

Supporting executable scientific workflows in a clustered Infrastructure: DARIAH.it and H2IOSC

Emiliano Degl'Innocenti¹, Francesco Pinna¹, Alessia Spadi¹, Federica Spinelli¹

¹Istituto Opera del Vocabolario Italiano del Consiglio Nazionale delle Ricerche (OVI-CNR)

Introduction

H2IOSC (Humanities and cultural Heritage Italian Open Science Cloud) is a project funded by the National Recovery and Resilience Plan (PNRR) in Italy, aiming at creating a federated cluster comprising the Italian branches of four European research infrastructures (RIs) - CLARIN, DARIAH, E-RHIS, OPERAS - operating in the Social Sciences and Humanities (formerly "Social and Cultural Innovation") sector of ESFRI (European Strategy Forum for Research Infrastructures).

The H2IOSC goal is to create a FAIR ecosystem to support the digital transition in the SSH and CH disciplines and foster interdisciplinary research, enabling close collaboration among researchers working with digital data, tools and methods in.

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 organization and representation framework within the cloud environment.

This involves setting up the physical infrastructure, which includes designing and building 8 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 that will ensure all the data gathered by the 4 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. In particular, DARIAH.it will implement a 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 for the above task. A significant aspect of this process is the semantization 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 resources alignment involves data cleaning, mapping and modelling, as well as the definition of standardized 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. Connected to this research and development activities, DARIAH.it will promote training, outreach, and dissemination efforts, as well as the funding of a Doctoral Scholarship in Digital Philology¹.

For the implementation of the executable workflows supporting the Scientific Pilots DARIAH.it created a set of core elements to manage the interaction of the selected services (i.e. the API manager) and to allow the actual execution of the services in a specific runtime environment (i.e. the AEON - dAriah sErvice Oriented iNfrastructure), a service-provision oriented infrastructure interacting with the Marketplace developed within the Federation and aligned with SSHOC and EOSC platforms.

By implementing this environment, DARIAH.it aims at providing an innovative and sustainable research ecosystem for humanities and human sciences research, as well as cultural heritage preservation. This will be achieved by bridging the gap in the transition to digital approaches in research activities, overcoming the obsolescence of current sector-specific research product systems, and maximizing the potential impact of research communities by developing innovative approaches.

Bridge the gap in digital humanities research

While traditional humanities are a well-known and established discipline, digital humanities (DH) have emerged as a significant field in the last 50 years. The definition of digital humanities as a discipline has a fundamental milestone in the 1960s, in the pioneering work of Jesuit Father Roberto Busa, that worked on Thomas Aquinas' texts using digital approaches². That experience defined the application of information technologies to the study of humanities as the core characteristic of the emerging discipline.

In the 1960s, discussions arose regarding the impact of large datasets on knowledge accumulation and management methods, leading to the development of data science or data-driven science. By the 1980s, numerous projects were utilizing substantial data and employing computer methodologies across various fields, including the humanities: disciplines such as linguistics, literature, and history were early adopters.

The internet phenomenon, marked by widespread access to vast amounts of information since the 1990s, further advanced this tendency in digital humanities: mass digitization efforts of books, documents, and resources resulted in unprecedented growth in their availability. The early 2000s witnessed a peak in this trend, ushering humanities in the full digital era. Consequently, contemporary studies cannot overlook the availability of digital services due to the abundance of data and tools in the digital realm surpassing those in the physical world. These trends are the same that shaped the internet ecosystem. Today, access to digital tools

¹ PhD in Filologia Romanza e Italiana Digitale (FROID): <https://www.sns.it/it/disciplinacorso-di-laurea/corso-phd/filologia-romanza-e-italiana-digitale-froid>

² <https://www.corpusthomisticum.org/>

and data is indispensable, 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 transitioned from experimental approaches to become a standard practice. The definition of workflows in digital humanities follows the development of the discipline from the analogic era to the digital era. Humanists can be said to have already employed workflows, intended 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. The shift from traditional humanities to digital humanities involves the integration of computational tools and methods into traditional humanistic research. This transition enables scholars to analyse vast amounts of digital data, uncover new patterns and insights, and create innovative forms of scholarly expression. Digital humanities expand the scope and depth of humanistic inquiry, challenging traditional methodologies³ and opening up new venues for research and collaboration (Biemann et al., 2014).

This apparently simple transition carries the need to reorganize well-established and successful procedures in humanities research. The rooting of research methods in a consolidated tradition makes the application of digitally enabled workflows even more complex than in other disciplines, at the same time the digital era forces this shift.

Bruno Latour⁴, a renowned French scholar, characterized in his lectures this digital transition as the "screentoria"⁵, a modern-day equivalent of the scriptorium where medieval monks contemplated knowledge in solitude: the scriptorium has evolved into a digital screen, reflecting the shift in its dimensions and distribution.

Research infrastructures (RI) emerge as a fundamental element of connection, facilitating knowledge transmission and societal development. While not inherently digital, they encompass all elements that enable individuals to thrive within society.

Within the humanities, RIs organize a vast, distributed, and often non-interoperable system of knowledge. This fragmentation arises from the use of different languages, necessitating an alignment of meaning for effective utilization by researchers and scholars. Reducing this fragmentation is essential for promoting the transition from data to knowledge, enabling the application of information extracted from data to address societal challenges and needs and

³ Father Roberto Busa too recommended to consider 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

⁴ For Latour's works on technology, science, and society see "We have never been modern" and "Reassembling the Social"

⁵ <http://www.bruno-latour.fr/node/563.html>

foster development and growth. Information retrieval plays a crucial role in producing intangible knowledge: from an initial knowledge base, new knowledge is created.

The digital ecosystem comprises resources made available by memory institutions, research performing organisations as well as from other actors in the GLAM (Galleries, Libraries, Archives, Museums) sector and individuals (citizen scientists). Fragmentation refers to the technological gap that hinders communication and knowledge production among these resources. Therefore, reducing fragmentation involves fostering the transition from data to knowledge, facilitating the flow of information.

To achieve this, semantic continuity must be ensured, translating concepts across disciplines to enable resources to discuss the same objects using a common language, regardless of their format, production method, or underlying technology.

Research infrastructures, considered as sociotechnical systems, are interfaces between humans and technologies: one of the fundamental objectives of DARIAH.it in H2IOSC is to bridge the gap between the two cultures (Wikipedia contributors, "The Two Cultures") - commonly described as hard sciences and humanistic knowledge. Hard sciences typically investigate the natural world, while humanities explore the human realm. The digital environment offers an opportunity for these different approaches to knowledge to converge, allowing researchers to transcend technological, disciplinary, and other barriers that may hinder the depth of their research.

Digital tools and methods can help reducing the barriers between tangible and intangible entities. For example, a book is both a physical object and a cultural artifact. Recognizing these entities as interconnected is one of the possibilities offered by the work of semantic alignment and data integration undertaken by DARIAH.it within the H2IOSC project, at the core of the digital philology scientific workflow implementation.

To maximise the impact of this development activity and promote access to innovative research environments that can revolutionize the current landscape, enabling the analysis of large datasets and fostering interoperability among platforms that were previously not interoperable, it's essential to involve all the relevant stakeholders, including researchers, interested communities and other potential beneficiaries.

Workflows for digital humanities

The concept of workflow comes from the industry as it was used to describe the flow of business activities. Later the concept was adopted for scientific endeavours and workflows became a component of the research infrastructure. Scientific workflows are increasingly being adopted in the Social Sciences and Humanities (SSH) sector to streamline and enhance research processes. By breaking down complex activities into smaller, manageable steps, workflows enable researchers to automate repetitive tasks, manage data more efficiently, and promote reproducibility. This approach is particularly beneficial for SSH researchers as it allows them to focus on higher-level analysis and interpretation, rather than getting involved with technical issues. Furthermore, sharing workflows through repositories, like those

provided by the SSHOC⁶ (Social Sciences & Humanities Open Cloud) and the H2IOSC projects, can foster collaboration, knowledge sharing, and the advancement of the field (Concordia et al, 2020).

In digital humanities' context, the interest in defining, composing, and executing workflows has evolved naturally, especially within the context of the SSHOC project, that built the [SSH Open Marketplace \(SSHOMP\)](#),⁷ a discovery portal that gathers and contextualises resources for Social Sciences and Humanities research communities. The primary objective of the SSHOC Marketplace was to establish a collaborative space where users can access and share digital tools and resources, fostering greater transparency and cooperation in research.

Today, the SSHOC Marketplace stands as a benchmark for digital research across Europe, driving 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 SSHOC Marketplace describe a research workflow as a sequence of steps that can be performed on research data throughout its lifecycle. Workflows can be executed using a variety of tools, resources and methods, and useful resources connected to each step.⁸

In the context of contemporary research, scientific workflows are employed by scientists 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 et al, 2020). It suffices to refer to [WfCommons](#)⁹: 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.

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

⁶ <https://sshopencloud.eu/>

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

⁸ For more details see: <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/>

of similar methods. Publishing datasets together with scripts or workflows is common practice, although it may not be sufficient to ensure reusability of data and reproducibility¹⁰. Workflows become increasingly complex while advancing research endeavours. During the design phase of the AEON platform, the runtime environment implemented by DARIAH.it to allow the execution of scientific workflows, the team defined different levels of complexity to be composed according to the flow to be performed:

- Unitary Workflow: simple tasks requiring a single action.
- 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 (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 (opposed to Generic workflows) are highly specialized, 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's 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 DHs, in order to have in the first instance tools that can streamline workflows in research and data analysis. Moreover, as DHs 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.

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

How to implement the workflow

Given the consideration expressed in the previous paragraph, DARIAH.it worked towards the definition of performing workflows into the H2IOSC project. A systematic approach was therefore applied to the development of a robust workflow that encompass also servification, virtualization, and remotization in research infrastructures. Within the H2IOSC federation 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 Federation can effectively transition to a more service-oriented and accessible model, enhancing collaboration and innovation within the research community.

Within this framework DARIAH.it wants to provide its users with a complete system to create and manage workflows. DARIAH-IT focuses on upgrading its current service provision capabilities to match the needs of the national digital humanities research community, by designing and implementing the AEON - dAriah sErvice Oriented iNfrastructure. The following paragraphs describe the implementation of this system.

DARIAH.it will make the services findable through the H2IOSC Marketplace (and the cooperating projects) and will offer actual services provision via the AEON platform. To achieve this goal, the DARIAH.it team focussed on a set of key activities, that included the AEON platform design and implementation; the existing services evaluation and refactoring; as well as the development of specific policies and guidelines to ensure effective interoperability and security for existing and newly created resources. The latter aspect will not be addressed in this paper but, due to the current situation, has become a major issue for DARIAH.it (Wikipedia contributors, "British Library cyberattack,") (Goodin, Dan. 2024)

(Bellini, Emanuele, and Emiliano Degl'Innocenti. 2024). The first step to implement this system is to define users and the actions they can undertake within the system.

AEON supports four main user roles, each with increasing levels of permissions and responsibilities: Basic User, Contributor, Curator, and Administrator.

Each role has specific functions. Basic Users can search and execute services and applications, create and manage workflows, publish and share their workflows, and participate in the community. Contributors can submit new services, update documentation, and provide technical support for the services they contribute. Curators, in addition to these functions, review and approve new submissions, manage the organisation of the catalogue, and generate reports on service usage. Administrators oversee the system's operation, handling user management, role assignment, system security, performance monitoring, log analysis,

and software updates.

The **Basic User** has access to the catalogue and public services, he can create personal descriptive workflows and is allowed to publish and share them. The **Contributor** role includes all the permissions of the Basic User, with the added possibility to submit approval new services or applications to the AEON's catalogue. Contributors are also responsible for updating the descriptions of existing services. The **Curator** builds upon the Contributor role, with the authority to review and approve new services or applications and manage categories and tags within the catalogue. Finally, the **Administrator** has full access to all system features, including the management of users and roles, configuration of system settings, performance monitoring, and ensuring security (Table 1 summarize the organization of the roles).

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• 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

Typical scenarios vary by role. For example, a researcher (Basic User) searches for relevant services in the catalogue, creates a workflow, runs it on its data, and optimises it as needed. A developer (Contributor) develops a new service, tests it, documents it, submits it for

approval, and provides user support after its inclusion in the catalogue. A museum curator (Basic User) searches for applications to create virtual exhibitions, customises them with relevant contents, tests them, and publishes them for visitors.

The Administrator monitors the system's performance, manages user access, reviews security logs, applies updates, and manages integrations with external systems.

Once the role is defined the process to create a new workflow can start by providing an accurate description of the scope and expected results of the process. An administrator has access to a "workflow manager" functionality. The workflow manager allows the administrators to create a new workflow or modify an existing one.

If the intention is to publish a descriptive workflow, a narrative description is sufficient, and the administrator can choose to save and publish the workflow on the platform. It's worth noting that, in the case of a descriptive workflow, involving services is not essential.

However, if the intention is to create an executable workflow, the selection of at least one service is necessary. If there is a single service, to complete the process, the administrator must associate a GUI (Graphic User Interface) with the service. In case the service doesn't own natively a GUI, a standard one will be provided by the workflow manager, basing on the service manifest file that contains information about which kind of computational and semantical input and output are expected (see next section). This mode of creating "atomic" workflows is, in fact, the process by which individual services are uploaded to AEON.

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.

Composer

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

- 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.
- 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.
- 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:

- 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.
- 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 manifest](#) to be associated with the workflow. Actually, a *manifest.json* file is a required configuration file for browser extensions and web applications that use WebExtension APIs. It is a JSON-formatted file that contains metadata about the extension, such as its name, version, and various functional aspects like background scripts, content scripts, and browser actions. The manifest can also include semantic information about the data used by the application or extension. This allows the browser or other systems to better understand and interact with the content and functionality provided by the extension.

There is obviously a predefined schema for these files, which also includes a brief and extended description of the service, characterization of the inputs and outputs, even at the semantic level.

The guidelines adopted in defining possible schemas for manifest files associated with services are those identified within H2IOSC WP6, task 6.1, and as such are common to all federation services.

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 input 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 files formats. The output part would be the collation table (downloadable), which may be provided in the form of XML files, or CSV, or similar formats.

WORKFLOW MANAGER

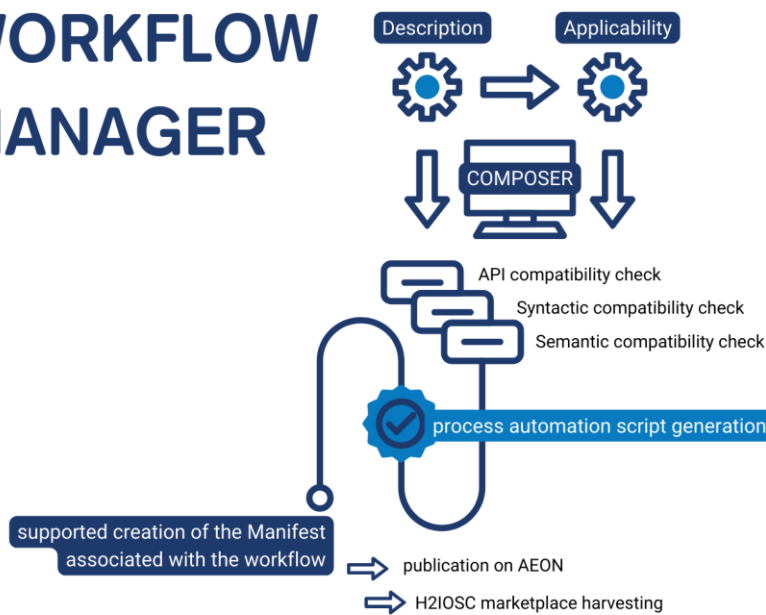


Image 1 workflow manager architecture

RESTORE¹¹ (smaRt accESs TO digital heRitage and memory), a project coordinated by the OVI CNR Institute (Istituto Opera del Vocabolario Italiano, Consiglio Nazionale delle Ricerche based in Florence), was selected as a use case for the AEON platform design. Among the project's goal there was the possibility to collect, manage and integrate data coming from different contexts and described according to different models, populating a SPARQL¹² database with a 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 the RESTORE project, 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. The project also created mapping schemas to transform the original data on resources stored inside the archives in semantic data.

These data needed to be processed in order to extract the relevant information and subsequently save it in a SPARQL database, making it accessible through simple queries for any research or study needs.

Schematically:

¹¹ RESTORE project, Higher Education intervention program (CNR4C), co-financed by the Tuscany Region CUP B15J19001040004 <<https://restore.oivi.cnr.it/>>

¹² 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/> .

```

'''
### 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 accessible
'''

```

This workflow becomes 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 need expressed by the project's stakeholders: simplify, speed up and supporting the data clean-up process across the entire data injection chain.

Examples

Here's a basic overview of what the previously presented pipeline may look like using Python with Apache Airflow¹³ 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 YAML¹⁴, 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)¹⁵. In this example, the placeholders for connecting and loading data into the SPARQL database

¹³ 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.

¹⁴ YAML (YAML Ain't Markup Language) is a human-readable data serialization format used for configuration files and data exchange.

¹⁵ 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.

are visible to track the steps, but typically a library like `SPARQLWrapper`¹⁶ is used to run queries against the endpoint.

```
```python
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')
```

---

<sup>16</sup> SPARQLWrapper is a Python library that abstracts the process of querying a SPARQL endpoint and processing the returned results.

```

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

```

```yaml
relevant_info: 'record' # Path in XML where relevant data is stored
fields:
  - id
  - title
  - description
  -
date
...

```

manifest_service_b.yaml

```

```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'
` ``
```

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<sup>17</sup>, or even entirely different frameworks and tools. Jupyter notebooks, for instance, offer a more interactive approach, allowing researchers to experiment with and visualize 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.

However, regardless of the specific tools used, the true value of these pipelines lies not just in the choice of technology but in their ability to meet the key requirements expressed by the project's stakeholders: simplifying the workflow, speeding up data processing, and streamlining the entire data injection chain. By automating tasks such as parsing, filtering, triplifying, and loading data, we significantly reduced manual intervention, minimizing the errors, and created a cleaner, more efficient process. Ultimately, the goal is to ensure that valuable data is made accessible and usable in a timely manner for any scholarly or research need—whether through Airflow, Jupyter, or another toolset entirely.

## Conclusions: Towards the EOSC

Research Infrastructures manage data assets that have the potential to produce new knowledge and promote innovation. As stated in ESFRI Whitepaper 2020 European Research Infrastructures are enablers, supporting the digital transition in different disciplines and research fields, promoting data-driven research and fostering the development and implementation of advanced solutions for scientific data management.

Due to the increasing use by large research communities, Research Infrastructures (RIs) become able to manage high-quality FAIR data that can be used for various purposes, including to help solving complex societal issues and advancing the Sustainable Development Goals (SDGs): “European RIs foster the definition, implementation and further development of advanced solutions for the effective provisioning and use of high-quality scientific data, with effective metadata descriptors, ease of access, interoperability and reusability, fully implementing the FAIR (Findable, Accessible, Interoperable, and Reusable) principles. [...] These efforts must be recognized, properly analysed and utilised to contribute to shaping the European Open Science Cloud system” (ESFRI White Paper 2020)

---

<sup>17</sup> Jupyter notebook is an open-source browser application that allows users to create and share documents that contain live code, visualizations, and narrative text (such as explanations and descriptions).

The European Open Science Cloud (EOSC) will leverage the experience and expertise of individual RIs as well as thematic clusters, to further enhance data sharing and collaboration among researchers. On the long run, EOSC aims at providing significant benefits to researchers, including improved data access, advanced e-infrastructure services, and the ability to address interdisciplinary challenges.

As stated in this paper the introduction of technological tools brought a substantial shift in the way of doing research, transitioning from traditional research to research assisted by digital tools. The advancement of technology is now bringing forward another shift that sees at its core the use of Artificial Intelligence in research.

Announcing the launch of the EOSC Node the European Commission stated: “The full deployment of the EOSC EU Node includes a robust set of services that address key challenges in modern research workflows, allowing users to operate efficiently within data-intensive environments” (European Commission, 2024)

This means that, going towards the European Open Science Cloud (EOSC) and the constitution of nodes, RIs will see a further shift of their role from data suppliers to primary users of machine enabled workflows. To ensure sustainability of workflows, given by their reproducibility and replicability, new aspects need to be defined: transparency of algorithms for the researchers, quality assessment of data, FAIR-by-design methodologies and Open Science principles should also be considered while developing workflows.

The transition towards data-driven science, facilitated by national (such as H2IOSC) and international platforms (like EOSC), has the potential to significantly impact scientific research practices. However, while this shift promotes open science principles and accessibility to data, it also raises concerns about potential disparities in resource access that needs to be addressed at European level.

By promoting the design and development of advanced technological solutions such as AEON, DARIAH.it wants to take a step towards a more efficient and collaborative future for digital humanities research. By providing a user-friendly platform for creating, managing, and sharing workflows, DARIAH.it supports its users in the creation of new research paths and promotes reproducibility and replicability for open research.

## Bibliography

- European Strategy Forum on Research Infrastructures (ESFRI). ESFRI WHITE PAPER. 2020. MAKING SCIENCE HAPPEN. A new ambition for Research Infrastructures in the European Research Area [https://www.esfri.eu/sites/default/files/white\\_paper\\_esfri-final.pdf](https://www.esfri.eu/sites/default/files/white_paper_esfri-final.pdf)

- Automated Research Workflows for Accelerated Discovery: Closing the Knowledge Discovery Loop. 2022. Washington, D.C.: National Academies Press. doi:10.17226/26532.
- Dr, Prof & Becker, Jörg & zur Muehlen, Michael. 2002. 'Workflow Application Architectures: Classification and Characteristics of Workflow-Based Information Systems'.  
[https://www.researchgate.net/publication/2404323\\_Workflow\\_Application\\_Architectures\\_Classification\\_and\\_Characteristics\\_of\\_Workflow-based\\_Information\\_Systems/citation/download](https://www.researchgate.net/publication/2404323_Workflow_Application_Architectures_Classification_and_Characteristics_of_Workflow-based_Information_Systems/citation/download)
- European Strategy Forum on Research Infrastructures (ESFRI, <https://www.esfri.eu/>, date of access: 01.02.2024).
- COUNCIL RECOMMENDATION on a Pact for Research and Innovation in Europe. 13701/21 EL/DOS/en ECOMP.3.B.
- MINISTERO DELL'ISTRUZIONE, DELL'UNIVERSITA' E DELLA RICERCA, DECRETO 9 maggio 2019 Ammissione del progetto «DARIAH-IT - Developing nAtional and Regional Infrastructural nodes of dAriaH in Italy» al finanziamento previsto dal decreto direttoriale 28 febbraio 2018. (Decreto n. 900/2019). (19A04611) (GU Serie Generale n.165 del 16-07-2019).  
[https://www.gazzettaufficiale.it/atto/serie\\_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2019-07-16&atto.codiceRedazionale=19A04611&elenco30giorni=true](https://www.gazzettaufficiale.it/atto/serie_generale/caricaDettaglioAtto/originario?atto.dataPubblicazioneGazzetta=2019-07-16&atto.codiceRedazionale=19A04611&elenco30giorni=true)
- Carbé, Emmanuela, Gabriele Lo Piccolo, Alessia Valenti, and Francesco Stella. *La Memoria Digitale: Forme Del Testo e Organizzazione Della Conoscenza. Atti Del XII Convegno Annuale AIUCD*. AIUCD, 63–64, 2023.  
<https://doi.org/10.6092/UNIBO/AMSACTA/7721>
- Wilkinson, Mark D., Michel Dumontier, IJsbrand Jan Aalbersberg, Gabrielle Appleton, Myles Axton, Arie Baak, Niklas Blomberg, et al. 2016. 'The FAIR Guiding Principles for Scientific Data Management and Stewardship'. *Scientific Data* 3 (1): 160018. doi:10.1038/sdata.2016.18
- Broeder, Daan, Maria Eskevich, and Monica Monachini, eds. 2020. *Proceedings of the Workshop about Language Resources for the SSH Cloud*. Marseille, France: European Language Resources Association, 1–4, 2020. <https://aclanthology.org/2020.lr4sshoc-1.0>.
- National Academies of Sciences, Engineering, and Medicine. 2019. Reproducibility and Replicability in Science. Washington, DC: The National Academies Press. doi:10.17226/25303
- 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)'. Schloss Dagstuhl – Leibniz-Zentrum für Informatik. doi:10.4230/DAGREP.4.7.80.
- 'The European Commission Announces the EOSC EU Node's Transition to Full Production | European Open Science Cloud - EU Node'. 2024. Accessed November 11.

<https://open-science-cloud.ec.europa.eu/news/european-commission-announces-eosc-eu-nodes-transition-full-production>.

- European Commission. Directorate General for Research and Innovation. 2020. *Supporting the Transformative Impact of Research Infrastructures on European Research: Report of the High Level Expert Group to Assess the Progress of ESFRI and Other World Class Research Infrastructures towards Implementation and Long Term Sustainability*. LU: Publications Office. doi:10.2777/34233
- SSHOC. 2022. 'SSHOC Legacy Booklet'. Zenodo. doi:10.5281/ZENODO.6394462.
- Barbot, Laure, Yoann Moranville, Stefan Buddenbohm, Klaus Illmayer, and Matej Ďurčo. 2020. 'MS42 Marketplace – Alpha Release', June. doi:10.5281/ZENODO.4585700.
- Wikipedia contributors, "The Two Cultures" Wikipedia, The Free Encyclopedia, [https://en.wikipedia.org/w/index.php?title=The\\_Two\\_Cultures&oldid=1248006694](https://en.wikipedia.org/w/index.php?title=The_Two_Cultures&oldid=1248006694) (accessed November 15, 2024).
- Wikipedia contributors, "British Library cyberattack," Wikipedia, The Free Encyclopedia, [https://en.wikipedia.org/w/index.php?title=British\\_Library\\_cyberattack&oldid=1255404045](https://en.wikipedia.org/w/index.php?title=British_Library_cyberattack&oldid=1255404045) (accessed November 15, 2024).
- Goodin, Dan. 2024. 'Archive.Org, a Repository of the History of the Internet, Has a Data Breach'. Ars Technica. October 10. <https://arstechnica.com/information-technology/2024/10/archive-org-a-repository-storing-the-entire-history-of-the-internet-has-been-hacked/>.
- 'Transitioning SSH European Research Infrastructures to Critical Infrastructure Through Resilience'. 2024 *IEEE International Conference on Cyber Security and Resilience (CSR)*, September, 801–6. doi:10.1109/CSR61664.2024.10679383.