Ref. Ares(2023)254755 - 13/01/2023

DESTINI

SMART DATA PROCESSING AND SYSTEMS OF DEEP INSIGHT

http://www.destini2020.eu

Deliverable D5.5

Report on Content for Training Sessions 2

Document details:

Editor:	CUT
Contributors:	UNIROMA, JADS
Date:	30/09/2022
Version:	V6.0

Document history:

Version	Date	Contributor	Comments
1.0	20/08/2022	A.S. Andreou S. Mappouras A. Christoforou P. Christodoulou M. Pingos	Initial document, structure and content
2.0	30/08/2022	A.S. Andreou S. Mappouras	Additions & Corrections
3.0	05/09/2022	A.S. Andreou Partners	First review
4.0	08/09/2022	S. Mappouras	Additions & Corrections
5.0	25/09/2022	A.S. Andreou	Final review
6.0	30/09/2022	Partners	Final review & approval

Contents

1. Introduction

1.1. Purpose

This document presents a detailed description of the outcomes obtained during the mobility period of Sapienza and JADSs ESRs that were relocated in Cyprus during August and September 2022. The ESRs were selected during the Satellite Mobility event of DESTINI at SummerSOC2022 as described in deliverable D5.3.

This deliverable is part of Work-Package 5 (WP5) that develops and executes a plan for joint education and training programmes between the members of the consortium. It also devises and applies a mobility programme to help attracting early-stage researchers within the consortium and beyond, describes training and mentoring activities offered to researchers and outlines incentives, like the announcement of PhD topics in high-demand and popular subjects, financial and local support for relocation.

Despite the short period of stay of the ESRs the collaborations that were established have shown to be very efficient, which is proved by the outcomes of the research collaborations. The collaborations established were mainly between ESRs from the three partner universities of the project, as anticipated, but also from other universities that mainly joined the satellite event at SummerSOC2022. In addition, collaboration was extended with additional contributors from the faculty of Computer Engineering and Informatics of CUT, such as the internationally recognized Digital Cultural Heritage Lab (DCH), or other colleagues that pursuit research in healthcare applications, specifically on the emergency cases handling in Cyprus using ICT, and/or early identification/treatment of stroke cases.

In terms of planning mobility time, we may consider the two phases of the mobility program as follows: In Phase A' (SummerSOC-2022 satellite event in Crete, Greece), twenty-five (25) ESRs joined the event for one (1) week, so a total of twenty-five (25) mobility weeks may be accounted for. In Phase B', eight (8) ESRs from the leading institutions relocated in Cyprus for a total of fifteen (15) weeks. Therefore, in total DESTINI's mobility program was executed for forty (40)

weeks. The planned mobility figures were ten (10) ESRs for a total of up to one hundred and twenty (120) weeks. Therefore, in plain numbers the ESRs that joined the mobility program was nearly a match with the planned figure, while a 33% of the planned time was achieved. Nevertheless, the intensified effort put in this smaller period and the actual outputs in terms of the level of research collaboration, as well as the number of topics investigated and number and variety of the ideas produced for new research and/or proposals to attract additional funding, exceeded expectations. Therefore, the consortium is fully satisfied with the outcomes of this work package.

It should also be noted the last day of the mobility program included presentations of the outcomes of the mobility program, which opened ways for future collaboration between the ESRs and the universities/groups that constitute the consortium of DESTINI.

1.2. Definitions, Acronyms, and Abbreviations

ESRs: Early-stage researchers

DCH: Digital Cultural Heritage

CUT: Cyprus University of Technology

ICT: Information and Communication Technology

JADS: Jheronymous Academy of Data Science

DL: Data Lake

BPM: Business Process Mining

1.3. Overview

The rest of the document is structured as follows:

Section 2 presents the agenda of the final event of DESTINI regarding the mobility program, which was organized and disseminated among the associates of the partner universities. Section 3 reports with details the ideas discussed during the mobility event. Section 4 presents the outcomes of the mobility program in general and the topics that the ESR identified, worked on and will collaborating on. Finally, section 5 concludes this document.

2. Dissemination Material

2.1. Event Agenda

DESTINI

CLOSING MEETING AGENDA

29/09/2022
Cyprus University of Technology

TIME	DESCRIPTION	FACILITATOR
09:00 - 09:45	Introduction	Andreas Andreou
09:45 - 10:15	Content fruition in Digital Humanities	Alberto Morvillo
10:15 - 10:45	Applying Process Mining to Human Daily Activities	Silvestro Veneruso
10:45 - 11:15	Data Lakes/Swamps and Process Mining	Dario Benvenuti
11:15 - 11:45	Digital transformation in manufacturing: towards the Industry 4.0 model	Flavia Monti
11:45 - 12:15	break	---

DESTINI
CLOSING MEETING AGENDA

29/09/2022
Cyprus University of Technology

TIME	DESCRIPTION	FACILITATOR
12:15 - 12:45	From Record Linkage to Knowledge Graphs: an application in the healthcare domain	Jerin Mathew
12:45 - 13:15	Business Process Management in a Hyper-Connected IoT World	Francesca De Luzi
13:15 - 13:45	DNA identification and classification, Forensic DNA analysis	Mirella Sangiovanni
13:45 - 14:15	Discovering Data Products and Domains	Stefan Driessen
14:15 - 14:45	Brainstorming	---
14:45 - 15:15	Closing Remarks	Andreas Andreou

3. Ideas discussed during the mobility event

3.1. Dario Benvenuti

Department / Division:	Sapienza University of Rome
Collaborator:	Michalis Pingos (PhD Candidate from CUT)
Dates:	18 August 2022 – 30 August 2022
Supervision:	Andreas Andreou Tiziana Catarci

Collaboration notes
In the following, we will look at how Process Mining applied on Big Data Pipeline could be paired with Metadata Enrichment using Blueprints in Data Lakes to generate interesting research directions. Blueprint Models theory wants to provide a solution for the standardization required by Process Mining (PM). Each entity of this standardization will consist in processes that generate logs that correlate and interact together chronologically and based on dependencies. The goal is to utilize a semantic metadata enrichment mechanism via Blueprints, which utilizes the 5 Vs and ontologies to assist data processing (storing and retrieval) in Data Lakes (DLs) with ponds and puddles, with emphasis on organizing and preparing data to facilitate PM.

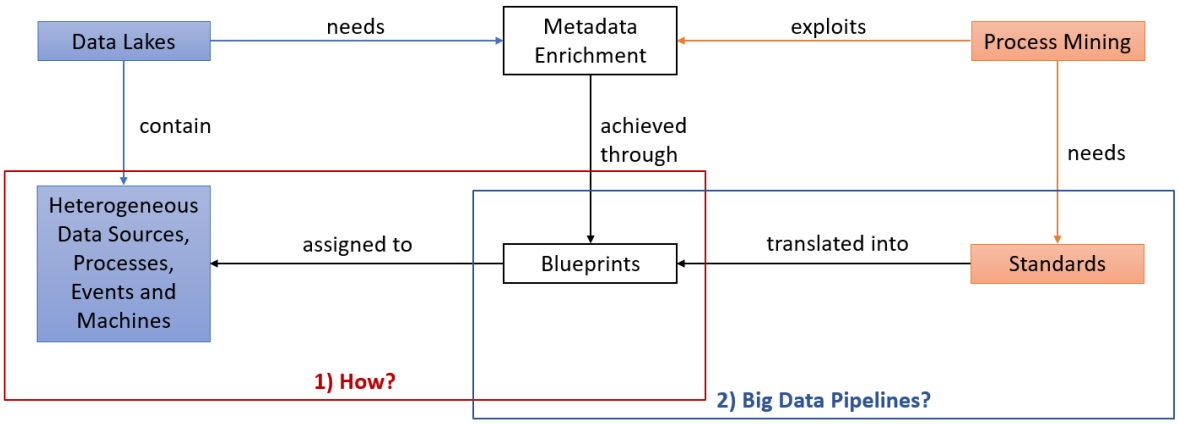


Figure 1: Analysis of the relationship between Data Lakes and Process Mining

But how do PM and DL really relate?

As can be seen in Figure 1, DLs need, by definition, metadata enrichment to keep data ordered, managed, and easily processed. While in PM there is no real need of having documentation and metadata, it can surely exploit such an enrichment in cases where the logs under investigation need some refinements before their ingestion.

In DLs, metadata enrichment can be achieved through the application of the Blueprint Models theory, and the standardization required by every PM technique can be easily translated into a blueprint.

Finally, these blueprints will be assigned to every component of the DL, including heterogeneous data sources, processes, events, and machines.

From this schema, two main questions arise:

It's clear from that having DLs with metadata enrichment through blueprints gives a lot of benefits, including process mining readiness, but there is not yet any guideline on how this scenario can be achieved in practice. In the work done, the main assumption is that the DL is already set and that blueprints are "magically" attached to data when it enters in the DL.

Standardization in PM can be specialized to exploit properties of a specific context, or simply to obtain more domain-specific insights while applying PM techniques like discovery or conformance checking. So how could we translate a specific standardization, developed ad-hoc for Big Data Pipelines, into blueprints?

From these two points, we can develop some ideas for collaboration.

3.2. Silvestro Veneruso

Department / Division:	Sapienza University of Rome

Collaborators	Michalis Pingos (PhD Candidate from CUT) Professor Efthyvoulos Kyriacou (CUT)
Dates:	23 August 2022 – 09 September 2022
Supervision:	Andreas Andreou Francesco Leotta

Collaboration notes

As part of my research, I work in the context of smart spaces, i.e., environments enriched with sensors and actuators which aim to provide automatic/semi-automatic services to its inhabitants. A smart space produces output data in the form of a sensor log, which consists in a series of raw measurements from the sensors within the smart environment itself. Such logs can be used as a testbed by researchers in the field. Researchers need open datasets to validate their approaches, ensuring reproducibility. During my PhD, I focused on developing unsupervised learning techniques to mine human habits. With the term habit we refer to the human routine. E.g, what does the user usually do in the morning after breakfast? Our goal is to provide autonomous mechanisms to segment logs (in particular we are interested in sensor logs) with the aim of reducing the human effort and avoiding annoying and sometimes imprecise training sessions. Furthermore, techniques from the Process Mining area could be applied to properly converted and segmented logs [1]. Process mining is a fairly recent research discipline which combines Data Mining techniques with techniques used in Business Process Management (BPM) [2], such as process discovery and process analysis. Its main goal is to extract meaningful information from event logs.

- In this context, collaboration can be established with Michalis Pingos and Professor Efthyvoulos Kyriacou onto combining the smart environments and Process Mining knowledge from my PhD to monitor elderly people with mental diseases using camera/video content to mine their daily habits and detect abnormal situations or anomalies to prevent accidents.
- During the brainstorming session has also emerged the will of dealing with data gathered from the Cyprus' emergency institution. In particular, they are collecting from emergency calls. Recently they also started tracking the timestamp related to each event. This kind of information could be used to mine process models describing how such an emergency system deals with emergency calls, e.g. car crashes.
- In particular, the proposed approach consists in 4 steps
    - Gather events related to emergency calls

- Properly convert the log into an event log
- Mine the related process model by applying Process Mining modeling's techniques
- Perform Process Management and Analysis on mined models.

- The details regarding the data provided from the emergency's institution must be clarified in a future call. In particular, we estimated it will take at least some months before acquiring such data.

[1] Leotta, M. Mecella, and J. Mendling, "Applying process mining to smart spaces: Perspectives and research challenges," in Advanced Information Systems Engineering Workshops, A. Persson and J. Stirna, Eds. Cham: Springer International Publishing, 2015, pp. 298–304.

[2] M. Dumas, M. La Rosa, J. Mendling, H. A. Reijers et al., Fundamentals of business process management. Springer, 2013, vol. 1.

3.3. Flavia Monti

Department / Division:	Sapienza University of Rome
Collaborator:	Andreas Christodoulou (Post-doc researcher at SEIIS) Spyros Loizou (PhD Candidate from CUT)
Dates:	08 September 2022 – 28 September 2022
Supervision:	Andreas Andreou Francesco Leotta

Collaboration notes
My research interests mainly focus on Smart Manufacturing (or Industry 4.0). Particularly, the application of innovative technologies, e.g., Artificial Intelligence and Computer Vision, to augment industrial machinery. Goal of my research is to improve production quality, reduce costs, increase machinery uptime, analyze failures and defects towards zero defect manufacturing.

Usage of FCM for evaluating a Maturity Model

Industry 4.0 has introduced a huge number of technologies which requires absolute knowledge to correctly be put into action. Because of the multitude of solutions and techniques, it is not easy for a company to schedule and plan the roadmap and the investments required to shift towards Industry 4.0. Moreover, it is not even straightforward for a company to understand its readiness as an Industry 4.0 player. The presence of a maturity model to assess the maturity and readiness of a company as an Industry 4.0 actor according to the complexity and the type of software and hardware installed and their usage, can bring great advantages. Companies can use it to evaluate their strengths and weaknesses, but, more importantly, they can use it to define a roadmap of investments to reach a higher "maturity level".

Part of my PhD work consists in the definition of a maturity model relying on Information Systems (IS) employed by a company. A set of relevant IS and technologies have been identified as having great impact on the transition towards Industry 4.0, and they were analyzed to define the structure of the maturity model.

Predictive Maintenance strategies in Maritime using Digital Twins

Objective

Development of a predictive maintenance strategy solution in maritime domain by means of Digital Twins

Overview

Nowadays, the maritime domain is undergoing a digital transformation by adopting digital technologies. Likewise Industry 4.0, maritime transformation promotes automation and digitization of processes and control operations.

Modern ships are getting more complex and intelligent. Indeed, the integration of sensors, data storage, computational power and connectivity, leads to powerful systems having the goal of reducing costs and improving business performances.

Digital twins are seen as a key enabler in the maritime domain to collect insights in each stage of the lifecycle for different purposes. Authors in [12] identified and highlighted several needs in the maritime domain such as remote inspection and maintenance assistance and remote monitoring of offshore assets.

[12] Taylor, N., Human, C., Kruger, K., Bekker, A., & Basson, A. (2019, October). Comparison of digital twin development in manufacturing and maritime domains. In International Workshop on Service Orientation in Holonic and Multi-Agent Manufacturing (pp. 158-170). Springer, Cham.

3.4. Jerin George Mathew

Department / Division:	Sapienza University of Rome
Collaborator:	Michalis Pingos (PhD Candidate from CUT)
Date:	08 September 2022 – 28 September 2022
Supervision:	Andreas Andreou Donatella Firmani

Collaboration notes

This research proposal focuses on extracting fine-grained information about patients (e.g. the entire history of medical examinations of a patient across hospitals included in the data lake) and building a knowledge graph using such data.

This KG can be used along any other data structure built on top of the data lake (e.g. the blueprint framework proposed in [1] to provide an additional interface or view of the data lake.

However, building such a KG is a challenging task for several reasons. Patients' data may span several hospitals and such EHRs may be noisy or incorrect (e.g. incorrect name, or age, etc.). Thus, a preliminary step is required to find records that refer to the same person and conciliate conflicting information related to the same patient.

In this scenario, state-of-the-art ER systems can be leveraged to find and cluster together EHR records that refer to the same person. Notably, several ER systems take as input a pair of records having the same exact schema (i.e. list of attributes). However, EHR records may have different formats depending on the specific hospital. Moreover, some records might not have an underlying schema at all, like doctor's notes or medical prescriptions.

Recent deep learning-based approaches have been successfully used in such complex scenarios. Such systems usually leverage large-scale neural networks that were previously trained on huge amounts of unstructured textual data (e.g. [9,10,11]) and are further trained for the ER task.

The training procedure requires a modest amount of pairs of textual descriptions along with a binary label to distinguish between matching and non-matching pairs of description. Notably, these record descriptions do not have a specific format (i.e. an underlying schema), so these deep learning-based ER systems can handle both structured and unstructured pairs of records.

[1] Andreou, Andreas, and Pingos Michalis. "A Data Lake Metadata Enrichment Mechanism via Semantic Blueprints." Evaluation of Novel Approaches to Software Engineering (2020)

[9] Vaswani, Ashish, et al. "Attention is all you need." Advances in neural information processing systems 30 (2017).

[10] Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." arXiv preprint arXiv:1810.04805 (2018).

[11] Sanh, Victor, et al. "DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter." arXiv preprint arXiv:1910.01108 (2019).

3.5. Alberto Morvillo

Department / Division:	Sapienza University of Rome
Collaborator:	Michalis Pingos (PhD Candidate from CUT) Spyros Loizou (PhD Candidate from CUT) Prof. Marinos Ioannides (DCH lab, CUT)
Date:	08 September 2022 – 28 September 2022
Supervision:	Andreas Andreou Massimo Mecella

Collaboration notes

My research topic involves cultural heritage fruition and, as a part of it, I'm currently working with knowledge graphs and geographic reference of a content.

A knowledge graph is a special kind of database which stores knowledge in a machine-readable form and provides a means for information to be collected, organized, shared, searched and utilized [1].

In a knowledge graph the data are represented by the connection between the concepts that define them. For instance, using the Resource Description Framework (RDF) [2] an "entity" can be described by "triples" of subject, predicate and object; so, a content could be described as:

"The Divine Comedy" – is a type of – poem

"The Divine Comedy" – has been written by – "Dante Alighieri"

Expanding this approach to the subject involved in the relations (in this example "Dante Alighieri"), a knowledge graph can be used to represent any kind of knowledge. Using them for the matter of research purposes could help to keep track of the links between each element, be it a concept, a location or a person, revealing unexpected connections during the exploration of contents and, thus, generating new ideas, which can be visually represented.

In Digital Humanities, where the data fruition is of primary importance, the use of knowledge graphs could give many advantages [3][4].

Research proposal no.1: Digitalization of a place of interest.

Target: Tourists, visitors and researchers of all ages and gender.

Setting: Any physical place that represents a place of interest (e.g.: monuments, small villages, museums) and does not already allow a digital fruition of his content.

Goal: Facilitate a place fruition in order to

- Increase tourism by providing easy-to-access content, which will also increase the accessibility of the place (e.g., people with disabilities, people who cannot reach the physical place, etc.)
- Facilitate the research by acquiring and managing detailed data of the place, which the researcher can explore using search engines or data navigation techniques.
- Preserve the physical place, by applying analysis and preservation techniques based on the data collected for digitization.

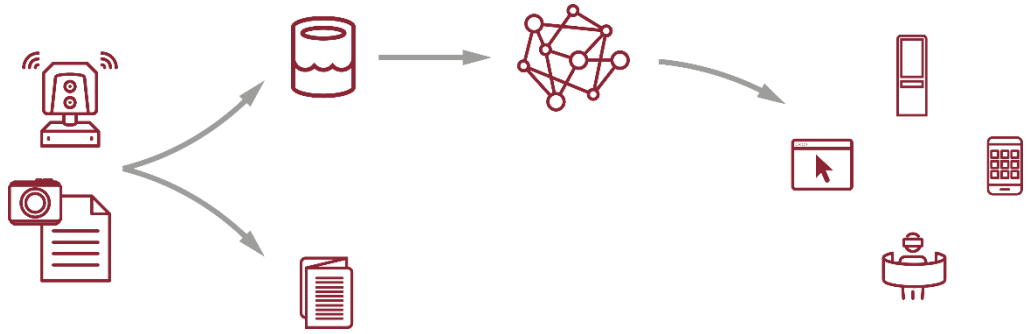
Proposed approach
The approach is divided into 3 steps:

1. Data acquisition
2. Data management
3. Data representation

For the data acquisition, digital 3D scanners (LIDAR), text recognition technologies (OCR) and manual acquisition, together with a metadata standardization technique [5] (like IIIF [6]) can be used. The acquired data could be stored in data lakes [7] to optimize their retrieval.

Once all the data are collected and stored, knowledge graphs could be used to organize them in a concept-oriented way in order to facilitate searches and navigation [3][4]

The data representation will apply a user-oriented approach, with the use of different media and interfaces specifically designed for a designed target. Brochures, multimedia kiosks, websites, mobile apps, VR devices and Braille stands are all examples of media and interfaces (both digital and not digital).



Research proposal no.2: Tracking of an artifact transaction using blockchains.

Target: Entities or companies that must manage artifact transitions.

Goal: Allow to track artifact transactions easily and securely.

Proposed approach

Blockchains are a decentralized approach to store information and are widely used in financial and business fields to avoid the use of a central Certification Authority to validate the transactions.

In respect to a centralized approach, in which a central entity manages all the transactions and stores all the information in a single place; the blockchains use a peer-to-peer approach and each party involved in a transaction has a copy of the data, while the validation of them is demanded by other parties.

Logistics and supply chains can take advantage of a distributed approach to store the product history [8] and in the food industry there are already some use cases [9].

The proposal for this research topic is to apply the same methodology to physical artifacts, like archaeological or fine arts, to track their transactions or logistics.

[1] “About DBpedia”, https://www.dbpedia.org/about.

[2] “RDF 1.1 Concepts and Abstract Syntax”, https://www.w3.org/TR/rdf11-concepts.

[3] B. Haslhofer, A. Isaac, R. Simon, “Knowledge Graphs in the Libraries and Digital Humanities Domain”, 2018, DOI: 10.48550/arXiv.1803.03198, http://arxiv.org/abs/1803.03198.

[4] E. Bernasconi, M. Ceriani, M. Mecella, “Exploring a text corpus via a knowledge graph”, in: IRCDL, 2021, pp. 91–102, http://ceur-ws.org/Vol-2816.

[5] E. Maravelakis, A. Konstantaras, A. Kritsotaki, D. Angelakis, and M. Xinogalos, “Analysing User Needs for a Unified 3D Metadata Recording and Exploitation of Cultural Heritage Monuments System”, Berlin, Springer Berlin Heidelberg, 2013, pp. 138—147, ISBN: 978-3-642-41939-3.

[6] “International Image Interoperability Framework (IIIF)”, https://iiif.io.

[7] M. Pingos, A. Andreou, “International Image Interoperability Framework (IIIF)”, 2022, DOI: 10.5220/0011080400003176

[8] E. Tijan, S. Aksentijević, K. Ivanić and M. Jardas, “Blockchain Technology Implementation in Logistics”, 2019, DOI: 10.3390/su11041185, https://www.mdpi.com/2071-1050/11/4/1185.

[9] “Birra Peroni e la tracciabilità in blockchain del malto 100% italiano: una storia di open innovation”, https://www.birraperoni.it/news-e-media-comunicati-stampa/birra-peroni-e-la-tracciabilita-in-blockchain-del-malto-100-italiano-una-storia-di-open-innovation.

3.6. Stefan Driessen

Department / Division:	Jheronymous Academy of Data Science (JADS)
Collaborator:	Michalis Pingos (PhD Candidate from CUT)
Date:	13 September 2022 – 16 September 2022
Supervision:	Andreas Andreou Willem-Jan van den Heuvel

Collaboration notes

<u>Proposal 1: Extending Metadata Blueprints for Data Meshes</u>

Short Overview

The proposal focuses on extending the current blueprint for a metadata ontology based on the 5 V's for big data. The blueprints could include new entities that are explicitly relevant for data products on data meshes or data markets.

1. Background and Problems Addressed

Data Lakes were proposed in the 2010's as architectures suitable for dealing with "Big Data" and for assisting organisations towards being data-driven. Current literature shows a trend towards more decentral data exchange architectures such as data markets and data meshes [1]. Data Meshes, in particular, attempt to address some of the shortcomings of monolithic data platforms such as data lakes [2, 3]. Creating proper data products puts requirements on metadata templates that are not yet addressed by existing approaches.

2. Proposed Work and Links to Undergoing Research

Currently, Michalis Pingos has proposed some blueprint for a metadata ontology based on the 5 V's that drive data lake development [4]. Since data meshes are also designed for handling big data, these metadata aspects should still be relevant. However, as indicated by Stefan's undergoing research collaborations with companies (e.g., Automotive and Telecom), data mesh imposes new requirements and restrictions on the metadata of their data products. One potentially interesting approach is

therefore to extended Michalis' templates to include new metadata that is required for data products such as data product owner, usage policy, Domain, etc. This ties into Stefan's current work of setting up a standardised metadata template for data products.

References

[1] S. Driessen, G. Monsieur, and W. V. D. Heuvel, "Data market design: A systematic literature review," IEEE Access, vol. 10, pp. 1–1, 2022.

[2] A. Loukiala, J. P. Joutsenlahti, M. Raatikainen, T. Mikkonen, and T. Lehtonen, Migrating from a Centralized Data Warehouse to a Decentralized Data Platform Architecture. Springer International Publishing, 2021, vol. 13126 LNCS, ¡u¿Cited for:¡br/¿¡/u¿- Monolithic solutions suck¡u¿¡br/¿¡/u¿- Existing solutions do not scale.¡br/¿¡br/¿¡u¿Useful for¡/u¿:¡br/¿- Existing solutions are central, monolithic and hard to scale.¡br/¿- Sources for definitions of data warehouse, data lake, data platform architecture, SOA, microservices.¡br/¿- Domain-Driven design pattern called bounded context.¡br/¿- Parallel between data mesh and micro services (MSA).¡br/¿- The central components (e.g., data lake are responsible for models and standards).¡br/¿- How to build templates.¡br/¿- The coupling between business domains and IT infrastructure is a challenge that can be overcome through decentralisation.¡br/¿- Scaling by only publishing metadata, without ever creating API's for consuming across domains!¡br/¿- Auditing by a central team. [Online]. Available: http://dx.doi.org/10.1007/978-3-030-91452-3 3

[3] Z. D. (Thoughtworks), "How to move beyond a monothilitic data lake to a distributed data mesh," 2019, ¡u¿Cited for:¡br/¿¡/u¿- Generalist data markets can exist as internal data markets when moving to a data mesh.¡br/¿- Challenge of monolithic approaches is scalability. [Online]. Available: https://martinfowler.com/articles/data-monolith-to-mesh.html

[4] M. Pingos and A. S. Andreou, "A data lake metadata enrichment mechanism via semantic... -google scholar," 2022. [Online]. Available: https://scholar.google.nl/scholar?hl=en&as sdt=0%2C5&q=A+Data+Lake+Metadata+Enrichment+Mechanism+via+Semantic+Blueprints&btnG=

<u>Proposal 2: Discovering Data Products and Domains</u>

Short Overview

The proposal focuses on addressing one of the largest challenges in transitioning towards a data mesh architecture: how to create domains and prioritise data products.

1. Background and Problems Addressed

Many (large) companies and organisations have initiatives to transition from their existing, monolithic, data platforms towards more decentral data exchanges, such as internal data markets or data meshes [1, 2]. One of the main challenges for this transition, in addition to the novelty of the concepts, is how to divide up the data landscape into domains and identifying data assets that should be turned into data products [3]. These organisational challenges are in fact often perceived to be more daunting than the technical challenges associated with data mesh design [4]. Methods for creating domains can

(presumably) be found in domain-driven design literature? Methods for prioritising data products are supposedly related to data product valuation techniques, which are currently lacking.

2. Proposed Work and Links to Undergoing Research

Based on Michalis' existing metadata framework it could be possible to establish a hierarchical structure within the domain assets. On the one hand primarily static metadata attributes such as origin and type of data can be used to establish domain boundaries that are neither too large nor too small. On the other hand, dynamic metadata on usage can be used to help organisations and companies prioritise which data assets to turn into data products first.

References

[1] A. Goedgebuure, "Data mesh: Systematic gray literature study, reference architecture, and cloud-based instantiation at asml," 2022. [Online]. Available: https://stefan-driessen.github.io/publication/data-mesh-systematic-grey-literature-study/

[2] R. Eichler, C. Giebler, C. Groger, E. Hoos, H. Schwarz, and B. Mitschang, "Enterprise-wide¨metadata management: An industry case on the current state and challenges," Business Information Systems, pp. 269–279, 2021, ¡u¿Useful for¡/u¿:¡br/¿- Challenging to motivate data providers for internal data markets.¡br/¿- Existing tools are insufficient¡br/¿- There is a lack of research for metadata management for internal data markets.

[3] S. Driessen, G. Monsieur, and W. V. D. Heuvel, "Data market design: A systematic literature review," IEEE Access, vol. 10, pp. 1–1, 2022.

[4] Z. Dehghani, Data Mesh: Delivering Data-Drien Value at Scale, 1st ed., M. Duffield, G. O'Brian, and B. Kelly, Eds. O'Reilly, 2022.

3.7. Francesca De Luzi

Department / Division:	Sapienza University of Rome
Collaborator:	Spyros Loizou (PhD Candidate from CUT)
Date:	23 September – 28 September
Supervision	Andreas Andreou Massimo Mecella

Collaboration notes

General objective

The general objective of my research is to devise conceptual and technical research solutions to effectively exploit IoT technologies towards the improvement of BPM practices in use within the enterprises, with the target to realize unused potential for BP optimization, by consequently boosting the overall enterprises' productivity.

Introduction

The main idea is to provide a systematic methodology through which it will be possible to identify critical situations in IoT-aware business processes, supporting at the same time the companies to diagnose issues in their organizational BPs and suggesting solutions for BP improvement, thus bringing strategic advantages against the competitors. To achieve this objective, we have defined a 3-steps methodology, which includes:

- Systematic analysis of the literature on combining BPM and IoT to investigate existing solutions and research channels.
- Definition of a case study and identification of relevant IoT-aware BPs from it and validation of the research solutions over them.
- Implementation and testing of the identified BPs with an existing BPM system, customized according to the research purposes.

Systematic literature review

The current research on the integration between BPM and IoT is still evolving. There are many studies that discuss the opportunities and challenges that such an integration can bring [C. Janiesch]. However, IoT technology is far to be integrated into BPM systems. For this reason, the first effort of my research has been focused on integrating the challenge of bridging the gap between IoT and BPM through the definition of a systematic literature review (SLR) to identify, select and critically appraise existing research solutions in this area. To conduct the SLR we have followed the guidelines, procedures, and policies proposed by Kitchenham in [Keele]. Accordingly, we have defined a review protocol that starts with the specification of the need to conduct a review and then we have formulated the research questions. Then, a search string was defined to identify all relevant literature from a selection of data sources. As a result, we obtained a list of papers that we evaluated according to the title and abstract, for inclusion in the set of candidate papers to be finally reviewed. Inclusion and exclusion criteria have been defined to assess each candidate paper. As a result of this stage, an initial set of potential primary studies was obtained. In addition, to reduce the probability of missing relevant studies, we have complemented the initial search with backward and forward snowballing.

Resilient-Aware Process Models

The widespread diffusion of Internet-of-Things (IoT) technologies is prompting organizations to rethink their business processes (BPs) towards incorporating the data collected from IoT devices directly into BP models for improved effectiveness and timely decision making. Nonetheless, IoT devices are prone to failure due to their limitations in terms of computational power and energy autonomy, leading to compromise the availability and quality of the collected data, with the risk to prevent the correct execution of the entire BP. To mitigate this issue, resilience is a feature that any data-aware BP should support at design-time, by focusing on the role of available - as an alternative to unreliable - data as a resource for increasing BP robustness to failures. In this paper, we formalize an approach for designing and evaluating resilient-aware BP models in BPMN (Business Process Modeling and Notation) through a maturity model that considers their degree of awareness through levels of resilience, which can be computed using the provided formalization. In addition, we show how to extend the metamodel of BPMN 2.0 to address the proposed resiliency levels, and we investigate the feasibility of the approach through a user evaluation [Agostinelli].

KPI-based process mining over structured and textual data

The control of the times, volumes and resources used in the treatment of judicial processes is a fundamental activity to obtain a better use of the same. In this regard, it is necessary to define Key Performance Indicators (KPIs) to monitor judicial processes. In this work, we performed a State-of-the-art analysis on KPI extraction methodologies, with a particular process perspective, together with the study of Natural Language Processing (NLP) and Process Mining techniques to extract relevant information.

- Analysis of some judgments and identification of a use case in the context of civil trial.
- Understanding of macro-phases, activities, and events over time.
- Use case modeling through the production of a BPMN scheme.
- Extraction of temporal information in the legal field.

- Mapping between the activities of the trial and the temporal information obtained from the judgments to identify new indicators of process performance.

The case study concerns the contentious divorce process. In detail, the BPMN scheme shows the phases of the process from which useful temporal information was extracted. It has thus become clear that the case is delayed in the phase of issuing the decree, since this time interval (between the filing of the appeal and the issuance of the decree by the president of the court) must be set at five days. To date, the research activity has focused on the extraction of temporal information to identify new KPIs. However, further factors for their identification will be investigated in the future [Bianchini].

Research Proposal #1

TARGET: Judge and other persons in the domain of justice (e.g. administrative and technical employee)

SETTING: E-justice

PROBLEM: Management of the backlog in justice processes, for example, the judge often fails to read all the papers carefully, or it is often necessary to differentiate the work immediately

GOAL: combine Digital Twins and process mining in a unified graphical and interactive dashboard to deliver support to the judge to write judicial documents and other tasks such as calendaring through a visual representation of business data and logs

ADVANTAGES: flexible and user-friendly dashboard, generic framework that can be able applied in every business and data context.

Research Proposal #2

TARGET: BP designer – domain expert

SETTING: IoT-aware BPs (e.g. smart manufacturing, logistics, smart health or whatever application domain)

PROBLEM: IoT data quality – due to their limitations in terms of computational power and energy autonomy, they often compromise the quality characteristics with the risk of preventing the correct execution of the entire BP.

GOAL: show how a Digital Twin of data can be used to validate its quality in terms of completeness, accuracy, timeliness, consistency, uniqueness, integrity, compliance

ADVANTAGES: automatically detection of faulty data and to correct measurement errors

References

- Agostinelli, S., De Luzi, F., di Canito, U., Ferraro, J., Marrella, A., & Mecella, M. (2022). A Data-Centric Approach to Design Resilient-Aware Process Models in BPMN. In International Conference on Business Process Management (pp. 38-54). Springer, Cham.

- C. Janiesch, A. Koschmider, M. Mecella, B. Weber, A. Burattin, C. Di Ciccio, G. Fortino, A. Gal, U. Kannengiesser, F. Leotta et al., “The internet of things meets business process management: A manifesto,” IEEE Systems, Man, and Cybernetics Magazine, vol. 6, no. 4, pp. 34–44, 2020.
- S. Keele et al., “Guidelines for performing systematic literature reviews in software engineering,” Technical report, Ver. 2.3 EBSE Technical Report. EBSE, Tech. Rep., 2007.

Under Review

Devis Bianchini, Carlo Bono, Alessandro Campi, Cinzia Cappiello, Stefano Ceri, Francesca De Luzi, Massimo Mecella, and Barbara Pernici, Pierluigi Plebani. xxxx. Challenges in AI-supported process analysis in the Italian judicial system: what after digitalization? Commentaries paper. ACM, NY

3.8. Mirella Sangiovanni

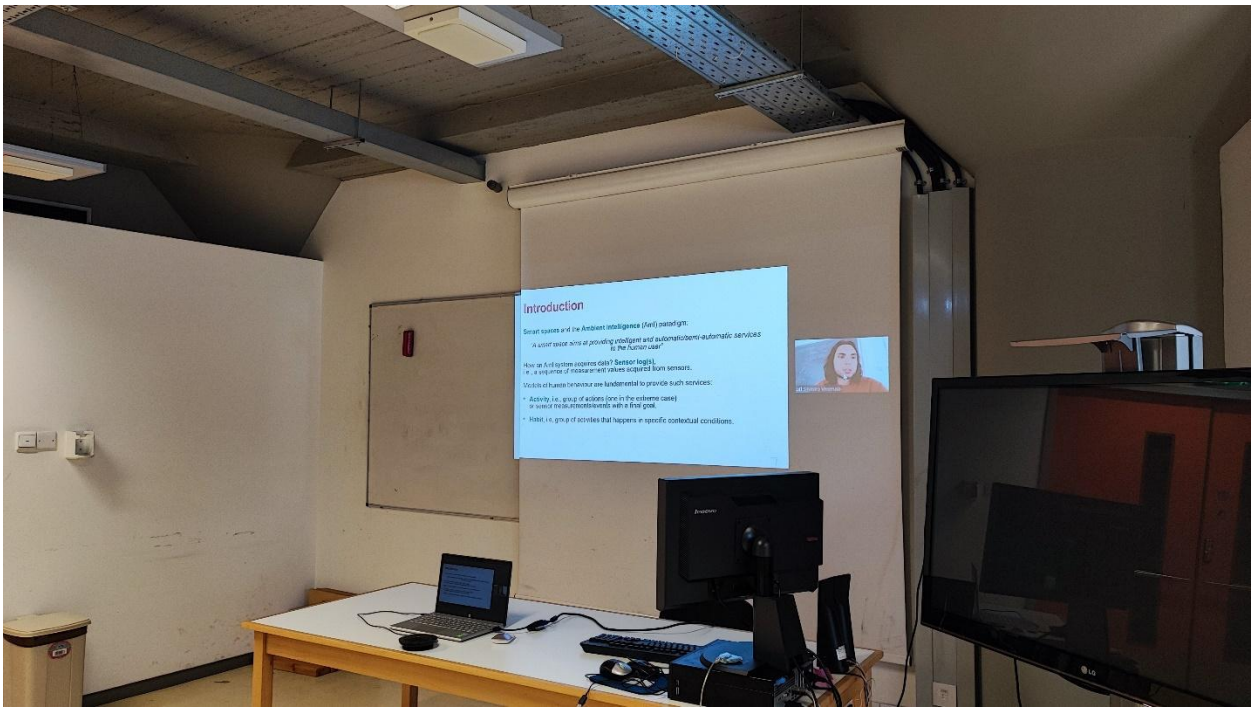
Department / Division:	Sapienza University of Rome
Collaborator:	Michalis Pingos (PhD Candidate from CUT) Prof. George Botsaris (Department of Agricultural Sciences, Biotechnology and Foor Science, CUT) Andreas Christoforou (Post-doc researcher at SEIIS lab)
Dates:	23 September 2022 – 30 September 2022
Supervision:	Andreas Andreou Willem-Jan van den Heuvel

Collaboration notes
Mirella has a background in biology, being a holder of BSc on biology with research interests on health, genetics, biodiversity, microbiology and computational biology. She is a member of the Biodiversity institute in Netherland, and currently researching on Statistical biology. The collaboration has been established on two topics: 1. With Professor George Botsaris and Michalis Pingos the common area of interest identified in researching on biodiversity and animals, specifically on using statistical methods to face the problem of contaminations and diseases of cow (Mastitis) and analyze the effects on people. Additionally, the idea of creating a monitoring system to get the results of the examinations faster, and predict the types of contamination has also been expressed. Lastly, this topic is currently postponed until an appropriate European call is found, in order to attract funding. 2. Collaboration between Andreas Christoforou has been established on Fuzzy Cognitive Maps (FCM) in order to be able to have a more holistic view on the data, from huge samples of data. Furthermore, another idea is FCM to be used to find correlations on complex data, and explore Forensics Data of NFI.

4. ESR presentation day on the DESTINI closing event

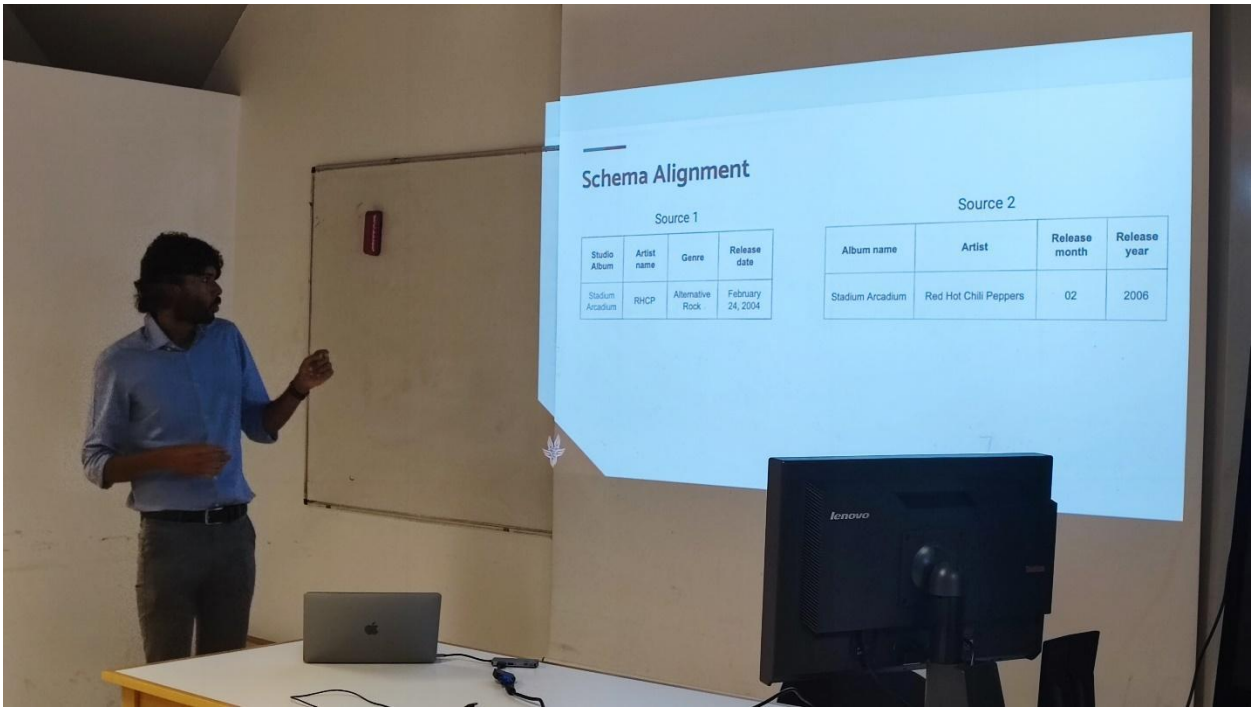
One of the two days of the DESTINI's closing event has been devoted to the presentations of the ESRs participating in the Mobility program that took place in Cyprus. The program being a crucial part of the project, had a whole day for demonstrating its achievements and showcasing the trainings that have been conducted, the mentoring activities and the brainstorming sessions, along with their outcomes.

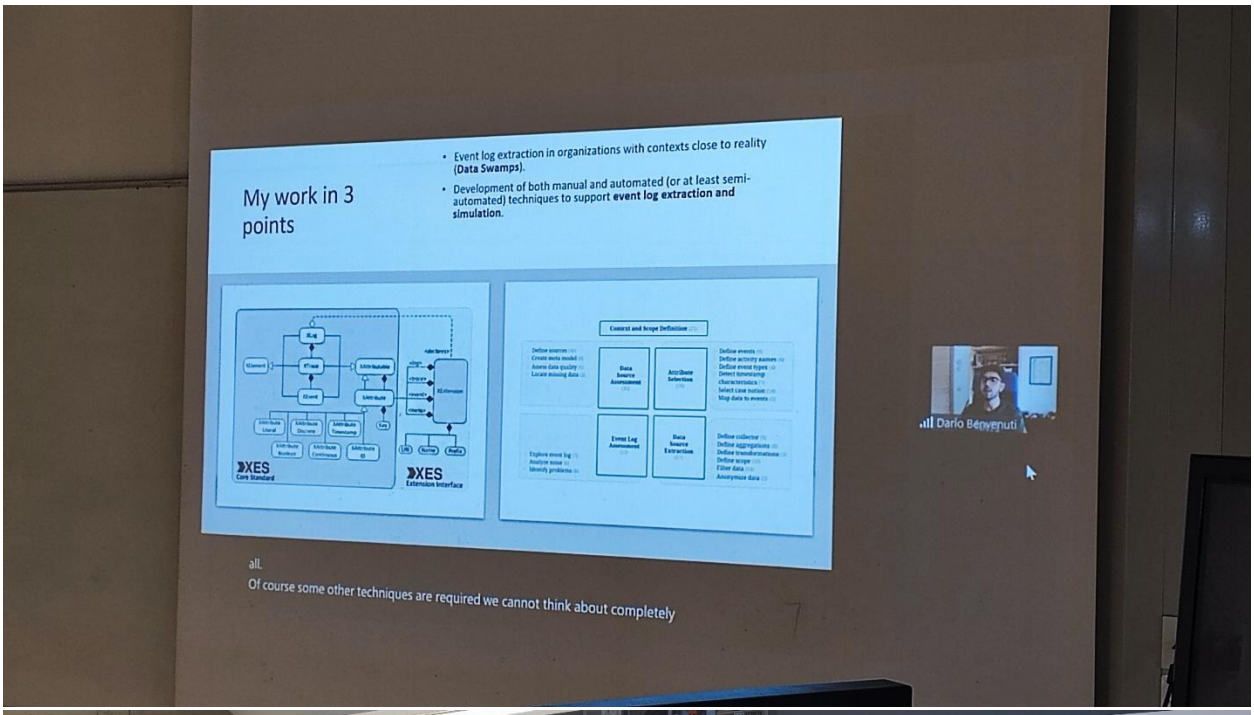

Digital Humanities: main challenges

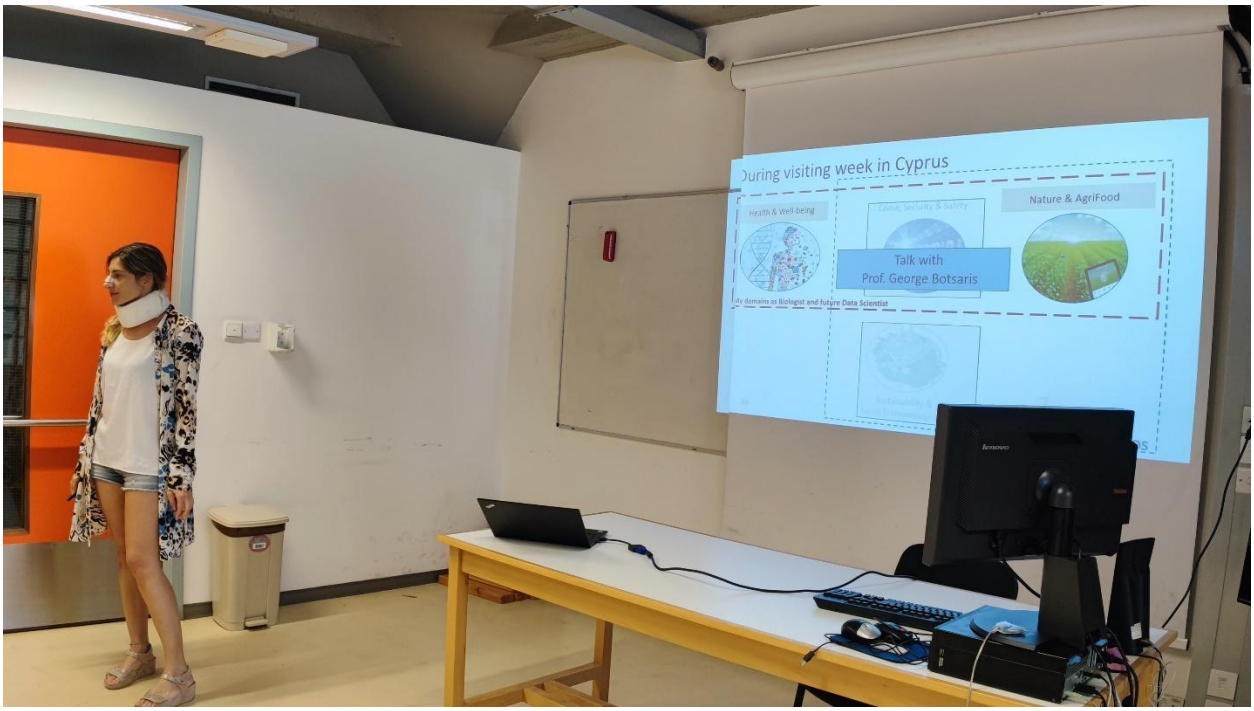

Introduction
Smart spaces and the Ambient intelligence (AmI) paradigm:
"A smart space aims at providing intelligent and automatic/semi-automatic services to the human user"
How an AmI system acquires data? Sensor logics,
i.e., a very small amount of measurement values acquired from sensors.
Models of human behaviour are fundamental to provide such services:
Activity, i.e., group of actions (one in the extreme case) or sensor measurements/events with a final goal.
Habit, i.e. group of activities that happens in specific contextual conditions.


Agenda
• Systematic Literature Review: IoT and BPM
• A Data-Centric Approach to Design Resilient-Aware Process Models in BPMN
• KPI-based process mining over structured and textual data


Quality assessment (of the production)
Detection of defects during the manufacturing process
Ensure quality assurance to the manufactured products


Schema Alignment
Source 1
Studio Album
Artist name
Genre
Release date
Stadium Arcadium
RHCP
Alternative Rock
February 24, 2004
Source 2
Album name
Artist
Release month
Release year
Stadium Arcadium
Red Hot Chili Peppers
02
2006


My work in 3 points
• Event log extraction in organizations with contexts close to reality (Data Swamps).
• Development of both manual and automated (or at least semi-automated) techniques to support event log extraction and simulation.
XES Core Standard
XES Extension Interface
Context and Scope Definition
Data Source Assessment
Attribute Selection
Event Log Assessment
Data Source Extraction
all.
Of course some other techniques are required we cannot think about completely
Dario Benvenuti


During visiting week in Cyprus
Health & Well-being
Nature & AgriFood
Talk with
Prof. George Botsaris

5. Conclusions

This deliverable is part of Work-Package 5 (WP5) that develops and executes a plan for joint education and training programmes between the members of the consortium. Its main outcome is to devise and apply a mobility programme to help attracting early-stage researchers within the consortium and beyond, and describe training and mentoring activities offered to researchers and outline incentives, like the announcement of PhD topics in high-demand and popular subjects, scholarships and local support for relocation.

In the context of the mobility program, which was launched during the SummerSOC2022 school-conference, selected ESRs were invited to relocate in Cyprus for several weeks and work closely with the group from CUT. Several mentoring and brainstorming activities were performed leading to identifying a variety of new ideas for research collaboration and for preparing proposals for attracting new funding in the future.

Although the covid-19 virus outbreak initially posed significant obstacles and created problems for the successful development of the mobility program, the consortium of DESTINI with huge and intense effort managed to make the program work and be very successful, as presented in the findings recorded in this deliverable.