Workshop on Big data analytics and AI for STI Policy

Programme

17 – 19 April 2024

Wednesday 17 April 2024

09h00-10h30 (1) General introductory presentations – overview of relevant projects, experiences and expertise, or future events at the two institutes

> ***Johann Mouton, Robert Tijssen*** and ***Rodrigo Costas*** - welcome and introduction to workshop *(duration: approx. 30 minutes)*

> ***Sergio Salles-Filho -*** general presentation on University of Campinas, Department of Science and Technology Policy, School of Applied Sciences and the Laboratory of Studies on Research Organization and Innovation - LabGEOPI *(approx. 30 minutes)*

> ***Nicholas Vonortas -*** general presentation on the InSySPo Project**:** scope, objectives and initiatives *(approx. 30 minutes)*

10h30-11h00: Break

11h00-12h30: (2) Data infrastructures: presentations with general information and introductory overviews (including CWTS representative)

> ***Nees Jan Van Eck*** *(head IT, CWTS, Leiden Univ.):* "Embracing openness and AI: new projects and developments in the CWTS data infrastructure" *(online presentation; approx. 45 minutes)*

> ***Herman Redelinghuys a****nd* ***Johann Mouton*** *"The CREST data infrastructure"* In this presentation, we will give a brief overview of the history and current status of our scientometric data-infrastructure. We will highlight some of the unique and proprietary databases that we have developed over time (*SA Knowledgebase* and *SA Theses database*) and how we have managed to create meaningful linkages between these databases and, for example, the Web of Science database that we manage. We will then show how – through these linkages – we have been able to produce more meaningful scientometric and bibliometric reports and address a whole host of new and interesting research questions *(duration: approx. 45 minutes)*

12:30-13:30: Lunch

13h30 – 15h00: (2) Data infrastructures: presentations and discussion on project applications and detailed issues (problems and challenges, options and opportunities)

> ***Alysson Mazoni*** *and* ***Rodrigo Costas*** on the InSySPo infrastructure and experience with publications, patents, altmetrics *(approx. 45 minutes)*

> ***Adriana Bin*** *and* ***Sergio Salles-Filho*** on the Research on Research project data infrastructure *(approx. 45 minutes)*

15h00 – 15h15: Break

15h15 – 16h00: *(3) A scientometric observatory: challenges and opportunities*

> ***Johann Mouton*** *and* ***Graeme Mehl*** "Constructing a national scientometric observatory" CREST will present a short demo of the current status of a scientometric observatory (the SASTI Observatory) that has been under development for the past eighteen months. We

will explain the thinking behind it, the expected value and benefit of the observatory to different stakeholders. But we will also go into some of the technical challenges that underlie the construction of the different sections of the Observatory and how these pull data in on hundreds of indicators from various bibliometric and other databases *(approx. 30 minutes, including discussion)*

17h00: Wine tasting and dinner at the Lanzerac Hotel (transport from guesthouses to Lanzerac will be arranged)

Thursday, April 18, 2024

09h00-10h30 (4) Presentations on ongoing or new research projects involving Big Data and AI elements

> ***Robert Tijssen*** How can evidence-informed STI policy benefit from AI? Pros and Cons of AI Tools as Policy Advisors *(approx. 30 minutes)*
>
> ***Alysson Mazoni*** on clustering experiences at Campinas (publications, Twitter clustering, Wikipedia, etc.) and use of LLM for labeling of clusters of publications *(approx. 30 minutes)*
>
> ***Roberto Marcondes*** *and* ***Jesus Mena-Chalco*** on Introduction on LLM models and multimodal AI, and "AI related works" *(approx. 30 minutes; possible online presentation by JMC?)*

10h30-11h00: Break

11h00-12h30: (5) Brainstorming on research and infrastructural collaborative opportunities, outlining strategies for funding opportunities, research agenda alignments, exchange programs, existing and new partnerships, etc.

> **Nick Vonortas/Rodrigo Costas** on "alternative data sources and analytics" (e.g. regional innovation engines, science communication, policy impact, funding and funding flows, 'multiversatory', open science, etc. *(approx. 30 minutes)*
>
> **Adriana Bin** on "Reviewers data and study of biases", pseudo-experimental settings, SDGs delineation *(approx. 30 minutes)*
>
> **Johann Mouton and Ahmed Hassan**: On getting the balance right between algorithmic bibliometric results and contextual knowledge: examples from a study on STI funding in Africa. *(approx. 30 minutes)*

12h30-13h30: *Lunch break*

(Afternoon free)

18h00*: Dinner at Decameron restaurant (Stellenbosch city center)*

Friday morning April 19, 2024

09h30 -10h30*:* (6) Preparing the groundwork for establishing a *Community of Practice* (CoP) around *Big Data and AI for STI policy* for Brazil and South Africa; declaration of intent to develop an action plan

> ***Rodrigo Costas:*** on Barcelona Declaration on open research information systems (*approx. 30 minutes*)

> **Sergio Salles-Filho** on CEPID main research lines (*approx. 30 minutes)*

> ***Robert Tijssen*** Studying the regional relevance of universities: an opportunity for 'big data' cooperation and joint institutional learning? (*approx. 30 minutes)*

10h30-11h00: Break

11h00 – 12h30*:* (7) Discussion and decision-making on organizational frameworks and strategies for further collaboration; selecting joint activities and projects for 2024/2025

(8) Road map for producing joint outputs in 2024: (a) *Discussion Paper* (for blog post and/or presentation at InSySPo conference); (b) *Project Report* for funders; (c) first outline for a joint *Research Proposal* to NRF and/or FAPESP regarding short-term and longer-term joint activities; (d) developing joint activities and producing outputs in 2025/2026

12h30: Lunch and departure

Participants

Isabel Basson is a postdoctoral research fellow at the Centre for Research on Evaluation, Science and Technology (CREST) at Stellenbosch University. She holds a Ph.D. in Science and Technology Studies and her dissertation focused on investigating the potential citation advantage of open access academic journal articles. She also holds an MPhil in Social Science Methods and a BA in Humanities. Previously, she was a postdoctoral fellow at the School of Library and Information Science (École de bibliothéconomie et des sciences de l'information) at the University of Montreal. There she continued her work in developing indicators while working on the IDRC (International Development Research Centre) project on Open access in Africa: Indicators, infrastructure, and policies. During her previous stay at CREST, she was involved in the State of the South African Research Enterprise report by investigating R&D indicators. Her current research interests are in scientometrics with a focus on open access journal publications.

Adriana Bin has a degree in Food Engineering from the State University of Campinas (2000), a master's degree (2004) and a doctorate (2008) in Scientific and Technological Policy from the same University and a post-doctorate at the Manchester Institute of Innovation Research (MIoIR) at the University of Manchester. She is a professor at the Faculty of Applied Sciences of the State University of Campinas (FCA/UNICAMP) in the area of administration, a full professor in the Postgraduate Programs in Administration and in Scientific and Technological Policy (DPCT/IG/UNICAMP), a collaborating professor at the Interdisciplinary Master's Degree in Applied Human and Social Sciences (FCA/UNICAMP) and associate coordinator of the Laboratory for Studies on Research and Innovation Organization (Lab-GEOPI). She works on research and extension projects in the areas of planning and management, evaluation and prospecting in science, technology and innovation.

Roberto Marcondes Cesar is full Professor at the University of São Paulo (USP) since 2008. He has a degree in Computer Science from the Universidade Estadual Paulista Júlio de Mesquita Filho (IBILCE - UNESP - 1992), a master's degree in Electrical Engineering from the State University of Campinas (UNICAMP - 1993) and PhD in Physics from the University of São Paulo (USP - 1997). He is currently a Full Professor at the Department of Computer Science-IME-USP. He is an Associate Researcher at InovaUSP. He was a member of the FAPESP Computing Area Coordination (2006-2012) and the Capes Triennial Assessment Committee (Computing (2007-2009; 2010-2012). He was Director of the eScience Research Center at USP and Head of the Department of Computer Science - IME-USP. He was a researcher at CTBE - CNPEM (2010-2012). He was an advisor to the Scientific Director and Program Coordinator of FAPESP's CEPIDs and CPEs (since 2006-

2023). He is a member of the State Academy of Sciences de SP (ACIESP). Presented invited talks at different conferences (SIBGRAPI 2011; SHAPES 2.0 - 2012; eSon - IEEE eScience 2013). Co-creator of scriptLattes. Has experience in the areas of Computer Science (mathematical modeling and engineering), with emphasis on computer vision, pattern recognition; image processing, bioinformatics and eScience (articles and citations at http://scholar.google.com/citations?user=yGsck-AAAAAJhl=en)

Rovashni Chetty is a Sociologist in the School of Social Sciences at the University of Kwa-Zulu Natal. Mrs. Chetty holds an LLM, and is currently pursuing a PhD in Community Development. With extensive experience in criminology, criminal justice, victimology, sociology, and community development, she serves as a lecturer in Sociology at the University of KwaZulu-Natal in Durban. Her research focuses on sociology, law, criminal justice, criminology, and victimology. She has published in these fields and presented papers at international and local conferences.

Rodrigo Costas is a senior researcher at the Centre for Science and Technology Studies (CWTS) at Leiden University (the Netherlands). Rodrigo is also an Extraordinary Associate Professor at the Centre for Research on Evaluation, Science and Technology (CREST) of Stellenbosch University (South Africa). He holds a PhD in Library and Information Science from the CSIC in Spain. His areas of expertise include the fields of information science, scientometrics, and social media metrics. At CWTS he leads the research line in 'altmetrics', focused on developing new theoretical and analytical approaches to study the interactions between social media and science. Some of his other topics of research include the development of advanced scientometric studies of individual scholars and the study of funding acknowledgments as a new source of scientometric data. In addition to his research activities, Rodrigo is also involved in various European and contract research projects.

Maoka Andries Dikotla is Associate Professor in the Department of Information Science at the University of South Africa. Prof Dikotla holds a bachelor's degree, Honours degree, a master's degree in information studies from the University of Limpopo, and a Doctor of Philosophy from the University of Fort Hare. He also holds a Postgraduate diploma in Higher Education from Durban University of Technology. Prof. Maoka Dikotla holds a position of an Associate Professor in the College of Human Sciences, Department of Information Science, at Unisa. Before joining Unisa, Dikotla worked as a Senior Lecturer in the Department of Information Studies at the University of Limpopo. He has extensive experience both as an LIS practitioner and as a LIS educator. As a LIS practitioner, Dikotla worked as a Librarian at Tshwane University of Technology, University of South Africa, University of Limpopo, Monash University of South Africa, and as an Information and Knowledge manager at the Limpopo Department of Public Works. Dikotla's area of interest includes records, information and knowledge management, school and public libraries, teaching and learning, information organisation and retrieval. Dikotla is a member of the Library and Information Association of South Africa (LIASA) in good

standing and was a chairperson of LIASA Limpopo Province in 2016-2018. He has presented at various conferences and has also published in a number of articles in accredited journals. As part of academic citizenship, Dikotla serves as a reviewer for various journals, including the South African Journal of Library & Information Science and Mousaion: South African Journal of Information Studies.

Marcel Dunaiski is a lecturer at the Computer Science Division at Stellenbosch University and researcher at the School for Data Science and Computation Thinking. He completed all his degrees at Stellenbosch University and his PhD with the title: Dissertation: Using test data to evaluate rankings of entities in large scholarly citation networks. His research falls into two overlapping fields: informetrics and natural language processing. In terms of informetrics, he works with publication data. More specifically, he is interested in any data that deals with scientific research output from journal articles to open-source software repositories on GitHub. His focus so far has been the evaluation of impact indicators (such as the h-index) and ranking algorithms (such as PageRank) on publication databases to understand how well and how fairly they rank researchers and papers. His current focus is on South African research output stemming from higher education institutions. In terms of natural language processing, he focuses on transformer-based large language models, in particularly in low-resource settings. He also tries and combines natural language processing techniques with bibliometric data in order to gain further insights from the data contained in articles and dissertations.

Kyle Ford is a senior research assistant and data analyst at CREST. He studied at the University of the Western Cape and completed his BA Honours in 2015. Kyle previously worked at the City of Cape Town as a development information analysis intern.

Ahmed Hassan is a postdoctoral researcher at the Center for Research on Evaluation, Science, and Technology (CREST). Dr. Hassan has a Ph.D. in computer science with a focus on artificial intelligence. He has an M.Sc and B.Sc. (Honours) in Mathematics. He received the Sudanese Minister of Higher Education Prize in Mathematics. He recieved the third prize of the Center of High-Performance Computing poster competition for PhD students in science and engineering in South Africa. Dr. Hassan was a postdoctoral researcher for the Research Chair on Machine Learning which is a partnership between the University of Pretoria and Multichoice. Dr. Hassan worked on the automated design of video quality assessment metrics and produced metrics that outperformed the state-of-the-art metric (VMAF) developed by Netflix. His work in video quality metrics was highlighted as one of the flagship projects of the Research Chair for the past 6 years. Dr. Hassan worked on hyper-heuristics and improved the performance of the best-known hyper-heuristic for the CHeSC International Research

Challenge hosted by the University of Nottingham and attracked considerable attention. He developed Dynamic Heuristic Sets (DHS) to improve the best-known hyper-heuristic for the challenge (here is the paper). DHS is available in github and the Javadoc can be found here

Mpho Mafata currently works as a post-doctoral fellow with the database team at the Center for Research on Evaluation Science and Technology (CREST). Previously, she was a joint post-doctoral fellow at the South African Grape and Wine Research Institute (SAGWRI), as well as the Stellenbosch University's School for Data Science and Computational Thinking. Her research is in the applications of Data Science in the fields: Analytical Chemistry, Biotechnology, Sensory evaluation, and Oenology. Currently, I work on transdisciplinary approaches for chemometric and sensometric data fusion. Her career goal is to extend her skills in data analysis to include more open-source software for wider applicability and independence. She aspires to continue to be an interdisciplinarian in integrating and interpreting data from different sources.

Alysson Mazoni is currently a post-doctoral research fellow in the Innovation Systems, Strategies and Policy project (InSySPo) in the Department of Science and Technology Policy, at University of Campinas (UNICAMP). He holds a MSc and PhD on Mechanical Engineering, both from UNICAMP, studying vibration control of structures and the application of artificial intelligence to mechatronic systems, respectively. He also works with processing and modelling for biomechanical signals and image techniques for fossil photogrammetry. His research at InSySPo considers the possible applications of big data to citation analysis and innovation impact studies.

Graeme Mehl is a software engineer and database specialist at CREST. He has extensive experience in building and maintaining database driven websites and mobile applications. He is currently managing the SASTI platform as well as some of the large databases hosted by CREST. Graeme holds a Mechanical Engineering degree and is currently completing his MPhil in science and Technology studies.

Johann Mouton is the director of the DSI-NRF Centre of Excellence for Scientometrics and STI Policy and and professor at CREST. He is on the editorial board of five international journals including *Science* and *Public Policy, Science, Technology and Society* and *Minerva*. He has authored or co-authored 10 monographs including *The practice of social research* (2002, with E Babbie), *How to succeed in your Masters and doctoral studies* (2001) and *Doctoral education in South Africa* (with Nico Cloete, 2015). He has also edited or co-edited 9 books, published 90 articles in peer reviewed journals and chapters in books, written more than 100 contract and technical reports and given more than 200 papers at national and international conferences and seminars. He has presented more than 60 workshops on research methodology, post-graduate supervision and bibliometrics and supervised more than 100 doctoral and master's students. He has received two prizes from the Academy for Science and Arts in South Africa including one for his contribution to the promotion of research methodology in South Africa. He established the African Doctoral Academy (ADA) in 2009. In 2012 he was elected to the Council of the Academy of Science of South Africa. In 2020 he was listed in the second edition of ASSAf's publication on *The legends of South African science*. He was lead or co-author on the following recent books and major reports – *Building a new cadre of researchers in South Africa* (2017), *The next generation of scientists in Africa (2018) scientists*; *The quality of South Africa's research publications* (2019), *The state of the South African research enterprise* (2019) and *A review of the NRDS and TYIP (2020)*. He is currently working on predatory publishing, funding of science in Africa, the mobility of South African doctoral graduates and the state of knowledge production at SA universities.

Herman Redelinghuys is the Database Manager at the Center for Research on Evaluation, Science, and Technology (CREST). He is an experienced database developer, software developer and Linux systems administrator. He manages, design, and implement database and user interface solutions within the CREST knowledge management group. He previously worked as an independent engineering consultant on various IT systems management and electronic communication projects. His current research interests include machine learning, metabolomic data analysis and natural language processing projects. He has a keen interest in agriculture and is a founding member of the South African Botanical Products Association. He holds a masters degree in Electrical Engineering.

Sergio Luiz Monteiro Salles Filho is an Agricultural Engineer who graduated from UFRRJ (1981), master in Agricultural Sciences from Unesp in Botucatu (1985) and a PhD in Economics from Unicamp (1993). He is Full Professor at the Department of Scientific and Technological Policy at the Institute of Geosciences at Unicamp. He was Superintendent of Operational Planning at FINEP in the period 2001-2003, Director of the Faculty of Applied Sciences at Unicamp from 2010 to 2013 and Director of the Institute of Geosciences at Unicamp between 2017 and 2021. He was a Visiting Researcher at the Manchester Institute of Innovation Research between 2013 and 2014. At Unicamp, he was a three-time recipient of the Zeferino Vaz Award (1998, 2001 and 2009), in recognition of his academic performance. He was deputy evaluation coordinator at FAPESP between 2010 and 2020. His research lines include economics, planning and management of science, technology and innovation. He is the author of more than 120 publications in journals, books and book chapters, having supervised 50 postgraduate students at the master's and doctoral levels. He coordinated more than 60 research and extension projects throughout his career. His highlights include his work to support the planning of public and private research organizations, work to evaluate the results and impacts of research and innovation policies and programs from funds, funding agencies and public and private organizations, and prospecting and decision support in science, technology and innovation

Philangani Sibiya (Phila) is from the University of South Africa (UNISA), College of Human Sciences in the Department of Information Science. He is a Senior Lecturer teaching information organisation courses, specifically, metadata standards in the said department. He specialize in research data management, digital scholarship, digitisation, indigenous knowledge systems scholarly communication, LIS education and training and job market. He holds the following qualifications: PhD Information Science (UNISA), Master of Arts (Information Science) and Bachelor of Arts in Library and Information Science (both from the University of Zululand).

Radhamany Sooryamoorthy is Professor of Sociology at the University of KwaZulu-Natal. He was the Acting Dean and Head of the School of Social Sciences at the University of KwaZulu-Natal and is a Research Fellow at the DSI-NRF Centre of Excellence in STI Policy, Stellenbosch University, South Africa. Previously he taught at institutions in India, Canada and Sweden. He is the author or editor of 22 titles including *Transforming Science in South Africa* (author), *Science, Policy and Development in Africa* (author), *Scientometrics for the Humanities and Social* Sciences (author), *Doctoral Training and Higher Education in Africa* (co-editor), *The Oxford Handbook of the Sociology of Africa* (co-editor), *Independent*

Africa, Dependent Science (author) and *Family Influence on Adolescent Sexual Behaviour* (co-author). He serves on the editorial boards of *Current Sociology*, *International Sociology*, and the *Journal of Comparative Family Studies.*

Robert Tijssen is a board member of SciSTIP. He is professor emeritus of Science and Innovation Studies at Leiden University (The Netherlands), an international research partners at the Centre for Global Higher Education (Oxford University, United Kingdom), and a research fellow at the LDE Center for Frugal innovation in Africa (an inter-university research consortium of Leiden University, Delft University of Technology and Erasmus University Rotterdam). His main interests concern the development and application of analytical frameworks and measurement models to assess general trends and patterns within the 'knowledge triangle' (higher education, scientific research, technological innovation), one of the main drivers of today's knowledge-based economies. Robert's current research agenda includes studies of use-inspired researchers, comparisons of research universities (innovation, entrepreneurship, local and regional socioeconomic impacts), and assessment of research excellence in science systems of countries in the global South. Robert's advisory work, stretching back more than 30 years, covers a wide range of consultancy activities. Dealing mostly with research management and science/innovation policy issues (both for public sector clients and private companies), the portfolio includes policy advice on quality assurance systems, indicator-based assessments of research programs, developing evaluation systems of regional innovation impact, as well as conducting sector-wide strategic analyses of R&D performance.

Nees Jan van Eck is senior researcher at the Centre for Science and Technology Studies (CWTS) at Leiden University in the Netherlands. He is doing research in the field of bibliometrics and scientometrics. His research focuses on the development of visualization tools and algorithms, mainly for analyzing the structure and development of science. Nees Jan is the main developer of VOSviewer and CitNetExplorer, two well-known software tools for visualizing bibliometric data. The VOSviewer software is frequently used in bibliometric studies. More than 20,000 publications have appeared in international scientific journals in which the software is employed. In addition to his work on bibliometric visualization, Nees Jan also focuses on the study of bibliometric data sources and the analysis of the full text of scientific publications. Nees Jan has published more than 80 publications. He is editor brief communications of the Journal of the Association for Information Science and Technology, he serves in the editorial boards of Journal of Data and Information Science and Quantitative Science Studies, and he served as an elected board member of the International Society for Informetrics and Scientometrics between 2015 and 2023. Nees Jan has been involved in various commercial bibliometric research projects and training courses in which the VOSviewer software plays a key role. He is also coordinator of the Information & Openness focal area of CWTS and the head of data science of CWTS, making him responsible for the entire data infrastructure of the center.

Marthie van Niekerk is the Centre Manager and manages all operational aspects, including financial forecasting and budget oversight, bursary administration, grants management and reporting. As part of the management group she contributes to the strategic oversight and core functions by supporting the Director. Marthie develops and maintains stakeholder engagement with collaborating universities and other relevant stakeholders. She completed a MPhil in Science and Technology Studies focusing on predatory open access journals.

Nick Vonortas is Professor of Economics and International Affairs at The George Washington University in Washington D.C. He is a faculty member of the Department of Economics, of the Institute for International Science and Technology Policy, and of the Trachtenberg School of Public Policy and Public Administration. He is currently the Director of the IISTP and the Director of PhD candidacy in Economics. Professor Vonortas holds a 'São Paulo Excellence Chair' in Technology and Innovation Policy at the University of Campinas, São Paulo, Brazil. He is also a Leading Research Fellow at the Institute for Statistical Studies and Economics of Knowledge, National Research University Higher School of Economics, Moscow, Russian Federation. Professor Vonortas is editor of the peer-reviewed journal Science and Public Policy. Professor Vonortas' teaching and research interests are in industrial organization, in the economics of technological change, and in technology and innovation policy and strategy. He specializes on innovation networks, investment under uncertainty, technology transfer, knowledge-intensive entrepreneurship, and R&D program evaluation. He has published widely on these issues. Professor Vonortas has held visiting appointments at several universities including the Tsinghua University (China), University of Lund (Sweden), Fundação Getulio Vargas (Brazil), University of Campinas (Brazil), Korea University (Republic of Korea), National Research University Higher School of Economics (Russian Federation), the University of Maastricht (Netherlands), Luigi Bocconi University (Italy), and the Athens University of Economics and Business and the National Technical University of Athens (Greece). He has served as a consultant to many government agencies in the United States, the European Union and member states, Latin America, East Asia as well as to several international organizations such as the World Bank, the Inter-American Development Bank, and the United Nations. Professor Vonortas holds a Ph.D. and M.Phil. in Economics from New York University (USA), a MA in Economic Development from Leicester University (UK), and a BA in Economics from the National and Kapodistrian University of Athens (Greece).