

Supporting executable scientific workflows in a clustered Infrastructure: DARIAH-IT and H2IOSC

Eseguire workflow scientifici con il supporto di un cluster di infrastrutture: DARIAH-IT e H2IOSC

Emiliano Degl’Innocenti,¹ Francesco Pinna,¹ Alessia Spadi,¹
Federica Spinelli¹

¹Opera del Vocabolario Italiano, Consiglio Nazionale delle Ricerche,
Florence, Italy

Abstract in English

Workflows have become essential in digital humanities, enabling the formalisation, automation, and reproducibility of complex research processes. As humanities increasingly adopt data-driven methodologies, workflows offer structured approaches to manage diverse data and tools while supporting transparency and collaboration. Recognising this need, DARIAH-IT has advanced research infrastructure development by focusing on workflow-based services within the H2IOSC (Humanities and Cultural Heritage Italian Open Science Cloud) project. DARIAH-IT leads the design and implementation of a national cloud system to support digital humanities research, ensuring interoperability, FAIR data practices, and semantic integration across disciplines. Central to this effort of DARIAH-IT within H2IOSC is AEON (dAriah sErvice Oriented iNfrastructure), a platform that enables the creation, execution, and management of scientific workflows. AEON integrates service provisioning, semantic validation, and runtime orchestration, supporting reproducible research and collaborative practices.

Keywords in English: Digital Humanities, Scientific Workflows, Research Infrastructures, FAIR Data Principles, Semantic Interoperability

Abstract in Italian

I workflow sono diventati sempre più rilevanti nell’ambito Digital Humanities, poiché permettono di formalizzare, automatizzare e rendere riproducibili processi di ricerca complessi. Con la crescente adozione di metodologie basate sui dati nelle scienze umane e sociali, i workflow offrono approcci strutturati per gestire dati e strumenti eterogenei, promuovendo al contempo trasparenza e collaborazione. Riconoscendo questa esigenza, DARIAH-IT ha contribuito allo sviluppo di infrastrutture di ricerca focalizzandosi su servizi basati su workflow all’interno del progetto H2IOSC (Humanities and cultural Heritage Italian Open Science Cloud). In questo contesto, DARIAH-IT ha come obiettivo l’implementazione di un sistema cloud nazionale a supporto della ricerca nelle Digital Humanities, garantendo interoperabilità, pratiche FAIR per i dati e integrazione semantica tra le risorse. Al centro di questo impegno di DARIAH-IT nell’ambito di H2IOSC si colloca AEON (dAriah sErvice Oriented iNfrastructure), una piattaforma che consente la creazione, l’esecuzione e la gestione di workflow scientifici. AEON integra l’erogazione di servizi, la validazione semantica e l’orchestrazione in fase di esecuzione, supportando la riproducibilità della ricerca e le pratiche collaborative.

Keywords in Italian: Digital Humanities, Workflow Scientifici, Infrastrutture di Ricerca, Principi FAIR, Interoperabilità Semantica

Acknowledgements

H2IOSC Project - Humanities and cultural Heritage Italian Open Science Cloud funded by the European Union NextGenerationEU - National Recovery and Resilience Plan (NRRP) - Mission 4 “Education and Research” Component 2 “From research to business” Investment 3.1 “Fund for the realization of an integrated system of research and innovation infrastructures” Action 3.1.1 “Creation of new research infrastructures strengthening of existing ones and their networking for Scientific Excellence under Horizon Europe” - Project code IR0000029 - CUP B63C22000730005. Implementing Entity CNR.

1 Introduction

Access to digital tools and data is now common, along with the opportunities offered by advanced technologies such as big data, artificial intelligence, and large language models. The mature phase of humanities studies, supported by the widespread availability of digital data, has also transitioned from experimental approaches to

standard practice¹. This widespread digitisation has a direct impact on scholarly methods in general, and more particularly on the workflows that are adopted to perform research tasks.

The definition of workflows in digital humanities (DH) follows the development of the discipline from the analogue era to the digital era. Humanists can be said to have already employed workflows, understood as a sequence of steps to produce a research result, in their endeavours; the new aspect is the application of digital tools to these steps. This transition enables scholars to analyse vast amounts of digital data, uncover new patterns and insights, and create innovative forms of scholarly expression (Biemann et al. 2014).

This apparently simple transition² carries the need to reorganise well-established and successful procedures in humanities research, since the rooting of research methods in a consolidated tradition makes the application of digitally enabled workflows complex.

A persistent challenge in the social sciences and humanities (SSH), and cultural heritage (CH) sectors is the fragmentation of data, tools, and practices. The diversity of languages, disciplinary traditions, metadata standards, and technological systems often hinders collaboration, data reuse, and large-scale analysis. This fragmentation not only limits interoperability among resources but also hinders the development of cohesive research environments. In this context, Research Infrastructures (RIs)³ can play a leading role in SSH research to coordinate distributed knowledge systems⁴, promote standardisation, and enable shared access to tools, services, and datasets.

1. Father Roberto Busa also recommended considering information technology not as a mere tool for efficiency but as a catalyst for innovation. By demanding new research strategies and a higher level of human engagement, it propels intellectual advancement beyond the limitations of traditional methods: “Mi preme fare ai giovani una raccomandazione: non mettete vino vecchio in otri nuovi, e tenete conto che l’informatica non è per fare le stesse ricerche di prima con gli stessi metodi di prima ma solo più velocemente e magari con meno lavoro umano. L’informatica obbliga a due cose: primo, all’invenzione di nuove strategie di ricerca, proporzionate alla possibilità di questo strumento; e, secondo, impegna a un lavoro umano più intenso, più condensato, a livelli umani superiori.” See: http://circe.lett.unitn.it/attivita/eventi/pdf_eventi/busa.pdf. Accessed 15 April 2024

2. French philosopher Bruno Latour, characterised in his lectures this digital transition as the “screentoria” (Latour 2014, 2005, 1993), a modern-day equivalent of the scriptorium where mediaeval monks contemplated knowledge in solitude: the scriptorium has evolved into a digital screen, reflecting the shift in its dimensions and distribution.

3. For a definition of *research infrastructure* see European Union 2013, Article 2 (6): “ ‘research infrastructures’ mean facilities, resources and services that are used by the research communities to conduct research and foster innovation in their fields. [...] “

4. Let us refer to a knowledge system as an organised structure of people, practices, data, tools, and institutions involved in producing, managing, and using knowledge. It includes both formal and informal mechanisms for knowledge creation, dissemination, and validation within a domain.

They are uniquely positioned to reduce conceptual and technical silos, foster semantic continuity (i.e., ensuring that different data and tools can refer to and operate on shared or compatible meanings across systems and disciplines), and support interdisciplinary research by translating and aligning diverse resources.

DARIAH-IT, the Italian node of the Digital Research Infrastructure for the Arts and Humanities (DARIAH ERIC), advanced this mission within the H2IOSC (Humanities and Cultural Heritage Italian Open Science Cloud) project.⁵ One of the objectives of DARIAH-IT within H2IOSC⁶ is precisely to bridge the gap between traditionally separate domains, such as hard sciences and humanities (Snow 1961), enabling deeper, interdisciplinary research. This objective is further exemplified by DARIAH-IT’s activities in developing integrated strategies for data management and semantic interoperability in the arts and humanities (Degl’Innocenti, Di Meo, and Spadi 2024). As part of this effort within H2IOSC DARIAH-IT leads the development of integrated solutions that support workflow management, interoperability, and FAIR data practices. To maximise this impact and foster interoperability and large-scale dataset analysis, it is essential to involve all stakeholders, including researchers, interested communities, and beneficiaries.

Building upon this broader perspective, the following section provides an overview of the specific context in which these efforts materialise, namely AEON (dAriah sErvice Oriented iNfrastructure), a cloud-based infrastructure developed within DARIAH-IT under the H2IOSC project. AEON’s primary goal is to support the execution and management of complex scientific workflows, enhancing interoperability and reproducibility in digital humanities research, while also exposing the full catalogue of DARIAH-IT services.

2 Context

H2IOSC is a project funded by the National Recovery and Resilience Plan (NRRP) in Italy (Presidenza del Consiglio dei Ministri 2021), aiming at creating a federated cluster comprising the Italian branches of four European research infrastructures (RIs) - CLARIN, DARIAH, E-RIHS, OPERAS - operating in the Social Sciences and Humanities sector of ESFRI (European Strategy Forum for Research Infrastructures). This initiative aligns with broader European strategies promoting data-driven research and fostering interdisciplinary collaboration (ESFRI 2020; Wilkinson et al. 2016).

5. <https://www.h2iosc.cnr.it/>

6. It’s one of the current main projects participated in by DARIAH-IT. At the beginning of Sec. 2, a detailed presentation is provided.

DARIAH-IT is in charge of the H2IOSC’s cloud system creation, leading the architecture design, coordinating the software and hardware implementation and elaborating the information organisation and representation framework within the cloud environment.

This involves setting up the physical infrastructure, which includes designing and building eight data centres across different locations and then connecting them, as well as developing a shared semantic framework to manage the project knowledge. This framework is basically a set of agreed-upon terms, relationships, and vocabularies ensuring that all the data gathered by the four research infrastructures (RIs) in H2IOSC can be understood and used effectively for research purposes. The federated cloud is needed to implement the primary objective of H2IOSC: enabling data-driven research activities and supporting the description and execution of scientific workflows, which are implemented through Scientific Pilots. Within the H2IOSC project, “Scientific Pilots” refers to structured and replicable research experiments designed to demonstrate the practical utility and robustness of workflows in addressing specific scholarly tasks. These pilots showcase the technological infrastructure capabilities in concrete scenarios, typically involving collaboration among researchers, data curators, and developers, and serve as practical benchmarks to validate the interoperability, reproducibility, and scalability of digital tools and workflows. In particular, DARIAH-IT will implement Scientific Pilots supporting the execution of a digital philology workflow. In order to achieve this, DARIAH-IT will undertake a set of preparatory activities, including the collection and evaluation of existing tools, datasets, and services that are relevant to the above task. A significant aspect of this process is the semanticisation of selected data and metadata to promote interoperability among heterogeneous resources (data coming from archives, libraries, museums and/or produced by researchers). This process of resource alignment involves data cleaning, mapping and modelling, as well as the definition of standardised vocabularies and ontologies, in synergy with the deployment of tools and workflows that are specifically designed to implement the semantic transformation. The Pilots, presented as platforms or hubs, integrate domain-specific services, workflows, and interfaces and are conceived as executable scientific workflows, combining different resources in specific computational chains.

For the implementation of the executable workflows supporting the Scientific Pilots, DARIAH-IT implemented

- a set of core elements to manage the interaction of the selected services (i.e., the API manager);
- a service-provision-oriented infrastructure that enables the actual execution of the services in a specific runtime environment (i.e., the AEON),

By implementing this environment, DARIAH-IT aims to provide an innovative, sustainable, and resilient research ecosystem for SSH research.

3 Workflows for digital humanities

In academic and industry settings, workflows represent organised sequences of tasks or operations. Within scientific research, the concept of scientific workflows specifically refers to formally defined computational processes that enable automation, reproducibility, and systematic execution of research steps (Atkinson et al. 2017). Thus, while “workflow” is a general concept, “scientific workflow” explicitly denotes computationally executable processes. By breaking complex research tasks into smaller, manageable steps, scientific workflows automate repetitive activities, enhance data management efficiency, and foster reproducibility, transparency, and replicability, all of which are particularly critical in data-intensive research (Gil et al. 2008; National Academies of Sciences, Engineering, and Medicine 2019; Concordia, Meghini, and Benedetti 2020).

To address the emerging needs related to the implementation of complex workflows, platforms have been developed across disciplinary domains. Among others, Galaxy (Giardine et al. 2005) has been successfully used in bioinformatics, while Kepler (Altintas et al. 2004; Ludäscher et al. 2009) is a relevant attempt known for its generic applicability across scientific disciplines. Within the SSH, relevant platforms include the CLARIAH Media Suite (Melgar-Estrada et al. 2019) and the SSH Open Marketplace (Barbot et al. 2020), a discovery portal that gathers and contextualises resources for the SSH research communities. The SSH Open Marketplace was developed within the SSHOC (Social Sciences and Humanities Open Cloud) project (SSHOC 2022), a European initiative aimed at building a cloud-based infrastructure for SSH research by integrating services, tools, and training resources. The primary objective of the SSH Open Marketplace is to establish a collaborative space where users can access and share digital tools and resources, fostering greater transparency and cooperation in research.⁷ Today, the SSH Open Marketplace drives digital transformation within the social sciences and humanities by providing access to a wide range of tools, data, services, and resources tailored to researchers and scholars in these fields, it also responds to the growing need for dedicated workflow management tools.

The Guidelines of the SSH Open Marketplace describe a research workflow as a sequence of steps that can be performed on research data throughout its lifecycle.

7. <https://marketplace.sshopencloud.eu/>

Workflows can be executed using a variety of tools, methods, and useful resources connected to each step.⁸

Scientific workflows are employed by researchers as a means of defining automated, scalable, and portable experiments: “A scientific workflow is a composition of interconnected and possibly heterogeneous scripts that are used in a scientific experiment” (Concordia, Meghini, and Benedetti 2020). It suffices to refer to WfCommons⁹: a tool that serves as a comprehensive framework aimed at advancing research and development in scientific workflows. It offers tools, datasets, and infrastructure to support the creation, simulation, and comparison of scientific workflow instances (Coleman et al. 2022).

These workflows streamline the research process by providing a structured approach to data management and analysis, allowing researchers to focus on their core scientific questions.

The formal description of an experiment as a workflow can improve the replicability and reproducibility of experiments. Replicability concerns the consistency of results obtained using the same data, computational steps, methods, code and analysis conditions. Reproducibility, on the other hand, concerns the consistency of results between different studies that attempt to answer the same scientific question. Reproducibility requires the use of original data and codes, while replication requires the collection of new data and the use of similar methods. Publishing datasets together with scripts or workflows is common practice, although it may not be sufficient to ensure the reusability of data and reproducibility.¹⁰ Workflows become increasingly complex while advancing research endeavours.

Considering the inputs provided by the literature and using the SSH Open Marketplace as a benchmark, the AEON platform design phase was informed by existing knowledge and best practices. When defining the runtime environment for executing scientific workflows, it became necessary to identify the different levels of complexity that typically characterise such workflows. The identified levels are:

- Unitary Workflow: simple tasks requiring a single action.

8. <https://marketplace.sshopencloud.eu/about/service>

9. Developed as an open-source platform, WfCommons addresses the complexities involved in running intricate workflows on distributed computing environments, such as cloud and high-performance computing (HPC) systems. For researchers in STEM fields focused on workflow management, WfCommons provides a stable foundation for testing, refining, and benchmarking workflow management solutions across a wide array of scientific applications. For more details on WfCommons see the official website <https://wfcommons.org/>

10. For a comprehensive reflection on the discussion around replicability and reproducibility of experiments, see <https://nap.nationalacademies.org/read/25303/chapter/6>

- Generic Workflow: general data management operations.
- Complex Workflow: multiple steps with specific requirements.
- Domain Workflow: tailored to a specific research domain.

Unitary workflows are the most basic type, representing simple tasks or processes that can be completed in a single step: examples include data ingestion, cleaning, or basic analysis. Generic Workflows (as opposed to domain-based workflows) are more comprehensive, outlining common data management operations such as data ingestion, transformation, and analysis. These workflows can be applied to various research projects, providing a foundational framework. Complex Workflows involve multiple steps, each requiring specific competencies, resources, and tools to be executed effectively. These workflows are often tailored to address complex research questions or projects. Domain Workflows (as opposed to Generic workflows) are highly specialised, focusing on a specific research domain (e.g, philology, arts, philosophy, etc.). They are designed to address specific needs or challenges within that domain, and they may combine multiple complex workflows to achieve the desired outcomes.

When a workflow (unitary, complex, generic or domain workflow) is completely automated (i.e., not requiring human intervention to be completed) it is called a Pipeline. In the field of Information Technology (IT), the concept of *pipelines* has become a fundamental tool for managing automated workflows. Pipelines refer to a series of interconnected processes through which data flows, transforming and refining information in a systematic and gradual manner. This model has been extensively developed, documented and refined over the years, enabling greater efficiency, accuracy and scalability in managing complex operations.

It therefore seems natural to draw inspiration from the IT world to introduce the concept of a pipeline in the SSH scientific domain, in order to have an instance of tools that can streamline workflows in research and data analysis. Moreover, as SSH increasingly involve the use of digital tools to process large amounts of textual, visual and multimedia data, pipelines can help automate repetitive tasks, improve data processing capabilities and foster collaborative research. By adopting pipeline strategies from IT, humanities scholars can benefit from established automation methods, reducing manual effort while ensuring the accuracy of tasks such as text analysis, metadata extraction and digital archiving.

The following section presents how DARIAH-IT concretely implemented these principles in the AEON platform through a systematic workflow development process.

4 Workflow implementation

In this context, DARIAH-IT worked towards the definition of performing workflows in the H2IOSC project. Within this framework, DARIAH-IT wants to provide its users with a complete system to create and manage workflows by developing AEON to upgrade its current service provision capabilities to match the needs of the national digital humanities research community.¹¹

A systematic approach was applied to the development of a robust workflow that also encompasses servification, virtualisation, and remotisation¹² in research infrastructures. To achieve this the following key stages were identified: i) assessment of existing tools and services to identify their suitability for transformation; ii) design and development of standardized interfaces and protocols for interoperability and integration; iii) implementation of virtualization and cloud-based platforms to provide scalable and accessible services; iv) development of user-friendly interfaces and workflows to facilitate seamless interaction for researchers; v) rigorous testing and quality assurance to ensure the reliability and performance of the services; and vi) continuous monitoring and evaluation to identify areas for improvement and adaptation to evolving needs. By implementing these key features, the research infrastructures involved in the Project can effectively transition to a more service-oriented and accessible model, enhancing collaboration and innovation within the research community (i.e. scholars, researchers, and practitioners in the SSH and CH domains) who interact with and benefit from shared resources and workflows.

DARIAH-IT will make the services findable through the H2IOSC Marketplace (and the cooperating projects) and will offer actual service provision via the AEON platform. In this Section, the set of activities undertaken to achieve this goal is presented, which include the design and implementation of the AEON platform, as well as the evaluation and refactoring of existing services. To strengthen the infrastructure, DARIAH-IT is also concerned about the development of specific policies and guidelines to ensure effective interoperability and security for existing and newly created resources, which is becoming increasingly important due to recent cyber attacks on major cultural heritage institutions.¹³

11. The AEON platform has not been made available for the general user. To know more about the service, visit DARIAH-IT and H2IOSC websites.

12. Servification refers to the transformation of resources into standardized, reusable services accessible via APIs; virtualisation involves abstracting physical resources into virtual environments for flexible allocation; remotisation enables remote access and management of resources through network-based platforms.

13. An illustrative example is the recent cyber attack on the British Library, which raised concerns about the resilience of cultural heritage institutions. See: *Financial Times*, *Cyber attack on British*

The first step to implement this vision is to define users and the actions they can undertake within the system.

AEON supports four user roles, each with clearly defined permissions and responsibilities summarised in Tab. 1. Workflow creation and management permissions begin with the Basic Users, who can create and manage personal workflows. They can also request that a personal workflow be added to the catalogue of exposed services, which will happen after validation and authorisation by administrator-level users. Administrators oversee all aspects, including the workflows created by other users, ensuring system security, performance, and proper configuration.

Typical usage scenarios depend on the user’s role. A researcher, typically a Basic User, searches the catalogue for relevant services, creates a workflow, executes it on specific data, and adjusts it as necessary. A developer, usually a Contributor, is responsible for developing new services, testing and documenting them, submitting them for approval, and providing user support once the services are included in the catalogue. A museum curator, also a Basic User, explores applications suitable for creating virtual exhibitions, customises them with appropriate content, tests the result, and publishes it for public access. The Administrator monitors system performance, manages user access, reviews security logs, applies software updates, and oversees integration with external systems.

Once a user’s role is defined, the workflow creation process can begin with a clear description of its scope and intended outcomes. Administrators have access to a dedicated functionality called the workflow manager, which enables them to create new workflows or modify existing ones.

If the purpose is to publish a descriptive workflow, a narrative account is sufficient, and the administrator may choose to save and publish the workflow on the platform. In such cases, the inclusion of services is not required.

Conversely, to create an executable workflow, at least one service must be selected. When only one service is involved, it must be associated with a graphical user interface (GUI). If the service does not include a native GUI, the workflow manager will provide a standard one, generated based on the service’s manifest (see Sec.4.1), which specifies the expected types of computational and semantic inputs and outputs. This process, which enables the creation of so-called atomic workflows, constitutes the standard method for integrating individual services into the AEON platform.

Library raises concerns over lack of UK resilience. Accessed 15 November 2024, <https://www.ft.com/content/642ee014-4768-4c65-b1ee-0d4f39a8a63d>. A comprehensive summary is also provided in Wikipedia contributors. *British Library cyberattack*. Wikipedia, The Free Encyclopedia. Accessed November 15, 2024, https://en.wikipedia.org/w/index.php?title=British_Library_cyberattack&oldid=1255404045.

Role	Permissions	Functions
Basic User	<ul style="list-style-type: none"> • Access to the catalogue and public services • Create, publish, and share personal workflows 	<ul style="list-style-type: none"> • Search and execute services and applications • Create and manage workflows • Publish and share workflows (under approval) • Participate in the community
Contributor	<ul style="list-style-type: none"> • All Basic User permissions • Submit new services or applications to the catalogue (subject to approval) • Update descriptions of existing services 	<ul style="list-style-type: none"> • Submit new services • Update documentation • Provide technical support for contributed services
Curator	<ul style="list-style-type: none"> • All Contributor permissions • Review and approve new services or applications • Manage categories and tags in the catalogue 	<ul style="list-style-type: none"> • Review and approve submissions • Organise catalogue content • Generate reports on service usage
Administrator	<ul style="list-style-type: none"> • Full system access • Manage users and roles • Configure system settings • Monitor performance and security 	<ul style="list-style-type: none"> • Oversee system operations • User management and role assignment • System security and performance monitoring • Log analysis and software updates

Table 1

For complex workflows that involve the automatic concatenation of multiple services, the administrator uses a dedicated functionality within the workflow manager, which we call the Composer. The following subsection details the Composer’s functionality.

4.1 Composer

The composer essentially performs three compatibility checks and generates the script that automates the process:

1. *Compatibility check between the input and output of the services in their sequence (I/O API)*
That is, for example, verifying that an API that receives as input an XML file can only be concatenated downstream of an API that produces an XML file.
2. *Syntactic compatibility check (metadata structure)*
That is, for example, verifying that an API that receives as input a JSON file responding to a certain schema can only be concatenated downstream of an API that produces a JSON file responding to the same schema.
3. *Semantic compatibility check*
Assuming that the structure of the metadata passed between one API and another is the same, it is also necessary that they are semantically compatible.

Once the previous checks have been passed, a script is generated to automate the process. The workflow is therefore articulated in:

- (a) **input**: the inputs necessary to activate the service are passed to the script. If there is only one service, the execution is therefore complete; otherwise, this operation is repeated for each service with the appropriate inputs.
- (b) **output**: presentation of the output of the script or the management of any error messages.

After saving any type of workflow (descriptive or executable), to proceed with the publication, the workflow manager manages and supports the administrator in compiling the JSON manifest¹⁴ to be associated with the workflow. A *manifest.json* file is a structured JSON document that describes the properties and configuration of a service within the AEON platform. It includes metadata such as the service name,

14. <https://developer.mozilla.org/en-US/docs/Mozilla/Add-ons/WebExtensions/manifest.json>

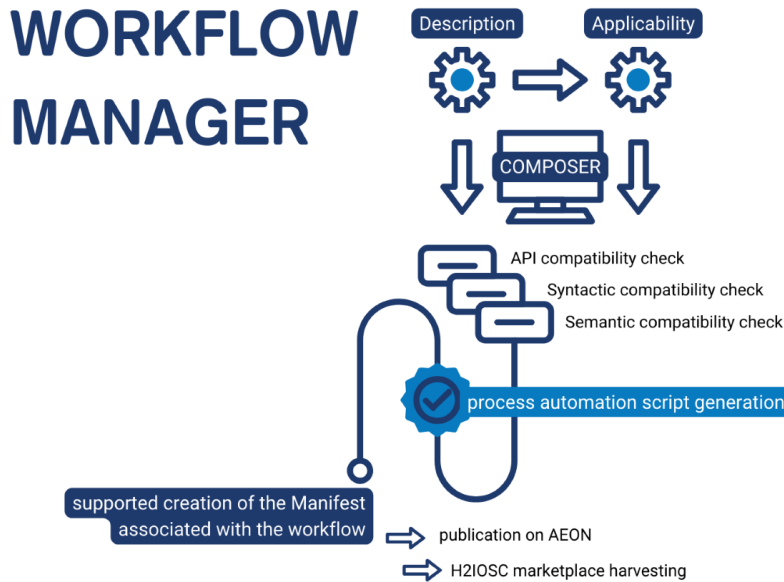


Figure 1: workflow manager architecture

version, input and output formats, and supported operations. This file enables the platform to interpret how a service should be executed, what data it requires, and how results should be presented. Additionally, it provides semantic annotations to ensure compatibility and interoperability across services. While the concept of a *manifest.json* file originates from browser extensions using WebExtension APIs, its adaptation in AEON follows a common practice in web services: using structured metadata files to describe how services behave, what inputs they require, and how outputs are managed. This approach ensures consistency, automation, and interoperability across different components of the platform.¹⁵

Based on this file, the administrator in the workflow manager determines the GUI. To give a simple example, if the service is a philological collation support tool, the inputs are the different witnesses, so in the GUI the input part will be represented by the upload of the resources in TXT, XML, or similar file formats. The output part would be the collation table (downloadable), which may be provided in the form of XML files, CSV, or similar formats. In the context of the AEON platform design,

¹⁵. AEON relies on a predefined schema for these manifest files to ensure consistency in service description. The guidelines adopted in defining possible schemas are those identified within H2IOSC WP6, task 6.1, and as such are common to all federation services.

a selection of heterogeneous data originating from multiple sources and described through diverse metadata models was used as a use case.¹⁶ The objective was to enable the possibility to collect, manage and integrate data coming from different contexts and described according to different models, populating a SPARQL¹⁷ database with a Linked Open Data (LoD) to be used for various research purposes (from conceptual browsing to interdisciplinary research, such as considering the concept of manuscript as defined and used in different research contexts and analysing the related resources).

More specifically, by supporting this use case, the DARIAH-IT team had the opportunity to work with data from archives, museums and research institutes, with the goal of creating a complete workflow for the management and integration of archival resources, belonging to different domains, with different kinds of resources belonging to other actors of the GLAM landscape.

To enable effective reuse, the data first needed to be parsed and cleaned to isolate the most relevant information. This information was then transformed—typically from XML or JSON formats—into a structured form such as CSV. The cleaned and structured data were subsequently converted into RDF triples and stored in a SPARQL-compliant database. This final step allows the data to be queried easily using semantic search techniques, making it accessible and usable for a wide range of research purposes. Schematically:

```
### service a
INPUT = XML or JSON; --> (parsed, selection of only the relevant info
by a manifest YAML file);
OUTPUT = CSV tables with desired info
```

```
### service b
INPUT = CSV tables; --> (selection of only the relevant columns by a
manifest YAML file); triplification (file TTL) of relevant info
following a schema given in another YAML file
```

```
### service c
INPUT = file TTL loaded into Virtuoso and/or graphDB - SPARQL
```

16. The dataset used in this example was produced within the project “smaRt accESs TO digital heRitage and memory” (RESTORE), coordinated by the Istituto Opera del Vocabolario Italiano of the CNR (Florence): RESTORE project, Higher Education intervention program (CNR4C), co-financed by the Tuscany Region CUP B15J19001040004, see <http://restore.oivi.cnr.it/>

17. SPARQL, for SPARQL Protocol and RDF Query Language, is a semantic query language for databases, able to retrieve and manipulate data stored in Resource Description Framework (RDF) format. For more details, see <https://www.w3.org/TR/sparql11-query/>

accessible

This workflow became a pipeline that was successfully used not only for importing and managing the data from the Archivio di Stato di Prato, but also for other XML data representing information from the Museo del Palazzo Pretorio di Prato, responding to the needs expressed by the project’s stakeholders, namely to simplify, speed up and supporting the data clean-up process across the entire data injection chain. To illustrate the practical application of the described pipeline, the following section presents a concrete example demonstrating how these services can be orchestrated in a real-world scenario.

4.2 Examples

This section provides a basic overview of what the previously presented pipeline may look like using Python with Apache Airflow¹⁸ tasks. In the following paragraph, the code snippet shows a simplified version of the process. Each step corresponds to one of the described services.

Key Points:

1. *Parsing XML/JSON and Extracting Data (service_a)*

This service reads XML or JSON, selects the relevant information based on a manifest defined in YAML,¹⁹ and outputs the information into CSV format.

2. *CSV to TTL Conversion (service_b)*

It processes the CSV to select relevant columns and converts the data into RDF triples (TTL format) according to a schema defined in a YAML manifest.

3. *Load TTL into SPARQL (service_c)*

The TTL file is then loaded into a SPARQL endpoint (Virtuoso or GraphDB).²⁰ In this example, the placeholders for connecting and loading data into the SPARQL database are visible to track the steps, but typically a library like ‘SPARQLWrapper’²¹ is used to run queries against the endpoint.

18. Apache Airflow is a Python-based platform for programmatically authoring, scheduling, and monitoring workflows. It is used for orchestrating and managing complex data pipelines and workflows.

19. YAML (YAML Ain’t Markup Language) is a human-readable data serialisation format used for configuration files and data exchange.

20. Virtuoso and GraphDB are open-source graph database management systems used for storing and querying highly interconnected data, such as in semantic web and linked data applications.

21. SPARQLWrapper is a Python library that abstracts the process of querying a SPARQL endpoint and processing the returned results.

See the relevant code snippets in Sec. A.²²

While the presented pipeline is implemented using Apache Airflow and Python, it is important to note that the same result could be achieved using other technologies, such as a Jupyter notebook,²³ or even entirely different frameworks and tools. Jupyter notebooks, for instance, offer a more interactive approach, allowing researchers to experiment with and visualise the data at each stage of the pipeline. This can be especially useful for smaller projects or in educational contexts where step-by-step exploration is key.

By automating tasks such as parsing, filtering, triplifying, and loading data, we significantly reduced manual intervention, minimised errors, and created a cleaner, more efficient process. Ultimately, the goal is to ensure that valuable data is made accessible and usable promptly for any scholarly or research need—whether through Airflow, Jupyter, or another toolset entirely. With these technical achievements in place, it is important to consider their broader implications for the evolving landscape of European research infrastructures.

5 Conclusions

The continuous advancement of digital tools and methods in the SSH has reshaped scholarly practices, demanding new infrastructures capable of supporting data-intensive, collaborative, and reproducible research. Within this evolving landscape, workflows have emerged as a fundamental mechanism for structuring and automating research processes, enabling scholars to engage meaningfully with complex research scenarios. Starting from the emergence of digital humanities as a data-intensive field, the discussion has outlined how workflows have become increasingly essential to formalising, automating, and reproducing complex research processes. Particular attention was paid to persistent challenges within the SSH sectors, namely fragmentation, heterogeneity of standards, and limited interoperability, and to the critical role of research infrastructures in addressing these limitations. Within this context, DARIAH-IT’s work in the H2IOSC project exemplifies a strategic effort to design and implement service-based, semantically enriched, and FAIR-compliant environments for digital research.

22. For a more complete example project, see TransfDAGExample on GitHub repository. Accessed 20 June 2025; example code for transfer learning on Directed Acyclic Graphs at <https://github.com/fpinnahub/TransfDAGExample>.

23. Jupyter notebook is an open-source browser application that allows users to create and share documents that contain live code, visualisations, and narrative text (such as explanations and descriptions). Official repository: <https://github.com/jupyter/notebook>

At the core of this effort stands the implementation of a service-oriented infrastructure designed to support the execution and management of scientific workflows across SSH domains. This infrastructure integrates resource discovery, semantic validation, service provisioning, and workflow orchestration into a unified system that improves the usability, transparency, and scalability of digital research practices. By offering a framework that enforces both syntactic and semantic compatibility across services, AEON fosters reproducible and reusable research. The implementation of practical pipelines, such as those for processing archival and museum data, demonstrates the feasibility of these principles and underscores the potential of structured workflows to bridge institutional and disciplinary boundaries.

The contribution outlined DARIAH-IT's strategy to advance workflows from a conceptual ideal to an operational standard in the digital humanities. This represents a significant step forward in addressing the methodological and infrastructural needs of the humanities in the digital age. In particular, the approach adopted by DARIAH-IT emphasises not only technical interoperability, but also semantic coherence and community alignment, establishing a foundation for long-term sustainability and scalability.

Future developments will focus on enhancing the maturity of the AEON infrastructure, extending its integration with federated marketplaces such as H2IOSC marketplace and SSH Open Marketplace, and supporting novel research scenarios involving AI-assisted and cross-disciplinary workflows. Additional attention will be directed toward implementing version control, data provenance tracking, and quality assurance mechanisms within workflow execution environments.

A Code snippets examples

Simplified code snippet example for Python Apache Airflow usage to create a DAG for workflow automation.

```
from airflow import DAG
from airflow.operators.python_operator import PythonOperator
from datetime import datetime
import pandas as pd
import xml.etree.ElementTree as ET
import json
import yaml
from rdflib import Graph, Literal, RDF, URIRef
```

```

# Define the DAG
default_args = {
    'owner': 'airflow',
    'depends_on_past': False,
    'start_date': datetime(2023, 1, 1),
    'retries': 1,
}

dag = DAG(
    'dh_pipeline',
    default_args=default_args,
    description='DH_Pipeline_Example',
    schedule_interval=None,
)

### service a: parse XML/JSON and output CSV ###
def service_a(**context):
    # Read manifest YAML file that contains rules for
    # selecting data
    with open('manifest_service_a.yaml', 'r') as file:
        manifest = yaml.safe_load(file)

    # Assume we have an XML or JSON input file
    input_file = context['params'].get('input_file')

    if input_file.endswith('.xml'):
        tree = ET.parse(input_file)
        root = tree.getroot()
        data = []
        for elem in root.findall(manifest['relevant_info']):
            row = {field: elem.find(field).text \
                    for field in manifest['fields']}
            data.append(row)
    elif input_file.endswith('.json'):
        with open(input_file, 'r') as f:
            json_data = json.load(f)
            data = [{field: entry[field] for field in \

```

```

        manifest['fields']}] for entry in json_data]

# Convert selected info to CSV
df = pd.DataFrame(data)
df.to_csv('/tmp/output_service_a.csv', index=False)

#### service b: CSV to TTL triplification ####
def service_b(**context):
    # Read manifest for columns and RDF schema
    with open('manifest_service_b.yaml', 'r') as file:
        manifest = yaml.safe_load(file)

    # Load CSV
    df = pd.read_csv('/tmp/output_service_a.csv')

    # Select only relevant columns
    selected_df = df[manifest['relevant_columns']]

    # Create RDF graph
    g = Graph()

    # Triplify data
    for _, row in selected_df.iterrows():
        subject = URIRef(
            f"{manifest['namespace']}/{row[manifest['id_column']]}"
        )
        for column, predicate in manifest['schema'].items():
            g.add((subject, URIRef(predicate), \
                Literal(row[column])))

    # Output TTL file
    g.serialize(destination='/tmp/output_service_b.ttl', \
        format='ttl')

#### service c: Load TTL into Virtuoso/GraphDB ####
def service_c(**context):
    ttl_file = '/tmp/output_service_b.ttl'

```

```

# This is a placeholder. The actual code would depend |
on your specific setup for Virtuoso/GraphDB
# It would typically involve connecting to the SPARQL |
endpoint and performing an INSERT or LOAD query

with open(ttl_file , 'r') as file:
    ttl_data = file.read()

# Example SPARQL Update query for loading into |
Virtuoso/GraphDB
sparql_update = """
INSERT DATA {
    %s
}
""" % ttl_data

# Assume connection to Virtuoso/GraphDB SPARQL endpoint
# You would use a package like SPARQLWrapper or a |
custom connection to load the TTL
# sparql_client.query(sparql_update)

print ("Loaded_TTL_into_SPARQL_database.")

#### Define Airflow tasks ####
task_a = PythonOperator(
    task_id='service_a',
    python_callable=service_a,
    provide_context=True,
    params={'input_file': '/path/to/input.xml'},
    # or input.json
    dag=dag,
)

task_b = PythonOperator(
    task_id='service_b',
    python_callable=service_b,
    provide_context=True,
    dag=dag,
)

```

```
)

task_c = PythonOperator(
    task_id='service_c',
    python_callable=service_c,
    provide_context=True,
    dag=dag,
)

# Task dependencies
task_a >> task_b >> task_c
```

Example YAML Manifest (for Service A and B)

manifest_service_a.yaml

```
relevant_info: 'record' # Path in XML where relevant
                        # data is stored

fields:
  - id
  - title
  - description
  - date
```

manifest_service_b.yaml

```
relevant_columns:
  - id
  - title
  - description
schema:
  id: 'http://example.org/id'
  title: 'http://purl.org/dc/elements/1.1/title'
  description: 'http://purl.org/dc/elements/1.1/description'
  date: 'http://purl.org/dc/elements/1.1/date'
namespace: 'http://example.org/resource'
id_column: 'id'
```

References

- Altintas, I., C. Berkley, E. Jaeger, M. Jones, B. Ludascher, and S. Mock. 2004. “Kepler: an extensible system for design and execution of scientific workflows.” In *Proceedings. 16th International Conference on Scientific and Statistical Database Management, 2004*. 423–424. <https://doi.org/10.1109/SSDM.2004.1311241>.
- Atkinson, Malcolm, Sandra Gesing, Johan Montagnat, and Ian Taylor. 2017. “Scientific workflows: Past, present and future.” *Future Generation Computer Systems* 75 (June): 216–227. <https://doi.org/10.1016/j.future.2017.05.041>.
- Barbot, Laure, Yoann Moranville, Stefan Buddenbohm, Klaus Illmayer, and Matej Ďurčo. 2020. *MS42 Marketplace – Alpha Release*. Zenodo. Milestone 42 of SSHOC project (alpha release of SSH Open Marketplace), June. <https://doi.org/10.5281/zenodo.4585700>.
- Biemann, Chris, Gregory R. Crane, Christiane D. Fellbaum, and Alexander Mehler. 2014. “Computational Humanities – Bridging the Gap between Computer Science and Digital Humanities (Dagstuhl Seminar 14301).” *Dagstuhl Reports* (Dagstuhl, Germany) 4 (7): 80–111. ISSN: 2192-5283. <https://doi.org/10.4230/DAGREP.4.7.80>. <https://drops.dagstuhl.de/entities/document/10.4230/DagRep.4.7.80>.
- Coleman, Tainã, Henri Casanova, Loïc Pottier, Manav Kaushik, Ewa Deelman, and Rafael Ferreira da Silva. 2022. “WfCommons: A framework for enabling scientific workflow research and development.” *Future Generation Computer Systems* 128:16–27. ISSN: 0167-739X. <https://doi.org/https://doi.org/10.1016/j.future.2021.09.043>. <https://www.sciencedirect.com/science/article/pii/S0167739X21003897>.
- Concordia, Cesare, Carlo Meghini, and Filippo Benedetti. 2020. “Store Scientific Workflows Data in SSHOC Repository” [in eng]. In *Proceedings of the Workshop about Language Resources for the SSH Cloud*, edited by Daan Broeder, Maria Eskevich, and Monica Monachini, 1–4. Marseille, France: European Language Resources Association, May. ISBN: 979-10-95546-43-6. <https://aclanthology.org/2020.lr4sshoc-1.1/>.
- Degl’Innocenti, Emiliano, Carmen Di Meo, and Alessia Spadi. 2024. “DARIAH.it: Data Integration Strategies and Solutions for Digital-Resources Management and Research in the Arts and Humanities.” Published in June 2024; Open Access under CC BY-NC-ND 4.0 :contentReference[oaicite:1]index=1, *Mimesis Journal* 13 (2): 119–134. ISSN: 2279-7203. <https://doi.org/10.13135/2389-6086/9920>.

- ESFRI. 2020. *ESFRI White Paper: Making Science Happen. A New Ambition for Research Infrastructures in the European Research Area*. Technical report. White Paper published 27 April 2020 by ESFRI :contentReference[oaicite:1]index=1. Brussels: ESFRI, April. https://www.esfri.eu/sites/default/files/white_paper_esfri-final.pdf.
- European Union. 2013. *Regulation (EU) No 1291/2013 of the European Parliament and of the Council of 11 December 2013 establishing Horizon 2020 – the Framework Programme for Research and Innovation (2014–2020) and repealing Decision No 1982/2006/EC*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:32013R1291>. Official Journal of the European Union, L347, 20.12.2013, pp. 104–173, December.
- Giardine, Belinda, Cathy Riemer, Ross C. Hardison, Richard Burhans, Laura Elnitski, Prachi Shah, Yi Zhang, et al. 2005. “Galaxy: a platform for interactive large-scale genome analysis.” Published online 16 September 2005; printed October 2005; PMID 16169926; PMCID PMC1240089 :contentReference[oaicite:1]index=1, *Genome Research* 15, no. 10 (October): 1451–1455. <https://doi.org/10.1101/gr.4086505>.
- Gil, Yolanda, Ewa Deelman, Mark Ellisman, Thomas Fahringer, Geoffrey Fox, Dennis Gannon, Carole Goble, Miron Livny, Luc Moreau, and James Myers. 2008. “Examining the Challenges of Scientific Workflows.” *Computer* 40 (January): 24–32. <https://doi.org/10.1109/MC.2007.421>.
- Latour, Bruno. 1993. *We Have Never Been Modern: Essai d’anthropologie symétrique*. Translated by Catherine Porter. Translated by Catherine Porter; original French edition published 1991 :contentReference[oaicite:1]index=1. Cambridge, MA, USA: Harvard University Press. ISBN: 0-674-94838-6.
- Latour, Bruno. 2005. *Reassembling the Social: An Introduction to Actor-Network Theory*. First edition published 2005; part of the “Clarendon Lectures in Management Studies” :contentReference[oaicite:2]index=2. Oxford & New York: Oxford University Press. ISBN: 978-0199256044.
- Latour, Bruno. 2014. *How Better to Register the Agency of Things: Ontology*. Tanner Lecture, Yale University. Delivered 27 March 2014; accessed today :contentReference[oaicite:3]index=3, March. <http://www.bruno-latour.fr/node/563.html>.

- Ludäscher, Bertram, Ilkay Altintas, Shawn Bowers, Julian Cummings, Terence Critchlow, Ewa Deelman, David De Roure, et al. 2009. “Scientific Process Automation and Workflow Management.” Chap. 13 in *Scientific Data Management: Challenges, Technology, and Deployment*, edited by Arie Shoshani and Doron Rotem, 467–508. Computational Science Series. Book chapter in the Computational Science Series; DOI and publisher confirmed :contentReference[oaicite:1]index=1. Boca Raton, FL: Chapman & HallCRC, December. <https://doi.org/10.1201/9781420069815-c13>.
- Melgar-Estrada, Liliana, Marijn Koolen, Kaspar Beelen, Hugo Huurdeman, Mari Wigham, Carlos Martinez Ortiz, Jaap Blom, and Roeland Ordeman. 2019. “The CLARIAH Media Suite: a Hybrid Approach to System Design in the Humanities.” In *Proceedings of the 2019 Conference on Human Information Interaction and Retrieval (CHIIR ’19)*, 373–377. March. <https://doi.org/10.1145/3295750.3298918>.
- National Academies of Sciences, Engineering, and Medicine. 2019. *Reproducibility and Replicability in Science*. A consensus study report defining reproducibility and replicability, with recommendations to improve research rigour and transparency :contentReference[oaicite:1]index=1. Washington, DC: The National Academies Press. ISBN: 978-0-309-48616-3. <https://doi.org/10.17226/25303>.
- Presidenza del Consiglio dei Ministri. 2021. *Piano Nazionale di Ripresa e Resilienza (PNRR)*. <https://www.governo.it/sites/governo.it/files/PNRR.pdf>. Versione trasmessa alla Commissione Europea il 30 aprile 2021, April.
- Snow, Charles P. 1961. *The Two Cultures and the Scientific Revolution*. The Rede Lecture. Cambridge: Cambridge University Press.
- SSHOC. 2022. *SSHOC Legacy Booklet*. Zenodo. Published on Zenodo covering SSHOC project outcomes. <https://doi.org/10.5281/zenodo.6394462>.
- Wilkinson, Mark D., Michel Dumontier, IJsbrand J. Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, and et al. 2016. “The FAIR Guiding Principles for Scientific Data Management and Stewardship.” Published 15 March 2016; PMC 4792175, *Scientific Data* 3 (1): 160018. <https://doi.org/10.1038/sdata.2016.18>.