

# **2017 ACM/IEEE Joint Conference on Digital Libraries (JCDL 2017)**

**Toronto, Ontario, Canada  
19 – 23 June 2017**



**IEEE Catalog Number: CFP17JCD-POD**  
**ISBN: 978-1-5386-3862-0**

**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:    | CFP17JCD-POD      |
| ISBN (Print-On-Demand): | 978-1-5386-3862-0 |
| ISBN (Online):          | 978-1-5386-3861-3 |

**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](mailto:curran@proceedings.com)  
Web: [www.proceedings.com](http://www.proceedings.com)

CURRAN ASSOCIATES INC.  
**proceedings**  
.com



# TABLE OF CONTENTS

|                                      |             |
|--------------------------------------|-------------|
| <b>Chairs' Welcome .....</b>         | <b>ix</b>   |
| <b>Conference Organization .....</b> | <b>xi</b>   |
| <b>Sponsors.....</b>                 | <b>xvii</b> |
| <b>Keynotes.....</b>                 | <b>xix</b>  |
| <b>Schedule.....</b>                 | <b>xxiv</b> |

## TECHNICAL PAPERS

|                                                                                                |           |
|------------------------------------------------------------------------------------------------|-----------|
| <b>Archival Crawlers and JavaScript: Discover More Stuff but Crawl More Slowly .....</b>       | <b>1</b>  |
| <i>Justin F. Brunelle, Michele C. Weigle and Michael L. Nelson</i>                             |           |
| <b>Building and Querying Semantic Layers for Web Archives.....</b>                             | <b>11</b> |
| <i>Pavlos Fafalios, Helge Holzmann, Vaibhav Kasturia and Wolfgang Nejdl</i>                    |           |
| <b>WikiM: Metapaths Based Wikification of Scientific Abstracts .....</b>                       | <b>21</b> |
| <i>Abhik Jana, Sruthi Mooriyath, Animesh Mukherjee and Pawan Goyal</i>                         |           |
| <b>Descriptor-Invariant Fusion Architectures for Automatic Subject Indexing.....</b>           | <b>31</b> |
| <i>Martin Toepfer and Christin Seifert</i>                                                     |           |
| <b>Identifying Important Citations Using Contextual Information from Full Text .....</b>       | <b>41</b> |
| <i>Saeed-UI Hassan, Anam Akram and Peter Haddawy</i>                                           |           |
| <b>Learning to Predict Citation-Based Impact Measures.....</b>                                 | <b>49</b> |
| <i>Luca Weihs and Oren Etzioni</i>                                                             |           |
| <b>Understanding the Impact of Early Citers on Long-Term Scientific Impact.....</b>            | <b>59</b> |
| <i>Mayank Singh, Ajay Jaiswal, Priya Shree, Arindam Pal, Animesh Mukherjee and Pawan Goyal</i> |           |
| <b>Matrix-Based News Aggregation: Exploring Different News Perspectives.....</b>               | <b>69</b> |
| <i>Felix Hamborg, Norman Meuschke and Bela Gipp</i>                                            |           |

|                                                                                                                                                                               |            |
|-------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>Quill: A Framework for Constructing Negotiated Texts - With a Case Study<br/>on the US Constitutional Convention of 1787 .....</b>                                         | <b>79</b>  |
| <i>Nicholas Cole, Alfie Abdul-Rahman and Grace Mallon</i>                                                                                                                     |            |
| <b>Realising a Layered Digital Library: Exploration and Analysis of the Live Music<br/>Archive through Linked Data.....</b>                                                   | <b>89</b>  |
| <i>Kevin R. Page, Sean Bechhofer, Gyorgy Fazekas, David M. Weigl<br/>and Thomas Wilmering</i>                                                                                 |            |
| <b>A Benchmark and Evaluation for Text Extraction from PDF .....</b>                                                                                                          | <b>99</b>  |
| <i>Hannah Bast and Claudius Korzen</i>                                                                                                                                        |            |
| <b>A Text Extraction Software Benchmark Based on a Synthesized Dataset .....</b>                                                                                              | <b>109</b> |
| <i>Kresimir Duretec, Andreas Rauber and Christoph Becker</i>                                                                                                                  |            |
| <b>Mathematical Document Categorization with Structure of Mathematical Expressions .....</b>                                                                                  | <b>119</b> |
| <i>Tokinori Suzuki and Atsushi Fujii</i>                                                                                                                                      |            |
| <b>Classifying Short Unstructured Data Using the Apache Spark Platform .....</b>                                                                                              | <b>129</b> |
| <i>Eduardo P. S. Castro, Saurabh Chakravarty, Eric Williamson,<br/>Denilson Alves Pereira and Edward A. Fox</i>                                                               |            |
| <b>On the Various Semantics of Similarity in Word Embedding Models .....</b>                                                                                                  | <b>139</b> |
| <i>Abel Elekes, Martin Schaefer and Klemens Boehm</i>                                                                                                                         |            |
| <b>Estimating the Date of First Publication in a Large-Scale Digital Library.....</b>                                                                                         | <b>149</b> |
| <i>David Bamman, Michelle Carney, Jon Gillick, Cody Hennesy<br/>and Vijitha Sridhar</i>                                                                                       |            |
| <b>The Lowest Form of Flattery: Characterising Text Re-Use and Plagiarism<br/>Patterns in a Digital Library Corpus.....</b>                                                   | <b>159</b> |
| <i>George Buchanan and Dana McKay</i>                                                                                                                                         |            |
| <b>Automating Data Citation: The eagle-i Experience.....</b>                                                                                                                  | <b>169</b> |
| <i>Abdussalam Alawini, Leshang Chen, Susan B. Davidson,<br/>Natan Portilho Da Silva and Gianmaria Silvello</i>                                                                |            |
| <b>Influence of Reviewer Interaction Network on Long-Term Citations:<br/>A Case Study of the Scientific Peer-Review System of the Journal<br/>of High Energy Physics.....</b> | <b>179</b> |
| <i>Sandipan Sikdar, Matteo Marsili, Niloy Ganguly and Animesh Mukherjee</i>                                                                                                   |            |
| <b>Discovering Scholarly Orphans Using ORCID .....</b>                                                                                                                        | <b>189</b> |
| <i>Martin Klein and Herbert Van de Sompel</i>                                                                                                                                 |            |
| <b>Building Entity-Centric Event Collections.....</b>                                                                                                                         | <b>199</b> |
| <i>Federico Nanni, Simone Paolo Ponzetto and Laura Dietz</i>                                                                                                                  |            |
| <b>Towards Building a Knowledge Base of Monetary Transactions from a News Collection .....</b>                                                                                | <b>209</b> |
| <i>Jan R. Benetka, Krisztian Balog and Kjetil Norvag</i>                                                                                                                      |            |

|                                                                                                                                     |            |
|-------------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>Local Memory Project: Providing Tools to Build Collections of Stories for Local Events from Local Sources .....</b>              | <b>219</b> |
| <i>Alexander C. Nwala, Michele C. Weigle, Adam B. Ziegler, Anastasia Aizman and Michael L. Nelson</i>                               |            |
| <b>What Is Part of That Resource? User Expectations for Personal Archiving .....</b>                                                | <b>229</b> |
| <i>Faryaneh Poursardar and Frank Shipman</i>                                                                                        |            |
| <b>A Portable Strategy for Preserving Web Applications Functionality .....</b>                                                      | <b>233</b> |
| <i>Weijia Xu, Maria Esteve, Deborah Beck and Yi-Hsuan Hsieh</i>                                                                     |            |
| <b>Client-Side Reconstruction of Composite Mementos Using ServiceWorker .....</b>                                                   | <b>237</b> |
| <i>Sawood Alam, Mat Kelly, Michele C. Weigle and Michael L. Nelson</i>                                                              |            |
| <b>HESDK: A Hybrid Approach to Extracting Scientific Domain Knowledge Entities .....</b>                                            | <b>241</b> |
| <i>Jian Wu, Sagnik Ray Choudhury, Agnese Chiatti, Chen Liang and C. Lee Giles</i>                                                   |            |
| <b>Smart Library: Identifying Books on Library Shelves Using Supervised Deep Learning for Scene Text Reading .....</b>              | <b>245</b> |
| <i>Xiao Yang, Dafang He, Wenyi Huang, Alexander Ororbia, Zihan Zhou, Daniel Kifer and C. Lee Giles</i>                              |            |
| <b>Impact of OCR Errors on the Use of Digital Libraries: Towards a Better Access to Information .....</b>                           | <b>249</b> |
| <i>Guillaume Chiron, Antoine Doucet, Mickael Coustaty, Muriel Visani and Jean-Philippe Moreux</i>                                   |            |
| <b>Information-Seeking in Large-Scale Digital Libraries: Strategies for Scholarly Workset Creation .....</b>                        | <b>253</b> |
| <i>David M. Weigl, Kevin R. Page, Peter Organisciak and J. Stephen Downie</i>                                                       |            |
| <b>Uncertainty about the Long-Term: Digital Libraries, Astronomy Data, and Open Source Software .....</b>                           | <b>257</b> |
| <i>Peter T. Darch and Ashley E. Sands</i>                                                                                           |            |
| <b>Towards Publishing Secure Capsule-Based Analysis .....</b>                                                                       | <b>261</b> |
| <i>Jaimie Murdock, Jacob Jett, Tim Cole, Yu Ma, J. Stephen Downie and Beth Plale</i>                                                |            |
| <b>Author Clustering Using SPATIUM .....</b>                                                                                        | <b>265</b> |
| <i>Mirco Kocher and Jacques Savoy</i>                                                                                               |            |
| <b>Retrieving and Combining Repeated Passages to Improve OCR .....</b>                                                              | <b>269</b> |
| <i>Shaobin Xu and David Smith</i>                                                                                                   |            |
| <b>CryptSubmit: Introducing Securely Timestamped Manuscript Submission and Peer Review Feedback Using the Blockchain .....</b>      | <b>273</b> |
| <i>Bela Gipp, Corinna Breitingner, Norman Meuschke and Joeran Beel</i>                                                              |            |
| <b>Citation Sentence Reuse Behavior of Scientists: A Case Study on Massive Bibliographic Text Dataset of Computer Science .....</b> | <b>277</b> |
| <i>Mayank Singh, Abhishek Niranjan, Divyansh Gupta, Nikhil Angad Bakshi, Animesh Mukherjee and Pawan Goyal</i>                      |            |

## POSTERS

|                                                                                                                                  |            |
|----------------------------------------------------------------------------------------------------------------------------------|------------|
| <b>A Machine Learning Approach to Evaluating Translation Quality .....</b>                                                       | <b>281</b> |
| <i>Brenda Reyes Ayala and Jiangping Chen</i>                                                                                     |            |
| <b>All for One and One for All: Reconciling Research and Production Values<br/>at the HathiTrust through User-Scripting.....</b> | <b>283</b> |
| <i>David Bainbridge and J. Stephen Downie</i>                                                                                    |            |
| <b>Big Brother Is Watching You - Now in a Doubleplusgood Way .....</b>                                                           | <b>285</b> |
| <i>Corey Sterling, Carlin St. Pierre and David Bainbridge</i>                                                                    |            |
| <b>Can Japanese Manga Be Automatically Classified from Public Library Holdings? .....</b>                                        | <b>287</b> |
| <i>Masaki Eto, Teru Agata, Noriko Sugie, Yasuharu Otani and Mari Agata</i>                                                       |            |
| <b>Cell-Based Provenance for Scientific Data.....</b>                                                                            | <b>289</b> |
| <i>Juyeong Park, Masatoshi Yoshikawa and Hiroyuki Kato</i>                                                                       |            |
| <b>Circulation Modeling of Library Book Promotions .....</b>                                                                     | <b>291</b> |
| <i>Robin Burke, Ana Lucic and John Shanahan</i>                                                                                  |            |
| <b>Data Discovery, Reuse, and Integration: The Perspectives of Natural Resources<br/>and Environmental Scientists .....</b>      | <b>293</b> |
| <i>Yi Shen</i>                                                                                                                   |            |
| <b>Expanding Access to Biodiversity Literature .....</b>                                                                         | <b>295</b> |
| <i>P. Randall</i>                                                                                                                |            |
| <b>Exploring Features for Predicting Policy Citations.....</b>                                                                   | <b>297</b> |
| <i>Christian Bailey, Bharat Kale, Jamieson Walker, Harish Varma Siravuri,<br/>Hamed Alhoori and Michael E. Papka</i>             |            |
| <b>Exploring Salient Thumbnail Generation for Archival Collections Online .....</b>                                              | <b>299</b> |
| <i>Bailey Jones and Robert Capra</i>                                                                                             |            |
| <b>Heterogeneous Resources Aggregation for Literature Usage Analysis in Academic Libraries .....</b>                             | <b>301</b> |
| <i>Jenq-Haur Wang, Yen-Sheng Chen, Jui-Wen Peng and Ing-Yi Chen</i>                                                              |            |
| <b>Impact of URI Canonicalization on Memento Count .....</b>                                                                     | <b>303</b> |
| <i>Mat Kelly, Lulwah M. Alkwai, Sawood Alam, Herbert Van de Sompel,<br/>Michael L. Nelson and Michele C. Weigle</i>              |            |
| <b>Information Needs for Anime Recommendation: Analyzing Anime Users' Online Forum Queries.....</b>                              | <b>305</b> |
| <i>Hyerim Cho, Marc L. Schmalz, Stephen A. Keating and Jin Ha Lee</i>                                                            |            |
| <b>Institutional Repository Digital Object Metadata Enhancement and Re-Architecting .....</b>                                    | <b>307</b> |
| <i>A. Saundry</i>                                                                                                                |            |
| <b>Linked Open Data (LOD) for Library Special Collections.....</b>                                                               | <b>309</b> |
| <i>Jacob Jett, Timothy W. Cole, Myung-Ja K. Han and Caroline Szylowicz</i>                                                       |            |

|                                                                                                                           |            |
|---------------------------------------------------------------------------------------------------------------------------|------------|
| <b>Modeling Complex Negotiations: The Quill Project .....</b>                                                             | <b>311</b> |
| <i>Olivia Griffiths, Nicholas Cole and Alfie Abdul-Rahman</i>                                                             |            |
| <b>Mr. DLib: Recommendations-as-a-Service (RaaS) for Academia .....</b>                                                   | <b>313</b> |
| <i>Joeran Beel, Akiko Aizawa, Corinna Breitingner and Bela Gipp</i>                                                       |            |
| <b>NCHC Blockchain Construction Platform (NBCP): Rapidly Constructing<br/>Blockchain Nodes around Taiwan .....</b>        | <b>315</b> |
| <i>Lo-Yao Yeh, Peggy Joy Lu and Jen-Wei Hu</i>                                                                            |            |
| <b>Open Peer Review in the Era of Open Science: A Pilot Study of Researchers' Perceptions .....</b>                       | <b>317</b> |
| <i>Manasa Rath and Peiling Wang</i>                                                                                       |            |
| <b>OXPath-Based Data Acquisition for dblp.....</b>                                                                        | <b>319</b> |
| <i>Christopher Michels, Ruslan R. Fayzrakhmanov, Michael Ley,<br/>Emanuel Sallinger and Ralf Schenkel</i>                 |            |
| <b>PIDs to Support Discovery and Citation: Persistent Identifier Service<br/>Design and Delivery at UNSW Library.....</b> | <b>321</b> |
| <i>Daniel Bangert and Maude Frances</i>                                                                                   |            |
| <b>Preservation Planning and Workflows for Digital Holdings at the<br/>Thomas Fisher Rare Book Library.....</b>           | <b>323</b> |
| <i>Jess Whyte</i>                                                                                                         |            |
| <b>Provenance Enriched PID Kernel Information as OAI-ORE Map Replacement<br/>for SEAD Research Objects .....</b>          | <b>326</b> |
| <i>Inna Kouper, Yu Luo, Isuru Suriarachchi and Beth Plale</i>                                                             |            |
| <b>Query Reformulation Patterns in Cross-Device OPAC Search .....</b>                                                     | <b>328</b> |
| <i>Dan Wu and Renmin Bi</i>                                                                                               |            |
| <b>Taking a Book off the Shelf in a Virtual Library .....</b>                                                             | <b>330</b> |
| <i>Sally Jo Cunningham, Bill Rogers and Jane Kim</i>                                                                      |            |
| <b>TouchDoc : A Tool to Bridge the Gap between Physical and Digital Libraries.....</b>                                    | <b>332</b> |
| <i>Nicolas Sidere, Cyrille Suire, Mickael Coustaty, Joseph Chazalon,<br/>Jean-Christophe Burie and Jean-Marc Ogier</i>    |            |
| <b>Towards a Cross-Canadian Digital Library Platform.....</b>                                                             | <b>334</b> |
| <i>Matthew Barry and Daniel Sifton</i>                                                                                    |            |
| <b>Towards a Ranking Model for Semantic Layers over Digital Archives.....</b>                                             | <b>336</b> |
| <i>Pavlos Fafalios, Vaibhav Kasturia and Wolfgang Nejdl</i>                                                               |            |
| <b>Using the Jupyter Notebook as a Tool for Open Science: An Empirical Study.....</b>                                     | <b>338</b> |
| <i>Bernadette M. Randles, Irene V. Pasquetto, Milena S. Golshan<br/>and Christine L. Borgman</i>                          |            |
| <b>WAIL: Collection-Based Personal Web Archiving.....</b>                                                                 | <b>340</b> |
| <i>John A. Berlin, Mat Kelly, Michael L. Nelson and Michele C. Weigle</i>                                                 |            |

## TUTORIALS

|                                                                           |            |
|---------------------------------------------------------------------------|------------|
| <b>Introduction to Digital Libraries .....</b>                            | <b>342</b> |
| <i>Edward A. Fox</i>                                                      |            |
| <b>Introduction to the Digital Public Library of America API .....</b>    | <b>344</b> |
| <i>Unmil P. Karadkar, Audrey Altman and Mark Breedlove</i>                |            |
| <b>Scholarly Data Mining: Making Sense of Scientific Literature .....</b> | <b>346</b> |
| <i>Horacio Saggion and Francesco Ronzano</i>                              |            |

## WORKSHOPS

|                                                                              |            |
|------------------------------------------------------------------------------|------------|
| <b>Rich Semantics and Direct Representation for Digital Collections.....</b> | <b>348</b> |
| <i>Robert B. Allen</i>                                                       |            |
| <b>Social Media and the Web of Linked Data.....</b>                          | <b>350</b> |
| <i>Diana Trandabat and Daniela Gifu</i>                                      |            |
| <b>Web Archiving and Digital Libraries (WADL).....</b>                       | <b>352</b> |
| <i>Edward A. Fox, Zhiwu Xie and Martin Klein</i>                             |            |

## PANEL

|                                                                                                                   |            |
|-------------------------------------------------------------------------------------------------------------------|------------|
| <b>"Can We Really Show This"?: Ethics, Representation and Social Justice<br/>in Sensitive Digital Space .....</b> | <b>354</b> |
| <i>Deborah Maron, Dorothy Berry, Fay Cobb Payton, Samantha Lakin<br/>and Erin White</i>                           |            |

|                           |            |
|---------------------------|------------|
| <b>Author Index .....</b> | <b>357</b> |
|---------------------------|------------|