DiVA
http://www.diva-portal.org

This is the published version of a paper presented at *EUROCC2 & HPC CHAIR @ U.ÉVORA, Évora, Portugal, 27-29 March, 2023*.

Citation for the original published paper:

Kamal, A M., Milrad, M., Mohammed, A T. (2023)
Code Read: Assessing the Programming Skills and Reflections among Digital Humanities Master Students (CT)
In: *Programming and Data Infrastructure in Digital Humanities: Book of Abstracts* (pp. 4-5).

N.B. When citing this work, cite the original published paper.

Permanent link to this version:
http://urn.kb.se/resolve?urn=urn:nbn:se:lnu:diva-135843

Programming and Data Infrastructure in Digital Humanities

Book of Abstracts

High Performance Computing Centre, University of Évora, Portugal
27-29 March 2023

SPONSORED BY EUROCC2 & HPC CHAIR @ U.ÉVORA

Contents

Opening Talk 1

Julia Flanders
Research Computing for the Rest of Us: Challenges of High-Performance Computing Infrastructure for the (Digital) Humanities 1

Session 1 - Programming in Digital Humanities 1

William Mattingly
Digital Humanities Ecosystem with Python & Machine Learning (IT) 1

Yannick Frommherz
„Hello Humanities!" – A Modular Python Programming Course Targeting the Specific Needs and Requirements of Humanities Students (CT) 1

Francisco Coelho
Digital Humanities with Julia - An Overview (CT) 2

Marcelo Milrad
Combining Programming Education and Computational Thinking in the Field of Digital Humanities (IT) 3

Ana Alexandra Silva
Using Virtual Reality in Vocational Language Teaching Programs (CT) 3

Ahmad Kamal
Code Read: Assessing the Programming Skills and Reflections among Digital Humanities Master Students (CT) 4

Nick Montfort
Exploratory Programming for Arts and Humanities (IT) 5

Session 2 - Data Infrastructure and Data Processing in Digital Humanities 7

Eero Hyvönen
How to Create and Use a National Cross-domain Ontology and Data Infrastructure on the Se-

mantic Web (IT) 7

Miguel Avillez
The High Performance Computing Chair and the Computational & Data Infrastructure Available for the European Digital Humanities (CT) 8

Francis Harvey
Strengthening Computation Skills, Strengthening Digital Humanities, Strengthening Data Infrastructures in Warsaw (CT) 8

Luís Trigo
People First - Testing Integrated Digital Research/teaching Concepts from the Ground up (CT) 9

Francesca Tomasi
Data Models and Knowledge Organization in Digital Humanities (IT) 10

Session 3 - Tourism Data Analytics and Artificial Intelligence 10

Nikolaos Stylos
Big Data Empowered Agility for Dynamic, Volatile, and Time-Sensitive Service Industries: The Case of Tourism Sector (IT) 10

Jaime Serra
How to Handle a Smart Tourism Design Process for Sustainable Destinations Based on Small and Big Data? Evidences from the PISTA Project (CT) 11

Jacques Bulchand-Gidumal
Why Travel and Tourism and its Rich Data Potential is a Great Field to Deploy AI (IT) 12

Session 4 - Data Modelling in History and Cultural Heritage 14

Fiona Mowat
Fostering Digital Transformation in the Common European Data Space for Cultural Heritage - Through Assessment, Measurement and Data Analysis (IT) 14

Ivo Santos
Extracting and Sharing Portuguese Archaeological Knowledge (CT) 14

Tiago Gil
Oxoce - Structured Thematic [Re]Search Engine (CT) 15

Helena Freire Cameron
How Different are Diachronic Spelling Portuguese Variants? the Jaccard Similarity in Historic Portuguese Texts (CT) 15

Tara Andrews
Modelling Historical Data in the RELEVEN Project (IT) 16

Session 5 - Language Processing and Text Analysis 17

Fernando Sanz-Lázaro
Making Readings Readable: a Two-Step Process to Processing Plays (IT) 17

Nuno Miquelina
AiBERTa - An European-Portuguese Language Model (CT) 17

Álvaro Piquero Rodríguez
Semantic Analysis from a Relational SQL Database: a Practical Example (CT) 18

Mariana Pereira
Data Visualization Applied to Glossaries: Exploring Typologies via Employing Echarts.js (CT) 19

Micaela Aguiar
Using Bert to Retrieve Academic and Scientific Language in Small and Large Corpora (CT) . . 19

Closing Talk 21

Maria Zozaya-Montes
Mind the Gap: Gender Bias and Women's Social Representation in AI and DHs 21

Opening Talk

Research Computing for the Rest of Us: Challenges of High-Performance Computing Infrastructure for the (Digital) Humanities

Julia Flanders

Northeastern University, US

Email: j.flanders@northeastern.edu

As humanities researchers increasingly join the user community for high-performance computing infrastructure, they face distinctive challenges in making use of those facilities, which were often designed for very different audiences. But the task of building and supporting research infrastructure that meets the evolving needs of humanists poses complexities as well. What challenges do research computing groups face in supporting these researchers? And how can libraries and digital humanities centers work in partnership with research computing units to train and support humanities users including faculty, students, and staff, across the spectrum of expertise? This presentation will consider those challenges through a set of brief case studies.

Day 1 - Programming in Digital Humanities

Digital Humanities Ecosystem with Python & Machine Learning (IT)

William Mattingly

Smithsonian Data Science Lab, US

Email: wma229@g.uky.edu

Modern artificial intelligence is radically changing all disciplines, including the humanities. Today, AI is largely rooted in machine learning or deep learning. A decade ago, researchers needed a strong background in statistics and computer science to apply machine learning to projects, but this is no longer the case. With the approachability of machine learning through Python frameworks like PyTorch, TensorFlow, Keras, FastAI, and spaCy, digital humanists are in a unique position. If we can learn Python (which is one of the easier programming languages to learn), we can apply machine learning to our projects with just a few lines of code. Given the approachability of Python and the effect machine learning is having on the humanities, it makes more sense for the humanist of today to learn to code than to not.

In this talk, we will be focusing on how humanists can benefit from programming regardless of their project or area of expertise. We will cover subjects ranging from Holocaust history to medieval Biblical studies. To do this, we will maintain a big-picture overview of the benefits of Python and how it can be applied to real humanities projects. We will also look at how programming can be applied to digital humanities methods, such as social network analysis, transcription, and searching a corpus. For this talk, I will be drawing from my experience at the Smithsonian Institution, the United States Holocaust Memorial Museum, the Bitter Aloe Project (which focuses on Apartheid South Africa), and medieval social networks. It is my hope that this talk gives listeners a sense of the benefits of programming and how to begin learning.

„Hello Humanities!" – A Modular Python Programming Course Targeting the Specific Needs and Requirements of Humanities Students (CT)

Yannick Frommherz, & Simon Meier-Vieracker

Technical University Dresden, Germany

Email: yannick.frommherz@tu-dresden.de

At the Chair of Applied Linguistics at TU Dresden, teaching programming skills is part of the basic humanities curriculum designed to prepare students for the increasingly digital world. Building on several years of experience in teaching programming to linguistics students, we are currently developing a modular programming course in Python addressing a broader community of humanities students and researchers, not least from the newly-established DH Master's at TU Dresden. Embedded in virTUos which explores digital teaching and learning, we are designing materials taking into account the specific needs and requirements of our target group: Humanists typically have no prior knowledge of programming, often little technical knowhow in general and sometimes even reservations about technology [1].

We are creating modular Jupyter notebooks which do not presuppose any specific knowledge. We distinguish a basic module from advanced modules. The former contains notebooks covering the basics of programming, relying on examples relevant to DH, i.e., focusing on text rather than numerical data like in many programming textbooks/tutorials. The advanced modules, then, deal with real use cases, both general ones (e.g., web scraping) and ones that are of interest for specific humanities subjects such as automated news factor analysis for communication studies. Thanks to the modular structure, students from diverse subjects and on different levels can learn programming skills that are relevant to them specifically. Being designed for asynchronous learning, students work on the notebooks whenever, wherever and at the pace they wish. As we will publish our materials as Open Educational Resources (meanwhile available here), lecturers can include our notebooks in their teaching both as-is, or easily extend them with modules building on the provided ones, but delving into further topics. The notebooks are complemented by videos, e.g., introducing students to algorithmic thinking (shown to increase learning success, [2]) and show-casing the iterative and rarely straightforward coding reality.

At TU Dresden, we have integrated the notebooks in a Blended Learning setup where students individually study the materials but still meet regularly to address issues as well as to cultivate an open-minded, frustration-tolerant attitude towards programming. Furthermore, we organize hackathons as collaborative learning has been shown to be beneficial in acquiring programming skills [3, 4, 5].

On our talk we want to show an exemplary Jupyter notebook, highlighting our strategies in training our specific target group, humanities students, in programming, as well as sharing insights from teaching using our materials in the last terms.

Digital Humanities with Julia - An Overview (CT)

Francisco Coelho

HPC Chair, University of Évora, Portugal & NOVA LINCS, Portugal

Email: fc@uevora.pt

Julia is a recent programming language, designed for scientific computing and guided by two key objectives: (1) to be easy to use and learn and (2) to be fast. Our aim is to try to motivate digital humanities researchers that

use other languages, such as Python or R, to look at Julia. We focus on the easy to use and learn objective by presenting a few selected examples to highlight how straightforward and expressive Julia is. The talk starts with structured data: records, collections and data frames. Then we follow to a single and short example of how to read data from a file into a dataframe, extract relevant records, perform basic statistics and plot a graph.

Combining Programming Education and Computational Thinking in the Field of Digital Humanities (IT)

Marcelo Milrad, Mohammed Ahmed Taiye, & Sepideh Tavajoh

Linnaeus University, Sweden

Email: marcelo.milrad@lnu.se

The field of Digital Humanities (DH) is still in its infancy, with multifaceted aspects that are very open for debate (Pavlidis et al., 2018; Luhmann & Burghardt, 2022). These debates have raised the interest for developing complementary knowledge using ingenuine human and computer interventions to solve DH-related problems. However, concerns about humanists shying away from considering DH as a discipline have arisen. This can be linked to the blurred boundaries that humanists are familiar with diverse methods that can differ from those applied in DH (Clement, 2016).

Unsurprisingly, computer interventions are not limited to understanding application packages, social media, and web browsing alone but also to learn how to program. More importantly, humanists are bound to ask some central questions before learning to program, like; Why should humanists consider learning to program? And why humanists should not rely alone on open-source (drag and drop) applications designed for ease of use? Existing studies have shown that programming allows humanists to perform both quantitative and qualitative data analysis on a large scale that may be impossible to carry out with other. Nevertheless, reports have shown that learning programming has low satisfaction, high-dropout rates, and can be challenging, especially for non-technical students (Marcolino & Barbosa, 2017)

With regards to the mentioned above challenges, it is pertinent to democratize knowledge and teach humanists programming through a "Thought processing" approach like Computational Thinking (CT) (Wing, 2011). Employing CT concepts in DH may help not to depend on or scavenge works from computer scientists and programmers. CT concepts have the potential help in providing cognitive activities by formulating a problem that admits computational solutions. However, little is known about the epistemological landscape and how humanists apply CT concepts in programming education.

The aim of this paper is to better understand how teaching and learning programming within DH can be improved based on the analysis of existing educational practices. Data collected from submitted assignments & essays from two cohorts of DH master's students from Linnaeus University were the sources for our exploration & analysis. Data was manually coded applying directed content analysis where a CT framework was used as predetermine categories for inquiry (Hsieh & Shannon, 2005). The categories showed strengths, opportunities, and weaknesses as insights for epistemological landscapes of learning programming by these DH master students. These insights can be used to revise, further develop and co-design programming & CT modules in this field.

Using Virtual Reality in Vocational Language Teaching Programs (CT)

Ana Alexandra Silva

University of Évora, Portugal

Email: aasilva@uevora.pt

It is believed that integrating Virtual Reality technology in Vocational Education and Training (VET) classes will increase learners' motivation by being exposed to real-life situations. Smart et al. (2007) define virtual reality as a system that aims to bring simulated real-life experiences by providing topography, motion and physics that give the user the illusion of being in another environment. Similarly, Kim (2005) explains that it consists in reproducing a synthetic experience that represents a virtual or illusory simulation context for the user. In other words, it is an immersive technology capable, not only of imitating real life but also of transporting users to another world where they can negotiate authentic interactions. The aim is to enrich and redefine learning experiences. Studies, ensure that the benefits, face to traditional learning methods, are many. According to Braga (2001), the immersion, interaction and involvement that characterize Virtual Reality, well-conducted in education, brings, among other benefits, "greater motivation of students, allows the learner to develop work at his own pace and stimulates the active participation of the student". In this sense, not only the research of Johnsen at. al. (2007) concluded that the use of VR-based environments facilitates learning, as well as Hassan (2003), explored the use of Virtual Reality in education as an additional tool in the process of cognitive development.

The project VR-VOLL (Virtual Reality for Vocationally Oriented Language Learning) applies an action-research approach to identify where and how VR is likely to add value to vocationally oriented language learning. The benefits to learners must be evaluated, as well as the practical implications for learning providers.

It is the aim of this presentation to disseminate the aims of this project, its results and its activities, such as to identify key language competencies for target vocational areas through a needs and situation analysis. After presenting the Project we will be specifically looking at the target language competencies that healthcare students at the University of Évora feel the need to have when entering the labour world.

Code Read: Assessing the Programming Skills and Reflections among Digital Humanities Master Students (CT)

Ahmad Kamal, Marcelo Milrad, & Mohammed Ahmed Taiye

Linnaeus University, Sweden

Email: ahmad.kamal@lnu.se

Teaching computer programming in the digital humanities (DH) presents a unique challenge given the different epistemologies at play in such a classroom. The following questions guide our efforts: How do students trained in the humanities respond to computational approaches? What should be the pedagogical priorities? Using the case study of our "Programming for Digital Humanities" master course and following a summary of the decisions leading to the course's present curriculum, this paper presents data derived from the analysis of student coding scripts alongside student reflections from the mentioned above on-line course.

Like the DH international master programme at Linnaeus University, the programming course in question is the product of an interdisciplinary team of computer scientists and humanities scholars. This, along with the

sensibilities of students with humanities backgrounds, meant confronting different learning cultures (humanities vs computer sciences), approaches to coding (instrumental programming vs exploratory programming), and learning outcomes (programming skills vs computational thinking).

The talk begins with a synopsis of the course's iterative development as it confronted the tensions described above. Following this, the paper analyses the data to examine students' programming strategies and their experience from the 2020 cohort. Text mining and process mining techniques were applied to the codes written by students. Alongside each code script students also submitted a report reflecting on how they approached the task, issues encountered, and any reference sources they used for "inspiration" when coding. Brennan & Resnick's (2012) framework for assessing computational thinking (CT) was applied to examine the performance of students across various programming tasks. The two dimensions of this framework – computational practices and computational perspectives – were adopted for this analysis, with six sub-themes in total mobilized for the content analysis (Hsieh & Shannon, 2005) of the data.

The CT framework was found to be insufficient and was extended with an additional sub-theme to computational practices: generalizing. Despite recognizing the shortcomings of their solutions and their own limits at discovering more optimal/elegant solutions, several students expressed satisfaction with their attempts, and define them as creative practices rather than an instrumental one.

The talk concludes on how the gained insights have been used to guide our current efforts that included the development of a series of supplementary lectures series (from a non-computer scientist perspective) on exploratory programming (Montfort, 2021) as well as to prime students in earlier, non-programming courses to begin familiarizing themselves with code in self-contained exercises.

Exploratory Programming for Arts and Humanities (IT)

Nick Montfort

Massachusetts Institute of Technology, US, & University of Bergen, Norway

Email: nickm@nickm.com

Exploratory programming involves writing computer programs as a method of inquiry or to enable creative computing. A typical software developer is working toward a known objective to implement a pre-existing design. An exploratory programmer, by contrast, may proceed in a bottom-up, open-ended fashion and discover new aspects of different domains, including those in the arts and humanities. The exploratory programmer uses some of the same methods of a data scientist, but may be also be exploring computation itself and its relationship to platforms and digital media history. Although this approach can be used by advanced programmers, it is also an excellent way to introduce learners to the capabilities of computing, as I present in a book that is now available in print and in three open-access digital formats. Those who learn to program in an exploratory mode can go on to not just design and manage computing projects, but to participate as programmers.

References:

1. Kery, M. B., & Myers, B. A. 2017, "Exploring Exploratory Programming" in 2017 IEEE Symposium on Visual Languages and Human-Centric Computing (VL/HCC), pp. 25-29, IEEE.
2. Marino, M. C. 2020, "Critical Code Studies". Cambridge: MIT Press.

3. Montfort, N. 2021, ”Exploratory Programming for the Arts and Humanities”, 2nd Ed. Cambridge: MIT Press.

4. Montfort, N., Baudoin, P., Bell, J., et al. 2014, ”10 PRINT CHR $(205.5+ RND (1));: GOTO 10”. Cambridge: MIT Press.

Day 2 -Data Infrastructure and Data Processing

How to Create and Use a National Cross-domain Ontology and Data Infrastructure on the Semantic Web (IT)

Eero Hyvönen

Aalto University, Finland

Email: eero.hyvonen@aalto.fi

This invited talk presents a model and lessons learned for creating a cross-domain national ontology and Linked (Open) Data (LOD) infrastructure. The idea is to extend the global, domain agnostic ”layer cake model” underlying the Semantic Web with domain specific and local features needed in applications. To test and demonstrate the infrastructure, a ”Sampo” series of LOD services and portals in use have been created in 2002-2022 that cover a wide range of application domains. They have attracted millions of users in total suggesting feasibility of the proposed model. This line of research and development is unique due to its systematic national level nature and long-time span of some twenty years.

To develop applications, the layer cake model is not enough: domain and application specific infrastructures based on shared W3C standards and best practices are needed, too. These can focus on specific domains, such as medicine, biology, cultural heritage, or geography on an international level. However, in practice one also has to deal with national level issues and data available that are represented using national languages, data models, vocabularies, and are created using conventions of local legacy systems. For example, Cultural Heritage (CH) data in different countries is often nationally specific calling for adapted local solutions for representing and using the data.

Most of the international infrastructure work is focused on collaborations on particular application domains. In contrast, this paper concerns the question: *”How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It on the Semantic Web”*.

This problem is addressed by presenting, discussing, and evaluating approaches and living laboratory experiments developed in Finland during 2002-2022. Presenting lessons learned in this endeavour is hopefully useful in a more general setting, as similar challenges are likely to be faced in other countries, too.

This work is carried out as part of the national Finnish FIN-CLARIAH program for research infrastructures (https://seco.cs.aalto.fi/projects/fin-clariah).

References:

1. Hyvönen, E. 2023, ”Digital Humanities on the Semantic Web: Sampo Model and Portal Series”. Semantic Web, IOS Press (in press)
2. Hyvönen, E. 2022, ”How to Create a National Cross-domain Ontology and Linked Data Infrastructure and Use It”, Semantic Web, IOS Press (under review)

The High Performance Computing Chair and the Computational & Data Infrastructure Available for the European Digital Humanities (CT)

Miguel Avillez

HPC Chair, University of Évora, Portugal & Technical University of Berlin, Germany

Email: mavillez@uevora.pt

The High Performance Computing Chair is a R&D infrastructure (based at the University of Évora), endorsed by Hewlett Packard Enterprise (HPE), and involving a consortium of higher education institutions, research centres, enterprises, and public/private organisations.

One of the key projects of the HPC Chair is the establishment of a computational and data storage & processing infrastructure for digital humanities in Europe. The digital humanities Chair members have access to state-of-the-art computing resources (using CPUs and GPUs), in particular the OBLIVION Supercomputer and the Artificial Intelligence VISION cluster, to run their applications and carrying out, e.g., data analytics, corpus analysis, natural language processing, space imagery of archeological sites, digital twins, etc., to a storage system to backup all their data, and to a data processing and visualization framework that receives the data in different formats, allow its processing and displays the results. In addition, the HPC chair also provides under an ERASMUS+ Advanced Computing Consortium (1.4 M€ for a 3 years period) funding for training, job shadowing, and observations periods in any European facility involved in Digital Humanities research, data infrastructure deployment, and programming. For the time being the HPC Chair focuses on the DHs fields of digital media art, heritage, literature & linguistics, and tourism. In this talk the frameworks for DHs' research and training programs set in place are reviewed.

Strengthening Computation Skills, Strengthening Digital Humanities, Strengthening Data Infrastructures in Warsaw (CT)

Francis Harvey[1,3], Dariuz Gotlib[2], Michał Wyszomierski[2], Marta Kużma[1], Adrian Warsiński[1], Wiesława Duży[1]

[1] *University of Warsaw, Poland*

[2] *Warsaw University of Technology, Poland*

[3] *Leibniz Institute for Regional Geography, Leipzig*

Email: f.harvey@uw.edu.pl

This talk presents a unique collaboration with students for developing digital humanities and contributing to data infrastructures in Poland. At the University of Warsaw (UW) and Warsaw University of Technology (WUT), we are embarking on a project to enhance student learning and draw on this to enrich the digital humanities and support relevant data infrastructures. In the project, students from WUT work with UW researchers on developing computational support for working with data from public institutions in Poland. The project, People, Places, Events, involves the development of knowledge graph approaches to expand possibilities for analysing historical sources and data. The project works with cultural heritage institutions and develops toolkits for interpretation and explanation. Students from WUT fill an important role in the project. They develop the conceptual capacities for analysing and utilising data from existing cultural heritage sources and provide working examples. The scope of their work includes the development of a property graph data model and property graph database with data, preparing sample queries to the database and creating a web-based application to present the result of predefined queries using a graph database. Work includes data entry and preparation, adding additional data,

import of data into a database, initial queries, use of NLP to enrich the database, further development of queries and data analysis methods and use of graph-powered machine learning and creation of a client/server application for preconfigured queries. The student's science club helps, and the five tasks become the foci of diploma theses. Researchers from UW support the students by providing data, providing additional historical data and supplemental data. After the project and theses are completed, the application will be hosted on a UW server and available for the public for further research and projects. The development of this student work strengthens the digital humanities in Poland. It also contributes to strengthening the academic Spatial Data Infrastructure (SDI). The project uses the CENAGIS geo-cyberinfrastructure - a computing centre and repository for geospatial data. These connections allow the ongoing project to use big data technology and innovative geospatial analysis tools.

People First - Testing Integrated Digital Research/Teaching Concepts from the Ground up (CT)

Luís Trigo[1,2]; Carlos Silva[1]; Vera Moitinho[2]; Diogo Marques[2]

[1] *Centre of Linguistics, University of Porto, Portugal*
[2] *Centre for Digital Culture and Innovation, Portugal*

Email: ltrigo@letras.up.pt

The Centre for Digital Culture and Innovation (CODA) started its activity at the beginning of the 2022/2023 academic year with the aim of supporting Digital Humanities (DH) development in the eight research organizations from the Faculty of Arts and Humanities of the University of Porto (FLUP). CODA comprises three multidisciplinary PhD researchers that develop their own projects and integrate existing projects to bring awareness to a comprehensive range of DH methods - from data collection based on real-world objects or texts to analytics and artistic data experimentation.

Like many other DH community members, CODA members share the idea that Digital Humanities should be a collaborative effort. This is also a practical concern, as they acknowledge that their coverage may be very limited regarding the dimension of their work. One of the first tasks was to map the needs and skills that each research centre could bring to a common pool for starting a DH community at FLUP.

CODA also envisioned the educational institution that framed the research centres as an advantage to implement DH as a convergence between machine and human computation. Thus, CODA selected a high-impact research project that could be used as a demonstration of DH methods' potential. Specifically, the Phonology lecturer from the Linguistics MsC course to implement a project-based learning methodology, where students would learn with data, contributing to a little-explored area - Portuguese Creole Studies. They also acquired data concepts that will be useful in their academic and professional lives. The collaboration also reinforces CODA's goals of fostering interdisciplinary studies and collaboration between FLUP research units - from linguistics to historical, geographical, ethnological, and genetic data.

Regarding infrastructure, we opted to use freely available web-based resources, having as reference the Minimum Viable Product approach. Beyond the static interaction through Moodle, shared Google Presentation and Sheets files were used for real-time collaboration in the classroom. Students learned, also by trial and error, how to clean and transform data while developing their algorithmic thinking. They also had the chance to work their data with basic python script through Google Colaboratory web service. Regarding Linked Open Data, students

were introduced to OpenRefine. Student engagement was a success not only in the work they produced and integrated on an open GitHub repository but also as a learning experience as the high results of the course in the pedagogical inquiries confirm.

Data Models and Knowledge Organization in Digital Humanities (IT)

Francesca Tomasi

University of Bologna, Italy

Email: francesca.tomasi@unibo.it

Semantic web technologies have had a strong impact on Digital Humanities (DH) methodologies in managing both document-centric and data-centric collections. One of the most interesting outputs in the DH domain is the ontologies reuse for producing Linked Open Data (LOD) for cultural heritage (CH) objects' description. Both structured data coming from traditional Libraries, Archives or Museums (LAM) database and non-structured or semi-structured data coming from the full-text of literary works have been the basis for the production of new scholarly research projects (collections, editions or archives) as LOD datasets. The data modelling phase has been the point of attention of these projects, since it represents the research questions formalization. Producing LOD on the basis of new data models, as a combination of different ontological approaches, is the way for DH to manage knowledge organization issues. Activities as indexing, content description, subjecting and classifying are the methodologies for enhancing CH data with Semantic Web techniques. In order to let DH people to adhere to this approach the important point is to provide as much documentation as possible, in order the ensure the maximum reusability not just of data, but of the whole process (data model, applications, tools and libraries). The idea is to provide scholars with all the steps of the workflow, to easily replicate the same project with different data. Some case studies, and especially the relevant documentation, will be introduced: Vespasiano da Bisticci as a scholarly digital edition (https://dharc-org.github.io/vespasiano-da-bisticci-letters-de/documentation/index.html), MythLOD as a data collection (https://dharc-org.github.io/mythlod/static/mima.html). Documentation become in these projects the methodology for expressing how to organize the knowledge emerging from data.

Day 2 - Tourism Data Analytics and Artificial Intelligence

Big Data Empowered Agility for Dynamic, Volatile, and Time-Sensitive Service Industries: The Case of Tourism Sector (IT)

Nikolaos Stylos

University of Bristol, UK

Email: n.stylos@bristol.ac.uk

Dynamic, volatile, and time-sensitive industries, such as tourism, travel and hospitality require agility and market intelligence to create value and achieve competitive advantage (Buhalis, 2020; Del Vecchio et al., 2018). Little research exists on the key drivers of big data (BD) use for dynamic, real-time and agile businesses (Mandal, 2019). Aim of the current research is to examine the influence of BD on the performance of service organizations and to probe for a deeper understanding of implementing BD, based on available technologies.

In this vein, an ethnographic study was conducted following an abductive approach. A primary qualitative research scheme was implemented with 35 information technology and database professionals participating in five online focus groups of seven participants each. Analytical themes were developed simultaneously with the literature being revisited throughout the study to ultimately create sets of common themes and dimensions.

The current research reveals that BD can help organizations build agility, especially within dynamic industries, to better predict customer behavioral patterns and make tailor-made propositions from the BD. An integrated BD-specific framework is proposed to address value according to the dimensions of need, value, time and utility.

This research adds to the developing literature on BD applications to support organizational decision-making and business performance in the tourism sector. This study responds to scholars' recent calls for more empirical research with contextual understanding of the use of BD to add value in marketing intelligence within business ecosystems. It delineates factors contributing to BD value creation and explores the impacts on the respective service encounters, as per Jiang & Stylos (2021) call for research.

References:

1. Buhalis, D. 2020, "Technology in tourism-from information communication technologies to eTourism and smart tourism towards ambient intelligence tourism: a perspective article", Tourism Review, 75(1), 267-272.
2. Del Vecchio, P., Mele, G., Ndou, V., & Secundo, G. 2018, "Creating value from social big data: Implications for smart tourism destinations", Information Processing & Management, 54(5), 847-860.
3. Mandal, S. 2019, "The Influence of Big Data Analytics Management Capabilities On Supply Chain Preparedness, Alertness And Agility: An Empirical Investigation", Information Technology & People, 32(2), 297-318.
4. Jiang, Y., & Stylos, N. 2021, "Triggers of consumers' enhanced digital engagement and the role of digital technologies in transforming the retail ecosystem during COVID-19 pandemic", Technological Forecasting and Social Change, 172, 121029

How to Handle a Smart Tourism Design Process for Sustainable Destinations Based on Small and Big Data? Evidences from the PISTA Project (CT)

Jaime Serra[1,2], Maria do Rosário Borges[1,2], Joana Lima[2] & Noémi Marujo[2]

[1] *HPC Chair, University of Évora, Portugal*
[2] *CIDEHUS, University of Évora, Portugal*

Email: jserra@uevora.pt

Recent literature on Smart Tourism based on Big Data, exerted that this topic is an emerging paradigm that is remodeling the theory and practice of tourism (Ardito et al., 2019). As stated by Xiang et al. (2021), Smart Tourism Design is "explicitly focused on the development of digital artifacts that support new and innovative processes, systems and experiences which can then be used to reshape tourism". In this context, the technological layer of smart tourism design is critical for the implementation of this data support decision in tourism destinations. In light of this, the PISTA project was conducted between the years of 2020 and 2022, which aimed to implement

the first Sustainable Smart Tourism platform, at a regional scale, in Portugal, specifically in the Alentejo region. This region has implemented the first sustainable tourism observatory (ASTO) in Portugal aiming to monitor the evolution of sustainable tourism development in Alentejo and since 2018 it is the first region of Portugal integrating the INSTO led by the UNWTO. Currently, ASTO is supporting the implementation of the first sustainable tourism smart tool (PISTA Digital), which aims to contribute to the involvement of all tourism agents in assessing the risks, costs, impacts and limits of regional tourism activity. It also aims to facilitate the identification of opportunities for innovation in the organizations and help identify better solutions for using the scarce resources, within the general principles of sustainable tourism development. Based on the support of a High Performance Computing infrastructure, a prototype was developed and tested with several public and private tourism agents. Data handling and analytics are core competences for the future of Smart Tourism tools as the first results of the PISTA project evidenced: there is real importance of a (real time) data-driven solution to support tourism strategic management decisions at a municipal level, as in terms of economic and social sustainability, municipalities are starting to use this information to support decisions of future investments.

References:

1. Ardito, L., Cerchione, R., Del Vecchio, P., & Raguseo, E. 2019, "Big data in Smart tourism: challenges, issued and opportunities", Current Issues in Tourism, 22 (15), 1805-1809.

2. Xiang, Z., Stienmetz, J., & Fesenmaier, D. R. 2021, "Smart Tourism Design: Launching the annals of tourism research curated collection on designing tourism places", Annals of Tourism Research, 86, 1-7.

Why Travel and Tourism and its Rich Data Potential is a Great Field to Deploy AI (IT)

Jacques Bulchand-Gidumal

Institute for Sustainable Tourism and Economic Development, University of Las Palmas of Gran Canaria, Las Palmas, Spain

Email: jacques.bulchand@ulpgc.es

Travel and tourism is one of the industries with the greatest potential for using big data and AI. There are multiple sources of data, and these data are generated in different geographic locations, at different points in time and by different companies involved in the tourism value chain. And the main management questions are still unsolved. This rich and complex environment creates an ideal field in which to deploy AI systems and applications. As we improve the databases and processing algorithms, the expected results and improvements make the travel and tourism industry one of the best places to experiment with big data and AI.

Tourism data comes from six main categories. First, user-generated content (UGC), which is content posted by users on social media such as social networks, discussion forums and review sites. Second, device data generated while the tourist is at the destination by phones, Wifi networks, RFID and Bluetooth, among others. Third, transaction data generated during travel preparation and while at the destination, with data from web searches, bookings and purchases. Fourth, data from traditional research methods such as questionnaires, in-depth interviews, Delphi methods and discussion groups. Fifth, data about the destination, such as events, weather conditions and traffic data. Finally, other data sources, such as public administration databases and open data that can be used to complement the previous categories.

The collection and aggregation of the aforementioned data would allow the use of AI to generate valuable insights for all stakeholders. However, the generation of such complex and rich databases is one of the main challenges currently faced by tourism. In this sense, up to now most of the research in tourism has been developed using only one or two sources. There are many studies that analyze data from TripAdvisor, and there are studies that combine data from mobile phones and spending patterns. In order for AI to develop and reach its potential in the travel and tourism industry, it is necessary to provide systems with good quality and diverse data. Without these types of databases, AI will not be able to develop its potential.

In fact, at the moment it can be seen that most applications in the industry fall short, providing very simple recommendations to potential tourists and showing a limited understanding of the real dynamics of the sector.

Day 3 - Data Modelling in History and Cultural Heritage

Fostering Digital Transformation in the Common European Data Space for Cultural Heritage - Through Assessment, Measurement and Data Analysis (IT)

Fiona Mowat

Europeana Foundation, The Netherlands

Email: fiona.mowat@europeana.eu

Digital transformation is an important aspect of the European strategy for data (2020) and will support interoperability in the Common European data spaces program launched in 2022, where it will become relevant to all European citizens, but how can we measure it and assess it? Based on data collected in the cultural heritage sector and focusing on self assessment methodologies, this paper will explore work undertaken by the Europeana Foundation and inDICEs partners on collecting data about digital transformation - especially discussing the ENUMERATE self assessment tool and related data. The paper will also outline directions for collaboratively collecting data, as well as choosing and developing indicators for digital transformation. How can data collected be used to further the digital transformation of cultural institutions and their audiences, and prepare them for participation in the future data spaces and a digital single market?"

Extracting and Sharing Portuguese Archaeological Knowledge (CT)

Ivo Santos

HPC Chair & CIDEHUS, University of Évora, Portugal

Email: ifs@uevora.pt

In Archaeology, like in other fields, research begins by a scientific question and by gathering all the available data, ranging from lists of artefacts and sites to 3D models and other forms of registering artefacts, monuments, and sites. Researchers gather the data from literature, contact with their peers, or from pre-existing information in various formats and multiple institutions. Despite some efforts, the community often works in isolation, limiting several stages of the scientific process.

To address this issue, this project aims to extract, organize, and share archaeological knowledge originated by the Portuguese archaeology community. We propose to apply Natural Language Processing (NLP) methodologies to extract archaeological information from a Portuguese Corpus. The outputs will be shared in an Open Access format, following the principles of FAIR data (Findable, Accessible, Interoperable, and Reusable).

Through the application of NLP, we will extract knowledge from a large volume of text, enabling the creation of a comprehensive database of archaeological information. By organizing the extracted information and making it available through Open Access, we aim to increase the accessibility and interoperability of data, enabling researchers to more easily collaborate, exchange knowledge and ideas, and ultimately to foster the development of Portuguese archaeology.

In summary, we propose to extract, organize and share archaeological knowledge from Portuguese context, by leveraging NLP methodologies and FAIR data principles. Although in the beginning, we believe that our initiative has the potential to contribute to Portuguese archaeology by promoting collaboration and knowledge exchange.

Oxoce - Structured Thematic [Re]Search Engine (CT)

Tiago Gil

University of Brasilia, Brazil

Email: tiagoluisgil@gmail.com

The purpose of this presentation is to show the functionalities of ”Oxoce”, an automated system for scanning, organizing, and structuring historical data. The aim of the system is to function as a search engine for a specific period and region. A historical search engine, better said. During the testing phase, the idea is to include only content about the history of colonial Brazil, with an emphasis on the 18th century, at least in its test version, incorporating, as the system is consolidated, other periods and geographies. The tool is capable of ”sweeping” books, articles, theses, and published historical sources, tracking down people's names, dates, places, themes (on two different levels), as well as bibliographical references, identifying the pages of the works where those data were mentioned. These data are all related and structured in an extensive database, allowing the search for specific people at specific times (and spaces) concerning specific topics. The system works, thus, as an internet search engine, but with the possibility of delimiting a historical period and a region and allowing extensive bibliographic surveys on specific periods, making it possible to separate, in the results references to works from the historical data itself. The name ”Oxoce” comes from the Yoruba god of excellent hunting, fishing, and plenty.

Developed in python language, ”Oxoce” is composed of several data collection modules. At this stage of development, the feeding of PDF files with scientific research content, such as books, articles, and chapters by columns, is done manually, observing each work with the code. Once a PDF file enters the system, it undergoes intense processing, which will “scan” the text searching for organized information.

How Different are Diachronic Spelling Portuguese Variants? the Jaccard Similarity in Historic Portuguese Texts (CT)

Helena Freire Cameron

Polytechnic Institute of Portalegre, Portugal & CIDEHUS, University of Évora, Portugal

Email: helenac@ipportalegre.pt

Spelling variants are part of the History of the Portuguese language across times. The same word can have many forms diachronically, due to linguistic or spelling causes, or just to the writer's or printer's lack of literacy.

This variation is a challenge regarding the computational processing of pre-contemporary texts, especially handwritten ones. Words have many spelling variants (we found a word with 20 variants in an 18th-century textual corpus), and they do not exist anymore in the current stage of the language. From the point of view of NLP, if the same word changes across time, we must perceive how similar variants are to the contemporary form.

In this communication, we present an approach of analysis using the Jaccard Similarity index. We constitute a study corpus with handwritten and printed texts from the 17th, 18th, and 19th centuries. We manually standardised them and calculated the similarity of pre-contemporary - contemporary variants pairs using the Jaccard Similarity in Julia Language. We analysed issues like the random use of capital letters until the 20th century, pseudo-etymological double consonants in 18th-century texts, and others. The similarity calculation helps to establish rules that assist modernisation systems that increase the precision of systems, aiming to have fully automatic systems able to standardise pre-contemporary texts and make them available to a broad public.

Modelling Historical Data in the RELEVEN Project (IT)

Tara Andrews

University of Vienna, Austria

Email: tara.andrews@univie.ac.at

The aim of the RELEVEN project is to cast a clearer light on the events of the "short eleventh century" (c. 1030–1095) and specifically to get a better understanding of the ways in which the Christian world was perceived by its inhabitants on the eve of the First Crusade, particularly in the eastern half of Christendom which was far more connected to the global networks of trade and idea exchange than the western half was at this time [1], and in the northern regions where Christianity was in the process of being adopted [2,3,4].

In this period, we encounter a whole range of contradictory opinions, uncertain facts and developing viewpoints that are crucial for reframing our understanding of the period. This leads us to the digital challenge: we must find a way to link and connect large amounts of disparate sorts of data, and specifically we need a way to express our collected knowledge about the eleventh century that allows us to incorporate and model different, and even conflicting, perspectives about the way this period was and is perceived among contemporary witnesses as well as later historians. Within the project we do this by moving from the usual model of "linked open data" to the idea of "linked open assertions", in which we make sure that no data point is divorced from the context – source and/or scholarly authority – in which it was produced. The STAR model (STructured Assertion Record) that we have developed [5,6] must also distinguish between "who made a claim" and "who recorded this claim"; interpretation of the source material, while distinct from acceptance of the source claims, is still a key part of the construction of assertions about the past.

In this talk I will present our model and its recent development with some examples to illustrate its use, as well as discuss our work on setting up adequate validation of the model through use of the Shapes Constraint Language (SHACL) on a Neo4J database. Our eventual aim is that the model should be useful far beyond the history of the eleventh century, and that we can contribute our work to a wider conversation about how to avoid divorcing data from the context of its creation.

References:

1 Shepard, J. 2017, "Storm Clouds and a Thunderclap: East-West Tensions towards the Mid-Eleventh Century" in Byzantium in the Eleventh Century: Being in Between, M. Whittow & M. D. Lauxtermann (Eds.), 127–53. Abingdon: Routledge.

2 Ivanov, S. 2009, "Religious Missions" in The Cambridge History of the Byzantine Empire, c. 500–1492, J. Shepard (Ed.), 305–32. Cambridge: Cambridge University Press.

3 Berend, N., Urbańczyk, P., & Wiszewski, P. 2013, "Central Europe in the High Middle Ages: Bohemia, Hungary and Poland, c.900–c.1300". Cambridge Medieval Textbooks. Cambridge: Cambridge University Press.

4 Bagge, S. 2014, "Cross and Scepter: The Rise of the Scandinavian Kingdoms from the Vikings to the Reformation". Course Book. Princeton: Princeton University Press.

5 Baillie, J., Andrews, T. L., Romanov, M., Knox, D., & Vargha, M. 2021, "Modelling Historical Information with Structured Assertion Records" in Digital History Berlin. Berlin: Hyphoteses.

6 Andrews, T. L., Ebel, C., & Deierl, M. 2022, "Gender Assignment as an Event: A Contemporary Approach to Adequately Depict Historical Gender Categories" in Digital Humanities 2022, 111–13. Tokyo, Japan.

Day 3 - Language Processing and Text Analysis

Making Readings Readable: a Two-Step Process to Processing Plays (IT)

Fernando Sanz-Lázaro

University of Vienna, Austria

Email: fernando.sanz-lazaro@univie.ac.at

Digitised texts intended for human reading are poorly suited for digital analyses. They require preparation and structuration before they can be the object of distant reading analyses. We have developed a two-step workflow that allows efficient processing of the texts, which shows its results in our extensive annotated tabular corpus and a number of XML-TEI encoded plays published. We discuss a heterogeneous human and technical ecosystem here, as each team member should be able to encode and analyse texts independently with minimum training and regardless of their computer specifications. Therefore, portability is an unavoidable prerequisite. According to this constraint, we relied on interpreted languages supported in a wide range of platforms, namely Python and R, to code our tools.

The first step is cleaning up and preparing the texts. Formats intended for reading, such as PDF or DOC, provide little information on the external structure of the play that the computer can understand, so they must be converted to plain text while keeping the structural marks the visual appearance conveys. This task requires some degree of human interaction and a platform that provides tools for text manipulation, such as complex substitutions and scripting capabilities. In this case, UNIX-like systems have proven to be much more efficient than other alternatives, provided they are equipped with the GNU flavour of the standard command line tools. However, manipulating text strings does not suffice as visual traits, such as typographical effects or a particular alignment, must also be translated into structural features. We have found that LibreOffice.org provides a suitable environment for addressing this kind of text manipulation.

Once the text structure is explicit, the second step can take place. The machine can assume most of the workload to extract the text for lexical analyses or encode it according to its dramatic features as a CSV or XML-TEI file. Text edition and command line substitutions are particularly in handy here, as minor corrections on the source texts or the output file are relatively frequent. Then, the verses may be metrically scanned. This task is very computing intensive and requires a reasonable amount of time, which can be reduced using a powerful CPU or a CUDA-capable GPU and specific binary libraries that compromise the portability. The outcome is data ready for anyone's statistics software of choice.

AiBERTa - An European-Portuguese Language Model (CT)

Nuno Miquelina, Paulo Quaresma, & Vitor Nogueira

University of Évora, Portugal

Email: d37384@alunos.uevora.pt

Building a language model needs a text dataset to train the model. Independently of the target language (or languages), the challenges are the same: gather text content, prepare the data for the training, train the model and evaluate the created model. In our case, the target is a European-Portuguese model and for that, we require Portuguese written text.

Arquivo.pt is a research infrastructure that allows searching and accessing Portuguese web pages archived since 1996. The main objective of this repository is the preservation of information published on the Web for research purposes. Arquivo.pt provides API services to pull those saved web pages. Nevertheless, since the archived content is from websites, it is embedded in HTML tags. Therefore, in order to obtain plain text, we needed to parse such content. Moreover, as a means to avoid repeated text in the dataset, we created a hash token for each text/phrase.

A dictionary was produced from the words extracted from the collected text. This dictionary allows converting the words (or parts of words) into numeric tokens. To train the model, the sentences are converted to token vectors (using the created dictionary), and applied to train the algorithm. This processing workflow was assembled with Python services, running autonomously from the web pages fetch until creating a phrase dataset ready to be processed by the training algorithm.

In order to address the computing power needed in the training process, we resorted to the University of Évora's High Performance Computing Center, more specifically, the Vision system. Up until now the model was trained with a subset of Portuguese written periodicals archived websites, and there is ongoing work to use more text content from Arquivo.pt.

After the model is trained, the next step will be to fine-tune it to perform natural language tasks like Part-of-speech (POS) tagging. This is ongoing work and aims to compare the performance of the POS with other models trained with Portuguese (Brazilian and European Portuguese) or with Multi-Language trained datasets.

Semantic Analysis from a Relational Sql Database: a Practical Example (CT)

Álvaro Piquero Rodríguez

Complutense University of Madrid, Spain

Email: alvaropiquero@ucm.es

The PhD project La imaginería en la poesía erótica de los Siglos de Oro (Imagery in the erotic poetry of Golden Age) was focused on the analysis of the vocabulary of more than five hundred Spanish sexual poems from the 16th and 17th centuries. With more than 1,300 words annotated throughout the study, the writing of the work required the development of a digital SQL relational database methodology to organise and retrieve the information in an orderly structure. The purpose of this contribution is to describe in detail its creation, its structure and its practical application in research, apart from demonstrate that SQL relational databases can be a suitable tool for the lexical-semantic analysis of literary texts. To sum up, this work aims to offer other researchers some theoretical basis for the development of similar projects in terms of theory or methodology.

References:

1. Navarro Colorado, B. 2015, "A computational linguistic approach to Spanish Golden Age Sonnets: metrical and semantic aspects", in Proceedibngs of the Fourth Workshop on Computational Linguistics for

Literature, Feldman, A., Kazantseva, A., Szpakowicz, S., & Koolen, C. (Eds.), Denver: Association for Computational Linguistics, pp. 105-113.

2. Piquero, Á. 2021 "La imaginería en la poesía erótica de los Siglos de Oro", Ph.D. Tesis. Complutense University of Madrid.

Data Visualization Applied to Glossaries: Exploring Typologies via Employing Echarts.js (CT)

Mariana Pereira, & Sílvia Araújo

University of Minho, Portugal

Email: maridoras79@gmail.com

Glossaries are a collection of terms and their definitions frequently used within a specific field or subject area (Grimaldi & Zanola, 2021). They provide a quick and easy reference for the reader to understand the meaning of technical or specialized terms used in a particular context. Glossaries vary from simple lists of terms and definitions to more complex sets that assimilate cross-referencing capabilities. Unlike taxonomies and thesauri, glossaries necessarily imply a hierarchical relationship between terms and alphabetical order (Navigli & Velardi, 2008).

This work seeks to connect information visualization with glossaries to develop interactive and instigating ways for viewers to access the content. The combination of both may result in an engaging resource for the community and a relevant tool to promote the dissemination of knowledge. The hierarchical data structure of glossaries is suitable for visualization in treemaps, layered icicles, or node-link diagrams; all these typologies have advantages and disadvantages, especially regarding the data exploration perspectives (Burch et al., 2020).

This work will explore the visualization of glossaries in a radial tree, sunburst, and icicle employing visual metaphors, zoom, filter, and other techniques. As a result of the hierarchical nature of the data, the possibility of exponential expansion of hierarchies represents a challenge in the constitution of the visualization, which may result in inefficiency in the hierarchical representation in space and the distribution of contents (Burch et al., 2020). The exploratory research found that the radial arrangement of trees provided the most appropriate layout for visualizing hierarchies due to the space-filling and leaf visibility of the chart.

References:

1. Burch, M., van de Wetering, H., & Klaassen, N. 2020, "Multiple linked perspectives on hierarchical data" in Proceedings of the 13th International Symposium on Visual Information Communication and Interaction, 1–8. New York: Association for Computing Machinery.

2. Grimaldi, C., & Zanola, M. T. (Eds.) 2021, Terminologie e vocabolari - Lessici specialistici e tesauri, glossari e dizionari (Vol. 129). Firenze: Firenze University Press.

3. Navigli, R., & Velardi, P. 2008, "From Glossaries to Ontologies: Extracting Semantic Structure from Textual Definitions" in Ontology Learning and Population: Bridging the Gap between Text and Knowledge, Paul Buitelaar (Ed.), 71–87. Amsterdam: IOS Press.

Using Bert to Retrieve Academic and Scientific Language in Small and Large Corpora (CT)

Micaela Aguiar, Sílvia Araújo, & José Monteiro

University of Minho, Portugal

Email: maguiar60@gmail.com

In this talk, we describe our work on using BERT (Bidirectional Encoder Representations from Transformers) (Delvin et al., 2019) to retrieve academic and scientific language from small and large corpora within the research project PortlinguE – Multilingual Portal for Specialized Languages: mining open data for cross-language information retrieval.

Our work with BERT was carried out to develop a search engine that can find scientific terminology in both Portuguese and English (a bilingual terminology search) and enable users to search an academic phrasebank we built without the need to browse through extensive lists of expressions and their functions (an academic writing assistant). To make this search engine available to a wider audience, we have developed a platform to support academic language (Bailey, 2018) and specialized languages (Afonso & Araújo, 2019) called Lang2science and we are excited to bring this resource to the academic community.

To achieve this, we used BERT in three distinct ways. First, we used it to find bilingual terminology in a large corpus of scientific abstracts retrieved from national repositories. As scientists, researchers, students, and translators face difficulty finding scientific texts aligned with their translations, our goal was to find translation equivalents in texts that are not translations of each other. We utilized two models pre-trained on Portuguese and English corpora to accomplish this task.

Second, we used BERT to identify similar academic language from a previously manually annotated corpus of academic expressions and their corresponding communicative functions, extracted from 40 scientific papers in European Portuguese. We aimed to semi-automatically expand our academic phrasebank by identifying similar expressions and their functions in a small corpus of 80 PhD theses and 80 master dissertations.

Finally, we used BERT to transform our academic phrasebank, which included phrase templates and in-context examples, into semantic vectors in order to match user queries to the appropriate expressions and functions. Students will have the ability to search for academic expressions, such as “in conclusion”, and receive results that correspond to the intended communicative function rather than relying solely on keyword matching. This empowers students with a limited academic vocabulary to expand their lexical and phraseological repertoire by exploring a wider range of academic language options.

Our work highlights the potential of deep learning models such as BERT for retrieving academic and scientific language from small and large corpora and we believe it contributes to the study of academic and scientific language using deep learning.

References:

1. Afonso, T., & Araújo, S. 2019 ”Abordagem heurística das linguagens de especialidade com recurso à linguística de corpus: caso de estudo em linguagem jurídica”. Polissema, Revista de Letras do ISCAP, 19, 9–34.

2. Bailey, R. 2018, ”Student writing and academic literacy development at university”, Journal of Learning and Student. Experience, 1: 7.

3. Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. 2019, ”BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding” in Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1, Burstein, J., Doran, C., & Solorio, T. (Eds.), Minneapolis: Association for Computational Linguistics, pp. 4171–4186.

Closing Talk

Mind the Gap: Gender Bias and Women's Social Representation in AI and DHs

Maria Zozaya-Montes

CIDEHUS, University of Évora, Portugal

Email: mzozayam@uevora.pt

”Mind the gap” succinctly warned of the dangers of falling on the London underground if the space between the platform and the train was not noticed, which could definitively distance the journey from the desired station. The expression has given titles to various metaphors associated with inequalities. It has also been applied to gender, from the sociology of the political vote (”gender gap”) or the difference in wages (”gender wage gap”). Here it is applied to Artificial Intelligence and Digital Humanities to prevent it from leading to unwanted places, avoiding perpetuating gender inequality.

The ”gender bias” expresses the biases that humans reproduce following the perception of gender. Gender is the social construction around capabilities and expectations attributed to a person depending on their sex (male/female). Gender representation will lead to assigning a place and a role in society, family, or work. Such gender preconceptions were forged and settled through practices and laws during the XIX-XX centuries, creating multiple social representations reinforcing them. Currently, they are projected in digital technologies. Feast (2019) recalls that this inherent human bias is transferred to AI as a pure social reflection. The bias ranges from voice recognition programs to the programming of algorithms to select workers (Smith & Russtagi, 2021). The best-known examples range from automatic translations to the reading of images or the automated chats of telephone operators and female figures who perform service work, such as the globalized Alexa or Siri (Feast, 2019), followed by various national variants (Ortiz, 2019).

This study draws the state-of-the-art on the forms of gender bias collected by recent research in AI. Secondly, it analyzes the ways of projection on HD and the models perpetuating gender biases. Thirdly, it searches for the spaces and phases where the possibilities of introducing bias lie (programming, language, design). The objective is to propose possible solutions to neutralize gender bias, proposing measures that meet the supranational or transnational policies that currently suggest ethical evaluation guides in AI (Hagendorff, 2020). Werker has highlighted that, although it is challenging to implement inclusive strategies, they must be established using the convergence between engineering and social sciences and that the more inclusive AI innovations are, the more significant impact they will achieve in society (2021).

References:

1. Bryden, J. 2017, ”Inclusive Innovation in the Bioeconomy: concepts and directions for research”, Innovation and Development, 7, 1-16.
2. Feast, J. 2019, ”4 Ways to Address Gender Bias in AI”, Hardvard Business Review.
3. Hagendorf, T. 2020, ”The Ethics of AI Ethics. An Evaluation of Guidelines”, Minds and Machines, 30, 99-120.
4. Foster, C., & Heeks, R. 2013, ”Conceptualising Inclusive Innovation: Modifying Systems of Innovation Frameworks to Understand Diffusion of New Technology to Low-Income Consumers”, European Journal of Development Research 25, 333–355.
5. Ortiz de Zárate Alcarazo, L. 2023, ”Sesgos de género en la inteligencia artificial”, Revista de Occidente., 502, 328-344.
6. Smith, G., & Russtagi, I. 2021, ”When good Algorithms go sexist: Why and How to Advance AI Gender Equity”, Stanford Social Innovation Review.
7. Werker, C. 2021, ”Gendered research: make women visible”, Delta Journalistic platform TU Delft.