1. Project Title & Acronym and Abstract

Title	*FAIR Data for Historical Games*
Acronym	*PLAYFAIR*
Abstract	The ERC-Digital Ludeme Project (DLP) is constructing a database of historical evidence for Ancient games, aiming at modeling the evolution of games throughout history. This database is unique in its scale, and its development is constrained by the unreliable nature of the data, lacking standards with other historical datasets. PLAYFAIR will apply FAIR principles to our dataset to maximise its usefulness and longevity, and explore the use of Semantic Web and Linked Data (LD) approaches for this purpose. We will connect our dataset — that is the world's most comprehensive dataset on Ancient games — with others, to make a universally FAIR to everyone, and find sources for additional data to complete our own set. PLAYFAIR will tackle the challenge of developing an LD workflow using CLARIAH, show the power of the Semantic Web to answering a research question, and enhance data published on the Web in any applications of digital humanities.
Target Start Date	01.09.2021
Target End Date	01.03.2021

2. Principal Investigator

Name	*Mr. Carlos Utrilla Guerrero*
Function	Research Data Scientist
Organization	Maastricht University (Institute of Data Science)
Address	Paul-Henri Spaaklaan 1, 6229 GT Maastricht
E-mail	c.utrillaguerrero@maastrichtuniversity.nl
Tel	00352661187604

3. Description of the Proposed Project

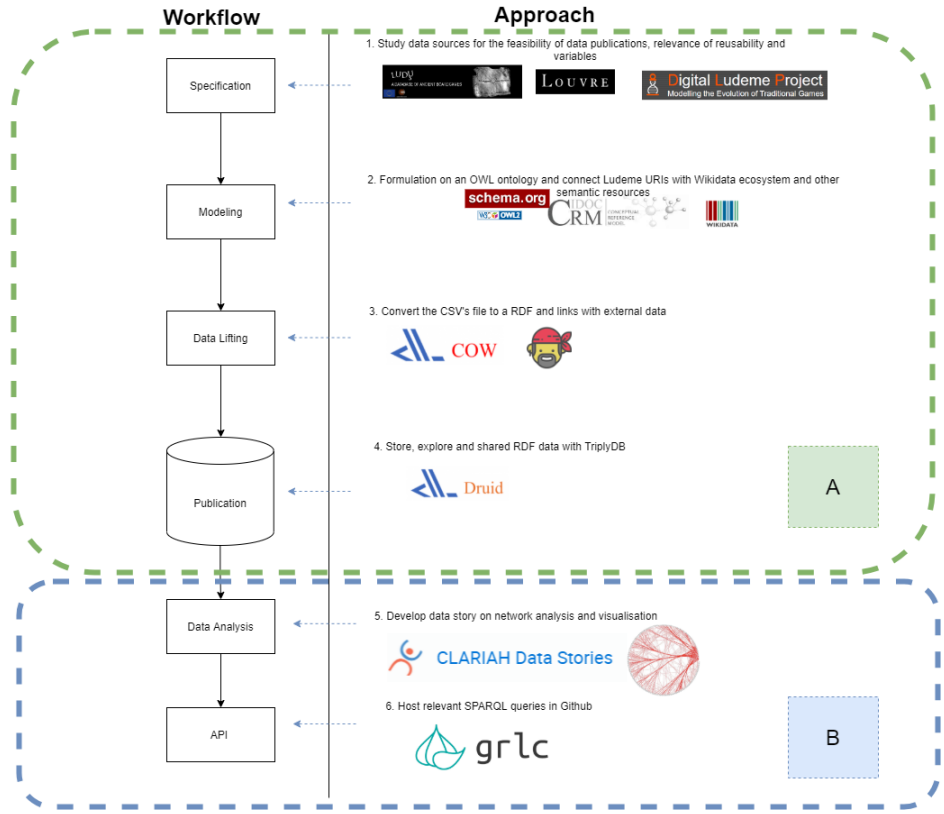
The world is awash with massive amounts of archaeological data, but researchers and data scientists are stymied in their ability to efficiently find, access, connect and reuse this digital gold. The issue arises when one needs to quantitatively describe and explain the major patterns of cultural variation across the history with what is called *data dragons*[1], with no connections to other databases. Archeological sources are abundant but isolated in museums and libraries. Towards addressing this problem, Semantic Web, Linked Open Data (LOD) and FAIR principles [1] were put by a scientific group that included Michel Dumontier, who is the Director of the Institute of Data Science (IDS) and supervisor of the main applicant. PLAYFAIR will address these challenges related to archeological data dragons by

[1] http://datadragon.link/

modeling semantic information and producing Linked Data according to the FAIR principles, resulting in a Knowledge Graph. PLAYFAIR will leverage existing CLARIAH infrastructure, and use existing tools and methods specifically for the conversion and storage of RDF triples.

Inspired by the activities from the guidelines [2] and open challenges pointed here out [3], we will perform the following activities according to our technological approach and workflow as depicted in **Figure 1**.

Figure 1: Technological approach and workflow for THEMIS Knowledge Graph generation



It involves mainly two parts: (A) RDF generation and (B) Data analysis and visualisation.

A: RDF Generation

The part A aims to materialize data as RDF for all those data resources conforming to a standard ontology. The activities that will be performed in this part are specification, modeling, data lifting and publication.

1. **Specification**

The first activity refers to data identification and selection. In this activity, several historical knowledge available on the Web will be analyzed in line with the feasibility of data publication and relevance to the research question. Below is a list **(Table 1**) of the initial datasets from digital projects and cultural institutes that will be published and annotated as part of this project.

Dataset	Description	Coverage	Dimension	Formats	Observations
Digital Ludeme	Digital database of traditional games	Ancient, Medieval, Modern	games, rulesets, ludemes, evidence	Mysql, csv, txt	1000
Locus Ludi	Geographical information system (GIS) database of ancient game boards	c. 800 BCE - 500 CE	games, period, location, material, rules	csv	506
Louvre	Collections databases of the Musee du Louvre and Musee National Eugene-Delacroix			csv	100
GeaCron	GIS database for geopolitical world maps	3000 BCE - to date	locations, trade routes, trade routes, expeditions		

Table 1: List of datasets and dimensions

Other list of collections will be considered to answer specific research question about Roman evidence:

- Database of Pestilence in the Roman Empire
- Stanford Geospatial network Model of Roman World
- Open Access Roman datasets

The datasets will be selected or discarded for transforming into LOD.

2. Modeling

In this activity, data is connected with vocabularies. These vocabularies will be directly either reused if there is a suitable one which encompasses the data, or developed, if there is not any available vocabulary that is suitable for all entities. Widely recognized vocabularies from culture heritage domain *e.g. CIDOC CRM, Linked Data Vocabularies or Getty* domain and beyond the scope of cultural and archeology semantic web resources *e.g. Schema.org and Wikidata* will be used to annotate, for example, rule set properties from game structure which are common in all datasets, such as identifiers, game type, locations and period. Developed vocabularies will be formulated to annotate data particularities from Ludemeplexes (see **Figure 2**)[2]. Following FAIR tenets, this developed vocabulary will be available online as human-readable documentation and in machine format (Ontology Web Language-OWL).

[2] Ludus latrunculorum game has this URI in Ludeme dataset: *https://ludii.games/identifier?Id=DLP.Games.4* while in wikidata is: *https://www.wikidata.org/wiki/Q1700280*

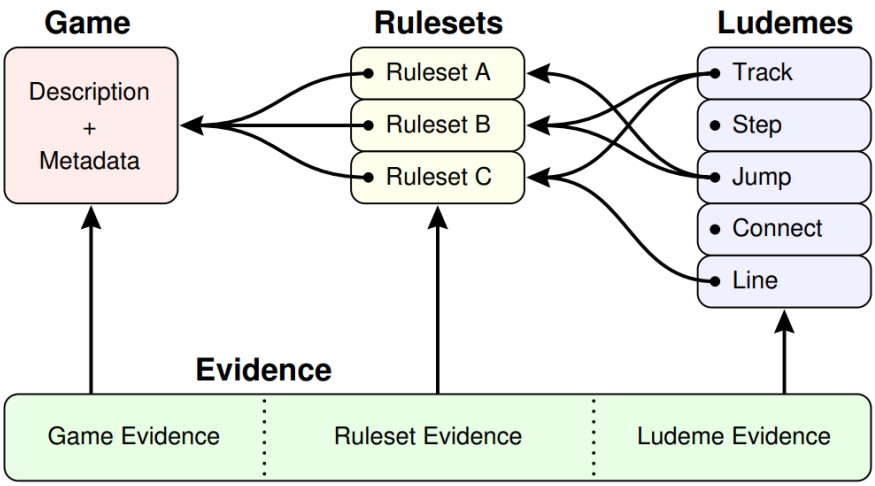


Figure 2. Overview of entities, properties and relationship between DLP Database [4]

3. Data Lifting

This activity aims to represent data in RDF. The data lifting involves two tasks: transformation and linking. In the first task, the datasets selected during the first activity are transformed according to the vocabulary specified in the modeling activity. This task will be implemented with using CoW if possible, or developing scripts using RDF Mapping Language if required. In the second task, external data sources are integrated in order to enrich the dataset. The only main external data source that will be integrated so far is GeaCron to interlink the evolution of games with any given key historical event [4].

4. Publication

This activity consists of data storage and publication. The resulting RDF will be hosted in a triplestore (RDF Database), aka Druid. Druid will allow us to expose a SPARQL endpoint to query our RDF triples. This technology will allow us to take advantage of linked data, because it enables more complex queries than those allowed by an API, for example how likely is that a type of game was played in periods of downs of the Roman civilization from ca. 200 BC to ca. 600 AD? This will be an exciting challenge to explore that will benefit from collaboration with economic historians.

B: Data Analysis and visualisation

This part B aims to implement a simple network analysis, plot graphs and make accessible the SPARQL queries and results in human *e.g. HTML* and machine-readable form *e.g. API*. It involves two activities: data analysis and API.

5. Data Analysis

For the first part, we aim to address this research question: What changes are visible over periods of crisis Roman socio-economy in the evolution of game rulesets? And how games are frequently transmitted between cultures? [4] We will aim to quantitatively identify such patterns using the most common type of network generated from our RDF graphs: co-presence networks. We will also explore the functionalities of storiesdatalegend to show connections between rule games-period and classify games based on historical and mathematical dimension.

6. API

The second activity concerning API, we will host the relevant SPARQL queries in GitHub with metadata to expose them as a grlc.

3.1 Research Question

A general research question is formulated for PLAYFAIR: *How do we bring the significant data set generated by Digital Ludeme Project (DLP) in line with existing datasets in similar Digital Humanities projects?* The above general research question can be divided into the following sub-questions. Using CLARIAH tools:

R1. Can we create a Knowledge Graph that helps communication between DLP and ERC Locus Ludi and other datasets?

R2. Can we convert other relevant external datasets to a form for possible inclusion in the DLP dataset?

R.3. Can we find relations across different datasets and entities? E.g., what changes are visible over periods of the Roman socio-economy crisis in the evolution of game rulesets?

3.2 CLARIAH Component(s)

Combining data science with digital humanities is a unique but essential element of our proposed project. The most important incentive to carry out this project is the ability to address research questions and queries across datasets in CLARIAH infrastructure. We aim to create a resource to help digital historians, data scientists run data analysis using datasets that are entirely FAIR.

This report [7] clearly highlighted datalegend as viable ecosystem for Linked Humanities and Archaeology Data. Our approach depicted in Figure 1 will be implemented to learn about CLARIAH infrastructure, how to use the system to publish our datasets, link them to existing vocabularies and other datasets, and thereby contribute to a growing collection of interlinked datasets. *R.1* and *R.2* are mostly concerning section A of **Figure 1** to assess the CLARIAH tools, whereas last *R.3*. aims to provide a specific case study to perform data analysis and create visualisation over the RDF triplestore to prove useful of the knowledge graph and methods selected (section B of **Figure 1**).

3.3 Description

The core aim of the PLAYFAIR is to make use of CLARIAH infrastructure for raising awareness of FAIR and applying semantic technologies to historical data. We have started testing the software and we have formulated the following functional specifications that will serve as basis for writing the User and Software Technical Specifications (**Table 2**).

UniqueID	Functional Requirement
IDS-01	CSVw converter should enable us to express the types of property.objects
IDS-02	CSVw converter should offer enough preprocessing functions e.g. generate the propertyURL
IDS-03	CSVw converter should define a type for each entity created
IDS-04	CSVw converter might include the option to create own functions
IDS-05	CSVw converter might generate two different subjects from the same object in one mapping file
IDS-06	Druid should storage more than 100,000 triples
IDS-07	Druid should provide means to query and visualize the data
IDS-08	Druid should allow to upload data publicly, with no restricting its access to users
IDS-09	Druid should allow the programmatically upload and insert RDF files via HTTP API
IDS-10	Druid should provide stable SPARQL endpoints
IDS-11	Druid should permit the implementation of complex SPARQL *e.g. network analysis and principal component analysis*
IDS-12	Druid should generate outputs in graph format

Table 2: List of unique identifier functional capabilities that should be met by CLARIAH tools

Dissemination and collaborations

We aim to publish our work in high impact journals such as Journal of Archaeological Open Data, and at top international conferences such as the World Wide Web Conference, the International Semantic Web Conference, the Computer Applications & Quantitative Methods in Archaeology or the Artificial Intelligence and Digital heritage (ARTIDIGH 2021). We will organize a workshop to share our experience with CLARIAH and semantic web in our national and international networks e.g. GO FAIR. We will use this project as a stepping stone for further collaborations in data science and digital humanities. For instance, Mr. Emonet is currently working in the Data Science Research Infrastructure (DSRI) in Maastricht University and will discuss the utilization of our platform to align with CLARIAH needs. We can provide also the DSRI and IDS capabilities to run CLARIAH services and help deploy it on kubernetes at the same time we contribute to documentation, code and bug reports.

3.4 Plan

The project will run from September 2021 until March 2022. **Table 3** shows the section, description and expected time for each deliverable. Section 1 focuses on the Design and implementation of Knowledge Graph as well as the data analysis for answering a domain specific research question. Section 2 focuses on documentation and dissemination of CLARIAH tools and experiences, aiming at providing a reproducible and extensible user and software report. Section 3 focuses more on the provision of permanent website and communication infrastructure for OWL ontology, which we will keep maintain after the project. Section 4 focuses on the publication and dissemination strategy for the project, including workshops and webinar to share our experience with the CLARIAH tools through our national and international network. Finally, Section 5 focuses on manage and keep track of list of bugs over the project in a reproducible and standardized manner.

Section	Description	Deliverable	M1	M2	M3	M4	M5	M6
1	**Design and implementation of Knowledge Graph**	**D1**						
1	Specification	**D1.1**	**D1.1**					
1	Modeling	**D1.2**			**D1.2**			
1	Data Lifting	**D1.3**				**D1.3**		
1	Publication	**D1.4**				**D1.4**		
1	Data Analysis	**D1.5**					**D1.5**	
1	API	**D1.6**					**D1.6**	
2	**User evaluation of the CLARIAH tools**	**D2**						
2	User requirement specification	**D2.1**						**D2.1**
2	Soft-type evaluation dimensions	**D2.2**						**D2.2**
3	**Documentation for the web platform**	**D3**						
3	OWL ontology	**D3.1**						**D3.1**
4	**Dissemination**	**D4**						
4	Workshops	**D4.1**						**D4.1**
4	Journals	**D4.2**						**D4.2**
5	**Management**	**D5**						
5	Quality Assurance	**D5.1**						**D8**

Table 3: Section, description and expected time for each deliverable

4. Deliverables and Milestones

The **table 4** below shows the deliverables for the different work packages and their tasks:

Deliverable	Task	Month	Type	Description
1	1	1	report/datasets	Report and datasets of FAIRified datasets for selected research question and projects
1	2	2	report	An OWL ontology to describe terms concerning ancient games
1	3	4	datasets	First version RDF representation of datasets
1	4	4	file	RDF into triple store
1	5	5	report / demonstration	Data Analysis notebook with SPARQL
1	6	5	web documentation	host in API
2	1	6	report	User technical requirements of overall CLARIAH in practical setting e.g. user story
2	2	6	report	Architectural and design requirements for CLARIAH tools
3	1	6	web and OWL documentation	Comprehensive documentation for the OWL ontology (for both users and developers)
4	1	6	report	Workshops around FAIR, Semantic web within archeological research field
4	2	6	publication	Publication on results of network analysis in conferences
5	1	6	Report	Final Historical Software Bug Report Github

Table 4: Deliverable, Tasks, Month, Type and Description

5. Expertise of the applicant

Carlos Utrilla Guerrero will take the lead of PLAYFAIR. He is a Data scientist with an MSc in Social Research Methods and Statistics. He actively participates in promoting FAIR via the Community of Data Driven Insights (CDDI), which is a university wide ambition to become the first FAIR University (DLP is part of the CDDI). Mr Utrilla also teaches data science and quantitative methods in different degree programs and workshops. PLAYFAIR will be conducted with the guidance of Co-Investigator Associate Professor Dr Cameron Browne, who is PI of the DLP, and IDS developer's from which the data science expertise will be provided.

6. Project budget details

PLAYFAIR will be conducted with the guidance of Dr. Browne. Also, a Postdoctoral Researcher on the DLP, Dr Walter Crist, will contribute on the implementation of suitable methodology for the research question. He is the resident expert on the contents of the DLP database, having manually accumulated and carefully curated the existing data. Dr. Crist is respected internationally as a leading authority in his field, is the author of the pre-eminent book and several articles on this topic, and has extensive contacts with relevant colleagues, authorities and institutions worldwide.

In addition, Mr Utrilla will tightly work with the Institute of Data Science developer's team members from the start to apply our expertise on Semantic Web technologies and FAIR services to the project, in order to fine tune CLARIAH tools. Together with Mr Emonet will devote great efforts to fully understand the functionalities that CLARIAH has, and propose technical improvements that completely statistifes needs and expectations from future users. **Table 5** shows the distribution of requested funding for this call.

	FTE	**costs**	**in-kind**	**Requested funding**
IDS (Mr Utrilla)	0.75	€5440		€ 5440
IDS (Mr Emonet)	0.4	€3450		€ 3450
DLP (Prof Dr Browne and Crist)		€3450		€ 3450
other project-related activities		€500 (workshops and webinars)		€ 500
Software and expertise services (CDDI)		€5420 (FAIR expertise, computing capacity)	€5420	€ 0
Total		**€ 17,840**	**€ 5420**	**€ 12840**

Table 5: Table budget details

7. Literature

[1]: Mark D. Wilkinson, Michel Dumontier, [..] Barend Mons: The FAIR Guiding Principles for scientific data management and stewardship: https://www.nature.com/articles/sdata201618

[2]: Florian Thiery: Sphere 7 Data: LOUD and FAIR Data for the Research Community : https://zenodo.org/record/2643470#.YHAKqOgzZZc

[3]: Albert Meroño-Peñuela, Ashkan Ashkpour , Marieke van Erp , Kees Mandemakers , Leen Breure , Andrea Scharnhorst , Stefan Schlobach , and Frank van Harmelen: Semantic Technologies for Historical Research (2014): A Survey http://www.semantic-web-journal.net/system/files/swj588_0.pdf

[4]: Samanth Subramanian: What we learn from one of the world's oldest board games (2019): https://www.newyorker.com/culture/culture-desk/what-we-learn-from-one-of-the-worlds-oldest-board-games

[5]: Matthew Stephenson, Walter Crist and Cameron Browne Digital Ludeme Project Database Guide (2020): https://ludii.games/downloads/DLP_Database_Guide.pdf

[6]: Cameron Browne: AI for Ancient Games. Report on the Digital Ludeme Project (2020): https://link.springer.com/content/pdf/10.1007%2Fs13218-019-00600-6.pdf%29

[7]: Rinke Hoekstra et al.: The dataLegend Ecosystem for Historical Statistics (2020): https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3180339