

« hedera » platform



Hugues Cazeaux

hugues.cazeaux@unige.ch

<https://orcid.org/0000-0002-5618-2670>

Head of e-Research group, IT Service, University of Geneva

Mathieu Vonlanthen

mathieu.vonlanthen@unige.ch

<https://orcid.org/0000-0002-8982-1432>

e-Research group, IT Service, University of Geneva

Abstract

« hedera » is an active research data management platform based on linked data principles developed by the University of Geneva. It is mainly dedicated to the preservation and dissemination of Linked Open Data. The main objective of « hedera » is to offer a centralized and unified system to avoid data fragmentation and data losses, often encountered by researchers working in siloed, purpose-built, small-scale infrastructures. Data can be imported from existing archives or from another active data platform, in two different forms: metadata import (XML, JSON, CSV) and import of research data files (e.g., images, sound, video, PDF files).

Keywords

Research Data Management, Open Linked Data, Interoperability, Preservation, Reuse, RDF, IIIF, SPARQL



Cet article est disponible sous licence [Creative Commons Attribution - Partage dans les Mêmes Conditions 4.0 International](https://creativecommons.org/licenses/by-sa/4.0/).

1. « hedera » platform

1.1. Origin

Research projects in digital humanities have limited funds available for a limited period of time. Yet, research groups work on their research topics for several years. They use different digital tools or software, create datasets, design websites to promote their work and data. At the end of the funding period, they are looking for ways to keep their results working. In 2018, the IT department at the University of Geneva started to work on these requirements. The beginning of the solution starts with the design of a proof of concept based on open-source software, applying digital humanities best practices: CIDOC-CRM (Bekiari et al., 2024), RDF data (Cyganiak et al., 2014), SPARQL queries (Harris & Seaborne, 2013), IIIF (Appleby et al., 2020)... Working with several pilot and representative projects, in collaboration with the chair of Digital Humanities at Geneva's Faculty of Letters, the proof-of-concept was built to validate the approach. It proved promising, and the decision was taken to develop the « hedera » platform in 2023.

1.2. Inception

Based on the needs of researchers and an analysis of existing projects in the field of digital humanities, the IT service of the University of Geneva began to think about and design the pilot to meet these requirements. The result is the development of a new platform, called « hedera ».

The origin of the name « hedera » comes from « Hedera Helix », the Latin name of the common ivy. This plant, which grows from node to node, evokes the RDF format whose knowledge graphs are articulated by interconnecting nodes. The rapidly climbing nature of the ivy reminds us of the living and growing nature of data. The choice of a plant to name this service is in line with other digital services from the institution, such as Yareta (<https://yareta.unige.ch>) used for long-term archiving and Baobab, a computing cluster; hedera is thus the new blossoming service within the UNIGE IT department's garden.

1.3. Overview

The objectives of the « hedera » platform are:

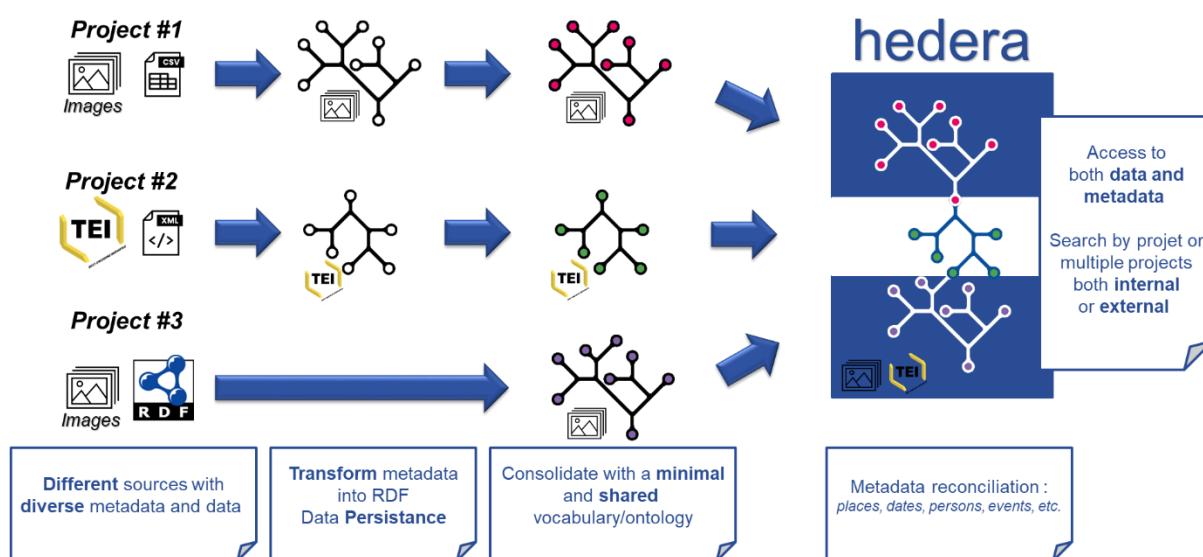
- To offer a centralized and unified platform for researchers
- To manage active data
- To implement best practices in dynamic data management
- To promote standard formats and open-source tools
- To comply with the Open Science and FAIR Principles

The main features of the platform are summarized with the following keywords:

- **Structure**
Organize metadata according to shared and well-established data models
- **Inter-contextuality**
Match metadata entities with corresponding entities across other datasets both within and outside the university

- **Open Access**
Easily browse metadata and data using Linked Open Data standards
- **Enrichment**
Enhance metadata with semantic relationships to make them more meaningful
- **Interoperability**
Integrate data with Yareta, swisscovery, and other systems, thanks to the IIIF & RDF frameworks

Another key value of the « hedera » platform is its capability of importing different kinds of source data, such as CSV, JSON or XML, or RDF data directly. The following schema shows different use cases.



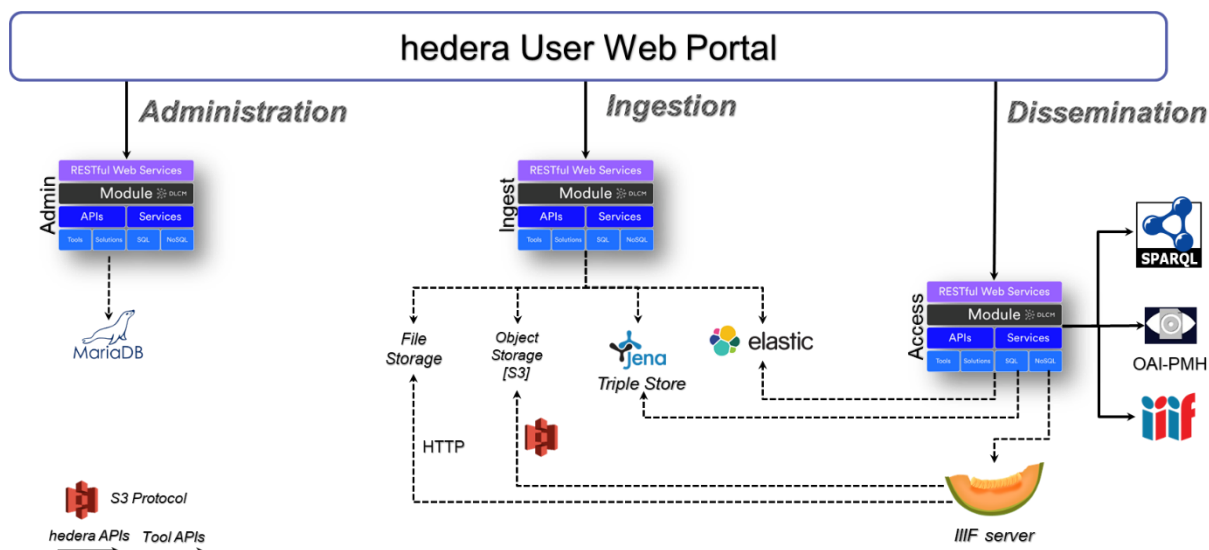
2. « hedera » standards

2.1. Resource Description Framework (RDF)

As a platform for linked data, the pivotal format of « hedera » is RDF, a standard defined by the W3C. RDF provides a standard way of describing the structure and semantics of data, promoting interoperability between different systems and datasets. This is crucial in research, where data often needs to be combined from various sources. RDF enables data to be linked in a semantically rich way, meaning that the relationships between them are explicitly defined. This makes the data more meaningful and valuable for researchers.

All imported metadata will be stored in their original format and converted to RDF using a mapping language such as RML (De Meester et al., 2024). The use of authority files such as VIAF or GeoNames and reference ontologies such as CIDOC-CRM (CIDOC Conceptual Reference Model, and RIC-O (Records in Contexts Ontology (Clavaud & Wildi, 2021), a new standard in RDF for archive description, which will eventually replace ISAD(G) (International Council on Archives, 2000), enables metadata graphs to be aligned between different projects. Each project manager can define a set of research object types aligned with ontology classes. Research objects are extracted from source metadata, assigned a unique identifier, and if necessary, linked to a research data file. Research objects can be searched using SPARQL, a standardized query language for RDF, but they are also indexed in a full-text indexer. This

makes it possible to benefit from the power of graph-based search provided by triple stores such as Fuseki or GraphDB, and also from the full-text search capability of indexers such as Elasticsearch or Solr:



Unlike traditional database formats, Open Linked Data allows greater flexibility in schemas, which can evolve over time or be extended. This is beneficial for research data, which can be dynamic and change as research progresses. Researchers can choose the ontologies they want for their data and metadata models, but for the domain considered in the context of « hedera », the platform currently recommends two ontologies: CIDOC-CRM and RIC-O. In addition to the recommended ontologies, the platform is open to the use of other ontologies.

2.2. SPARQL Protocol and RDF Query Language

« hedera » enables researchers to perform complex queries with SPARQL to retrieve and manipulate data from multiple projects and sources. Once imported, data can be disseminated via SPARQL endpoints (i.e., web services that enable the execution of SPARQL queries against RDF), OAI-PMH protocols (Open Archives Initiative, 2015), IIIF protocols and conventional file downloads. Thanks to these open standards and protocols proposed by « hedera », the « FAIRness » of semantic datasets is improved.

SPARQL endpoints allow researchers to execute queries on the metadata graph of their own projects, and with SPARQL federated queries it is also possible to execute queries on multiple projects. The existence of authority files common to several projects, such as VIAF, enables researchers to discover new facts about their own research objects scattered across datasets produced by other researchers. To enable researchers without SPARQL knowledge to query the data, a prompt-based interface will be provided in the future using a Large-Language Model (LLM) to translate natural language into SPARQL. In this way, a request to the LLM APIs with a specially designed meta-prompt will associate a SPARQL query with each natural language prompt.

Another aim of the « hedera » platform is to offer a data catalog of all research data produced by an institution. The digital humanities community was chosen as a pilot area because of the common requirements of different research groups for the use of open (meta)-data formats

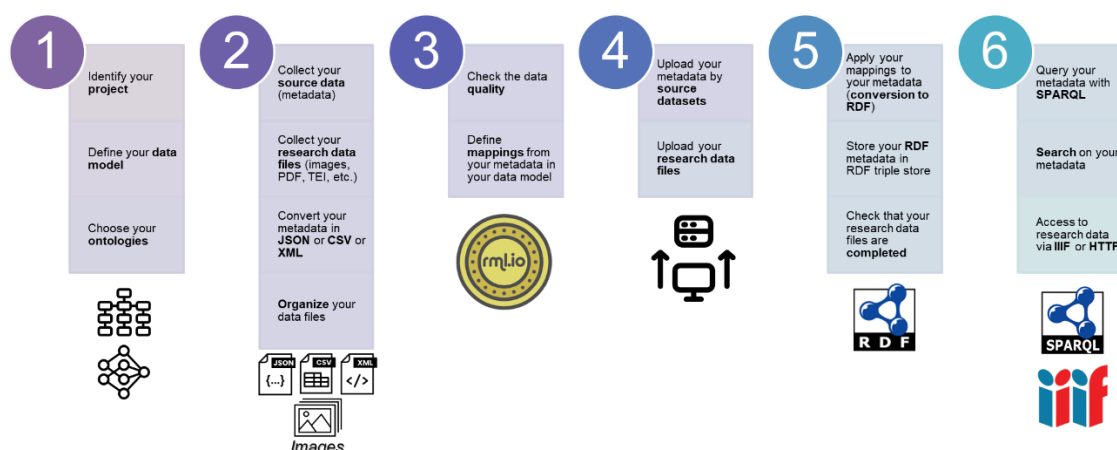
and standard protocols. Published research objects referenced in a project will be disseminated via this catalog using OAI-PMH. On the DLCM repository side, OAI-PMH provides another catalog for archived datasets.

2.3. International Image Interoperability Framework (IIIF)

The IIIF protocol has become the de facto standard for disseminating collection of images with their metadata. This open standard is commonly used in digital humanities to compare editions of manuscripts or museum pieces, which could be hosted in different institutions. In the « hedera » platform, IIIF manifests (a metadata JSON file representing a digitized object with the image list and their metadata) are generated dynamically from the metadata graph linked to the project. This way, for each project a SPARQL request is created to select the metadata required to populate a predefined manifest template. This means that researchers no longer must worry about generating IIIF manifests.

IIIF is widely adopted in the digital community for its interoperability allowing researchers to compare images coming from different datasets. IIIF also enables annotation of image regions and integration of metadata in image viewers such as Mirador.

3. Main steps to import data in the « hedera » platform



3.1. Step #1 – Identification

The first step in using « hedera » is to identify the **research project** by choosing a name, a description and all the information needed to describe it. A project logo can also be defined.

Next, a **data model** must be defined, identifying and naming the different types of research objects. And, for each object type, properties and their format must be defined to describe the object type in detail.

Once the object properties have been identified, the researcher must find if existing **ontologies** could match to the data model. It is recommended to use existing ontologies to facilitate interoperability. If there is no ontology close enough to the data model, it is also possible to create its own ontology.

Once all that information is provided, « hedera » can be configured.

3.2. Step #2 - Collecting data

At this stage, the researcher must identify and collect all data of his/her project. There are two kinds of data:

▪ Source data

- Represents the research object **metadata** and must correspond to the chosen data model.
- Must be converted into one of the supported formats: **JSON** or **CSV** or **XML** or **RDF**
- Must be organized into groups, which represent a batch, called **source dataset**.

If the metadata volume is large (up to 200MB), it is recommended to place it in several files.

▪ Research data files

- Represent the research object **item** itself.
- Can be images, PDF, TEI (Text Encoding Initiative Consortium, 2025), etc.
- Must be organized into folders to manage them more easily.

3.3. Step #3 – Import preparation

At this stage, the researcher must prepare the data import by checking its **quality**. This is very important to guarantee the efficiency of the imports. For example, some checks to do consist in:

- Ensuring the metadata values are consistent with the property format.
- Checking and fixing typos
- Verifying the encoding format; Unicode is recommended

The second task is to prepare for data conversion in the case of non-RDF data. « heder » embeds the RDF Mapping language (RML) to address this task. So, for each format and type of source data, **mapping rules** must be defined in an RML file to establish a correspondence between the metadata and the data model. Some examples of RML files can be provided.

3.4. Step #4 – Uploading data

This step corresponds to the uploading of all data into « heder ». First, the **research data files** must be uploaded by respecting the folder organization and by specifying the type (IIIF, IIIF manifest, TEI or WEB). All data loading operations must be completed without error for the process to continue.

Next, the **source datasets** must be created according to the group organization. For each source dataset, the source data files must be uploaded. Some APIs can be used to upload large volumes.

3.5. Step #5 – Processing data

Once all the data is uploaded, the processing step can be launched. There are three stages involved during this step:

1. RDF conversion

- a. For each **JSON/CSV/XML** source dataset file, the corresponding mapping rules (described within an RML file) must be applied to convert the metadata into RDF.
- b. For all source dataset files, the file must be transformed into a new RDF file with « hedera » identifiers, created or assigned during this transformation.
- c. *All actions must be completed without error to go to the next stage.*

2. RDF storage

- a. Each generated RDF file is stored in the « hedera » triple store.
- b. A generated RDF file can be replaced or removed at this stage.
- c. *All actions must be completed without error to go to the next stage.*

3. IIIF resource generation

- a. Once all RDF files are in the triple store, the IIIF resource generation can be started.
- b. This process will generate IIIF manifests and IIIF collections based on the project configuration.

3.6. Step #6 – Accessing data

Once the five previous steps have been completed, the project data is ready to be accessed. Metadata, that is, RDF data, are searchable using **SPARQL** queries. The research data files are in turn accessible through the IIIF protocol, using compatible clients like Mirador or the HTTP protocol.

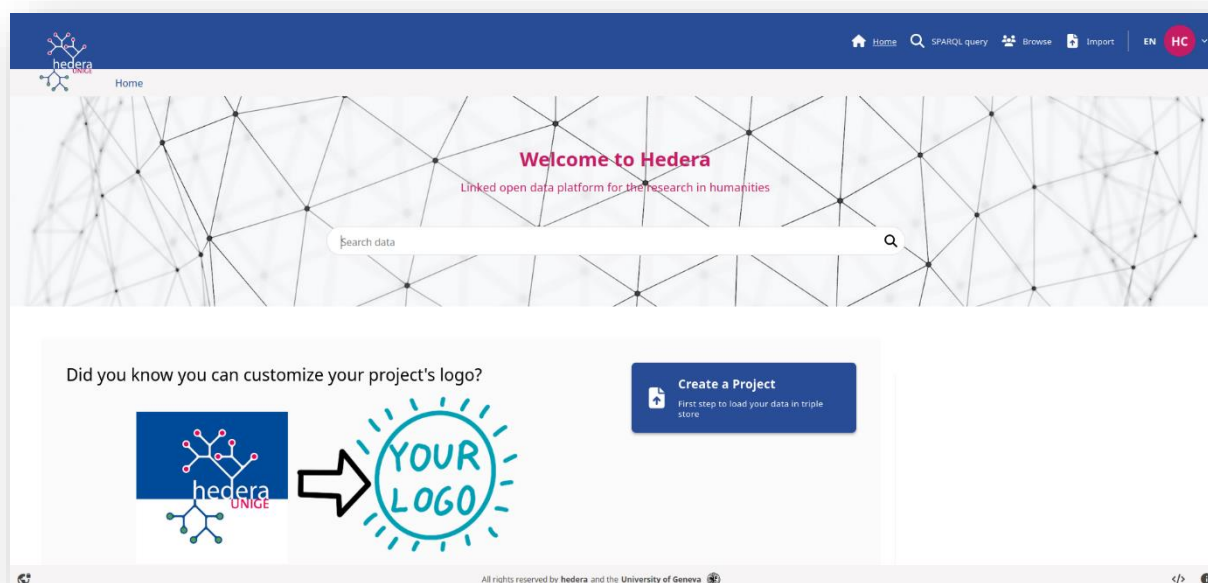
More generally, the « hedera » portal can **browse** the various projects and **search** the content based on metadata.

4. « hedera » portal

This last section of this paper illustrates the web user interface or user portal for the various operations described above.

4.1. The « Home » section

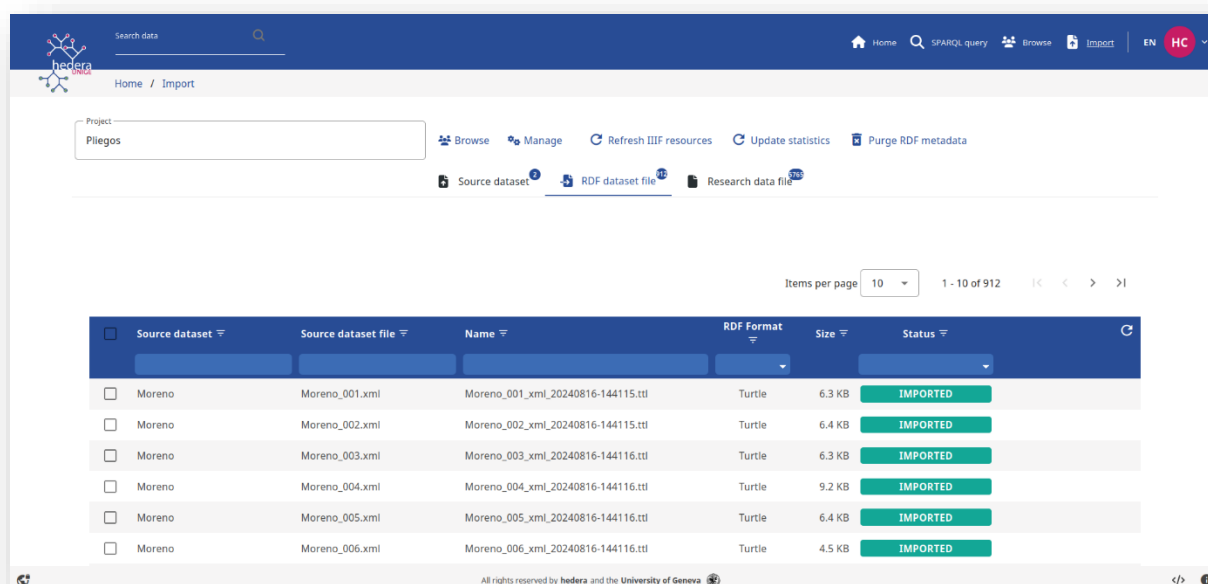
This screenshot represents the « hedera » home page, where users can choose the « browse » menu to access data or the « import » menu to ingest data, based on their rights.



4.2. The « Project Import » section

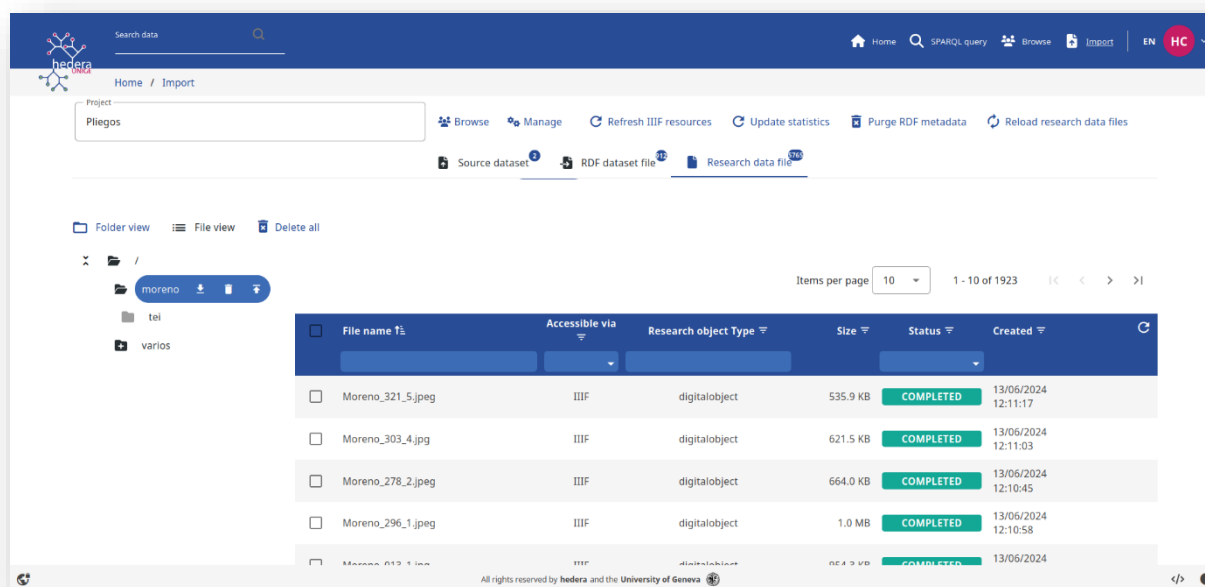
This section corresponds to the project import section, where the researchers organize, upload and manage project data by type: source data or datasets, RDF dataset files and research data files.

4.2.1. The « RDF dataset file » tab of « Pliegos » project



On this panel, which displays excerpts from the Pliegos project (Carta, 2024), the researcher manages the import of the source datasets. He/she can check the status of each imported dataset, verify that data conversion has been executed correctly, and check that RDF result metadata is stored in the triple store.

4.2.2. The « Research data file » tab of « Pliegos » project



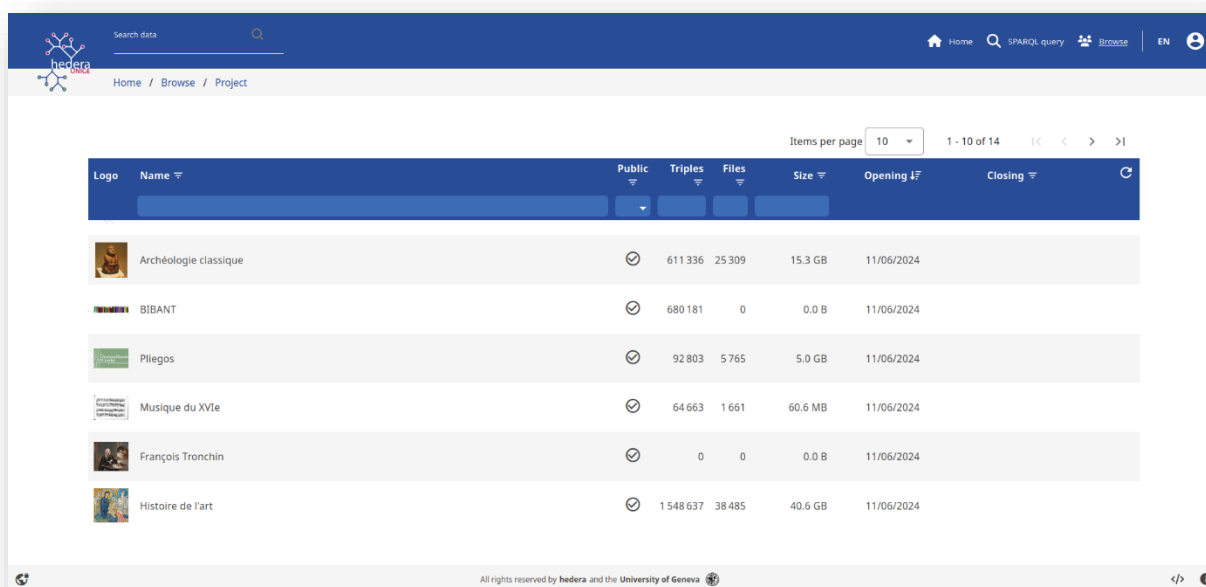
The research data file panel lets the user manage the data files if they are correctly uploaded and stored in the right folder structure. There is also a preview panel to verify that the content is the one expected and to view associated metadata.

4.3. The « Project Browse » section

The « project browse » section enables users to access, search or browse their projects. For public projects, access is open to the public. For non-public projects, the section is secured by a login, and users only have access to projects for which they have a specific role.

In this part of the user portal, users can browse projects and their different types of data: research objects, research files, IIIF manifests or IIIF collections. The following screenshots are just a few examples.

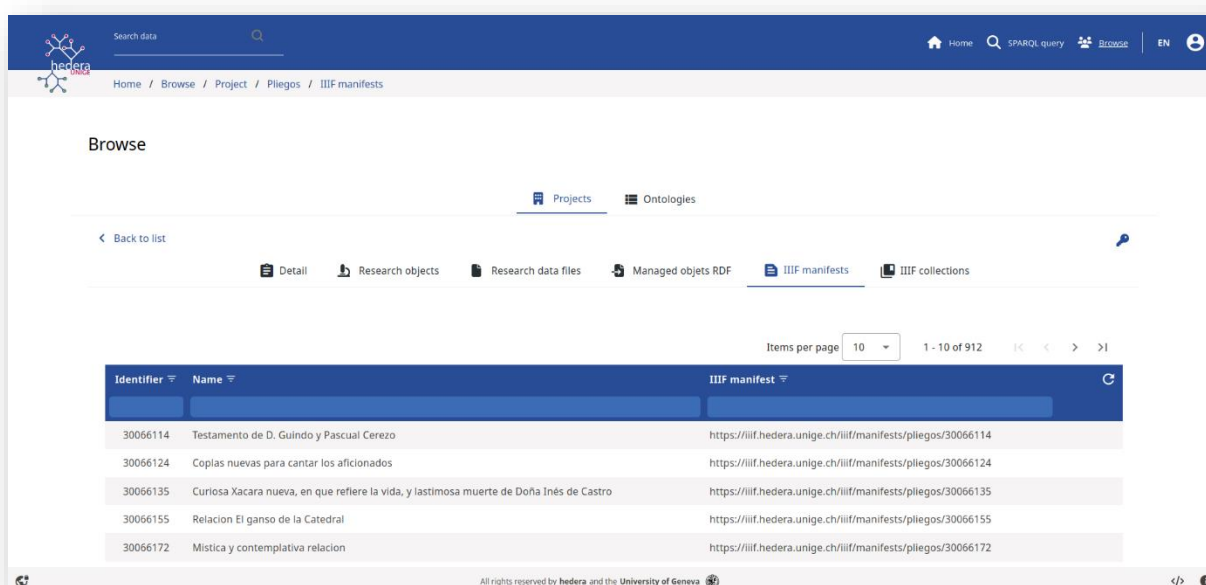
4.3.1. The « Project List » section



Logo	Name	Public	Triples	Files	Size	Opening	Closing
	Archéologie classique	✓	611 336	25 309	15.3 GB	11/06/2024	
	BIBANT	✓	680 181	0	0.0 B	11/06/2024	
	Pliegos	✓	92 803	5 765	5.0 GB	11/06/2024	
	Musique du XVIIe	✓	64 663	1 661	60.6 MB	11/06/2024	
	François Tronchin	✓	0	0	0.0 B	11/06/2024	
	Histoire de l'art	✓	1 548 637	38 485	40.6 GB	11/06/2024	

Based on user permissions, users can list and browse their authorized projects and can see the project statistics: research data files and RDF triples' number, the total volume of the data and the opening/closing dates.

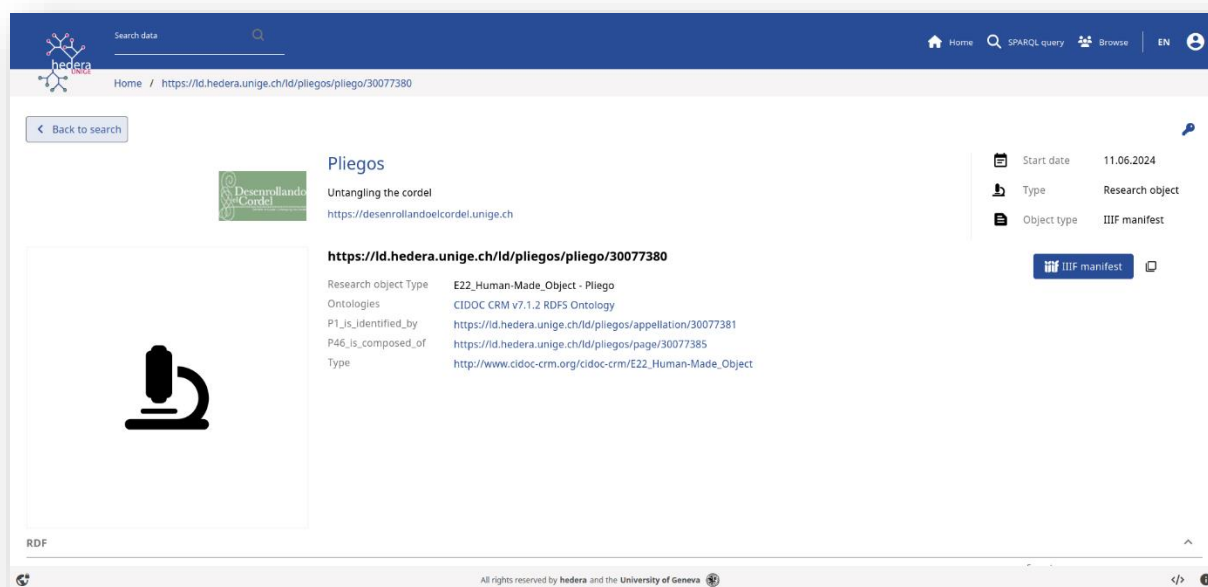
4.3.2. The « Project Details » section



Identifier	Name	IIF manifest
30066114	Testamento de D. Guindo y Pascual Cerezo	https://iif.hedera.unige.ch/iif/manifests/pliegos/30066114
30066124	Coplas nuevas para cantar los aficionados	https://iif.hedera.unige.ch/iif/manifests/pliegos/30066124
30066135	Curiosa Xacara nueva, en que refiere la vida, y lastimosa muerte de Doña Inés de Castro	https://iif.hedera.unige.ch/iif/manifests/pliegos/30066135
30066155	Relacion El gancho de la Catedral	https://iif.hedera.unige.ch/iif/manifests/pliegos/30066155
30066172	Mistica y contemplativa relacion	https://iif.hedera.unige.ch/iif/manifests/pliegos/30066172

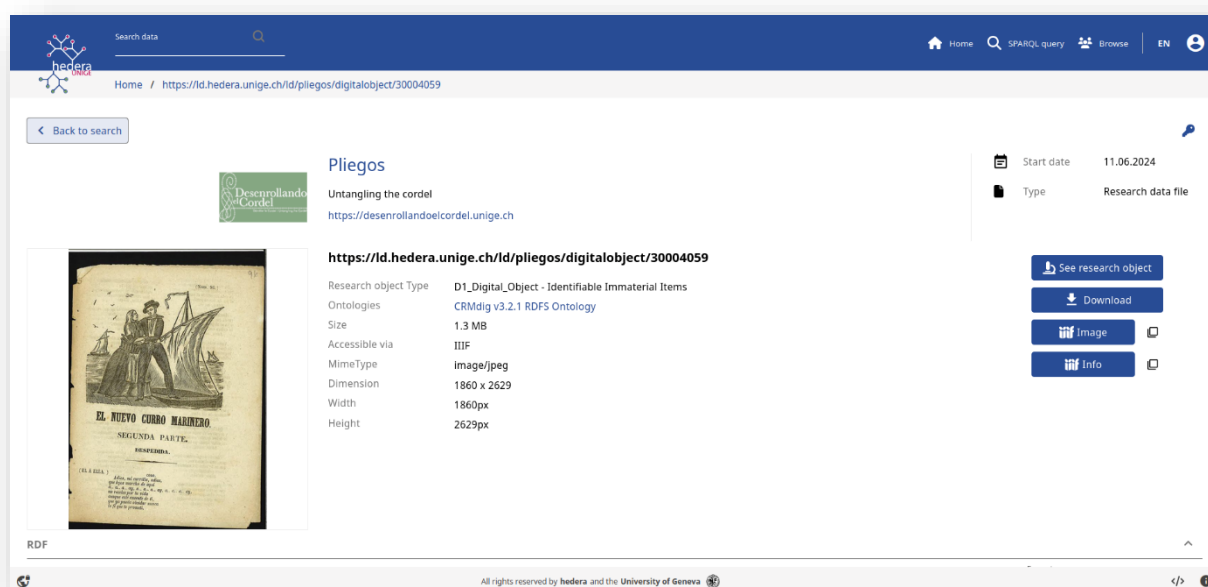
The panel has several tabs. These tabs are organized by project and list all related information: project details, research objects, research data files, managed RDF objects, IIF manifests and IIF collections.

4.3.3. The « Project IIIF Manifest » section



The IIIF manifest detail panel is an important screen to share the project data through the IIIF protocol. There is also the direct link of the manifest to open it with an IIIF-compatible viewer, like Mirador or Universal Viewer.

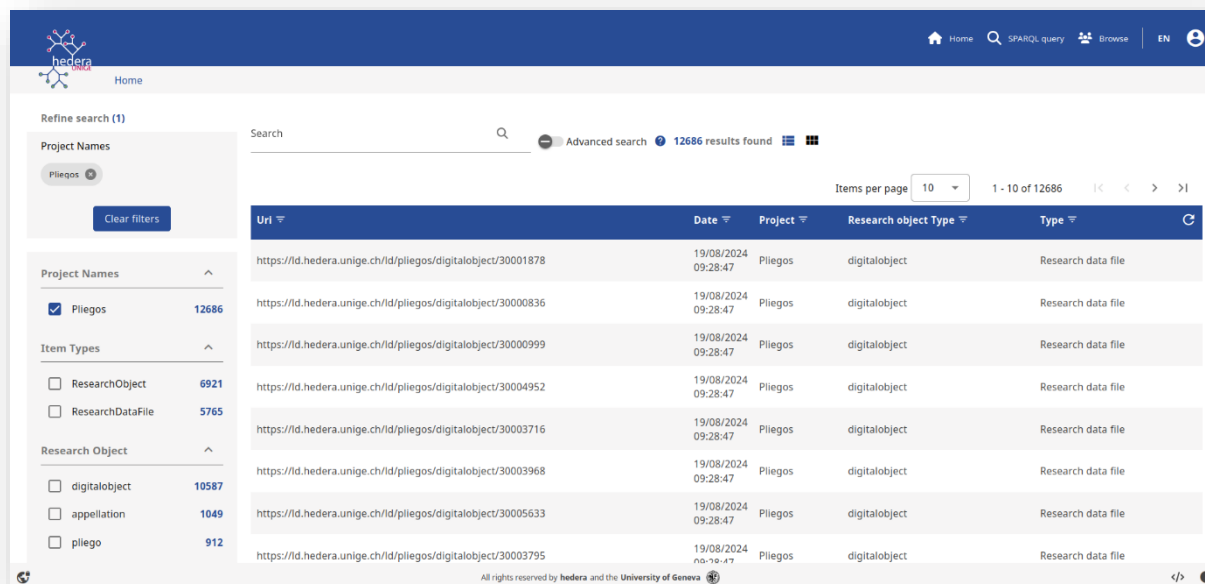
4.3.4. The « Project research file detail » section



The “research data file detail” lists the file’s metadata. A preview is available if the format is supported, which is generally the case for images. Direct links are also provided for downloading, viewing with IIIF, sharing metadata with IIIF and direct access to the details of the linked research object.

4.3.5. The « Project search » section

The « hedera » portal features a section for searching for data within a given project, with category filtering and keyword search capabilities. This search is intuitive for users with no knowledge of RDF or SPARQL.



5. Conclusion

The « hedera » platform represents an innovative solution for dynamic data management in digital humanities and other research fields. By adopting open standards such as RDF, SPARQL, and IIIF, it ensures data interoperability, semantic enrichment, and accessibility. Its focus on adherence to the FAIR principles, combined with its ability to integrate with existing tools such as Yareta and swisscovery, makes it a robust and future-proof platform.

Thanks to its structured workflow—from project setup and data collection to processing and access—« hedera » provides researchers with the tools to efficiently manage, analyze, and share their data. By facilitating the creation of knowledge graphs and fostering collaboration, « hedera » is positioned as an essential resource for advancing open science and supporting interdisciplinary research.

Bibliography

Appleby, M., Crane, T., Sanderson, R., Stroop, J., & Warner, S. (2020). IIIF presentation API 3.0. In *International image interoperability framework* (Version 3.0.0). <https://iiif.io/api/presentation/3.0/>

Bekiari, C., Bruseker, G., Canning, E., Doerr, M., Michon, P., Ore, C.-E., Stead, S., & Velios, A. (Eds.). (2024). *Definition of the CIDOC conceptual reference model* (Version 7.1.3). CIDOC CRM Special Interest Group. https://cidoc-crm.org/sites/default/files/cidoc_crm_version_7.1.3.pdf

Carta, C. (Ed.). (2024). *Démêler le cordel (2020-2024)*. <https://desenrollandoelcordel.unige.ch>

Clavaud, F., & Wildi, T. (2021). ICA Records in Contexts-Ontology (RiC-O): A Semantic Framework for Describing Archival Resources. In *Linked Archives 2021: Proceedings of Linked Archives International Workshop 2021 co-located with 25th International Conference on Theory and Practice of Digital Libraries (TPDL 2021)* (pp. 79–92). <https://enc.hal.science/hal-03965776>

Cyganiak, R., Wood, D., & Lanthaler, M. (Eds.). (2014). RDF 1.1 Concepts and Abstract Syntax. In *W3C Recommendation*. <https://www.w3.org/TR/rdf11-concepts/>

De Meester, B., Heyvaert, P., & Delva, T. (2024). *RDF Mapping Language (RML): Unofficial Draft* (A. Dimou & M. Vander Sande, Eds.). <https://rml.io/specs/rml/>

Harris, S., & Seaborne, A. (2013). SPARQL 1.1 Query Language. In *W3C Recommendation*. <https://www.w3.org/TR/sparql11-query/>

International Council on Archives. (2000). *ISAD(G): General international standard archival description* (2nd ed.). https://www.ica.org/app/uploads/2024/01/CBPS_2000_Guidelines_ISADG_Second-edition_EN.pdf

Open Archives Initiative. (2015). *The Open Archives Initiative Protocol for Metadata Harvesting*. <https://www.openarchives.org/OAI/openarchivesprotocol.html>

Text Encoding Initiative Consortium. (2025). *TEI P5: Guidelines for Electronic Text Encoding and Interchange* (Version 4.9.0). Text Encoding Initiative Consortium. <https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf>