MaDiH (مديح): Mapping the Digital Cultural Heritage in Jordan Project

Datasets Identification and Publication Protocol

Authors	Alessandra Esposito, Shatha Mubaideen, Pascal Flohr, James Smithies, Fadi Bala'awi
Date and version	28 October 2020, version 2
Contact(s)	madih.info@gmail.com
Funder & Scheme	AHRC Newton Fund

Related documents:

- Instance of MaDiH on CKAN
- MaDiH (مديح) website

Table of Contents:

1. Overall Research Outline

The research aims to:

- Identify and document datasets related to Jordanian cultural heritage held in countries inside and outside Jordan.
- Contribute to the long-term sustainable development of Jordan's digital cultural heritage, identifying key systems, datasets, standards, and policies, and aligning them to government digital infrastructure capabilities and strategies.
- Assist the Department of Antiquities in Jordan in their planning processes, help product development teams develop their systems, facilitate the aggregation of valuable datasets held in disparate repositories, and ensure data generated from research activity is properly stored and widely accessible.
- Identify infrastructural gaps and opportunities for further development, including system development, data aggregation, online learning.
- Engage in practical prototyping to ensure analysis and lessons learned are cost-effective and aligned to real-world scenarios.

In order to achieve these goals, one of the main outputs of MaDiH (مديح) is the creation and publication of a publicly available prototype data repository. On the one hand, the repository will fill an urgent need for a holistic view of cultural heritage datasets held across and outside Jordan. On the other hand, it will present a basis for future system design, data aggregation and integration (including the ability to cross-search existing databases), and product development.

This document presents the processes and choices involved in the creation of the MaDiH (مديح) repository. The repository is built using CKAN, an open source data publishing tool for collection of data.

2. Values and Principles

Aware of the General Data Protection Regulation (GDPR), MaDiH (مديح) subscribes to King's College London Statement on Use of Personal Data in Research. The information collected by MaDiH (مديح) is already available online for public access and the source of that information is always mentioned.
The nature of personal data is restricted to:

- Names and titles of the individuals that have created or maintain a dataset, and/or contributed to it
- Their institutional affiliation
- Their work/business email address (when available)

Where this information is not already available online, MaDiH (مديح) has made arrangements to contact the author/maintainer of the dataset to ensure permission, as detailed in 5. Dataset Identification and Publication Workflow.

MaDiH (مديح) does not collect any sensitive data.

If you wish for your name and/or your email address to be removed from the MaDiH (مديح)'s repository alongside the relevant dataset(s), please contact the MaDiH (مديح) team at madih.info@gmail.com. We will assess your request, consulting experts where required in relation to publicly funded content. If it is agreed the information should be removed the record will be changed to 'removed upon request of the author/maintainer'.

MaDiH (مديح) works together with the main heritage institutions in Jordan to ensure constant communication with the main Jordanian stakeholders. The workshop conducted at the British Institute in Amman (29th October 2019) gathered the national and international cultural heritage research and institutional community with the aim of opening a conversation on the needs and methods of collecting data about datasets collecting Jordanian cultural heritage.

3. Datasets Identification and Publication Team and Assignments

The MaDiH (مديح) Datasets Identification and Publication team consists of a Project Manager (1.0 FTE) and a Research Assistant (0.60 FTE) based at the CBRL in Amman, and a Research Associate based at the King's Digital Lab (0.65 FTE).
The team is coordinated by a consultant based in the UK and technically supported by the KDL Deputy Director and Senior Analyst.

The CKAN instance of MaDiH (مديح) has been developed by King's Digital Lab, with the involvement of the Principal Research Software Engineer and a Senior Research Software Engineer.

The Jordan-based Project Manager and Research Assistant focus on recording the datasets held by Jordanian institutions, in addition to the organization of the project workshops in Jordan to build networks and support the datasets identification process.
The UK-based Research Associate is responsible for recording datasets held outside Jordan and for the first drafts of all Datasets Identification Documents (i.e. Datasets Identification Template, Vocabulary, and this Protocol).

Once the Datasets Identification and Publication phase started, the Project Manager, the Research Associate, and the Research Assistant have focused on recording of the datasets in the CKAN instance of MaDiH.

Any matter related to the Identification and Publication processes is discussed in weekly meetings with the Jordanian PI based at the Hashemite University and the UK PI based at King's College London, as well as to the project consultants, the Director of the Council for British Research in the Levant (CBRL Amman) and the Research Associate at the Endangered Archaeology in the Middle East and North Africa (EAMENA, University of Oxford).

4. Testing and Development Phases

A preliminary assessment of potential datasets conducted between August and September 2019 showed a complex dataset landscape consisting of a variety of information (e.g. text, drawings, pictures, digital outputs like 3D models, GIS-based datasets) expressed in both physical and digital formats and available online (sometimes partially) or offline.

To best capture this varied landscape, the Dataset Identification and Publication team has first piloted a Datasets Identification Template in a Google Sheets document shared with the whole team to record a sample of datasets associated with a Vocabulary to ensure standardisation of metadata. After the Template and the Vocabulary had been reviewed by the rest of the project team and modified as needed, the Template was implemented on CKAN and the pilot datasets records uploaded.

The CKAN prototype was piloted by the team to ensure usability in data entry tasks and improve user experience, especially in regards to 'search' function and language accessibility. Any dataset identified after the CKAN piloting phase has been recorded exclusively on the CKAN instance.

5. Datasets Identification and Publication Workflow

The Datasets Identification Methodology is based on the following workflow:

1. Identification of datasets held inside and outside Jordan (see also 6. Datasets Identification Sources).
2. Entering the datasets into CKAN with the visibility set to 'private' (i.e. not publicly visible).
3. Weekly team meetings to update the project team and partners on the research progress and alert them of any risks and issues. The weekly meeting also functions as validation of the datasets identified by the research team prior to publication on CKAN, following the 'Publication specs in CKAN' below:
    - Public datasets in CKAN, containing public information and/or approved by the project team and dataset owner.
    - Private datasets in CKAN for cases where approval from the owner is pending or missing information about who to contact, or if the owner has declined participating in the project.
4. Dataset visibility changed to 'public' on the CKAN instance.
5. Communication and dissemination via the project website and social media accounts (Facebook: @MaDiHJO, and Twitter: @madih_info).
6. Assessing requests of taking down datasets or removing work/business email addresses.

Communication with Dataset Owners.

One of the main aims of MaDiH (مديح) is to engage with the wider research community, including the investigators and/or holders of the recorded datasets, to ensure permission and apply correct licensing of use.

However, given the variability of the datasets landscape as well as the time constraints of the Datasets Identification and Publication phase of the project, the team has prioritised contacting institutions and individuals whose datasets are not currently available online or in digital format. These conditions largely affect the datasets held inside Jordan. The MaDiH researchers in Jordan contacted and sometimes visited local institutions, like museums, which enabled inclusions of their datasets on the MaDiH instance.

Crucial to this phase of the project was also the network established at and after the Project Launch Lecture and the Workshop conducted at the British Institute in Amman (28th-29th October 2019).

6. Datasets Identification Sources

The Datasets Identification relies on:

- Online repositories
- Online library catalogues
- Online museum catalogues
- On-site visits to museums/institutions in Jordan
- On-site visits to museums/institutions outside of Jordan
- Consultants' knowledge and network
- Wide-spectrum internet search
- Word of mouth
- Publications
- Contacts established during MaDiH workshops and at conferences

7. Study Population and Sampling

The project aims to survey the dataset landscape to perfect research tools to enable a thorough mapping of the Jordanian cultural heritage datasets in MaDiH Phase 2, with possible engagement in digitising projects.

It will record datasets relevant to the geographic boundaries of contemporary Jordan, regardless of where those datasets are held, up to modern day.

Table 1. Priority given to the recording of datasets in MaDiH (مديح).

Type of Dataset	Priority	Complexity
Digital online datasets	High	Low
Digital offline datasets	Low	Medium
Analogue datasets	Low	High (approval needed)

8. Datasets Identification Tools

The data was initially collected in a Datasets Identification Template, a Google Sheets document hosted on the project's shared GDrive. The template was based on a pilot CKAN MaDiH Data Template developed by the KDL team, with the addition of 2 custom fields:

- Dataset Time Period
- Data Type

Both the custom fields are associated with a closed list of values. The Dataset Time Period list is based on the period list used in MEGA-Jordan, which was very slightly adapted following EAMENA. The **Data Type list** was based on the Metadata Schema for the Description of Research Data Repositories and the DataCite Metadata Schema Documentation for the Publication and Citation of Research Data and populated with FISH Terminologies and the Getty Categories.

To ensure consistent data quality, standardized vocabularies are used which in the fields are enforced through drop-down menus. Free text is allowed in all other fields, but the standardized vocabulary and detailed guidelines are adhered to when entering data in these fields.

The MaDiH Vocabulary is based on the expected types/features of the dataset landscape and inspired by glossaries used by EAMENA and other heritage institutional projects (Arches and FISH).

The Dataset Identification Template has then been implemented in the CKAN instance of MaDiH (Figure 1).

CKAN metadata	Description
Title	Name of the dataset as indicated by its creators. If an acronym, also indicate what it stands for.
URL (ckan)	URL identifying the dataset in the MaDiH repository, e.g. data.kdl.kcl.ac.uk/dataset/eamena
Project URL	Main URL of the dataset, i.e. its Home page.

Project Principal and Co-Investigators	Names of the dataset PIs and Co-Is.
Project Team	Names and titles of the team members accredited in the project.
Project Start Date	Start date of the project expressed in the format **yyyy/mm/dd.**
Project End Date	End date of the project expressed in the format **yyyy/mm/dd.**
Project Funder(s)	Names of the individuals, institutions, and/or organisations that funded the project.
Description	One paragraph providing an overview of the content of the dataset.
Tags	1. Digital or Analogue 2.Online or Offline 3. Geographical provenance of the records: Region / Governorate / Site name in Arabic and English; 4. pre-1750 and/or post-1750 5. Artefact type 6. inside_Jordan **OR** outside_Jordan 7. dataset_location_country (e.g. dataset_location_France) 8. Dataset language (e.g. dataset_language_Arabic) 9. Heritage Type_Tangibe-Movable_Third level (e.g. Archaeological Site, Object, etc)
Licence	Licence associated with the CKAN data.
Organisations	Organisation or institution that owns or holds the dataset.
Visibility	Set as '**Private**' or '**Public**' depending on whether the dataset is published on CKAN.
Project Status	Development status of the project to be marked as: '**Completed**' if the project has ended, or '**Ongoing** ' if the project is still under development.
Source	Source of the dataset (usually the project main website URL).
Author	Name of the person/people or the organisation that produced the dataset.
Author email	Work/Business Email contact to enquire about the data recorded in the dataset.
Maintainer	Name of the person/people or the organisation responsible for maintaining the dataset, indicating the postal address (if known). It could be the same as the author.
Maintainer Email	Email contact to enquire about the digital output of the dataset.
Dataset Time period	Expressed in cultural period, as found on MEGA-Jordan
Data Type	Type of dataset expressed according to the MaDiH (مديح) authority list.

Figure 1. MaDiH (مديح) Dataset Identification Template.

9. Data storage

The Datasets Identification Template is saved on the project's shared GDrive and is backed on Google servers.

The CKAN repository is hosted on KDL / King's College London servers, and will be supported for two years after the close of the project in 2021. KDL will facilitate migration to Jordanian servers at an appropriate time, with a view to long-term hosting there. The hope is that this preliminary scoping work will lead to follow-on funding to either maintain the resulting repository over the long term, or scale it into a production-grade repository for future generations.

The MaDiH documentation website is hosted on https://reclaimhosting.com/. After the end of the project, it will be either converted to a static site and hosted on KDL servers long-term, or passed to Jordanian partners for longer term hosting.

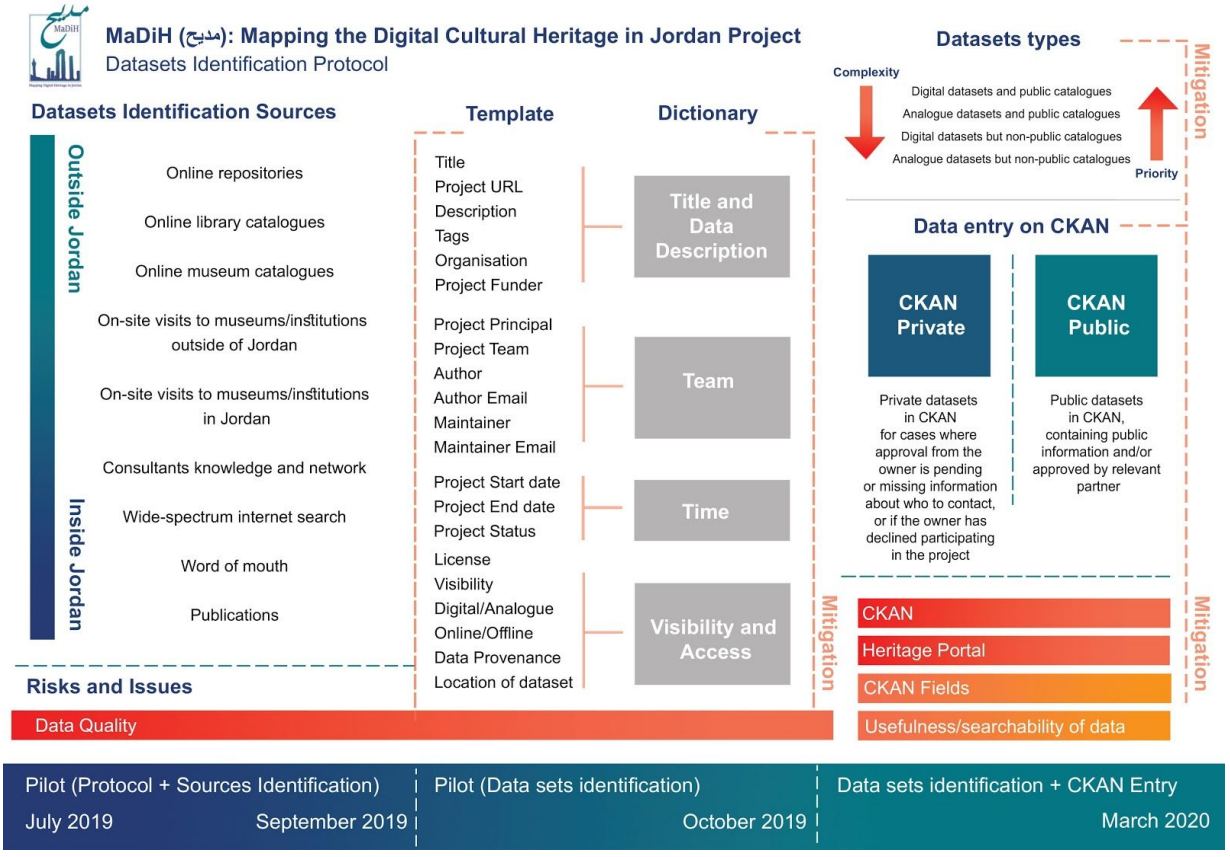

Fig. 2 Dataset Identification Protocol Infographic.