Ελλάδα 2.0
ΕΘΝΙΚΟ ΣΧΕΔΙΟ ΑΝΑΚΑΜΨΗΣ ΚΑΙ ΑΝΘΕΚΤΙΚΟΤΗΤΑΣ

Με τη χρηματοδότηση της Ευρωπαϊκής Ένωσης
NextGenerationEU

Call SUB 1.1 Research Excellence Partnerships

HAR.S.H: Hardware-Aware extReme-scale Similarity search

Proposal ID ΥΠ3ΤΑ-0560901

Deliverable D1.1

Data Management Plan

June 2026

Table Of Contents

D1.1 - 1 Introduction

D1.1 - 1.1 Information about the project

HAR.S.H aims to address major challenges in the large-scale processing of time series collections derived from real-world applications. Large-scale time series data collections are now present in nearly every scientific and societal domain. HAR.S.H will design and implement an extensive suite of algorithms, data structures, and mechanisms to tackle the scalability problem in analyzing vast volumes of time series data, leveraging modern and emerging hardware technologies. The algorithms, data structures, and mechanisms developed will form a robust library, ensuring their easy and efficient use across a wide range of applications. Specifically, HAR.S.H aims to:

- design and develop a new generation of algorithms and data structures that enable efficient parallel/distributed similarity search in large time series collections,
- leverage modern hardware technologies by studying their impact on the performance and scalability of such software,
- enable analysis on multimodal data, including text, images, and video, through integrations using deep learning models.

Pilot Applications - HAR.S.H will demonstrate the value of the technology it produces through the following three pilot applications:

- **Pilot Application 1 –** Similar document and file search. This application focuses on finding similar documents within large-scale document databases.
- **Pilot Application 2 –** Photo analysis for enhanced travel profiles. This application focuses on analyzing photographic content to enrich travel profiles for personalized travel recommendation systems. The main goal is to maximize visitor satisfaction with a travel destination.
- **Pilot Application 3 –** Public opinion analysis application. This application aims to manage the complexity and cost involved in monitoring and categorizing content from social media, providing valuable insights into public opinion across various contexts.

To meet the needs of the above applications, HAR.S.H will innovate in the following areas:

- **Compact and descriptive representation of multimodal data**. HAR.S.H aims to develop efficient techniques for handling multimodal data, particularly images, video, and natural language. The project will explore various data sources especially images, video, and natural language text that can be integrated in an end-to-end manner.
- **Robust mechanisms for high-performance processing of large-scale time series collections**. HAR.S.H will enhance performance and robustness in similarity query answering over large-scale time series collections by:
  1) leveraging the full computational power of modern platforms,
  2) developing processing mechanisms that are hardware-aware to minimize cost and enable fast parallel and distributed execution and,
  3) devising resilient techniques that support thread failures and allow rapid recovery of computation after system-wide crashes. HARSH will primarily focus on emerging technologies in memory, synchronization, and communication, and will investigate how exploiting such technologies can impact time series data processing.

D1.1 - 1.2 Data Management Plan – Principles and Goals

The Data Management Plan of HAR.S.H. aligns with the **FAIR** principles to promote open science and ensure public investment in research outcome. The **FAIR model** for Data Management Plans (DMPs) refers to a set of guiding principles that aim to make data **Findable, Accessible, Interoperable, and Reusable**. These principles were first introduced in 2016 by Wilkinson et al. [1], and they are widely adopted in scientific research, especially in contexts that emphasize **open science** and **data stewardship**.

When applied to **Data Management Plans**, the FAIR model helps ensure that data collected or generated during research is managed in a way that maximizes its long-term value and usability.

The FAIR Principles Breakdown are described below with indicative features that deliver the necessary properties:

1. Findable

- Data and metadata are assigned a **globally unique and persistent identifier** (e.g., DOI).
- Data are described with **rich metadata**.
- Metadata clearly and explicitly include the identifier of the data they describe.
- Data and metadata are **registered or indexed** in a searchable resource (like a repository or catalog).

2. Accessible

- Data are retrievable by their identifier using a **standardized communication protocol** (e.g., HTTP, FTP).
    - The protocol is open, free, and universally implementable.
    - The protocol allows for **authentication and authorization**, if necessary, according to data sharing and exchange rules
- Metadata remain accessible even when the data are no longer available.

3. Interoperable

- Data use a formal, accessible, shared, and broadly applicable **language** for knowledge representation (e.g., RDF, XML, JSON).
- Data use vocabularies that follow FAIR principles.
- Data include **qualified references to other data** (e.g., links to related datasets or publications).

4. Reusable

- Data are richly described with a **plurality of accurate and relevant attributes**.
    - Data are released with a **clear and accessible data usage license** and **use agreement**
    - Data are associated with **detailed provenance**.
    - Data meet domain-relevant **community standards**.

The abovementioned principles that align to the FAIR model will ensure that HARSH data:

- **Can be discovered** and cited by others (Findable)
- **Can be accessed** and retrieved by others with clear conditions (Accessible)
- **Can be combined and used** with other datasets (Interoperable)
- **Can be reused**, even long after the project ends (Reusable)

The purpose of the Data Management Plan (DMP) is to provide an analysis of the main elements of the data management policy that will be used by the partners with regard to all the datasets that will be collected and generated by the project. The DMP contributes to save time and effort, makes research process easier, helps to validate if the necessary support is considered, and enables to make sound decisions.

The DMP supports project partners to:

- understand the data and use it when needed; describe the handling of research data during and after the end of the project; how data will be shared or made open, how will be curated or preserved
- ensure continuity if project staff leave or new researchers join;
- avoid unnecessary duplication e.g. re-collecting or re-working data;
- contribute to more collaboration and advances research;
- increase visibility and impact;
- manage citations of other researchers on the project's data
- describe methodology and standards where they are applied

The document also provides the dataset templates for reporting and updating purposes, which shall be used to monitor the data set collected or generated during the project. The information updates will be included in future versions of the deliverable.

The present DMP has been developed based on the EC guidelines on data management in Horizon 2020 [2], the guidelines from Digital Curation Centre (DCC) [3], the Open Access (Open Science) policy lines from the EC.

D1.1 - 2 Internal and Public Communication Tools

The official means for communication between the members of HAR.S.H. are the Zoom conference tool and the publicly available Microsoft Teams platform.

HAR.S.H. members participate in teleconferences which are performed using Zoom [4]. Zoom offers free web conferencing, supporting screen and desktop sharing. Specifically, Zoom supports high-definition video with integrated audio, real-time content sharing (e.g. documents, agendas, notes), recordings, the possibility of participation to meetings through mobile devices (e.g. cellular phones), and security and strict access control policies.

Sometimes, instant messaging tools may be used as the fastest way to communicate for discussing specific topics. Teams [5] is probably one of the most widely used applications that provides instant messaging facilities, combined with voice and video support. It also provides file sharing and screen sharing. HAR.S.H. members may use Teams or other instant messaging applications to communicate with each other on specific topics.

*Meetings can be recorded **only** after being notified by all participants and upon their prior approval given by them after they are notified; recorded material will be available on the file sharing mechanism where meeting minutes, presentations and related material is kept.*

D1.1 - 2.1 File Sharing

In terms of file sharing, services such as Github and Microsoft OneDrive are utilized.

Shared or public source code is/will be stored in Github [6]. Github is a version control system, so it provides features to access and examine previous versions of the material uploaded on it, update to previous versions of files and directories, as well as of the metadata that accompanies them. With a simple interface, the Github repository can be checked out and existing working copies can be restored exactly as they were at any date in the past.

Sharing of presentation slides, documents, images and shared source code among the partners will be stored in the cloud using the Microsoft OneDrive [7] service. Cloud storage space is provided to the project through the OneDrive service under the Office365 tool suite.

The use of these repositories will ensure the long-term availability of the material even after the end of the project.

D1.1 - 2.2 HARSH Web Page and Social Networks

An Internet web page has been developed for HARSH. Its main goal is to diffuse the HARSH objectives and results as widely as possible, including public deliverables and open access resources, throughout interesting stakeholders and beyond. This web site can be found at: http://harsh-project.eu/ .

Figure 1: HAR.S.H. Web Page

Figure 1 shows the initial page of the web site of HARSH. All deliverables of type PUBLIC that will be produced in the context of the project will be uploaded on the web site. Additionally, the web site provides access to press releases, presentations, videos, posters, and other material related to the project.

HARSH has a Facebook page: https://www.facebook.com/harshproject

Figure 2: HAR.S.H. Facebook Page

HARSH has presence at LinkedIn: https://www.linkedin.com/company/108218606/ .

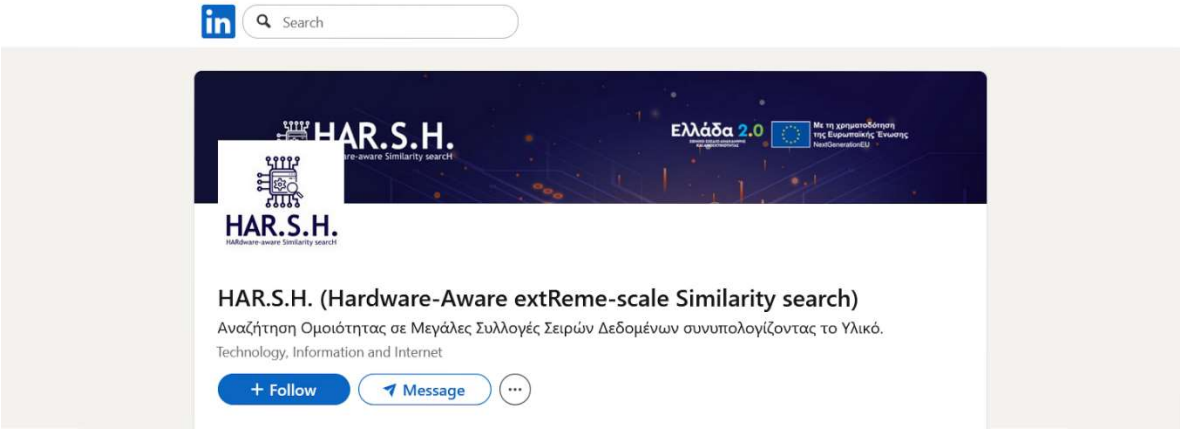

Figure 3: HAR.S.H. LinkedIn Group

D1.1 - 3 Data Management in HAR.S.H.

The following categories of datasets are identified within the HAR.S.H. project:

- **Evaluation datasets:** Collections of data series used for the evaluation of the implementations and for the validation of the results described in the publications.
- **Scientific publications:** Papers that describe the research work within HAR.S.H.
- **Other material:** Leaflets, documentation, dissemination material, deliverables, slides, etc.

This section provides a description of the data that will be collected and/or generated during the project for the relevant work package/task along the progress stages of the project. The description also includes information about the type/format, volume, ownership of the data, etc. The following tables present the data sets.

D1.1 - 3.1 Evaluation Datasets

In this section, we present the datasets (synthetic and real) that will be used in the HAR.S.H. project.

Data Set Name	HARSH_Random_Dataset
Data Set Type	Collection of randomly generated data series.
Data Set Description	This is a set of synthetic datasets with sizes from 50 million to 200 million data series composed of random walks of length 256. Each data point in the data series is produced as $x_i + 1 = N(x_i,1)$, where N (0,1) is a standard normal distribution.
Reuse & Sharing	The synthetic data generator code is included in the source code we will make available. Thus, it will be possible to be used by individuals for verification, further analysis and publications.
Ownership / Intellectual Property	Not applicable
Scope (WP & Task)	WP 3 Task 3.2 3.3
Volume	10GB – 300GB
Channel/Medium	Local File System
Usage	Hardware-aware and robust big-data analysis, similarity search
Confidentiality	Not applicable
Standards and Metadata	Binary files containing float numbers
Storage and Backup	The synthetic data generator code is included in the source code we will make available.
Security / Privacy	Not applicable
Contact	Themis Palpanas, themis@mi.parisdescartes.fr

Data Set Name	HARSH_Seismic_Dataset
Data Set Type	Publicly available data series collection describing seismic activity.
Data Set Description	The seismic dataset, Seismic, was obtained from the IRIS Seismic Data Access archive [8]. It contains seismic instrument recordings from thousands of stations worldwide and consists of 100 million data series of size 256.
Reuse & Sharing	The data are currently (publicly) available at: https://ln5.sync.com/dl/0b8135230/39vxx8su-tkfi7t2s-dgsvh8rp-k8ixcs8p?sync_id=12175200430004
Ownership / Intellectual Property	Not applicable
Scope (WP & Task)	WP 3 Task 3.2 3.3
Volume	100GB

Channel/Medium	Local File System
Usage	Local File System
Confidentiality	Hardware-aware and robust big-data analysis, similarity search
Standards and Metadata	Binary files containing float numbers
Storage and Backup	Storage in local file system Backup in Sync.com cloud storage: https://ln5.sync.com/dl/0b8135230/39vxx8su-tkfi7t2s-dgsvh8rp-k8ixcs8p?sync_id=12175200410004
Security / Privacy	Not applicable (data publicly available)
Contact	Themis Palpanas, themis@mi.parisdescartes.fr

Data Set Name	HARSH_Astro_Dataset
Data Set Type	Publicly available data series collection representing celestial objects.
Data Set Description	The astronomy dataset, Astro, represents celestial objects and was obtained from [9]. The dataset consists of 100 million data series of size 256.
Reuse & Sharing	The data are currently (publicly) available at: https://ln5.sync.com/dl/0b8135230/39vxx8su-tkfi7t2s-dgsvh8rp-k8ixcs8p?sync_id=12175200410004
Ownership / Intellectual Property	Not applicable
Scope (WP & Task)	WP 3 Task 3.2 3.3
Volume	265GB
Channel/Medium	Local File System
Usage	Hardware-aware and robust big-data analysis, similarity search
Confidentiality	Not applicable
Standards and Metadata	Binary files containing float numbers
Storage and Backup	Storage in local file system Backup in Sync.com cloud storage: https://ln5.sync.com/dl/0b8135230/39vxx8su-tkfi7t2s-dgsvh8rp-k8ixcs8p?sync_id=12175200410004
Security / Privacy	Not applicable (data publicly available)
Contact	Themis Palpanas, themis@mi.parisdescartes.fr

Data Set Name	HARSH_Judicial Data

Data Set Type	• Criminal decisions • Minutes of meetings • Criminal courts • Civil decisions • Commercial courts
Data Set Description	Archive of the Three-Member Court of Appeal of Nafplio - Criminal decisions Consists of 60 volumes of proceedings and criminal decisions of adjudication of criminal acts concerning thefts, embezzlements, crimes against property and property, bodily harm, homicides, etc. 1985-1989 Archive of the Three-Member Court of Appeal of Nafplio for Minors 1948 -1993 4 volumes - Consists of 4 volumes of criminal decisions of adjudication of criminal acts concerning juveniles concerning thefts, bodily harm, homicides, etc. Archive of the Five-Member Court of Appeal of Nafplio (Criminal) Consists of 30 volumes of proceedings. It contains criminal decisions, minutes, indexes, provisions for the trial of criminal acts relating to fraud, forgery, embezzlement, bribery, appropriations against the public, etc. The archive is classified into 4 sub-archives: 1) Three-member, 2) Five-member, 3) Seven-member, 4) Parliamentary Department. The first three sub-archives are classified into series: 1) Decisions, 2) Minutes, 3) Indexes, 4) Minorities, while the Parliamentary Department sub-archive is classified into two series: 1) Parliamentary, 2) Provisions. 1846-1959 Corinth Criminal Court Archive Contains decisions and minutes of sessions of the Corinth Criminal Court adjudicating criminal acts related to thefts, homicides, bodily harm, etc. 1903- 1959 178 volumes Athens Court of Appeal Archive Political Decisions Consists of 3306 volumes concerning political decisions of appeals of civil disputes related to labor, inheritance and property issues, leases, evictions, unions, adoptions, divorces, alimony, exercise of parental responsibility, etc. 1837-1980 Archive of the Athens Court of First Instance Political decisions civil disputes Consists of 1816 volumes concerning political decisions of civil disputes regarding labor, inheritance and property issues, leases, evictions, unions, adoptions, divorces, alimony, exercise of parental responsibility, etc. 1835-1920 Heraklion Commercial Court Archive

	Contains political decisions and minutes of the Heraklion Commercial Court regarding lawsuits, sales, corporate and commercial agreements, transactions, bankruptcies, etc. and minutes of the Heraklion Misdemeanor Court 1882 - 1892 Tripoli Criminal Court Archive Consists of 365 volumes with decisions and minutes of meetings of the Tripoli Criminal Court adjudicating criminal acts regarding thefts, homicides, bodily harm, etc. 1835-1917 Archives of the Criminal Court of Kalamata Consists of 420 volumes with decisions and minutes of sessions of the Criminal Court of Kalamata adjudicating criminal acts related to thefts, homicides, bodily harm, etc. 1855 - 1940
Reuse & Sharing	Legal restrictions (personal/sensitive data or other) should be considered
Ownership / Intellectual Property	State document - Legal restrictions (personal/sensitive data or other)
Scope (WP & Task)	WP4, Task 4.1, Task 4.2
Volume	> 20 million pages (binded into volumes) of 500 kb each one. Each decision needs 10-20 pages (~ 1 million cases)
Channel/Medium	DB, EAD-XML data structures, Disk for Images, TXT local File System
Confidentiality	NDA in place between AMS and ΓΑΚ. Need to set an NDA between UoC and ΓΑΚ
Standards and Metadata	EAD XML (ISAD-G, ISARR) CPF... - Metadata per folder (authority, number of pages, and a general case history and timeline
Storage and Backup	ΓΑΚ data center. It will migrate to G-cloud. A staging server and repository should be considered
Contact	AMS, Andreas Tsigris

Data Set Name	HARSH_Scraped_News
Data Set Type	Collection of articles, scraped from various Greek news sites.
Data Set Description	This dataset was obtained by scraping the HTML code of publicly available articles of Greek News sites. It includes the title, the description, the full text and the thumbnail of each article, as well as some metadata such as author and publish date. Articles that are referenced in other articles are linked as relative.
Reuse & Sharing	The news scraped code is included in the source code we will make available. Thus, it will be possible to be used by individuals for verification, further analysis and publications.

Ownership / Intellectual Property	Greek news sites
Scope (WP & Task)	WP4, Task 4.1, Task 4.2
Volume	Hundreds of thousands of articles with their respective thumbnail images.
Channel/Medium	Greek news sites RSS feeds
Usage	Fine-tuning of embeddings and use in application of NEUROLINGO.
Confidentiality	Not applicable
Standards and Metadata	RSS feeds, XML & HTML The metadata that accompany the dataset consists of the id and the title and the link of each article, as well as the description, the keywords and the category given by the authors. Also, we keep the author's name, the site where the article was retrieved and its published time.
Storage and Backup	Local DB & file system.
Security / Privacy	Not applicable
Contact	Christos Tsalidis, tsalidis@neurolingo.gr Ioannis Stamatopoulos, stamatop@neurolingo.gr

Data Set Name	HARSH_Searchculture_2D_Graphics
Data Set Type	Collection of images, such as photographs, maps, engravings, scanned press, letters, wishing cards etc. with their metadata.
Data Set Description	This dataset was obtained by scraping the Searchculture.gr site and getting publicly available images considered as items in the Searchculture.gr repository along with their RDF metadata info.
Reuse & Sharing	The data are currently (publicly) available at: https://www.searchculture.gr/aggregator/portal/advancedSearch/search?portalSearches%5B0%5D.ektTypes=http%3A%2F%2Fsemantics.gr%2Fauthorities%2Fekt-item-types%2Fdisdiastata-grafika&_portalSearches%5B0%5D.ektTypes=1&_portalSearches%5B0%5D.strictPeriods=on&page.page=1&resultsMode=GRID&sortResults=SCORE

Ownership / Intellectual Property	https://www.searchculture.gr
Scope (WP & Task)	WP4, Task 4.1, Task 4.2
Volume	291.634 images with their RDF/XML.
Channel/Medium	https://www.searchculture.gr
Usage	Meeting the project's testing needs. Enriching Neurolingo technology for search in content repositories
Confidentiality	Not applicable
Standards and Metadata	Linked Data, RDF/XML, W3C, SKOS and Europeana Data Model (EDM) The dataset is accompanied by RDF/XML metadata that includes the title, contributor, description, and type of each item. It also provides information about the location where the item was found or depicted, as well as the time associated with it. Additionally, the metadata lists relevant subjects, along with the provider and the collection to which the item belongs.
Storage and Backup	Local DB & file system.
Security / Privacy	Not applicable
Contact	Christos Tsalidis, tsalidis@neurolingo.gr Ioannis Stamatopoulos, stamatop@neurolingo.gr

Data Set Name	HARSH_Wikidata_GeoNames_Images
Data Set Type	Collection of images related to geographical names.
Data Set Description	This dataset was obtained by scraping the HTML code of publicly available images related to the collection of geographic terms curated by the EKT found in Wikidata. The dataset is a collection of images from Wikidata along with their RDF metadata info for the related geographic terms from the EKT Vocabulary. Also, extra text content metadata is kept for each geographical name from https://www.wikivoyage.org/ with tourist info.
Reuse & Sharing	The data are currently (publicly) available at: https://www.wikidata.org https://www.semantics.gr/authorities/vocabularies/geonames-places-earth/vocabulary-entries https://www.wikivoyage.org/
Ownership / Intellectual Property	https://www.wikidata.org https://www.semantics.gr/authorities/vocabularies/geonames-places-earth/vocabulary-entries

	https://www.wikivoyage.org/
Scope (WP & Task)	WP4, Task 4.1, Task 4.2
Volume	13.768 geographical names of EKT Vocabulary with at least one image in Wikidata
Channel/Medium	https://www.wikidata.org https://www.semantics.gr/authorities/vocabularies/geonames-places-earth/vocabulary-entries
Usage	Meeting the project's testing needs. Enriching Neurolingo technology for search in content repositories (tourism info - mention of toponyms).
Confidentiality	Not applicable
Standards and Metadata	Linked Data, RDF/XML, W3C, SKOS The dataset includes metadata detailing the name of the geographical term depicted in each image, as well as the title of the image.
Storage and Backup	Local DB & file system.
Security / Privacy	Not applicable
Contact	Christos Tsalidis, tsalidis@neurolingo.gr Ioannis Stamatopoulos, stamatop@neurolingo.gr

Data Set Name	HARSH_Tourist_Thematic_Data (Open FlickR)
Data Set Type	A multimodal dataset composed primarily of image data enriched with structured metadata. It falls under the category of real-world, user-generated content, making it highly suitable for machine learning tasks that require diversity in visual and contextual information. The dataset includes high-resolution photographs along with associated textual annotations such as titles, tags, and user-generated captions. Additionally, many images contain geolocation data (latitude and longitude), timestamps, and usage rights (e.g., Creative Commons licenses). This combination of visual and textual information classifies the dataset as image-text paired data, ideal for image classification, image-text retrieval, multimodal embedding, and tourism theme profiling based on user-generated content.
Data Set Description	The Flickr dataset comprises user-contributed photographs from the Flickr platform, enriched with valuable metadata such as titles, tags, geolocation (latitude and longitude), capture dates, and user information. It is widely used in computer vision and multimedia research due to its diverse and real-world content spanning global locations and themes. Through the Flickr API, developers and researchers can access millions of images under varying Creative Commons

	licenses, allowing filtered searches based on keywords, themes (e.g., "Greece", "beach", "mountain"), and spatiotemporal parameters. This makes it an ideal source for building datasets related to tourism, cultural exploration, and content-based recommendation systems.
Reuse & Sharing	The reuse and sharing of Flickr data depend primarily on the license associated with each image, as designated by the content uploader. Flickr supports a range of licenses, including various types of Creative Commons (CC) licenses that allow for reuse under specific conditions. Many Flickr images are released under open licenses, such as: CC BY (Attribution required) CC BY-SA (Attribution + ShareAlike) CC0 (Public domain) These images can be: Used for training ML models Shared in datasets Included in academic publications (with attribution where needed)
Ownership / Intellectual Property	Ownership and Intellectual Property Rights (IPR) of Flickr data are governed by the following key principles: Photographs uploaded to Flickr remain the property of their original creators. The user (photographer) retains full copyright unless they explicitly release it under a Creative Commons or other license. All images are protected by copyright law by default. Flickr acts as a hosting platform, not the rights holder. Use of Flickr content in machine learning or publications requires compliance with the license selected by the uploader. License Type and the linked IPR Implications are mentioned below: All Rights Reserved - Cannot reuse or share without explicit permission from the creator Creative Commons (e.g. CC BY, CC0) - Can reuse under specified terms (e.g., attribution, non-commercial, share-alike) Public Domain (CC0) - Freely reusable without restriction
Scope (WP & Task)	WP4, Task 4.1, Task 4.2
Volume	13,000 images and 13,000 textual descriptions/captions, including 211,225 words and 22 different captions, with a total size of approximately 1.5GB.
Channel/Medium	https://tgmstat.medium.com/explore-the-flicker-8k-dataset-b9230715fd42

Confidentiality	Not applicable
Standards and Metadata	Flickr images are accompanied by metadata, and there are some standardized structures and vocabularies used in handling and interpreting this metadata — especially when integrating Flickr data into research or machine learning pipelines. Metadata Type and the relevant description is as below: EXIF (Exchangeable Image File Format) - Captured from the camera (date, time, location, device, exposure, etc.) User-Generated Metadata - Title, description, tags, safety level, content type System Metadata - Flickr ID, owner ID, upload date, license, visibility (public/private) Geolocation Dat - Latitude, longitude, accuracy Licensing Info License type (e.g., CC BY, CC0, all rights reserved) Standards in use are: EXIF, Standard camera metadata format embedded in images (ISO 12234) IPTC, Often used for professional photography (e.g. caption, author, location, copyright) Dublin Core, Used in some integrations for interoperability with digital libraries WHen storing Flickr image metadata, we will mirror EXIF + user-level metadata into a schema (e.g., JSON or database). The YFCC100M dataset includes over 100 million Flickr images with standardized metadata CSV and licensing field
Storage and Backup	Local DB & file system.
Security / Privacy	Not applicable
Contact	Vassilis Spitadakis, spitad@mcbs.gr Antonis Karagiannakis, akar@mcbs.gr

D1.1 - 3.2 Scientific Publications, posters, presentations, keynotes

During the period of time that HAR.S.H will run a number of papers (at least 5) will be produced. 6 posters will be created and presented in well-known venues. The table below provides information about these papers and posters and will be updated within the course of the project.

Type	Name	Description	Sharing	Archiving and Preservation

Poster	HARSH_poster_GEC_2025_1	Poster at GEC 2025 (Concurrent Binary Search Trees Supporting Split and Join)	Available on HAR.S.H. website for public access	https://zenodo.org/records/15772010
Poster	HARSH_poster_GEC_2025_2	Poster at GEC 2025 (Concurrent Double-Ended Priority Queues)	Available on HAR.S.H. website for public access	https://zenodo.org/records/15772127
Poster	HARSH_poster_40_years_uoc	Poster at 40 years UOC event	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2025/07/csd-40years-poster-A0-portrait-Fatourou.pdf
Paper	HARSH_paper_VIS2025	Publication accepted at VIS2025 (TiVy: Time Series Visual Summary for Scalable Visualization	Available on HAR.S.H. website for public access	https://helios2.mi.parisdescartes.fr/~themisp/publications/vis25-tivy.pdf
Paper	HARSH_paper_DISC2025	Publication accepted at DISC 2025 (PIPQ: A Strict Insert-Optimized Concurrent Priority Queue)	Available on HAR.S.H. website for public access	https://arxiv.org/pdf/2508.16023v1
Paper	HARSH_paper_VLDB2025	Publication accepted at VLDB 2025 (The LAW theorem: Local Reads and Linearizable Asynchronous Replication)	Available on HAR.S.H. website for public access	https://law-theorem.com/
Poster	HARSH_poster_LAW	Poster at VLDB 2025	Available on HAR.S.H. website for public access	https://law-theorem.com/
Poster	HARSH_poster_csd_reunion	Poster at CSD Classes of ‘85/’95/’05	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2025/10/HARSH_poster_csd_reunion.pdf

Paper	HARSH_paper_PPoPP25	Paper presented at PPoPP25 (Aggregating Funnels for Faster Fetch&Add and Queues)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2411.14420
Paper	HARSH_paper_DISC2025_BA	Paper presented at DISC 2025 (Concurrent Double-Ended Priority Queues)	Available on HAR.S.H. website for public access	https://arxiv.org/pdf/2508.13399
Paper	HARSH_paper_OPODIS2025	Paper presented at OPODIS 2025 (Recoverable LockFree Locks)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2512.09710
Paper	HARSH_paper_PPoPP2026_1	Paper accepted at PPoPP 2026 (Sharded Elimination and Combining for Highly-Efficient Concurrent Stacks)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2601.04523
Paper	HARSH_paper_PPoPP2026_2	Paper accepted at PPoPP 2026 (Concurrent Balanced Augmented Trees)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2601.05225
Paper	HARSH_paper_CVPR2026	Paper accepted at CVPR 2026 (The More, the Merrier: Contrastive Fusion for Higher-Order Multimodal Alignment)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2511.21331
Poster	HARSH_poster_SPTDC2026_1	Poster presented at SPTDC 2026 (The More, the Merrier: Contrastive Fusion for Higher-Order Multimodal Alignment)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/sptdc2026-merrier.pdf
Poster	HARSH_poster_SPTDC2026_2	Poster presented at SPTDC 2026 (PPMR: Persistent Polymorphic Memory Resources)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/sptdc2026-ppmr.pdf

Poster	HARSH_poster_SPTDC2026_3	Poster presented at SPTDC 2026 (ATROPOS: Benchmarking Concurrent Persistent Algorithms under Failures)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/sptdc2026-atropos.pdf
Poster	HARSH_poster_SPTDC2026_4	Poster presented at SPTDC 2026 (PFRESH: A persistent Lock Free Data Series Index)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/sptdc2026-pfresh.pdf
Poster	HARSH_poster_SPTDC2026_5	Poster presented at SPTDC 2026 (Concurrent Double-Ended Priority Queues)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/sptdc2026-depqs.pdf
Paper	HARSH_paper_ICML2026	Publication accepted at ICML 2026 (ParisKV: Fast and Drift-Robust KV-Cache Retrieval for Long-Context LLMs)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2602.07721
Journal	HARSH_journal_TKDE2026	Journal accepted at IEEE TKDE 2026 Journal (PDET-LSH: Scalable In-Memory Indexing for High-Dimensional Approximate Nearest Neighbor Search with Quality Guarantees)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2603.24920
Paper	HARSH_paper_PVLDB2025	Paper accepted at PVLDB 2025 (MS-Index: Fast Top-k Subsequence Search for Multivariate Time Series under Euclidean Distance)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2512.14723
Paper	HARSH_paper_KDD2025	Paper accepted at ACM SIGKDD 2025 (Evaluating and Generating Query Workloads for High	Available on HAR.S.H. website for public access	https://helios2.mi.parisdescartes.fr/~themisp/publications/kdd25-hephaestus.pdf

		Dimensional Vector Similarity Search)		
Journal	HARSH_journal_PACMMOD2026_1	Journal accepted at PACMMOD 2026 (TaCo: Data-adaptive and Query-aware Subspace Collision for High-dimensional Approximate Nearest Neighbor Search)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2603.24919
Journal	HARSH_journal_SIGMOD2026_2	Paper accepted at SIGMOD/PODS 2026 (DARTH: Declarative Recall Through Early Termination for Approximate Nearest Neighbor Search)	Available on HAR.S.H. website for public access	https://arxiv.org/abs/2505.19001

Table: Scientific Publications and Posters

The next table summarizes information on presentations given in the context of HAR.S.H..

Type	Name	Description	Sharing	Archiving and Preservation
Slides	HARSH_slides_mexico_summer_school	Course at DC Mexico Summer School	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2025/08/FatourouSlides-Mexico.pdf
Slides	HARSH_slides_schools	Presentation at Department of Computer Science, UoC School Visits	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/school_visits.pdf
Slides	HARSH_slides_researchersnight	Presentation at Researcher's Night Event	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/researcher.pdf
Slides	HARSH_slides_researchersnightchatlab	Presentation at Researcher's Night Event (ChatLab)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/chatlab.pdf
Slides	HARSH_slides_Elim	Presentation at PPoPP 2026 (Sharded Elimination and	Available on HAR.S.H.	https://harsh-project.eu/wp-

		Combining for Highly-Efficient Concurrent Stacks)	website for public access	content/uploads/2026/06/sec.pdf
Slides	HARSH_slides_AugBalTrees	Presentation at PPoPP 2026 (Concurrent Balanced Augmented Trees)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/AugBalTrees.pdf
Slides	HARSH_slides_DEPQ_SPTDC2026	Presentation at SPTDC 2026 (Concurrent Double Ended Priority Queues)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/DEPQ.pdf
Slides	HARSH_slides_DISC2025	Presentation at DISC 2026 (Concurrent Double Ended Priority Queues)	Available on HAR.S.H. website for public access	https://harsh-project.eu/wp-content/uploads/2026/06/DEPQ.pdf
Slides	HARSH_slides_Simons_Talk	Talk at Simons Institute (Combining and Aggregating for Fast Synchronization)	Publicly available on HAR.S.H. website	https://harsh-project.eu/wp-content/uploads/2026/06/CombiningAggregatingFunnels.pdf
Videp	HARSH_video_Simons_Talk1	Talk at Simons Institute (Parallel Data Series Indexing and Similarity Search on Modern Hardware)	Publicly available on Simons Institute	https://simons.berkeley.edu/talks/panagiota-fatourou-university-crete-forth-2025-11-07
Video	HARSH_video_Simons_Talk2	Talk at Simons Institute (Combining and Aggregating for Fast Synchronization)	Publicly available on YouTube	https://www.youtube.com/watch?v=8-eZIZfoEzQ

Table 1: HAR.S.H. slides and videos.

The above two tables are updated periodically in order to include new publications, posters and presentations given about the project.

Apart from the above material, HAR.S.H. deliverables are considered to be scientific reports of the work performed during the project. Several deliverables will be publicly available and posted on the HAR.SH. website. The following table lists all deliverables based on their dissemination level.

Type	**Deliverable**	**Dissemination Level**	**Sharing**
Deliverable	D1.1,D1.2, D1.3,D2.1, D2.2,D3.1,	PU	Available on HAR.S.H. website for public access.

	D3.2,D5.2, D6.1,D6.3, D6.4		
Deliverable	D4.1,D4.2, D5.1,D6.2	PR	

`Table 2: List of deliverables based on their dissemination level

The public deliverables generated by the project, as well as, the data linked to them will be shared via the following tools for dissemination:

- data archives;
- journals (self-archiving) and publications' data banks;
- available online on the project website and websites of the partners;
- available in social media at the project level;
- available informally between researchers on a peer-to-peer basis;
- available through the organisation of conferences and workshops.

Potential users will be able to find the data through the means defined above as well as by using common web search engines since the project's website will be Search Engine Optimized (SEO).

Specific deliverables related to ongoing research pending peer-review publication are currently excluded and will be made open-access following the acceptance of the associated papers.

D1.1 - 3.3 Other Material

Type	Name	Description	Sharing	Archiving and Preservation
Press Release	HARSH_press_release	HARSH press release	Available on HAR.S.H. website for public access	https://www.uoc.gr/wp-content/uploads/2025/07/HARSH-MAY-2025-PressRelease-v8-2.pdf
Press Release	SPTDC_press_release	SPTDC press release	Available on HAR.S.H. website for public access	https://www.forth.gr/en/news/show/&tid=3198
Article	HARSH_HIPEAC_article	Hipeac Article	Available on HAR.S.H. website for public access	https://www.hipeac.net/news/magazine/7174/hipeacinfo-78/#/

D1.1 - 4 Open Access Pilot

HAR.S.H. ensures open access to all scientific information produced during the project. In this way, HAR.S.H. maximizes the potential of sharing the knowledge created throughout the project's lifetime. Specifically, in addition to providing access to publications via the project's website, the papers will be stored in the arXiv and/or zenodo online repository.

Software: HAR.S.H. will make the software library it develops available through Github.

Publications at conferences and journals: Most of the conference and journal publications in the context of HAR.S.H. will be published by IEEE and ACM, which are world-leading providers of publishing services for research papers. Both of them support flexible author rights policies. For instance, they allow authors who publish in their journals or conferences to share their research by posting a free draft copy of their article to non-commercial repositories and/or websites. All publications produced in the context of HAR.S.H. acknowledge project funding.

HAR.S.H. is in compliance with the requirement for Open Access publication, following the guidelines[1] presented by the European Commission. In terms of open access, the project will make available public results and publications through the following repositories:

The project website: https://harsh-project.eu , which has sections for uploading and sharing the outcomes of the project that are of type public and the HAR.S.H. publications. The following information will be made available on the website:

- news & articles
- paper publications
- Deliverables with public dissemination level
- YouTube channel, to be used for uploading videos featuring HAR.S.H. information.

D1.1 - 5 Conclusions

The data management is a "living" document that defines the framework and policies for the data handling during and after the project. The document presents the data that will be collected and generated during the project, describing both the internal as well as the external mechanisms, for storing, processing and accessing the project's data. This information is dynamic and is expected to be enriched with the progress of the project.

D1.1 – 6 References

[1] M. D. Wilkinson et. al., "The FAIR Guiding Principles for scientific data management and stewardship.," 2016.

[2] "Guidelines onFAIR Data Management in Horizon 2020," [Online]. Available: https://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-data-mgt_en.pdf.

[1] Guidelines on Open Access to Scientific Publications and Research Data in Horizon 2020 http://ec.europa.eu/research/participants/data/ref/h2020/grants_manual/hi/oa_pilot/h2020-hi-oa-pilot-guide_en.pdf

[3] S. Jones, "How to Develop a Data," [Online]. Available: https://www.dcc.ac.uk/sites/default/files/documents/publications/reports/guides/How%20to%20Develop.pdf.

[4] [Online]. Available: https://www.zoom.com/.

[5] [Online]. Available: https://www.microsoft.com/en-us/microsoft-teams/group-chat-software.

[6] [Online]. Available: https://github.com/.

[7] [Online]. Available: https://www.microsoft.com/el-gr/microsoft-365/onedrive/online-cloud-storage.

[8] "I. R. I. for Seismology with Artificial Intelligence,". Available: https://ds.iris.edu/data/access/

[9] S. Soldi, V. Beckmann, W. H. Baumgartner, G. Ponti, C. R. Shrader, P. Lubiński, H. A. Krimm, F. Mattana and J. Tueller, "Long-term variability of AGN at hard X-rays", in *Astronomy & Astrophysics*, 2014.