'registri' in Datini collection
itemHeader.update(
{'segnatura_registri_1': '<container type="%numero un%">#',
 'segnatura_registri_2': '<container type="%numero sott%">#',
 'segnatura_inserto': '<container type="inserto">#',
 'segnatura_busta': '<container type="busta">#'})

# e9b 'Segnatura codice'
itemHeader.update({'segnatura_codice': '<num type="chiave">#'})
```

Fig. 3: Custom parser from XML to CSV, the script which defines EAD csv final structure

Import of triplestore data (OpenLink Virtuoso)

The data are then loaded into a database (Triplestore) specialized in managing RDF triples that creates specific data “graphs”.

VIRTUOSO Universal Server[8] was chosen to manage this aspect of the project. VIRTUOSO automatically provides an endpoint to query data using the SPARQL language. One of the most powerful features of SPARQL is to allow multiple graphs querying, and data aggregation based on the query logic and variable binding.

[8] “VIRTUOSO Universal Server: https://virtuoso.openlinksw.com/” [date of access: 12.07.2021]

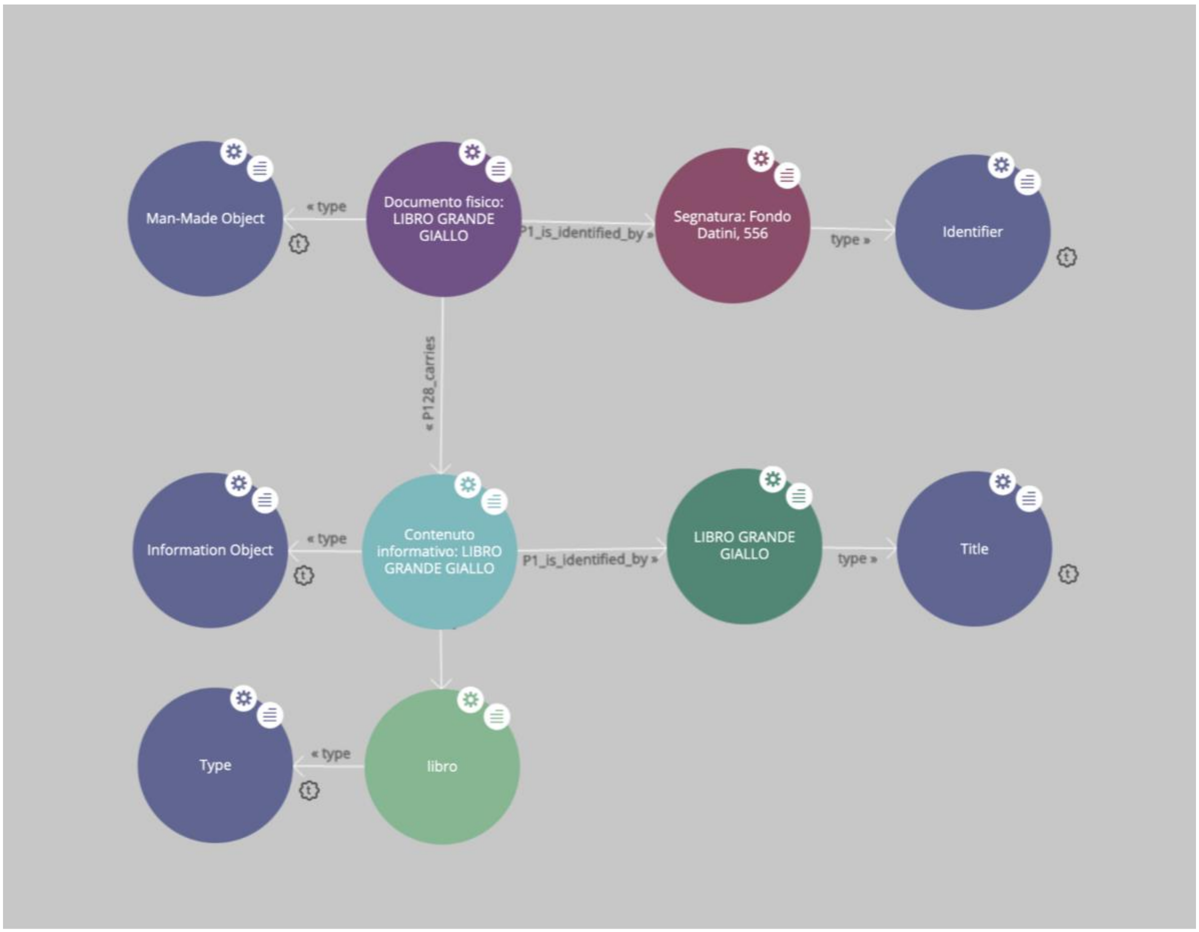


Fig. 4: LodLive view of the archival resource named "Libro Grande Giallo"

Visualization of semantic data (Faceted browser, LodLive)

The resources were connected to endpoints configured within an application for data visualization, allowing end users to switch from one source to another by exploiting the interconnection capabilities inherent in Linked Data[9].

Means of verification

The developed platform is open for testing as a public website (still in Alpha release) at the link below[10]. The SSHOC platform was built on top of the work done within the RESTORE project, also coordinated by the CNR-OVI Institute and co-funded by the Regione Toscana, in the context of the call POR-FSE 2014-20[11], working on data and resources provided by the "Datini" use case, aiming at the reconstruction -

[9] "LodLive - RESTORE: http://dev.restore.ovi.cnr.it/lodlive/" [date of access: 12.07.2021].

[10] "RESTORE platform - Homepage: http://restore.ovi.cnr.it" [date of access: 12.07.2021].

[11] "Programma operativo regionale (Por) del Fondo sociale europeo (Fse): https://www.regione.toscana.it/por-fse-2014-2020" [date of access: 13.07.2021].

through archival, museum and textual resources provided by the GLAM partners involved - of the history of the city of Prato starting from its Medieval roots.

In the "Datini" use case the history of the city of Prato is represented by a central XIV century figure, that of the famous merchant Francesco di Marco Datini (ca. 1335 - 1410), and embodied in his cultural legacy, his family and his *entourage*. Scientifically curated datasets related to the Datini pilot and similar sources were made available for data integration by the RESTORE project.

The overall approach, the software components and the technological solutions developed for the Datini use case are to be seen as original achievements and by-products provided by SSHOC T9.4, to the RESTORE project.

The development of the T9.4 pilot has, in fact, also benefited from the collaboration with DARIAH-ERIC (ESFRI Landmark for the humanities and social sciences) and E-RIHS (ESFRI project for heritage science) as the main stakeholders for the Digital Humanities and Heritage Science research communities, through the collaboration and knowledge sharing with several EU-funded or co-funded projects, such as PARTHENOS[12] and IPERION-HS[13]. In particular, the collaboration with the E-RIHS DIGILAB[14] Working Group (DWG) will be crucial for the final release of the integration platform, supporting datasets regarding the diagnostic analyses made on physical objects (such as the documents, artworks and texts, see paragraph 6, below).

Explanation on delay in achieving the Milestone

WP9 has experienced delays regarding deliverables and milestones due to the COVID-19 situation and limited personnel in the WP leading institution. The COVID pandemic has caused considerable delay in the finalisation of this Milestone 49 too. Whereas the construction of the platform was completed on time, its testing depended on the platform having ingested enough digitised information. The progress of digitising materials has been hit hard by the fact that the partners in this work had been closed altogether because of COVID.

With the suspension of activities and on-site work made by the data providers, which had a central role in the project, getting to the completion of MS49 had been only possible from their re-opening onwards.

Conclusions and next steps

The development of the tool has been achieved at the time of writing. The tool will be released in September 2021. Furthermore, the planned workflow for: content processing, enrichment, mapping, and conversion to CIDOC CRM (RDF generation); for dynamic mapping of data and metadata standards to

[12] "PARTHENOS Joint Resources Registry: http://146.48.123.109/" [date of access: 13.07.2021].

[13] "Integrated Platform for the European Research Infrastructure: http://www.iperionhs.eu/about/" [date of access: 13.07.2021].

[14] "E-RIHS Digilab: http://www.e-rihs.eu/ercim-news-digital-humanities/" [date of access: 13.07.2021].

CIDOC CRM; and for re-encoding said data and metadata standards in the diverse existing standards (RDF, SQL, JSON) is almost completed. A further step in the workflow and methods implementation is that of managing disambiguation and clarifying the hierarchies which need to be represented in each knowledge sector by designing and creating thesauri for the elements expressed by data mapping.

The platform future development will focus on the integration of the pilot knowledge base with datasets coming from mobile labs (i.e., MOLAB) such as those commonly in use in conservation labs and institutes internationally (RX, XRF IMAGING, Infrared thermography, XRD - data in the XRDML format-, multispectral information in the ICCD-based format, colorimetry, photographs, radiographs, 3D models, multispectral and raw OCT, TIFF, Raman spectra, and so on), developing a cross-disciplinary information integration system.

Acronyms table

AAT	Art & Architecture Thesaurus, https://www.getty.edu/research/tools/vocabularies/aat/
CIDOC - crm	Comité International pour la Documentation - conceptual reference model, http://www.cidoc-crm.org/
CKAN	Comprehensive Knowledge Archive Network, https://ckan.org
CNR – OVI	Consiglio Nazionale delle Ricerche - Istituto Opera del Vocabolario Italiano, http://www.ovi.cnr.it/
CSV	Comma-separated values
DARIAH-ERIC	Digital Research Infrastructure for the Arts and Humanities, https://www.dariah.eu/
DBMS	Database management system
EAD EAC-CPF ISAAD-G ISAAR-CFP	Encoded Archival Description Encoded Archival Context-Corporate Bodies, Persons and Families General International Standard Archival Description International Standard Archival Authority Record for Corporate Bodies, Persons and Families
EOSC	European Open Science Cloud - https://ec.europa.eu/info/research-and-innovation/strategy/strategy-2020-2024/our-digital-future/open-science/european-open-science-cloud-eosc_en
E-RIHS	European Research Infrastructure for Heritage Science - http://www.e-rihs.eu/
ESFRI	European Strategy Forum on Research Infrastructures, https://www.esfri.eu/
FAIR	Findability, Accessibility, Interoperability, and Reusability - https://www.go-fair.org/fair-principles/
GETTY RI	The Getty Research Institute, https://www.getty.edu/research/
GLAM (Institutions)	Galleries - Libraries - Archives - Museums
GOGS	GO (programming language) Git Service, https://gogs.io/

ICCD	(Italian) Institute for Cataloguing and Documentation
IPERION - HS	Integrating Platform for the European Research Infrastructure - https://www.iperionhs.eu/
JP(E)G	Joint Photographic Experts Group
JSON	JavaScript Object Notation
MOLAB	Mobile LABoratory, http://www.iperionch.eu/molab/
MySQL	Relational database management system
(raw) OCT	raw Optical Coherence Tomography
PDF	Portable Document Format
R	The Comprehensive R Archive Network, https://cran.r-project.org/
RDF	Resource Description Framework
RX, XRF	X-Ray, X-ray fluorescence spectroscopy or X-ray fluorescence
SPARQL	SPARQL Protocol and RDF Query Language
SQL	Structured Query Language
SSHOCro	SSHOC Reference ontology, https://zenodo.org/record/3744861#.YNCJoJMzZQI
TEI	Text Encoding Initiative
TGN	Thesaurus of Geographic Names (Getty Institute), https://www.getty.edu/research/tools/vocabularies/tgn/
TIFF	Tagged Image File Format
ULAN	Union List of Artists' Names (Getty Institute), https://www.getty.edu/research/tools/vocabularies/ulan/
XML	Extensible Markup Language
XRD, XRDML	X-ray diffraction, Scan raw data
X3ML	3M (Mapping Memory Manager), https://www.ics.forth.gr/isl/x3ml-toolkit