DIGITAL HERITAGE (2025)
S. Campana, D. Ferdani, H. Graf, G. Guidi, Z. Hegarty, S. Pescarin, and F. Remondino (Editors)

Towards a Scientometric Understanding of Cultural & Digital Heritage: Multi Source Data Integration Pipeline & EC Funding Trends

Walter Ehrenberger Sander Muenster

FSU Jena

Abstract

We introduce an ELT pipeline and a data model that integrates, sanitizes, unifies, and enriches multiple data sources to enable quantitative analysis of cultural and digital heritage. This results in 374,998 research outputs, 45,725 institutions and 19,558 projects after deduplication, including other entities and metadata. We developed scientometric use cases tailored for researchers and policy makers, and implemented a preliminary version of them in an interactive web app prototype. Using basic keyword filtering to identify relevant fields in our curated dataset, our analysis reveals that Digital Heritage funding surged 355% compared to Cultural Heritage's 137% growth (2015-2024), with Italy emerging as the leader in both fields. Computer Science dominates Digital Heritage (averaging 60% of funding), while Cultural Heritage maintains broader disciplinary distribution. Economics/Business show remarkable growth in both fields, suggesting increasing commercialization focus. These initial findings, as well as the use cases presented in the prototype, demonstrate the pipeline's potential while highlighting the critical need for sophisticated topic modeling and classification systems as well as further enrichment of the data to unlock deeper scientometric insights.

Keywords: Scientometric Analysis, Cultural & Digital Heritage, Multi Source Data Pipeline, Funding Trends

Acknowledgments: The research upon which this paper is based has received funding from the European Union's research and innovation program Horizon Europe under the grant agreement No. xxxxxxxxx.

Acronyms & Definitions

CH Cultural Heritage, a multidisciplinary field
Core Core Data Model
DH Digital Heritage, discussable a subfield of CH
EC European Commission
ELT Extract Load Transform Data, to develop data pipelines
ORM Object Relational Mapping, to simplify communication between programs and databases
RO Research Output
SME Small and Medium Enterprises

1. Introduction

Cultural heritage (CH) encompasses artifacts, monuments, buildings, and sites with symbolic, historic, artistic, and scientific significance, as defined by UNESCO [UNE09]. Meanwhile, Digital Heritage (DH) comprises computer-based materials of enduring value requiring active preservation for future generations [UNE20]. Both concepts represent what humans find meaningful and wish to preserve as legacy—from ancient texts like "The Epic of Gilgamesh" to modern digital phenomena. Much as UNESCO defines heritage as "our legacy from the past, what we live with today, and what we pass on to future generations" [UNE23], CH and DH serve as cultural anchors connecting past insights with future innovation.

Researchers and policy makers facing challenges in CH and DH fields require data-driven insights for decision-making. Without robust scientometric analysis, identifying emerging research areas, tracking collaboration networks, or allocating resources effectively becomes nearly impossible. Publications and project documentation constitute essential components for research advancement, yet analyzing these vast collections requires sophisticated data integration approaches. By examining publication patterns, citation networks, and funding distributions, scientometrics enables monitoring field development, tracking emerging trends, and assessing collaboration effectiveness—ultimately supporting informed policy decisions and strategic research directions in CH and DH domains.

The code for the pipeline can be accessed at https://github.com/VieVaWaldi/dh_pipeline/, and the website at https://heritagemonitor.org/.

DOI: 10.2312/dh.20253381

2. State of the Art

In the following, the State of the Art and current gaps in scientometric analysis are examined.

2.1. Literature Review: Scientometric studies using Cordis Projects in CH/DH

For our analysis, we conducted a search for studies performing scientometric analysis or gathering data in CH and DH using Cordis research projects. Our query, "scientometric" AND "cultural heritage" AND "cordis" (2020-present), aimed to discover works utilizing Cordis data for scientometric analysis in cultural heritage. Notably, this search yielded only 16 results, indicating limited application of Cordis data for scientometric analysis in this field.

After applying inclusion criteria requiring 1) usage of Cordis for scientometric analysis and 2) focus on Cultural Heritage, merely 3 papers proved relevant. Google Scholar provided these results on May 4th, 2025, sorted by relevance.

Kokkala presents a master thesis constituting the sole direct analysis of Cordis projects in CH/DH matching our query parameters. Her work employs systematic scoping review methodology across three Cordis datasets, emphasizing digital technology implementation in cultural heritage. While providing basic mapping, it lacks advanced scientometric methodologies necessary for comprehensive field analysis [Kok24].

Münster et al. utilize Cordis specifically for policy trend analysis rather than comprehensive scientometric integration. Most significantly, they explicitly acknowledge that an "up-to-date investigation on cultural heritage as a scholarly field is currently missing" and identify a substantial gap in systematic field mapping approaches [MUUA21].

Ariza et. al. conduct a scientometric study that references Cordis but doesn't analyze its data directly. Instead, they employ alternative scientometric methods and highlight literature fragmentation as a significant research gap in the field [AC*24].

Key findings from our literature review reveal that only 2 of 16 papers actually analyze Cordis data for cultural heritage applications. No identified studies integrate multiple data sources beyond Cordis, with existing work predominantly focusing on basic mapping rather than advanced scientometric analysis. A clear gap exists for the comprehensive pipeline approach our work provides, particularly regarding multi-source data integration and enrichment capabilities.

While our literature review yielded valuable insights regarding the specific intersection of scientometrics, cultural heritage, and Cordis data usage, we acknowledge several methodological constraints. Our highly focused search query produced a narrow result set (16 papers), which effectively demonstrates the gap for our specific pipeline approach but may not comprehensively capture all relevant work in adjacent areas.

2.2. Gaps in the State of the Art

Current literature reveals several critical gaps in CH and DH research that our scientometric approach aims to address:

Field Fragmentation According to [AHG*22], significant fragmentation exists within the cultural heritage sector, weakening collaborative potential and efficient resource allocation. Scientometric analysis could map existing networks between institutions, providing evidence for strategic partnership development.

Cross-Sector Barriers Despite recognition of digitization as an emerging focus area, [AHG*22] identify persistent barriers to effective cross-sector partnership. Comprehensive data analysis could uncover which sectors successfully collaborate and which remain isolated, creating a roadmap for targeted integration efforts.

Restricted Plurality Established power structures and conventional approaches limit innovation and perspective diversity within the field. Quantitative scientometric analysis could demonstrate which voices dominate discourse and which remain underrepresented, supporting field democratization [AHG*22].

Methodology Standardization Needs Münster highlights the absence of common methodological frameworks, attributable to the field's interdisciplinary nature. Scientometric approaches could identify successful methodological patterns that might serve as best practices or de facto standards [Mü17].

Obsolete Field Mapping Most notably, Münster concluded that "an up-to-date investigation on cultural heritage as a scholarly field is currently missing." Fresh scientometric analysis would identify key actors, emerging topics, and research trends, helping researchers locate experts and knowledge gaps [Mü17].

3. Research Questions

Following our analysis of the current state of the art, several key questions emerge which will guide our development of improved infrastructure for scientometric analysis in CH and DH fields.

3.1. Primary Research Question

Our investigation centers around one main research question:

> How can multiple data sources from the fields of CH and DH be unified through a pipeline to reveal scientometric insights, and what kind of enrichment and modelling is necessary to make these insights useful?

3.2. Case Study Research Questions

For empirical validation, we examine two specific case studies:

- **Case Study 1:** How do funding patterns for CH and DH differ across EU countries (2015-2024), and what methodological challenges emerge when integrating multiple data sources for scientometric analysis?
- **Case Study 2:** How have EC funding patterns evolved across fields and subfields in CH and DH (2015-2024), and what does this reveal about shifting research priorities and methodological requirements for scientometric analysis?

Moving forward, our research addresses the primary question through careful examination of both case studies and detailed discussion of implemented use cases.

4. Methodology: Core Data Model, Use Cases & Pipeline

Following sections examine the unified data model our pipeline creates, look at the Use Cases this model enables and finally explore the architecture and methodology of the pipeline to create the data.

4.1. The Core Data Model

Source data is stored in our database in its original schema, enabling individual analysis and fast iterative updating of the core data model. Our approach is designed after the ELT principle which first loads source data in its original schema into the database before transforming it into the core data model used for analysis. Consequently, the database must provide namespaces and tables matching the original data schema of the sources. Chapter 4.3 will define the ELT process more thoroughly.

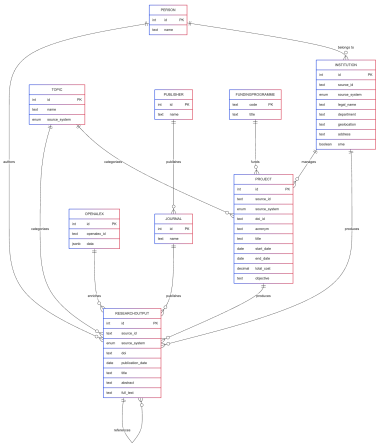

Figure 1: *Simplified core data model schema showing main entities and their relationships.*

4.1.1. Main Entities (3 core entities)

- **Projects:** Have start/end dates, research objectives, funding amounts, and work under a FundingProgramme. Projects include multiple partner institutions (one as coordinator), generate Research Outputs and are classified into Topics.
- **Institutions:** Research organizations (universities, SMEs, collaborators) with a legal name, and geographical locations. Institutions may work on a project and receive funding for that project. Institutions also produce Research Outputs.
- **Research Outputs (ROs):** Publications (various types) or administrative documents, published under a Journal and can also be classified into Topics. ROs are self-referential to create citation networks.

4.1.2. Other Entities

- **Persons:** Connected to multiple entities like ROs and Institutions.
- **Links:** Connections to multiple entities like ROs, Projects and Institutions.
- **FundingProgramme:** Framework Programme under which projects are funded and operate with Call identifiers.
- **Journal:** Publishing venue for Research Outputs.
- **Topic:** Classification system for projects and ROs into domain/field/subfield and topic.

4.1.3. Enrichment Entities

Each enrichment type has its own entity and is used to supplement missing information about the core entities.

- **CrossRef:** Enrichment for ROs (bibliography, affiliations, citations).
- **OpenAlex:** Enriches publications with citations and impact metrics.
- **MapBox:** Retrieves missing geographical information for institutions.

4.1.4. Metadata and Junctions

Each relation between entities is modelled with a junction table which may have more information like the funding per institution for a project or the position of the authors writing a paper. When an entity referenced by a junction table is deleted, the corresponding junction table records are automatically removed via cascade delete.

All entities also have various meta data (like, EC contribution to the project, DOIS, various IDs etc) which are left out for simplification.

4.1.5. Indexes

Indexes help find information in a database more quickly, by sorting and organizing it. We use the following main indexes:

- **B-tree indexes:** For quick searches and exact matches (find things by name, ID, dates, etc.)
- **GIN indexes:** For keyword search in text (search for words anywhere in text fields), e.g., to find all matching abstracts using keywords.
- **Trigram indexes:** For finding similar text (spot duplicates even when there are typos), important for deduplicating data.

4.1.6. Provenance

- Automatic tracking of when data was last updated (each row has a timestamp that updates automatically)
- Source tracking for imported data (records which source system data came from and its original ID)

Given this core data model we can create use cases for a scientometric analysis.

4.2. Developing Use Cases for Scientometric Analysis in CH / DH

Following feedback gathered by Sander Münster across five years of stakeholder interviews, several key requirements emerged from researchers and policy makers in both fields. Below are seven scientometric use cases designed to address these needs, highlighting how each might benefit key stakeholders while supporting our main research question about useful scientometric insights.

1. **Mapping key actors to topics:** Identifies leading researchers or departments through analysis of high-impact publications and citation counts within specific topics. For researchers, facilitates discovery of potential collaborators working in similar areas. For policy makers, enables connections with subject experts when designing new funding programmes.
2. **Understanding institutional contributions:** Examines institutional research outputs and impact within specific topics, revealing strengths and areas of expertise. Assists researchers in positioning themselves for funding by leveraging institutional strengths. Enables policy makers to make evidence-based decisions about funding distribution to institutions with proven expertise.
3. **Discovering collaboration networks:** Maps existing partnerships and reveals new collaboration opportunities through analysis of joint publications and co-funded projects. Helps researchers identify potential collaborators. Supports policy makers in addressing collaboration gaps through targeted funding programmes.
4. **Analysing topic structure in DH and emerging trends:** Provides comprehensive overviews of field structure and identifies emerging research directions through topic analysis and trend detection. Helps researchers "spot the curve" and identify declining research areas. Assists policy makers in addressing current research gaps and supporting emerging areas.
5. **Visualising funding patterns:** Creates representations of funding distribution across geography, topic, institution type, and other dimensions. Helps policy makers ensure equitable resource distribution and identify funding imbalances.
6. **Identifying institutions with successful track records:** Reveals best practices and success patterns by analyzing institutions that consistently win grant applications. Helps researchers better tailor grant applications. Enables policy makers to understand why specific institutions might be underperforming.
7. **Assessing impact of funding programmes:** Analyzes programme effectiveness by tracking outcomes like publications, collaborations, and follow-up funding. Supports policy makers in making data-driven decisions about continuing, modifying, or discontinuing funding programmes.

In section 5.2, using the web app prototype, an attempt is made to visualize these Use Cases and determine the future directions needed to apply them.

4.3. Methodology for the ELT Pipeline

We are using the ELT pattern for our pipeline, which serves as a data processing approach that enables organizations to first preserve complete source data within their data warehouse before performing transformations, allowing for greater flexibility and comprehensive data lineage tracking. According to research in advanced data integration technologies, the Extract, Load, Transform (ELT) pattern represents a paradigm shift in data warehouse architecture where raw source data is loaded directly into the target system before transformation occurs [SA22].

For our pipeline architecture, ELT offers three key advantages over traditional ETL: First, by loading raw data directly into dimensional staging tables, we maintain a complete audit trail of source transformations, enabling reconstruction of historical business rules if needed. Second, leveraging the target database's native SQL processing eliminates costly intermediate transformation servers, reducing infrastructure overhead while utilizing our warehouse's optimized columnar storage for faster analytical queries. Third, this pattern supports iterative schema development, allowing us to refine our core data model based on raw data patterns rather than being constrained by predetermined ETL transformations that may sacrifice valuable data attributes [SA22].

4.3.1. Data Sources

Our data pipeline integrates multiple heterogeneous research information systems, each providing unique perspectives on scientific outputs and cultural heritage projects. The ELT architecture enables flexible adaptation to new sources while preserving original data structures, allowing us to iterate the core data model without losing valuable attributes. The pipeline also enables usage of different queries for the sources using a configuration, which results in a new extraction process.

ArXiv The ArXiv platform serves as a pioneering open-access repository offering more than two million scholarly articles across eight subject areas, with community-based curation ensuring quality control [ArX24]. The system retrieves entries (publications) with structured metadata including titles, abstracts, authors, categories, and publication dates. Query used to fetch entries: All: computing AND (all: humanities OR all:heritage).

CORE CORE represents the world's largest collection of open-access research papers, aggregating over 200 million full-text documents from repositories and journals worldwide [COR24], the full-text documents being a major strength. Query used to fetch entries: ((computing AND cultural) OR (computing AND heritage)).

CORDIS The Community Research and Development Information Service maintains Europe's primary repository for EU-funded research projects from FP1 to Horizon Europe, encompassing comprehensive project metadata, deliverables, and funding information [Com24]. CORDIS's XML-based data model provides extensive relational information including project timelines, participant organizations, and budget allocations as well as topic modeled using EuroSciVoc. Query used to fetch entries: cultural OR heritage.

OpenAIRE OpenAIRE operates as a comprehensive research infrastructure connecting repositories, publications, and funding information across European research ecosystems [Ope24]. The platform's hierarchical funding structure captures EC program relationships, while its research output linking enables tracking of deliverables and publications. OpenAIRE's strength in connecting research outputs to funding sources provides crucial context for understanding EU research investment impacts. Query used to fetch entries: cultural OR heritage.

The pipeline currently processes four primary sources with varying checkpoint mechanisms and rate limitations, adaptable through configuration changes. Query patterns remain flexible, allowing refinement without architectural modifications. Future expansion will include Web of Science integration to align with methodologies employed in related literature on digital heritage research.

4.3.2. Pipeline Architecture and Workflow

Prior to describing pipeline architecture details, key definitions are presented below to make this section more accessible for readers without technical background.

Orchestration System: Apache Airflow controls data flow in the pipeline, chaining different jobs together in logical order, enabling weekly runs and improving monitoring by tracking failure points with direct access to logs.

Job: A single unit of work in the orchestration system (e.g., extract data from cordis until the next checkpoint, deduplicate core publications).

Raw data: CSV, JSON, XML or other source files directly received from external sources.

ORM: Object Relational Mapping simplifies raw data ingestion into databases. An ORM Model may represent a cordis research project or an arxiv author with metadata. Additionally, this enables extension to various database systems when needed. Models map directly to database tables that must exist before ORM models can be used.

Session: A database connection synchronizing ORM models with the database, enabling batch operations and transaction management. Models added to a session receive database-assigned IDs and become database rows when a batch is committed.

Batch: A collection of multiple rows, documents or models for efficient database communication instead of processing individual items.

Namespaces: Separate database schemas for source tables (CORDIS, ArXiv) versus unified core tables. Logical separation exists between source tables like open aire projects, cordis projects and unified core projects.

Source tables: Raw data preserved in original schema according to ELT principle.

Core tables: Unified data model designed for scientometric analysis and further processing of CH and DH data.

4.4. Architecture Overview

Apache Airflow orchestrates our pipeline as a DAG (Directed Acyclic Graph), enabling weekly automated execution with dependency management and comprehensive monitoring. Each job follows defensive programming patterns: abort on error, rollback to last checkpoint, and retry after fix. Centralized logging and configuration files manage source queries adaptively, allowing for flexible adjustment as needed. Figure 2 provides an overview of the pipeline.

4.4.1. Extract Phase

Initial extraction fetches and stores raw data from sources using configurable queries. Each source employs checkpoint-based incremental extraction, using various markers like total sorted hits, latest mentioned dates, or file modification timestamps to retrieve only new data. Special handling exists for in-place updates (e.g., CORDIS projects get updated over time, new ROs get added, project status changes). Basic transformations ensure consistent formatting for dates and other structured fields.

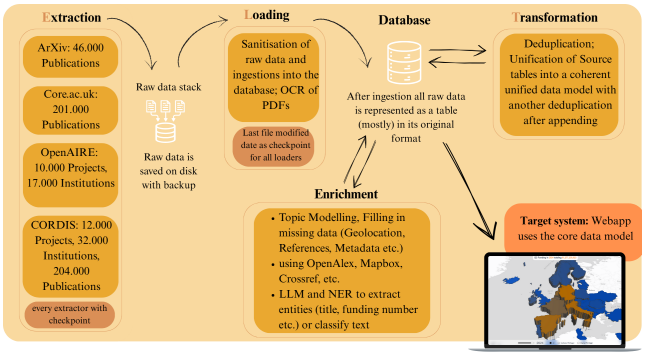

Figure 2: *Schematic overview of the pipeline work flow. Depicted tasks are orchestrated with Airflow.*

4.4.2. Raw Data Storage

All extracted data undergoes preservation on our university cluster with redundant backup systems. A dedicated file analysis module examines structure to determine file types, sizes, and counts for comprehensive overview. Key pattern extraction from JSON/XML aids schema understanding, critical for creating appropriate ORM translators.

4.4.3. Loading Phase

During loading, raw data undergoes sanitization and ingestion into database source tables. Processing begins with a database session that transforms raw files first into universal dictionaries, then into SQLAlchemy ORM models via an efficient file walker generator. Memory efficiency comes from yielding documents sequentially rather than loading all simultaneously, ensuring scalability for large document sets though preventing timestamp-based sorting.

Batch processing and committing mechanisms (n documents per batch) optimize database communication. Each source requires a custom ORM Translator that extracts dictionary keys/values and maps them to appropriate ORM models matching source schemas. A get-or-create pattern searches for existing rows using unique identifiers like titles or original IDs, preventing within-source duplicates.

Checkpoint mechanisms use file modification timestamps (UNIX epoch) to ensure later data loading processes only handle new or modified documents. Notable limitation: checkpoints can only be stored after process completion since documents aren't sorted by modification date. On error, entire sessions roll back for data integrity.

PDF processing employs concurrent OCR for documents lacking metadata, extracting full text for analysis. While multiple ORM libraries could be used, PyPDF2 has performed well even with complex 2-column documents.

4.4.4. Data Normalization

Every field undergoes careful type casting and normalization:

- Null-safe type conversion: All parsers defensively convert inputs

to target types (bool, int, float, string), returning None for invalid inputs
- Core string normalization: Leading/trailing whitespace removal, control character handling, space normalization
- Name preservation: Special characters (hyphens, apostrophes, periods) remain intact while normalizing spacing
- Content structure: Paragraph breaks maintained, internal spacing normalized, excess line breaks removed
- URL preservation: Minimal cleaning applied to maintain exact formats
- Date standardization: ISO format parsing with timezone handling
- Geolocation extraction: Support for various coordinate pair formats

4.4.5. Performance Optimization

Database performance relies heavily on indexing strategies:

- B-tree indexes accelerate exact/range searches (IDs, dates, names)
- GIN indexes enable powerful text search capabilities
- Trigram indexes facilitate fuzzy matching and duplicate detection

4.4.6. Source to Core Transformation

Moving from source tables to our core model involves SQL-based transformation with incremental processing using checkpoint timestamps. ON CONFLICT handlers manage updates to existing records, while entity resolution patterns handle related objects (people, topics, links). For safety, batch processing uses individual transactions.

4.4.7. Deduplication Strategy

Deduplication occurs at multiple levels:

- Source-level: Within each source via get-or-create patterns
- Core-level: Multi-pass approach for research outputs

First-pass matching uses exact DOI comparison, identifying 3,112 papers with reused DOIs. Second-pass processing employs fuzzy matching on normalized titles using trigram indexes and PostgreSQL similarity functions (0.9 threshold). Materialized views pre-process normalized titles and lengths, comparing only records within 5 characters difference. Our strategy identified 73,690 duplicate publications across sources.

Notable exclusions: 71,959 CORDIS documents remain unprocessed—50,026 PDFs requiring bibliometric enrichment for title extraction, plus 21,933 documents with generic titles. Final dataset currently comprises 374,998 research outputs, including these unprocessed documents. After deduplication.

Concerning projects, our pipeline integrated 12,792 CORDIS and 10,263 OpenAIRE projects for a total of 23,055 projects before deduplication. Using similar fuzzy matching methods as with research outputs, 3,497 duplicate projects were identified, resulting in 19,558 unique projects in the core data model. For institutions, 32,518 CORDIS and 17,470 OpenAIRE entities yielded 49,988 institutions total, with 4,263 duplicates detected and removed, leaving 45,725 deduplicated institutions in the core model.

4.4.8. Enrichment Pipeline

Missing metadata undergoes extraction through:

- LLM-based processing for titles, dates, and funding numbers (tested on samples)
- API integrations: OpenAlex (citations, impact), MapBox (geolocation), CrossRef (DOI metadata)

As case example, CORDIS had 6,875 institutions without geolocation data, with 6,647 locations successfully retrieved using MapBox API based on metadata like city, address, and legal name. Some source data contained inaccuracies (e.g., Dresden University mislabeled in Hamburg). Boolean tracking fields prevent redundant enrichment, while batch processing ensures API efficiency.

4.4.9. Error Handling and Orchestration

Robust error handling includes checkpoint mechanisms across all processes, transaction-based operations with rollback capability, defensive programming patterns (abort-fix-retry), and comprehensive logging. Weekly automated execution relies on Apache Airflow for coherent workflow orchestration with clear dependencies:

Extract individual sources → Analyze file contents → Load into database → Transform into core tables → Deduplicate → Enrich

Pipeline execution leverages our university's computing cluster with Slurm for job distribution, enabling efficient scaling and concurrency across processing nodes.

5. Results

In this the data is more closely examined, the Use Cases are visualized using the webapp and finally the Case Study RQs are answered.

5.1. Examining the data

Following deduplication, our pipeline generated a comprehensive dataset comprising 374,998 research outputs, 23,055 projects and 45,725 institutions. Metadata coverage varies significantly across source systems as shown in Tables 1 and 2.

Table 1: *Research Output Identifiers by Source System*

Source	Total ROs	DOI (%)
CORDIS	204,372	36.93
COREAC	201,116	32.43
ARXIV	46,312	16.94

Table 2: *Research Output Content Coverage by Source System*

Source	Language (%)	Abstract (%)	Full Text (%)
CORDIS	0.00	24.32	24.48
COREAC	56.83	85.76	99.41
ARXIV	0.00	100.00	95.84

Notable observations include COREAC's high full text coverage (99.41%) compared to CORDIS (24.48%) which soley comes from

the parsed PDFs, while only COREAC contains language code information across its collection. ARXIV provides complete abstract coverage but exhibits limited DOI availability (16.94%), leaving gaps for enrichment using CrossRef.

Beyond research outputs, our unified data model cataloged 1,980 distinct topics, 697,904 unique persons, 44,259 unique journals, and 891,010 links between entities. Many of these connections form essential networks for understanding collaboration patterns in CH and DH fields.

For effective analysis, missing metadata requires structured enrichment strategies, particularly for CORDIS documents where significant gaps exist in both DOI and abstract coverage. Such enrichment would enhance capabilities for advanced scientometric analysis of both CH and DH fields.

5.2. Discussion of the Webapp prototype and the scenarios

In Chapter 4.2, we developed seven scientometric use cases to address the needs of researchers and policy makers in the cultural and digital heritage fields. These use cases range from mapping key actors and institutional contributions to analyzing funding patterns and impact assessment. The following web app prototype scenarios implement initial versions of several of these use cases, demonstrating current capabilities while highlighting areas for further development.

As of May 2025 the scenarios displayed in the prototype do not yet include research outputs, as scalability has to be ensured first. This means that only Cordis and OpenAire data is used in the application. The web app itself is developed in Next.js, react and uses the GPU-powered framework Deck.gl for visual data analysis of geographical data, which efficiently supports millions of data points. This enables a geographical view of the scenarios.

5.2.1. Interactive Filtering Capabilities

Our web application prototype offers filtering options to enable simple data exploration across multiple dimensions. Users can refine their analysis through the following parameters:

- **Geographical filtering** via country selection
- **Programme-based filtering** through funding programmes and codes
- **Hierarchical topic filtering** with main, secondary, and tertiary topics (currently representing a combination from Cordis and OpenAire without formal hierarchy)
- **Temporal filtering** by selecting specific project date ranges
- **Semantic filtering** using keyword search with logical AND operators across titles, acronyms, and objectives (enabling proxy field identification through queries like “cultural & heritage” or “digital & heritage”)
- **Institutional filtering** based on Small and Medium Enterprise (SME) status

5.2.2. Scenario 1: Basic Project Overview

The first scenario provides a geographical visualization of projects across Europe. All funded projects are displayed as interactive dots positioned at their coordinator’s location, with the dataset spanning from 1957 to 2030.

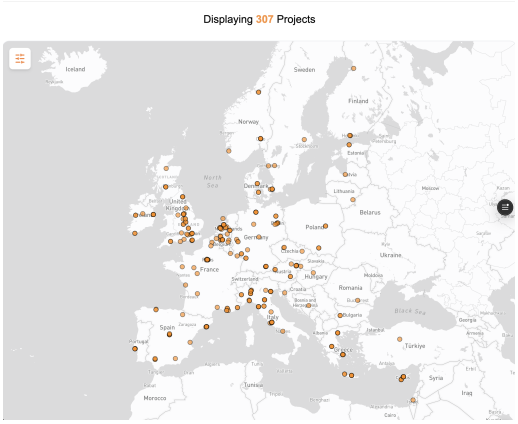

Figure 3: *Geographical distribution of DH projects across Europe, with each point representing a project’s coordinating institution.*

As shown in Figure 3, this visualization offers spatial insights into funding distribution patterns. The interface enables users to explore basic project metadata (title, acronym, duration, funding programme) and coordinator information through interactive elements. By combining keyword filtering with the geographical view, researchers can conduct preliminary exploration of topic distribution patterns.

This scenario partially addresses the "Analysing Topic structure in DH and emerging Trends" use case through basic keyword and date filtering capabilities. However, its utility is currently limited by the absence of robust topic classification systems and temporal analysis tools that would enable more sophisticated trend identification.

5.2.3. Scenario 2: Basic Institution Overview

The second scenario focuses on visualizing the institutional landscape, displaying all participating organizations as interactive elements on the map.

This visualization reveals the spatial distribution of research organizations involved in CH and DH projects. Users can access basic institutional metadata and filter organizations by SME status, facilitating the identification of key institutional actors by location and name.

The scenario supports two primary use cases from our framework: "Understanding institutional contributions" and "Identifying institutions with successful track records" by enabling geographical analysis of institutional participation. However, several limitations constrain its current utility: the absence of keyword filtering for institution-specific projects, lack of topic classification to reveal institutional research strengths, limited metadata regarding specialization or departmental structure, and missing connections between institutions and their research output impact metrics.

5.2.4. Scenario 3: Funding Visualization

Financial distribution analysis forms the core of our third scenario, which displays funding allocations as variable-height columns on the geographical map. Each column's height reflects the relative amount of funding received by the entity.

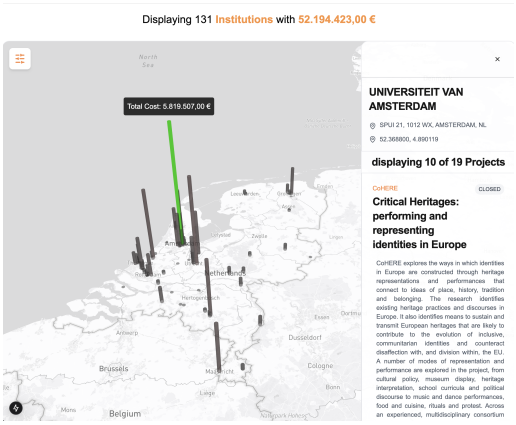

Figure 4: *Example of Cultural Heritage funding visualization in the Netherlands, with column heights representing funding amounts.*

Looking at Figure 4, users gain a spatial understanding of funding distribution across Europe. A key feature allows switching between project-level and institution-level funding views. For project visualization, columns appear at collaborator locations with heights representing total project funding. When viewing institutional funding, columns show aggregate funding received across all projects by each organization. Such representation facilitates identification of funding hotspots and regional disparities.

Within our scientometric framework, this scenario addresses two important use cases: "Visualising Funding Patterns by Projects and Institutions" and "Assessing impact of funding programmes" through geographical funding analysis.

Key limitations affect comprehensive analysis in this scenario. Many projects lack individual institution funding breakdowns, excluding them from institution-level views. An overrepresentation issue emerges in France, where numerous projects are associated with a single entity (Centre National de la Recherche Scientifique CNRS), creating visual distortion. Moreover, the scenario lacks topic-based funding analysis for field/subfield comparisons and misses connections between funding and research outputs or impact metrics.

5.2.5. Scenario 4: Collaboration Analysis

For analyzing institutional partnerships, this scenario displays organizations as vertical columns on the map, with heights representing the deduplicated count of collaborators across all projects.

Geographical analysis becomes possible through this visualization, which highlights institutional collaboration density and reveals spatial patterns of partnership through arc representations.

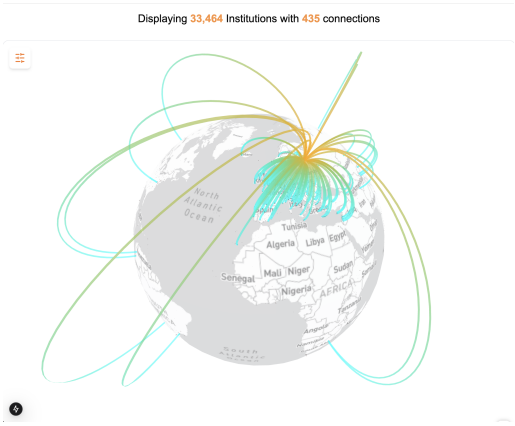

Figure 5: *Visualization of collaboration network density for University of Riga, showing connection arcs to partner institutions.*

Height variations of the columns visually encode the volume of unique collaborators for each institution.

Key use cases addressed include "Discovering collaboration networks between institutions" and "Understanding institutional contributions" by mapping existing partnership structures. However, current implementation faces several constraints: absence of keyword filtering for institutional projects, reliance on raw collaboration counts without quality or impact metrics, missing field-level or topic-based filtering capabilities for more granular analysis, and insufficient network visualization tools for analyzing multiple clusters or input-output relationships. Additionally, missing connections to research outputs and impact metrics limit comprehensive evaluation of collaboration effectiveness.

5.3. Analysis of EC Funding Trends per Country, Year and Topic

A fifth scenario implements a basic funding trend analysis that addresses both the "Visualising Funding Patterns" and "Analysing Topic structure and emerging Trends" use cases established in Chapter 4.2. Multi-dimensional analysis of funding distribution across geographical, temporal, and thematic dimensions is enabled by this scenario, providing insights for policy makers and researchers interested in funding priorities and shifts over from 2015 to 2025.

Offered are three key visualization modes: country-level funding extrusions adjustable by year (2015-2025), field/subfield/topic rankings per selected year, and country rankings for selected fields. By using the proxy field filtering for the keywords "cultural heritage" and "digital heritage", this scenario enables a comparative analysis of these fields funding trajectories. However, to effectively use the EuroSciVoc topics only Cordis data was used in this scenario. Downloaded as a JSON file can be the displayed data.

5.3.1. Cultural vs. Digital Heritage Funding (2015-2024)

Our analysis reveals divergent funding trajectories between CH and DH fields. While both fields experienced substantial growth, DH

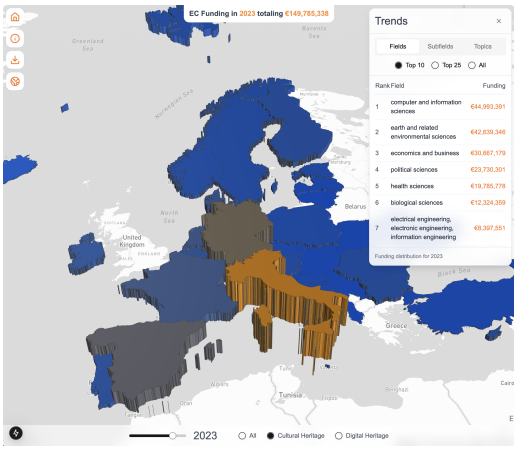

Figure 6: *Visualization of funding trends for CH projects across European countries for 2023.*

funding increased dramatically by 355% (€20M in 2015 to €91M in 2024), significantly outpacing Cultural Heritage's 137% growth (€76M to €180M) during the same period. An increasing prioritization of digital approaches within heritage research is reflected by this acceleration.

Italy maintains a dominant position in both fields, consistently leading funding acquisition in CH (ranking first in 8 of 10 years) and frequently leading in Digital Heritage as well. United Kingdom shows a notable decline in both fields following Brexit, with CH funding dropping from €23M in 2016 to €7M in 2024. Meanwhile, Spain has emerged as a major recipient, particularly in DH with a remarkable 20-fold increase from €0.6M in 2015 to €12M in 2024.

5.3.2. Field-Level Analysis

Disciplinary analysis reveals distinct priorities between the fields. Computer Science dominates DH funding (ranking first in 7 of 10 years and averaging approximately 60% of funding), indicating the field's strong technological orientation. In contrast, CH maintains a broader disciplinary distribution with Computer Science leading in 5 of 10 years but sharing prominence with Political Sciences and Economics/Business in different periods.

A notable trend across both fields is the dramatic growth of Economics/Business funding, particularly from 2020-2024, suggesting increasing emphasis on commercialization and business applications. Even more striking is the emergence of Artificial Intelligence as the top-funded subfield in both fields by 2024, despite having minimal or no funding in 2015-2016.

5.3.3. Limitations and Data Challenges

Effectively demonstrated is the pipeline's capability for multi-dimensional trend analysis, but this also highlights several methodological challenges in scientometric analysis of these fields:

- Keyword-based field filtering provides only an approximation without proper topic modeling
- Projects with multiple topic classifications can result in funding being counted multiple times
- EuroSciVoc's hierarchical classification (with 6 levels) requires careful mapping decisions
- Over 10,000 OpenAIRE projects contain relevant data but use incompatible topic classification
- Institution-level funding data granularity is limited to 2015 onwards in CORDIS

Directly connected to our Main Research Question are these limitations by demonstrating both the current capabilities and remaining challenges in developing a scientometrically useful data pipeline for CH and DH research.

6. Discussion

Analysis of our preliminary findings reveals expected limitations in the current implementation. While basic geographical and funding distributions demonstrate proof-of-concept for the pipeline's functionality, deeper scientometric insights remain constrained by insufficient data enrichment. Most critically, our results highlight a substantial need for robust topic modeling systems capable of accurately classifying projects and research outputs into hierarchical field structures. Such classification would enable more sophisticated analysis of CH and DH subject matter evolution, institutional specializations, and funding patterns across subfields. Furthermore, connections between projects, institutions, and research outputs require additional enrichment through citation metrics and impact data to support meaningful scientometric evaluation. All these requirements directly address our main research question by demonstrating what enrichment and modeling processes are necessary to transform raw integrated data into actionable scientometric insights for both fields.

6.1. Limitations and Future Directions

Our analysis reveals several key limitations of the current pipeline and dataset that must be addressed in future research. Most prominently, the dataset exhibits a strong European Union centricity, making findings primarily applicable to EU research and policy makers rather than reflecting global trends in CH and DH. Broader international data inclusion would significantly enhance the representativeness of the analysis.

Furthermore, focusing primarily on international EU projects creates an inherent bias toward cross-border collaboration while potentially missing important local and national research trends. National projects could be incorporated, although data fragmentation and inconsistent funding information present significant challenges for integration.

Concerning data quality, our current collection represents a superset that likely contains relevant CH and DH content mixed with substantial irrelevant material requiring better filtering mechanisms. Keyword-based approaches for publications and projects offer immediate improvement possibilities, but robust topic modeling represents a critical next step.

Fields (CH / DH) and specialized topics currently lack coherent organization, with identical concepts often appearing under different terminology across sources, time periods, and geographical

contexts. Adopting established ontologies from sources like OpenAlex or UNESCO, combined with assigning classifications via Large Language Models or keyword significance measures could address this issue. Methods such as Latent Dirichlet Allocation (LDA) could discover abstract topics across the corpus. Particularly important is recognizing that collaboration networks and most analyses fundamentally depend on well-defined topic structures.

Major limitations in the current prototype implementation include:

- Absence of research outputs with impact metrics
- Lack of sophisticated topic modeling
- Missing temporal analysis capabilities
- Insufficient collaboration modeling requiring comprehensive redesign

Looking forward, several promising directions exist for pipeline enhancement:

- Additional source integration, particularly Web of Science (frequently referenced in related literature)
- Enhanced data enrichment through:
- Addressing missing department-level data for institutions, potentially leveraging CrossRef information from research outputs
- Incorporating researcher feedback on prototype functionality and usability

Broadly speaking, scientometric approaches face inherent limitations regarding representation, as major search engines frequently fail to index research from developing nations, potentially marginalizing minority perspectives. Since our pipeline relies heavily on these same search engines, it likely inherits similar biases that must be acknowledged when interpreting results.

7. Conclusion

Our research has established essential infrastructure for scientometric analysis of CH and DH. First, we defined comprehensive use cases for scientometric analysis tailored specifically for policy makers and researchers, addressing critical gaps in understanding these multidisciplinary fields. Building on these requirements, we identified and integrated relevant data sources that collectively capture the complex landscape of heritage research.

Contributions include a robust data pipeline that extracts and sanitizes multiple heterogeneous sources, unifying them into a coherent analytical framework. After deduplication, our dataset comprises approximately 374,000 research outputs, 23,055 projects, and 45,725 institutions with associated metadata in CH and DH. Enrichment processes using various APIs and libraries enhanced data quality, particularly for geographical information where we successfully retrieved missing geolocation data for over 6,600 institutions.

Geographical visualizations implemented in our web application prototype demonstrate practical applications of the use cases, enabling spatial analysis of funding distribution, institutional contributions, and collaboration networks. Most notably, our analysis of EC funding trends reveals significant patterns: DH funding increased by 355% compared to CH's 137% growth between 2015-2024, with Computer Science maintaining dominance in DH while CH preserves broader disciplinary distribution.

Future work must focus on developing sophisticated topic modeling, enabling more nuanced analysis of collaboration networks and disciplinary trends. By establishing this foundation, our pipeline creates new possibilities for evidence-based decision making in heritage research and policy development.

References

[AC*24] ARIZA-COLPAS P. P., ET AL.: Sustainability in hybrid technologies for heritage preservation: A scientometric study. *Sustainability 16*, 5 (2024), 1991. URL: https://www.mdpi.com/2071-1050/16/5/1991, doi:10.3390/su16051991. 2

[AHG*22] ALMEVIK G., HANNES A., GOLFOMITSOU S., KARATAS K., LINDBLAD L., MARTIGNONI S., MIGNOSA A., VIO T.: *Who is not a stakeholder in cultural heritage?* Work package 4 report, CHARTER Consortium, 2022. URL: https://charter-alliance.eu/wp-content/uploads/2022/10/D4.1Report-Who-is-not-a-stakeholder-in-cultural-heritage-_V2.pdf. 2

[ArX24] ARXIV: Arxiv.org e-print archive, 2024. URL: https://arxiv.org/. 4

[Com24] COMMUNITY RESEARCH AND DEVELOPMENT INFORMATION SERVICE: Cordis - eu research results, 2024. URL: https://cordis.europa.eu/. 4

[COR24] CORE: Core - aggregating the world's open access research papers, 2024. URL: https://core.ac.uk/about. 4

[Kok24] KOKKALA V.: *Digital technologies for cultural heritage: An exploratory study of EU-funded research projects*. Master's thesis, University of Borås, 2024. URL: https://www.diva-portal.org/smash/get/diva2:1930503/FULLTEXT01.pdf. 2

[MUUA21] MÜNSTER S., UTESCHER R., ULUTAS AYDOGAN S.: Digital topics on cultural heritage investigated: how can data-driven and data-guided methods support to identify current topics and trends in digital heritage? *Built Heritage 5* (2021), 25. URL: https://link.springer.com/article/10.1186/s43238-021-00045-7, doi:10.1186/s43238-021-00045-7. 2

[Mü17] MÜNSTER S.: A survey on topics, researchers and cultures in the field of digital heritage. *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences IV-2/W2* (2017), 157–162. doi:10.5194/isprs-annals-IV-2-W2-157-2017. 2

[Ope24] OPENAIRE: Openaire - open access infrastructure for research in europe, 2024. URL: https://www.openaire.eu/. 4

[SA22] SINGHAL B., AGGARWAL A.: Etl, elt and reverse etl: A business case study. In *2022 Second International Conference on Advanced Technologies in Intelligent Control, Environment, Computing and Communication Engineering (ICATIECE)* (2022), IEEE, pp. 1–5. doi:10.1109/ICATIECE56365.2022.10046997. 4

[UNE09] UNESCO INSTITUTE FOR STATISTICS: *2009 UNESCO Framework for Cultural Statistics*. Tech. rep., UNESCO, 2009. URL: https://uis.unesco.org/en/glossary-term/cultural-heritage. 1

[UNE20] UNESCO: Concept of digital heritage, 2020. URL: https://webarchive.unesco.org/web/20230616073538/https://en.unesco.org/themes/information-preservation/digital-heritage/concept-digital-heritage. 1

[UNE23] UNESCO: World heritage, 2023. URL: https://www.unesco.org/en/world-heritage. 1