Report on the 1st Training School on Domain Specific Systems for Information Extraction and Retrieval (DoSSIER 2022)

Michail Salampasis  
International Hellenic University  
Thessaloniki, Greece  
msa@ihu.gr

Florina Piroi  
TU Wien  
Vienna, Austria  
florina.piroi@tuwien.ac.at

Allan Hanbury  
TU Wien  
Vienna, Austria  
allan.hanbury@tuwien.ac.at

Abstract

The DoSSIER project, an European Training Network, was kicked-off in late 2019. The first PhD candidate recruiting activities took place at the beginning of the Covid-19 pandemic and the first years of the project lived in the on-line universe. Naturally, this posed certain challenges to the cohesion of our network. With some delay due to the lock-downs and travel restrictions, our first Training School was successfully organized in September 2022 and it constituted one of the few occasions we had where the members of DoSSIER network met in person. This report gives an account on the first Summer School organized by DoSSIER, for the DoSSIER Early Stage Researchers, a school set in the picturesque village of Olympiada, which gave amply opportunity of interaction between students and senior researchers in the Information Retrieval domain.

Date: 25–30 September, 2022.
Website: https://dossier-project.eu/1st-dossier-training-school.

1 Introduction

DoSSIER is an EU Horizon 2020 ITN/ETN on Domain Specific Systems for Information Extraction and Retrieval. DoSSIER will elucidate, model, and address the different information needs of professional users.

DoSSIER groups its research activities into three general areas (Applications, Methods, and Models), pertaining to three domains: Innovation, Health, and Legal/Law. These feed into each other to generate new hypotheses, identify new experimental procedures, and bring about a better understanding of knowledge and information needs, and the processes by which the two interact. The programme is implemented via 15 Early Stage Researcher (ESR) projects. Eight of the 15 projects clearly identify one of the three application areas while the remaining seven projects provide horizontal links across domains and contribute to the harmonization of observations,
cross-fertilisation of ideas, and identification of synergies between research groups, domains, and topics.

The results of the DoSSIER research will provide the vital know-how and tools to the professional search industry. DoSSIER aims to make an impact on professional search processes and systems through research in all three areas. An additional key objective is the education and training of Early Stage Researchers. Specifically, within the DoSSIER network we train a new generation of scientifically-principled, creative, entrepreneurial and innovative researchers with the academic and industrial experience necessary to make a significant impact on professional search in Europe, and hence on the European economy. DoSSIER fosters excellence by structuring research and doctoral training to lead to a professional certification of the competences: research knowledge and intellectual abilities; research personal effectiveness; research governance and organization (including ethics and sustainability); and researcher engagement, influence and impact.

In this context, from 25 to 30 of September 2022, in collaboration with the Intelligent Systems Laboratory of the International Hellenic University, DoSSIER has organized in Olympiada (Chalkidiki, Greece) the 1st DoSSIER Training School as an internal training event. The small seaside village of Olympiada is a popular summer destination in the larger area known as Chalkidiki and mount Athos. Only half a kilometre from the current settlement of Olympiada is the location of Ancient Stageira, birthplace of Aristotle.

The school was intended for PhD students working in the 15 DoSSIER projects. The project partners collaborated on designing the most appropriate set of lectures answering both to the students’ expressed requests for specific training topics and to the DoSSIER project aims for researcher training. The 1st DoSSIER Training School was planned as a week-long event consisting of a series of lectures and seminars. The aim of the school was to give grounding in core research topics (e.g. IR Experimentation), but also to provide training in other subjects relevant to their research activities (e.g. Licensing practices, Search in Industry settings).

2 Lectures

The program started with two lectures focusing on the main topics of IR experimentation. Then it moved to User Studies, Information Interaction, Interfaces and Evaluation, and Search Systems for Clinical Decision Support. The next days contained lectures focusing on Challenges of Enterprise Search, Reproducibility in IR and Code Development in Industry settings. Finally, the school was completed with lectures on paths to a productionised service and licensing practices. Each of the twelve lecturers had two to three hours for presenting their topics. A short description of all lectures follows:

- **Modelling and Analyzing Information Retrieval System Performance: From Foundations to Measurement Issues.**
  
  by Nicola Ferro, University of Padua, Department of Information Engineering

  Starting with a brief background on the foundations of Information Retrieval Evaluation, Nicola introduced advanced IR topics for retrieval systems evaluations, including the notion of hypothesis testing and types of errors, as well as approaches on how to deal with multi-
ple comparisons and Family-wise Error Rate (FWER). The lecture also included modelling system effects using Analysis of Variance (ANOVA), i.e. one-way ANOVA, and a detailed explanation on how to incorporate system and topic effects (two-way ANOVA) to obtain a more reliable comparison across different retrieval systems.

• **Towards Better IR Experimentation: Validity and Reproducibility.**
  by Norbert Fuhr, *Universitat Duisburg-Essen, Informatik und Angewandte Kognitionswissenschaft*

Norbert’s talk pointed out common mistakes found in Information Retrieval publications in terms of validity and reproducibility. He explained that validity could be placed into two categories: (1) internal validity: the degree to which a study minimizes systematic error, and (2) external validity: the degree to which the results of a study can be generalized. Reproducibility means the results are reproducible independently. In other words, an experimental result is not fully established unless it can be independently reproduced. Based on his presentation, a summary of common mistakes in IR are: hypotheses are often formulated after the experiment, significance tests frequently ignore the multiple comparisons problem, MRR and ERR evaluation metrics violate basic requirements for a metric, MAP is based on unrealistic assumptions, effect sizes are ignored, reproducibility of the experiments might be nearly impossible, and sometimes authors claim proof by experimentation while proofs are about universally valid statements. Finally, Norbert outlined the possible solutions for overcoming the common mistakes such as Bonferroni correction for multiple comparisons, computing Effect Size, adding Metadata into the raw data, defining workflows in terms of rules, and Containerization with Docker.

• **User Studies.**
  by Katriina Bystrum, *OlsoMet, Faculty of Social Sciences.*

Katriina’s presentation covered three main parts: “who is a user”, “how are user’s information needs and their information sources related in a workplace”, and “why do a user study in a workplace”. Katriina discussed some central concepts for a user study in a workplace setting, such as user, context, information source, and information activities (e.g., information needs, information searching, information seeking). In small groups, the audience discussed how to define the information needs for different types of human users. According to Katriina, information need(ing) is seen as an activity alongside information seeking and searching. Katriina ended the presentation by providing a user study pre-design guide where she shows three study modes: Experimental (pseudo-contextual; often system-centred), Semi-naturalistic (semi-contextual; often person-centred), and Naturalistic (fully contextual; often ethnographic). She explained the aims, goals, methods, impact, challenges, and ethics for each mode.

• **Code Development in Industrial Settings.**
  by Antonis Makropoulos, *contextflow*
Antonis’ lecture presented the challenges of developing a search system for clinical decision support. He presented the example of contextflow, a startup working with radiology images to identify diseased areas in the lungs. He emphasised the need for such systems in the daily work of a radiologist. The presentation, then, focused on the steps necessary to move from the research ideas and results to a commercial product from the perspective of R&D. The first stages of each project should identify available models, datasets and performance metrics on which one could evaluate their solutions. Method development and testing should involve thorough optimisation and documentation and testing for both single performance metrics and user satisfaction of the end-to-end system. He also emphasised the importance of adhering to good coding practices like unit tests, code reviews, and following coding standards (e.g. Python’s PEP-8). Each idea in the presentation was supported by examples from the development of the contextflow SEARCH Lung CT software.

• **Information Interaction, Interfaces and Evaluation.**  
  by Elaine Toms, University of Sheffield, Management School.

This lecture was a continuation of Katriina Bystrom’s lecture in which Elaine defined the conceptual foundations underpinning information search/retrieval. The lecture had two parts. In the first part, Elaine discussed information interaction, interfaces and integrating search. She talked about the context in which knowledge workers operate and defined how to integrate information interaction into knowledge work. Elaine further discussed the importance of graphical user interfaces and provided concrete design principles to designing an interface which take into consideration cognitive tasks and tools. The second part of the lecture focused on the evaluation of the graphical user interfaces. Elaine discussed the landscape of interface evaluation and different aspects of an interface that we can evaluate and then talked about the different elements of interface evaluation. Finally, all the steps involved in designing a solid user study from the data collection to the evaluation were outlined. The talk was then closed with the audience having a discussion about how they would apply the theory they learned to their own research projects.

• **Challenges of Enterprise Search.**  
  by Udo Kruschwitz, Universitat Regensburg, Information Science.

This lecture started with a brief introduction to various types of searches such as desktop search, site search, blog search, database search, expertise search, etc. Various enterprise search examples and problems related to it were discussed. The lecture continued to mention differences between enterprise and web search, consisting mainly of differences with respect to the source of data, to smaller search scale, to lack of interconnected documents. Among the challenges of designing an Enterprise Search system Udo referred to the fact that in non-enterprise settings content is often created for dissemination purposes and not necessarily for answering possible future information needs. Other challenges include the handling of documents with different access restrictions (which a retrieval system may find but not be able to display in the result list), or the handling of redundant, outdated, or trivial data.
• **A New Generation of Neural Search and Knowledge Discovery Tools.**
  by [Jakub Zavrel](https://www.zetaalpha.com), *Zeta Alpha, Amsterdam.*

In his presentation, Jakub made the claim that large pretrained language models will augment human cognition for making decisions. He presented how Zeta Alpha brings this philosophy into reality with their knowledge discovery platform for the scientific domain. In this platform they leverage state-of-the-art neural search to create a contextual knowledge discovery system. He presented use cases of the Zeta Alpha platform where these novel methods are used to overcome known problems of traditional retrieval models, like lexical gap, bridging them with semantic understanding and handling of complex, relational and multilingual queries. Further, he presented their research for domain adaptation of neural search models with unsupervised data generation.

• **Designing Search @Spinque.**
  by [Roberto Cornacchia](https://www.spinque.com), *Spinque.*

“It is time to design your own search engine.” That initial quote in Roberto’s lecture outlines, in a few words, Spinque’s primary mission. Spinque is a spin-off company that leverages clients’ unique expertise to create the best search solutions for them. In his presentation, Roberto describes how they achieve their goal by analysing the different product design phases. By integrating data from multiple sources, they represent implicit information in a knowledge graph. As outlined in several parts of the lecture, an essential part of product design is dedicated to understanding clients’ needs and search tasks before designing the necessary search solutions. The latter involves relevance estimation and is implemented using their state-of-the-art column store database technology. Having outlined these components, Roberto presented two search applications related to cultural heritage and e-commerce, along with a hands-on session that allowed him to provide further details of the product design. Concluding his lecture, he mentioned the importance of the human factor in designing tailored search solutions for complex search problems.

• **Reproducibility in IR for Evolving Document Corpora.**
  by [Andreas Rauber](https://www.tuwien.ac.at), *Technische Universitat Wien, Institute of Information Systems Engineering.*

The subject of the lecture given by Andreas was the need for reproducibility in IR. Indeed, even if a study has a very high scientific value, it is difficult to integrate it in the larger context of research if it is not reproducible. For instance, if documents are changed or deleted, it would affect the ranking list, even if the documents are not present in the ranking. As for user studies, it should be possible to reproduce the experiments with different users. The solution would be to create algorithms that would be robust on changing data or having a different group of participants. One of the solutions would be dynamic data citation, which consists of working the data which contains time of access information, enhancing the means of access to the data (queries) with time stamping and hashing the value computed over the data for local storage. If the data changes, it is necessary to identify the queries
which are affected. Also, column store systems allow tracking of changes in the documents, along with their insertion to and removal from the collections.

- **Licensing Practices in Open Science and LCT tool.**  
  by Marianna Katrakazi, **Legal associate, ATHENA RC.**

Marianna talked about copyright and intellectual property licensing in the Open Science context. She introduced the Licence Clearance Tool, developed within the NI4OS project, that aims to assist researchers in selecting the appropriate licences for sharing their code or data. During this talk, the audience also discussed the researcher’s interests on project outcome licensing, and elucidated how to approach intellectual property licensing schemes from the resource creator roles. Given the rapid and wide distribution of new knowledge published openly and the adopted approaches for sharing outcomes of science, it is important to be aware of the legal implications of this practice. Legal boundaries allow us to protect our work and define the limits of the usage for outcomes of our research. We should ask ourselves what we want to share and how much others could and should use what we share. Since, in general, access to this specific information is limited, Marianna presented the License Clearance Tool, which could help by bringing awareness of, for example, the implications of choosing a specific licence and how licences are compatible between them.

- **From an Experiment to Production.**  
  by Petr Knoth, **Open University, Big Scientific Data and Text Analytics Group**

In his lecture, Petr introduced CORE.uk which emerged from his PhD work and is now one of the world’s largest aggregator of Open Access research papers retrieved from Open Access repositories, with a team of 12 people actively maintaining it. CORE.uk’s mission is to aggregate all Open Access research worldwide and deliver unrestricted access for all. In the process of aggregating content, CORE.uk relies on open protocols, such as OAI-PMH, and has a 6-stage pipeline to create a final index of Open Access resources. The pipeline’s first component downloads the metadata of Open Access articles. According to Petr, CORE.uk now scans over 10,000 data providers in over 90 languages which gives CORE.uk over 218 million metadata records, of which around 28 million include full-text Open Access publications, stored by CORE.uk. CORE.uk has around 30 million monthly active users that access the repository in 4 different uses: 1) content search and discovery, recommendations; 2) access provision to open access content in a machine readable format doing all the heavy lifting with the help of AI, which involves harmonizing metadata, cleaning and creating a structured representation of scholarly data, accessible through CORE.uk’s ”Raw Data Services; 3) own repository management, through a specially designed dashboard, available to academic subscribers, with overviews of repository harvesting status; 4) OIA resolution of metadata record identifiers, harvested and automatically registered as globally unique persistent identifiers (PIDs).
3 Conclusions

Thirteen PhD students attended the school and also discussed and coordinated their efforts to develop a search engine prototype for domain-specific tasks in a scheduled session. The organization of the panel on the topic of “Do we need academia in AI?” in the Agora of Ancient Stageira, birthplace of Aristotle, has been an inspiring moment to bridge ideas on ethics and research from the times of the ancient Greeks to the present society.

The lectures took place in the local municipal cultural center. There were interesting discussions during and after each lecture, and intense individual interactions between trainees and lecturers, which led to ideas for further research collaborations and research internships. The trainees had the opportunity to get to know experienced researchers working on the same, similar or complementary research areas.

Participants and lecturers made very positive comments regarding the location and organization of the training school. The location, a small village, offered an ideal environment where all the participants were focused on the training events. During the lectures the students participated actively by asking questions, which led to extensive off-line discussions with the lecturers during the coffee breaks, lunches, or dinner time. The organized social events allowed all participants to get to know each other and establish friendships.

As the summer school reached its end, after five days of intense attendance and tutoring, the general feeling was very positive and pleasant, making the 1st DoSSIER Training School a very successful event.
Acknowledgements

This event was supported by the DoSSIER project, an EU Horizon 2020 project under the Marie Skłodowska-Curie grant agreement No 860721. Many thanks also to the mayor of Olympiada for offering the local cultural center as the venue to host the lectures and to the local community for supporting the event in general. Finally, thanks must go to the DoSSIER ESRs for contributing the paragraphs with their views and comments about the lectures presented.