

2017 IEEE 13th International Conference on e-Science (e-Science 2017)

**Auckland, New Zealand
24 – 27 October 2017**



**IEEE Catalog Number: CFP1706A-POD
ISBN: 978-1-5386-2687-0**

**Copyright © 2017 by the Institute of Electrical and Electronics Engineers, Inc.
All Rights Reserved**

Copyright and Reprint Permissions: Abstracting is permitted with credit to the source. Libraries are permitted to photocopy beyond the limit of U.S. copyright law for private use of patrons those articles in this volume that carry a code at the bottom of the first page, provided the per-copy fee indicated in the code is paid through Copyright Clearance Center, 222 Rosewood Drive, Danvers, MA 01923.

For other copying, reprint or republication permission, write to IEEE Copyrights Manager, IEEE Service Center, 445 Hoes Lane, Piscataway, NJ 08854. All rights reserved.

****** This is a print representation of what appears in the IEEE Digital Library. Some format issues inherent in the e-media version may also appear in this print version.***

IEEE Catalog Number:	CFP1706A-POD
ISBN (Print-On-Demand):	978-1-5386-2687-0
ISBN (Online):	978-1-5386-2686-3

Additional Copies of This Publication Are Available From:

Curran Associates, Inc
57 Morehouse Lane
Red Hook, NY 12571 USA
Phone: (845) 758-0400
Fax: (845) 758-2633
E-mail: curran@proceedings.com
Web: www.proceedings.com

CURRAN ASSOCIATES INC.
proceedings
.com

2017 IEEE 13th International Conference on eScience

e-Science 2017

Table of Contents

Message from the General Chairs.....	xiv
Message from the Program Committee	
Chairs.....	xv
Message from the Publications Chair.....	xvi
Organizing Committee.....	xvii
International Steering Committee	
and Advisory Committee.....	xviii
Program Committee.....	xix
Subreviewers.....	xxi

Paper Session 1A: Deep Learning: From Dentistry to Satellites

Towards a Fully Automated Diagnostic System for Orthodontic Treatment in Dentistry	1
<i>Seiya Murata, Chonho Lee, Chihiro Tanikawa, and Susumu Date</i>	
An Unsupervised Deep Learning Approach for Satellite Image Analysis with Applications in Demographic Analysis	9
<i>Jessica Block, Mehrdad Yazdani, Mai Nguyen, Daniel Crawl, Marta Jankowska, John Graham, Tom DeFanti, and Ilkay Altintas</i>	
Developing a Volunteer Computing Project to Evolve Convolutional Neural Networks and Their Hyperparameters	19
<i>Travis Desell</i>	
A Mobile Application for Plant Recognition through Deep Learning	29
<i>Min Gao, Yang Lin, and Richard O. Sinnott</i>	

Paper Session 1B: Parallel Computing and Information Extraction

Data Provenance for Multi-Agent Models	39
<i>Delmar B. Davis, Jonathan Featherston, Munehiro Fukuda, and Hazeline U. Asuncion</i>	
Accelerating Exact Protein Structure Alignment with Graphics Processors	49
<i>Yishui Wu, Shuang Qiu, and Qiong Luo</i>	
Real-Time Data Analysis and Autonomous Steering of Synchrotron Light Source Experiments	59
<i>Tekin Bicer, Doga Gursoy, Rajkumar Kettimuthu, Ian T. Foster, Bin Ren, Vincent De Andrede, and Francesco De Carlo</i>	
SELFIE: Self-Aware Information Extraction from Digitized Biocollections	69
<i>Icaro Alzuru, Andréa Matsunaga, Maurício Tsugawa, and José A.B. Fortes</i>	

Paper Session 2A: Cloud Computing

Experiences with DERIVA: An Asset Management Platform for Accelerating eScience	79
<i>Alejandro Bugacov, Karl Czajkowski, Carl Kesselman, Anoop Kumar, Robert E. Schuler, and Hongsuda Tangmunarunkit</i>	
dSpark: Deadline-Based Resource Allocation for Big Data Applications in Apache Spark	89
<i>Muhammed Tawfiqul Islam, Shanika Karunasekera, and Rajkumar Buyya</i>	

Paper Session 2B: Crowdsourcing

Toward Using Citizen Scientists to Drive Automated Ecological Object Detection in Aerial Imagery	99
<i>Connor Bowley, Marshall Mattingly III, Andrew Barnas, Susan Ellis-Felege, and Travis Desell</i>	
Towards a Hybrid Human-Computer Scientific Information Extraction Pipeline	109
<i>Roselyne B. Tchoua, Kyle Chard, Debra J. Audus, Logan T. Ward, Joshua Lequieu, Juan J. De Pablo, and Ian T. Foster</i>	

Paper Session 3A: Workflow and Frameworks

Provenance-Based Scientific Workflow Search	119
<i>Amani Abu Jabal, Elisa Bertino, and Geeth De Mel</i>	
Task-Based Budget Distribution Strategies for Scientific Workflows with Coarse-Grained Billing Periods in IaaS Clouds	128
<i>Muhammad Hafizhuddin Hilman, María Alejandra Rodríguez, and Rajkumar Buyya</i>	

WorkflowHunt: Combining Keyword and Semantic Search in Scientific Workflow Repositories	138
<i>Juan Sebastian Beleno Diaz and Claudia Bauzer Medeiros</i>	
A Mathematical Programming- and Simulation-Based Framework to Evaluate Cyberinfrastructure Design Choices	148
<i>Zhengchun Liu, Rajkumar Kettimuthu, Sven Leyffer, Prashant Palkar, and Ian Foster</i>	

Paper Session 3B: People, Birds, and Frogs

In Pursuit of the Wisest: Building Cost-Effective Teams of Experts	158
<i>Yashar Najafloo and Kris Bubendorfer</i>	
Clustering and Visualization of Long-Duration Audio Recordings for Rapid Exploration Avian Surveys	168
<i>Philip Eichinski and Paul Roe</i>	
An Investigation into Acoustic Analysis Methods for Endangered Species Monitoring: A Case of Monitoring the Critically Endangered White-Bellied Heron in Bhutan	177
<i>Tshering Dema, Liang Zhang, Michael Towsey, Anthony Truskinger, Sherub Sherub, Kinley, Jinglan Zhang, Margot Brereton, and Paul Roe</i>	
Multi-Label Classification of Frog Species via Deep Learning	187
<i>Jie Xie, Rui Zeng Changliang Xu, Jinglan Zhang, and Paul Roe</i>	

Paper Session 4A: Applications: From Humanities to Fruit Flies

Experimental Humanities: An Adventure with Lovelace and Babbage	194
<i>David De Roure and Pip Willcox</i>	
Mid-level Image Representation for Fruit Fly Identification (Diptera: Tephritidae)	202
<i>Matheus Macedo Leonardo, Sandra Avila, Roberto A. Zucchi, and Fabio A. Faria</i>	
Ten Principles for Creating Usable Software for Science	210
<i>Lavanya Ramakrishnan and Daniel Gunter</i>	

Paper Session 4B: Computing Platforms

SNIC Science Cloud (SSC): A National-Scale Cloud Infrastructure for Swedish Academia	219
<i>Salman Toor, Mathias Lindberg, Ingemar Falman, Andreas Vallin, Olof Mohill, Pontus Freyhult, Linus Nilsson, Martin Agback, Lars Viklund, Henric Zazzik, Ola Spjuth, Marco Capuccini, Joakim Möller, Donal Murtagh, and Andreas Hellander</i>	
A Platform for the Analysis of Qualitative and Quantitative Data about the Built Environment and Its Users	228
<i>Mike Simpson, Simon Woodman, Hugo Hiden, Sebastian Stein, Stephen Dowsland, Mark Turner, Vicki L. Hanson, and Paul Watson</i>	

A Metropolitan Area Infrastructure for Data Intensive Science	238
<i>David Abramson, Jake Carroll, Chao Jin, and Michael Mallon</i>	

Paper Session 5A: Repositories

ScienceDB: A Public Multidisciplinary Research Data Repository for eScience	248
<i>Li Chengzan, Hou Yanfei, Li Jianhui, and Zhang Lili</i>	
Improving Small File I/O Performance for Massive Digital Archives	256
<i>Hwajung Kim and Heonyoung Yeom</i>	
Towards a Human-Machine Scientific Partnership Based on Semantically Rich Research Objects	266
<i>Jose Manuel Gomez-Perez, Raul Palma, and Andres Garcia-Silva</i>	

Paper Session 5B: Workflow and Workloads

Evaluating Distributed Execution of Workloads	276
<i>Matteo Turilli, Yadu Nand Babuji, Andre Merzky, Ming Tai Ha, Michael Wilde, Daniel S. Katz, and Shantenu Jha</i>	
A Computational Pipeline for the IUCN Risk Assessment for Meso-American Reef Ecosystem	286
<i>Hoang Anh Nguyen, Lucie Bland, Tristan Roberts, Siddeswara Guru, Minh Dinh, and David Abramson</i>	
High-Throughput Computing on High-Performance Platforms: A Case Study	295
<i>Danila Oleynik, Sergey Panitkin, Matteo Turilli, Alessio Angius, Sarp Oral, Kaushik De, Alexei Klimentov, Jack C. Wells, and Shantenu Jha</i>	

Paper Session 6A: Processing Pipelines: From Plants to Gravitational Waves

Change Frequency Heatmaps for Temporal Multivariate Phenological Data Analysis	305
<i>Greice Cristina Mariano, Natalia Costa Soares, Leonor Patricia Cerdeira Morellato, and Ricardo Da Silva Torres</i>	
Adaptive Lossy Compression of Complex Environmental Indices Using Seasonal Auto-Regressive Integrated Moving Average Models	315
<i>Ugur Cayoglu, Peter Braesicke, Tobias Kerzenmacher, Jörg Meyer, and Achim Streit</i>	
Similarity Projection: A Geometric Measure for Comparison of Biological Sequences	325
<i>Lawrence Buckingham, Timothy Chappell, James M. Hogan, and Shlomo Geva</i>	

BOSS-LDG: A Novel Computational Framework That Brings Together Blue Waters, Open Science Grid, Shifter and the LIGO Data Grid to Accelerate Gravitational Wave Discovery	335
<i>E.A. Huerta, Roland Haas, Edgar Fajardo, Daniel Katz, Stuart Anderson, Peter Couvares, Josh Willis, Timothy Bouvet, Jeremy Enos, William T.C. Kramer, Hon Wai Leong, and David Wheeler</i>	

Paper Session 6B: Visualization: From Networks to Archeology

TaRDIS, a Visual Analytics System for Spatial and Temporal Data in Archaeo-Related Disciplines	345
<i>Daniel Kaltenthaler, Johannes-Y. Lohrer, Ptolemaios D. Paxinos, Daniel Hämmerle, Henriette Obermaier, and Peer Kröger</i>	
Iterative Design and Evaluation of Regulatory Network Visualisation at Scale	354
<i>Samuel Thomas Smith, James Michael Hogan, Xin-Yi Chua, Margot Brereton, Daniel Johnson, and Markus Rittenbruch</i>	
Monitoring Water Quality in the Great Lakes Leveraging Geo-Temporal Cyberinfrastructure	364
<i>Indira Gutierrez-Polo, Yan Zhao, Shannon Bradley, Eugene Roeder, Michelle Pitcel, Kristin TePas, Paris Collingsworth, and Luigi Marini</i>	
Sciunits: Reusable Research Objects	374
<i>Dai Hai Ton That, Gabriel Fils, Zhihao Yuan, and Tanu Malik</i>	

Paper Session 7A: Stream Processing: From IOT to Synchrotron

Online Decision-Making Using Edge Resources for Content-Driven Stream Processing	384
<i>Eduard Renart, Daniel Balouek-Thomert, Xuan Hu, Jie Gong, and Manish Parashar</i>	
Energy-Efficient Dynamic Scheduling of Deadline-Constrained MapReduce Workflows	393
<i>Tong Shu and Chase Q. Wu</i>	
Collaborative Reuse of Streaming Dataflows in IoT Applications	403
<i>Shilpa Chaturvedi, Sahil Tyagi, and Yogesh Simmhan</i>	

Paper Session 7B: Infrastructure

Adapting Enterprise Architecture for eScience	413
<i>Richard Palmer, Kheeran Dharmawardena, and Hamish Holewa</i>	
Dynamic Network Bandwidth Resizing for Big Data Applications	423
<i>Fábio Diniz Rossi, Guilherme Da Cunha Rodrigues, Rodrigo N. Calheiros, and Marcelo Da Silva Conterato</i>	

Poster Papers

A Framework for Computing Artistic Style Using Artistically Relevant Features	432
<i>Catherine A. Buell, William P. Seeley, and Ricky J. Sethi</i>	
A Novel Data Mining Testbed for User Centred Modelling and Personalisation of Digital Library Services	434
<i>Maram Abdulrahman Almaghrabi and Girija Chetty</i>	
Accelerating Genome Sequence Alignment on Hadoop on Lustre Environment	436
<i>Eun-Kyu Byun, Junehawk Lee, Seok Jong Yu, Jae-Hyuck Kwak, and Soonwook Hwang</i>	
Determining the Function of Political Tweets	438
<i>Erik Tjong Kim Sang, Herbert Kruitbosch, Marcel Broersma, and Marc Esteve Del Valle</i>	
Efficient Program Analyses Using Deductive and Semantic Methodologies	440
<i>Ganesh Selvaraj, Gerald Weber, and Christof Lutteroth</i>	
Ensuring the Quality of Research Objects in the Earth Science Domain	442
<i>Andres Garcia-Silva, Raul Palma, and Jose Manuel Gomez-Perez</i>	
Geoscience Cyberinfrastructure in the Cloud: Data-Proximate Computing to Address Big Data and Open Science Challenges	444
<i>Mohan Ramamurthy</i>	
Hunting Data Rogues at Scale: Data Quality Control for Observational Data in Research Infrastructures	446
<i>Gilberto Pastorello, Dan Gunter, Housen Chu, Danielle Christianson, Carlo Trotta, Eleonora Canfora, Boris Faybishenko, You-Wei Cheah, Norm Beekwilder, Stephen Chan, Sigrid Dengel, Trevor Keenan, Fianna O'Brien, Abdelrahman Elbashandy, Cristina Poindexter, Marty Humphrey, Dario Papale, and Deb Agarwal</i>	
Introducing the Physiome Journal: Improving Reproducibility, Reuse, and Discovery of Computational Models	448
<i>David Phillip Nickerson and Peter John Hunter</i>	
Massive OceanColor Data Processing and Analysis System: TuPiX-OC	450
<i>Jung-Ho Um, Sunggeun Han, Hyunwoo Kim, and Kyongseok Park</i>	
Research Objects for Interworkability among Global Environmental and Geophysical Data	452
<i>Jose Manuel Gomez Perez, Chuck Meertens, Fran Boler, Henry Loescher, Christine Laney, Daniel Crawl, and Ilkay Altintas</i>	
Semantic Technologies and Text Analysis in Support of Scientific Knowledge Reuse	454
<i>Andres Garcia-Silva, Raul Palma, and Jose Manuel Gomez-Perez</i>	
Software Defined Cyberinfrastructure for Data Management	456
<i>Ryan Chard, Kyle Chard, Steve Tuecke, and Ian Foster</i>	

Spartan and NEMO: Two HPC-Cloud Hybrid Implementations	458
<i>Lev Lafayette and Bernd Wiebelt</i>	
Supporting Research and Operational Earth Science Portals through ROHub	460
<i>Raul Palma, Marcin Krystek, Jose M. Gomez-Perez, Andres Garcia-Silva, Sergio Ferraresi, Simone Mantovani, Serena Avolio, and Sergio De Gioia</i>	
Towards a Data-Centric Research and Development Roadmap for Large-Scale Science User Facilities	462
<i>E. Wes Bethel</i>	
How NeSI Helps Users Run Better and Faster on New Zealand's Supercomputing Platforms	465
<i>Alexander Pletzer, Wolfgang Hayek, Chris Scott, Brian Corrie, and Georgina Rae</i>	
Consideration of Improvement in the Educational Activities Using Web-Based Computational Science Platform	467
<i>Gi-Myeong Ryu, Jung-Hun Shin, Jeong-Cheol Lee, Sik Lee, and Kum-Won Cho</i>	
A Large Scholarly Corpus: A Bird's-Eye View	469
<i>Yashar Najafloo and Kris Bubendorfer</i>	

Workshop on Sustainable Software for Science: Practice and Experiences (WSSSPE5.2)

Design of a Smart Contract Based Autonomous Organization for Sustainable Software	471
<i>Alper Alimoğlu and Can Özturan</i>	
Science Gateways Incubator: Software Sustainability Meets Community Needs	477
<i>Sandra Gesing, Michael Zentner, Juliana Casavan, Betsy Hillery, Mihaela Vorvoreanu, Randy Heiland, Suresh Marru, Marlon Pierce, Nayiri Mullinix, and Nancy Maron</i>	
Understanding Software in Research: Initial Results from Examining Nature and a Call for Collaboration	486
<i>Udit Nangia and Daniel S. Katz</i>	
Artifact Evaluation: Is It a Real Incentive?	488
<i>Bruce R. Childers and Panos K. Chrysanthis</i>	

Safe Data Paradigms and Platforms: Enabling Collaborative Analysis of Sensitive Data

A Comparative Evaluation of Blockchain Systems for Application Sharing Using Containers	490
<i>Adam Brinckman, Donal Luc, Jarek Nabrzyski, Gary L. Neidig, Joel Neidig, Tyler A. Puckett, Swapna Krishnakumar Radha, and Ian J. Taylor</i>	
Safe Collections and Stewardship on Cloud Kotta	498
<i>Yadu N. Babuji, Kyle Chard, Eamon Duede, and Ian Foster</i>	
Safe Double Blind Studies as a Service	504
<i>Tyler J. Skluzacek, Suhail Rehman, and Ian Foster</i>	
ERMRest: A Collaborative Data Catalog with Fine Grain Access Control	510
<i>Karl Czajkowski, Carl Kesselman, and Robert Schuler</i>	
A Review of Privacy and Consent Management in Healthcare: A Focus on Emerging Data Sources	518
<i>Muhammad Rizwan Asghar, TzeHowe Lee, Mirza Mansoor Baig, Ehsan Ullah, Giovanni Russello, and Gillian Dobbie</i>	

BigDig: High Throughput Digitization for Natural History Collections

Mass Digitization of Individual Pinned Insects Using Conveyor-Driven Imaging	523
<i>Riitta Tegelberg, Jere Kahanpää, Janne Karppinen, Tero Mononen, Zhenzhe Wu, and Hannu Saarenmaa</i>	
Automatic Reconstruction of Polygon Triangulation for Mounted Skeleton Point Cloud	528
<i>Pierre-Yves Gagnier, Herbert Maschner, Auréliane Gailliègue, Loïc Norgeot, Charles Dapogny, Lionel Revéret, and Anick Abourachid</i>	
3D Scientific Visualisation of 19th Century Glass Replicas of Invertebrates	533
<i>Dante Abate, Sorin Hermon, Stefania Lotti, and Gianna Innocenti</i>	
Designing a High-Throughput Pipeline for Digitizing Pinned Insects	542
<i>Mark Hereld, Nicola J. Ferrier, Nitin Agarwal, and Petra Sierwald</i>	
Shape-from-Shifting: Uncalibrated Photometric Stereo with a Mobile Device	551
<i>Chia-Kai Yeh, Fengqiang Li, Gianluca Pastorelli, Marc Walton, Aggelos K. Katsaggelos, and Oliver Cossairt</i>	
Towards High-Throughput 3D Insect Capture for Species Discovery and Diagnostics	559
<i>Chuong Nguyen, Matt Adcock, Stuart Anderson, David Lovell, Nicole Fisher, and John La Salle</i>	

Environmental Computing Workshop (ECW)

Modelling the Coverage of Dipterocarp Trees in Central Visayas, Philippines	561
<i>Adrian Jose Sabado, Geoffrey Solano, Marilou Nicolas, Riza Batista-Navarro, Roselyn Gabud, and Vincent Hilomen</i>	

International Workshop on Workflow Science (WoWS)

Toward Prioritization of Data Flows for Scientific Workflows Using Virtual Software Defined Exchanges	566
<i>Anirban Mandal, Paul Ruth, Ilya Baldin, Rafael Ferreira Da Silva, and Ewa Deelman</i>	
On Analytics of File Transfer Rates over Dedicated Wide-Area Connections	576
<i>Satyabrata Sen, Nageswara S.V. Rao, Qiang Liu, Neena Imam, Rajkumar Kettimuthu, and Ian Foster</i>	
Deep Learning on Operational Facility Data Related to Large-Scale Distributed Area Scientific Workflows	586
<i>Alok Singh, Eric Stephan, Malachi Schram, and Ilkay Altintas</i>	
Supporting Data-Driven Workflows Enabled by Large Scale Observatories	592
<i>Ali Reza Zamani, Moustafa AbdelBaky, Daniel Balouek-Thomert, Ivan Rodero, and Manish Parashar</i>	
Author Index	596