D1.1 - Report on 1st Year Management and Project Activities

Project: NEREO - Neural Information Retrieval and NLP Systems
Grant Agreement: PRIN 2022
Deliverable ID: D1.1
Work Package: WP1 - Management and Project Activities
Due Date: M12
Lead Beneficiary: UNIPI

---

1. Executive Summary

This report summarizes the management and project activities carried out during the first year of the NEREO project (Months 1-12). The project has successfully initiated its research lines in Upstream IR Systems (WP2) and Cascading IR/NLP Systems (WP3), achieving significant milestones in reproducibility, robustness, and efficiency. Management activities ensured smooth coordination among partners, adherence to Open Science principles, and effective dissemination through the project website and scientific publications.

2. Management Activities (Year 1)

During the first year, the following management activities were conducted in compliance with PRIN guidelines:

2.1 Project Initiation and Coordination

- **Kick-off Meeting:** Held at the project start to define the roadmap, assign tasks, and establish communication channels.
- **Regular Progress Meetings:** Bi-monthly online meetings were organized to monitor WP progress, discuss technical challenges, and align research efforts.
- **Internal Communication:** Set up of shared repositories (GitHub) and communication platforms (e.g., Slack/Teams) for daily collaboration.

2.2 Administrative and Financial Management

- **Financial Monitoring:** Continuous tracking of budget expenditure, ensuring alignment with the financial plan.
- **Timesheets and Reporting:** Collection of timesheets and periodic activity reports from all involved personnel to justify personnel costs.
- **Recruitment:** Coordination of hiring procedures for PhD students and Post-docs funded by the project.

2.3 Data Management and Open Science

The project strictly adheres to the principles of Open Science to maximize the impact, reproducibility, and reuse of its research outcomes.

- **Data Management Plan (DMP):** An initial Data Management Plan was established at Month 3. This living document outlines the strategies for data handling throughout the project lifecycle. It ensures that all research outputs—including datasets, source code, and experimental results—are managed in compliance with the **FAIR** (Findable, Accessible, Interoperable, Reusable) principles:

    - **Findable:** All publications and datasets are assigned Persistent Identifiers (DOIs). Code releases are tagged with specific version numbers on GitHub.
    - **Accessible:** We prioritize "Open by Default." Source code is hosted on public GitHub repositories. Preprints of all publications are archived on arXiv or similar repositories to ensure immediate access.
    - **Interoperable:** Data and metadata are structured using standard community formats (e.g., JSON, TREC formatting) to facilitate integration with existing tools and benchmarks.
    - **Reusable:** Software is released under permissive open-source licenses (e.g., MIT, Apache 2.0). Datasets, where generated, are shared under Creative Commons licenses (CC-BY).
- **Open Access Strategy:** The consortium is committed to providing full Open Access to scientific publications.
    - **Green Route:** Immediate self-archiving of the author's accepted manuscript (post-print) in institutional repositories (e.g., IRIS) and subject-based repositories (arXiv) upon acceptance.
    - **Gold Route:** Prioritizing publication in fully Open Access journals or hybrid journals where funds allow, ensuring immediate availability of the Version of Record.
- **Research Data & Reproducibility:** A core focus of Year 1 (WP2) was on *reproducibility* (e.g., the PLAID study). To support this, we created specific "Reproducibility Packages" for our papers, containing the exact code snapshots, hyperparameter configurations, and scripts needed to replicate the results reported in deliverables and publications.

2.4 Dissemination and Communication

- **Project Website:** Delivery of the project website (D4.1) at M1, hosted at https://fabsilvestri.github.io/nereo.github.io/, serving as the main hub for public dissemination.
- **Social Media Strategy:** We have developed a comprehensive "Social Media Procedure" (see `Project Documents/social_media.md`) to standardize the project's dissemination efforts. This document defines the setup, branding, and content workflow for future project communications on platforms like LinkedIn and X.

---

3. Project Activities (Year 1)

Research activities in the first year focused on laying the foundations for advanced neural IR and NLP systems.

3.1 WP2: Upstream IR Systems

Activities and alignment with NEREO Objectives:

- **Robustness in Sequential Recommender Systems:** In *"Investigating the Robustness of Sequential Recommender Systems Against Training Data Perturbations"* (ECIR 2024), we conducted a comprehensive analysis of how sensitive Transformer-based models are to training data corruption. This work directly addresses **Objective O2** (*Modelling and managing damaging documents*), as "poisoned" data is a specific form of the "negatively relevant" information described in the proposal. By defining defensive strategies against targeted attacks, we are progressing towards mitigating Limitation **L2** (Lack of proper modelling of negatively relevant documents).
- **Graph Neural Networks & Cold Start:** Our work on *"Mitigating Extreme Cold Start in Graph-based RecSys through Re-ranking"* (CIKM 2024) proposes a novel re-ranking strategy that leverages topological features to improve recommendations. This contributes to **Objective O3** (*Optimising the IR output*) by demonstrating how upstream retrieval (or recommendation) can be refined before the

final stage, ensuring that even in data-sparse scenarios (cold start), the system provides useful candidates.
- **Theoretical Foundations & Stability:** We advanced the theoretical understanding of ranking stability with *"Finite Rank-Biased Overlap (FRBO)"*. This directly supports **Objective O1** (*Development of a novel evaluation system*), as it provides a new metric to assess the consistency of IR systems—a crucial step for evaluating cascading systems where ranking instability can propagate errors downstream.

3.2 WP3: Cascading IR/NLP Systems

Activities and alignment with NEREO Objectives:

- **Retrieval-Augmented Generation (RAG) & The Power of Noise:** The study *"The Power of Noise: Redefining Retrieval for RAG Systems"* (SIGIR 2024) constitutes a major breakthrough for **Objective O3** and **Objective O1**. It systematically analyzes the interaction between the IR component and the NLP generator, revealing that "irrelevant" (noise) documents can sometimes benefit generation. This challenges the traditional binary view of relevance (addressing **L1** and **L3**) and lays the groundwork for the "noise-aware" retrieval strategies envisioned in the proposal.
- **Large Language Models for Italian (DanteLLM):** In *"DanteLLM: Let's Push Italian LLM Research Forward!"*, we address **Objective O4** (*Develop new tools for enabling cascading IR/NLP applications*). By releasing a foundational Italian model, we are providing the "experimental environment" (R2) and tools necessary for researchers to build cascading systems that operate natively in Italian, fostering the adoption of these technologies in local contexts.
- **Explainability & Counterfactuals:** The work on *"Human-in-the-Loop Personalized Counterfactual Recourse"* aligns with the project's broader impact goals (aligned with **Horizon Europe's** focus on "human-centric AI"). It ensures that the "reasoning" part of the cascading system is interpretable and actionable, addressing the need for trustworthy AI systems that do not just retrieve and generate, but also explain their decisions to the user.

---

4. Scientific Publications (Year 1)

The following publications were produced during the first year of the project (2023-2024), acknowledging PRIN NEREO support.

2024

Journals

- **A topological description of loss surfaces based on Betti Numbers.**
  Maria Sofia Bucarelli, Giuseppe Alessio D'Inverno, Monica Bianchini, Franco Scarselli, and Fabrizio Silvestri.
  Neural Networks, 178, 106465.
  DOI: 10.1016/j.neunet.2024.106465

Conferences

- **Mitigating Extreme Cold Start in Graph-based RecSys through Re-ranking.**
  Alessandro Sbandi, Federico Siciliano, and Fabrizio Silvestri.
  CIKM 2024.
  DOI: 10.1145/3627673.3680069

- **DanteLLM: Let's Push Italian LLM Research Forward!**
  Andrea Bacciu, Cesare Campagnano, Giovanni Trappolini, and Fabrizio Silvestri.

LREC/COLING 2024.
URL

- **Investigating the Robustness of Sequential Recommender Systems Against Training Data Perturbations.**
  Filippo Betello, Federico Siciliano, Pushkar Mishra, and Fabrizio Silvestri.
  ECIR 2024.
  DOI: 10.1007/978-3-031-56060-6_14

- **Finite Rank-Biased Overlap (FRBO): A New Measure for Stability in Sequential Recommender Systems.**
  Filippo Betello, Federico Siciliano, Pushkar Mishra, and Fabrizio Silvestri.
  IIR 2024.
  URL

- **Rethinking Relevance: How Noise and Distractors Impact Retrieval-Augmented Generation.**
  Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri.
  IIR 2024.
  URL

- **Learning with Noisy Labels through Learnable Weighting and Centroid Similarity.**
  Farooq Ahmad Wani, Maria Sofia Bucarelli, and Fabrizio Silvestri.
  IJCNN 2024.
  DOI: 10.1109/IJCNN60899.2024.10650366

- **Robust Solutions for Ranking Variability in Recommender Systems.**
  Bonifacio Marco Francomano, Federico Siciliano, and Fabrizio Silvestri.
  RecSys 2024 (RobustRecSys Workshop).
  URL

- **The Power of Noise: Redefining Retrieval for RAG Systems.**
  Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri.
  SIGIR 2024.
  DOI: 10.1145/3626772.3657834

- **Personalized Audiobook Recommendations at Spotify Through Graph Neural Networks.**
  Marco De Nadai, Francesco Fabbri, Paul Gigioli, Alice Wang, Ang Li, Fabrizio Silvestri, Laura Kim, Shawn Lin, Vladan Radosavljevic, Sandeep Ghael, David Nyhan, Hugues Bouchard, Mounia Lalmas, and Andreas Damianou.
  WWW 2024.
  DOI: 10.1145/3589335.3648339

- **Human-in-the-Loop Personalized Counterfactual Recourse.**
  Carlo Abrate, Federico Siciliano, Francesco Bonchi, and Fabrizio Silvestri.
  xAI 2024.
  DOI: 10.1007/978-3-031-63800-8_2

2023

- **Adversarial Data Poisoning for Fake News Detection: How to Make a Model Misclassify a Target News Without Modifying it.**
  Federico Siciliano, Luca Maiano, Lorenzo Papa, Federica Baccini, Irene Amerini, and Fabrizio

Silvestri.
PKDD/ECML Workshops 2023.
DOI: 10.1007/978-3-031-74627-7_44

---

5. Conclusion and Next Steps

The first year of the NEREO project has been highly productive, meeting all administrative obligations and producing high-quality research outputs. The successful setup of the project infrastructure and the commencement of core research lines in WP2 and WP3 provide a strong foundation for the second year.

Priorities for Year 2:

- Finalize the CoDIME and ECLIPSE models (WP2).
- Deepen the integration of IR/NLP through advanced cascading techniques (WP3).
- Develop and release the open-source prototype (WP4, D4.2).
- Continue dissemination through high-impact publications and the project website.