Quantifying Research Integrity

Synthesis Lectures on Information Concepts, Retrieval, and Services

Editor
Gary Marchionini, ***University of North Carolina at Chapel Hill***

Synthesis Lectures on Information Concepts, Retrieval, and Services publishes short books on topics pertaining to information science and applications of technology to information discovery, production, distribution, and management. Potential topics include: data models, indexing theory and algorithms, classification, information architecture, information economics, privacy and identity, scholarly communication, bibliometrics and webometrics, personal information management, human information behavior, digital libraries, archives and preservation, cultural informatics, information retrieval evaluation, data fusion, relevance feedback, recommendation systems, question answering, natural language processing for retrieval, text summarization, multimedia retrieval, multilingual retrieval, and exploratory search.

Quantifying Research Integrity
Michael Seadle
2016

Web Indicators for Research Evaluation: A Practical Guide
Michael Thelwall
2016

Trustworthy Policies for Distributed Repositories
Reagan W. Moore, Hao Xu, Mike Conway, Arcot Rajasekar, Jon Crabtree, and Helen Tibbo
2016

The Notion of Relevance in Information Science: Everybody knows what relevance is. But, what is it really?
Tefko Saracevic
2016

Dynamic Information Retrieval Modeling
Grace Hui Yang, Marc Sloan, and Jun Wang
2016

Learning from Multiple Social Networks
Liqiang Nie, Xuemeng Song, and Tat-Seng Chua
2016

Scholarly Collaboration on the Academic Social Web
Daqing He and Wei Jeng
2016

Scalability Challenges in Web Search Engines
B. Barla Cambazoglu and Ricardo Baeza-Yates
2015

Social Informatics Evolving
Pnina Fichman, Madelyn R. Sanfilippo, and Howard Rosenbaum
2015

On the Efficient Determination of Most Near Neighbors: Horseshoes, Hand Grenades, Web Search and Other Situations When Close Is Close Enough, Second Edition
Mark S. Manasse
2015

Building a Better World with Our Information: The Future of Personal Information Management, Part 3
William Jones
2015

Click Models for Web Search
Aleksandr Chuklin, Ilya Markov, and Maarten de Rijke
2015

Information Communication
Feicheng Ma
2015

Social Media and Library Services
Lorri Mon
2015

Analysis and Visualization of Citation Networks
Dangzhi Zhao and Andreas Strotmann
2015

The Taxobook: Applications, Implementation, and Integration in Search: Part 3 of a 3-Part Series
Marjorie M.K. Hlava
2014

The Taxobook: Principles and Practices of Building Taxonomies, Part 2 of a 3-Part Series
Marjorie M.K. Hlava
2014

Measuring User Engagement
Mounia Lalmas, Heather O'Brien, and Elad Yom-Tov
2014

The Taxobook: History, Theories, and Concepts of Knowledge Organization, Part 1 of a 3-Part Series
Marjorie M.K. Hlava
2014

Children's Internet Search: Using Roles to Understand Children's Search Behavior
Elizabeth Foss and Allison Druin
2014

Digital Library Technologies: Complex Objects, Annotation, Ontologies, Classification, Extraction, and Security
Edward A. Fox and Ricardo da Silva Torres
2014

Digital Libraries Applications: CBIR, Education, Social Networks, eScience/Simulation, and GIS
Edward A. Fox and Jonathan P. Leidig
2014

Information and Human Values
Kenneth R. Fleischmann
2013

Multiculturalism and Information and Communication Technology
Pnina Fichman and Madelyn R. Sanfilippo
2013

Transforming Technologies to Manage Our Information: The Future of Personal Information Management, Part II
William Jones
2013

Designing for Digital Reading
Jennifer Pearson, George Buchanan, and Harold Thimbleby
2013

Information Retrieval Models: Foundations and Relationships
Thomas Roelleke
2013

Key Issues Regarding Digital Libraries: Evaluation and Integration
Rao Shen, Marcos André Gonçalves, and Edward A. Fox
2013

Visual Information Retrieval using Java and LIRE
Mathias Lux and Oge Marques
2013

On the Efficient Determination of Most Near Neighbors: Horseshoes, Hand Grenades, Web Search and Other Situations When Close is Close Enough
Mark S. Manasse
2012

The Answer Machine
Susan E. Feldman
2012

Theoretical Foundations for Digital Libraries: The 5S (Societies, Scenarios, Spaces, Structures, Streams) Approach
Edward A. Fox, Marcos André Gonçalves, and Rao Shen
2012

The Future of Personal Information Management, Part I: Our Information, Always and Forever
William Jones
2012

Search User Interface Design
Max L. Wilson
2011

Information Retrieval Evaluation
Donna Harman
2011

Knowledge Management (KM) Processes in Organizations: Theoretical Foundations and Practice
Claire R. McInerney and Michael E. D. Koenig
2011

Search-Based Applications: At the Confluence of Search and Database Technologies
Gregory Grefenstette and Laura Wilber
2010

Information Concepts: From Books to Cyberspace Identities
Gary Marchionini
2010

Estimating the Query Difficulty for Information Retrieval
David Carmel and Elad Yom-Tov
2010

iRODS Primer: Integrated Rule-Oriented Data System
Arcot Rajasekar, Reagan Moore, Chien-Yi Hou, Christopher A. Lee, Richard Marciano, Antoine de Torcy, Michael Wan, Wayne Schroeder, Sheau-Yen Chen, Lucas Gilbert, Paul Tooby, and Bing Zhu
2010

Collaborative Web Search: Who, What, Where, When, and Why
Meredith Ringel Morris and Jaime Teevan
2009

Multimedia Information Retrieval
Stefan Rüger
2009

Online Multiplayer Games
William Sims Bainbridge
2009

Information Architecture: The Design and Integration of Information Spaces
Wei Ding and Xia Lin
2009

Reading and Writing the Electronic Book
Catherine C. Marshall
2009

Hypermedia Genes: An Evolutionary Perspective on Concepts, Models, and Architectures
Nuno M. Guimarães and Luís M. Carrico
2009

Understanding User-Web Interactions via Web Analytics
Bernard J. (Jim) Jansen
2009

XML Retrieval
Mounia Lalmas
2009

Faceted Search
Daniel Tunkelang
2009

Introduction to Webometrics: Quantitative Web Research for the Social Sciences
Michael Thelwall
2009

Exploratory Search: Beyond the Query-Response Paradigm
Ryen W. White and Resa A. Roth
2009

New Concepts in Digital Reference
R. David Lankes
2009

Automated Metadata in Multimedia Information Systems: Creation, Refinement, Use in Surrogates, and Evaluation
Michael G. Christel
2009

Quantifying Research Integrity

Michael Seadle

ISBN: 978-3-031-01178-8 paperback
ISBN: 978-3-031-02306-4 ebook

DOI 10.1007/978-3-031-02306-4

A Publication in the Springer series
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES

Lecture #53
Series Editor: Gary Marchionini, *University of North Carolina at Chapel Hill*
Series ISSN
Print 1947-945X Electronic 1947-9468

Quantifying Research Integrity

Michael Seadle
Humboldt-Universität zu Berlin

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #53

ABSTRACT

Institutions typically treat research integrity violations as black and white, right or wrong. The result is that the wide range of grayscale nuances that separate accident, carelessness, and bad practice from deliberate fraud and malpractice often get lost. This lecture looks at how to quantify the grayscale range in three kinds of research integrity violations: plagiarism, data falsification, and image manipulation.

Quantification works best with plagiarism, because the essential one-to-one matching algorithms are well known and established tools for detecting when matches exist. Questions remain, however, of how many matching words of what kind in what location in which discipline constitute reasonable suspicion of fraudulent intent. Different disciplines take different perspectives on quantity and location. Quantification is harder with data falsification, because the original data are often not available, and because experimental replication remains surprisingly difficult. The same is true with image manipulation, where tools exist for detecting certain kinds of manipulations, but where the tools are also easily defeated.

This lecture looks at how to prevent violations of research integrity from a pragmatic viewpoint, and at what steps can institutions and publishers take to discourage problems beyond the usual ethical admonitions. There are no simple answers, but two measures can help: the systematic use of detection tools and requiring original data and images. These alone do not suffice, but they represent a start.

The scholarly community needs a better awareness of the complexity of research integrity decisions. Only an open and wide-spread international discussion can bring about a consensus on where the boundary lines are and when grayscale problems shade into black. One goal of this work is to move that discussion forward.

KEYWORDS

research integrity, plagiarism, data falsification, image manipulation, grayscale decisions, research fraud, detection tools, plagiarism tools, forensic droplets, Retraction Watch, Office of Research Integrity, HEADT Centre

To my wife Joan

Contents

Preface

Research integrity problems are nothing new.[1] While student misconduct is an important and much discussed issue, this lecture focuses on integrity issues involving the scientific and scholarly record. Frauds, mistakes, and retractions undermine the credibility of both the sciences and the humanities. Today, for example, elements of the popular press are skeptical about atmospheric change and global warming because they doubt the reliability of the theories and data that undergird the claims. Skeptics can point to retractions as evidence that science is flawed. The need for discovering and preventing research malpractice grows as the quantity of publication grows.

Retractions are in themselves not bad, because they are evidence of a self-policing process for the scholarly record. Nonetheless a different kind of danger exists when self-policing becomes a vigilante process with no clear rules or limits. As chair of my university's Commission on Research Malpractice,[2] I have encountered a wide range of accusations, some based on personal animus, some based on unrealistic standards of perfection, and many where the commission itself had to establish rules. The problem is that our scholarly measures for research integrity lack precision. The goal of this lecture is to suggest more accurate and more appropriate metrics to improve the self-policing process. I choose the term "quantification" for this because the goal is to approach a more formal and more mathematical accuracy in making judgments, even though missing information and unclear circumstances often reduce actual quantification to no more than a goal.

This work can also be read as a form of history of science that looks at the flaws and failings of the scholarly process. As scholars we often wish to believe that we create knowledge by building on discoveries from the past, but it may be more realistic to add that the discovery process also involves exposing frauds, mistakes, and other forms of intentional and unintentional error. Sometimes this is viewed critically, because people forget that scientists are mere humans working at a particular time in a particular place, as Steven Shapin explains in "Lowering the tone in the history of science" [Shapin, 2010]. The examples in this lecture come mostly from recent years, in part because new cases keep coming to light, but also to show how current the problem is. This lecture does not speculate about why well-trained scholars engage in malpractice. The reasons are too complex for simple answers like greed or peer pressure or cultural tolerance. It would take a more extensive work to gain a real understanding of the full range of social, psychological, and cultural factors involved in research integrity violations.

The risk in writing any lecture about research integrity issue is that the author must be absolutely scrupulous. This is particularly true in the area of plagiarism. Some plagiarism hunters regard paraphrasing even with a proper reference as a violation. For that reason I have chosen to

[1] See the literature review in Chapter 2.

[2] In German: Kommission zur Überprüfung von Vorwürfen wissenschaftlichen Fehlverhaltens.

quote directly from sources, rather than to summarize or paraphrase the information they provide. This also has the benefit of letting readers judge the sources for themselves.

Confidentiality and anonymity are important issues for any commission investigating research integrity cases, because the mere suggestion of malpractice can destroy careers. In this work I have used only cases where the accusations are publicly available. For cases that are very current and potentially undecided, I have deliberately left out the full names and used only abbreviations. Links in the references do, however, point to the actual accusations, which include the names of those involved. This offers a modest if imperfect balance between transparency and confidentiality. Readers who genuinely want to get at the names in publicly available materials may do so, but they must make an extra effort. One option would be to write a purely theoretical work about how research integrity issues should be measured, but that would perpetuate the existing tendency toward oversimplification. The messy empirical details of the cases matter in understanding how and how well quantification works.

Michael Seadle
Berlin, December 2016

Acknowledgments

First and foremost I would like to acknowledge the contributions of my wife, Prof. Dr. Joan Luft, for her suggestions, proofreading, and loving support. I could not have completed this work without her.

I would also like to thank my staff, Melanie Rügenhagen, Dr. Thorsten Beck, Ulrike Stöckel, and Stephanie van de Sandt for their encouragement and support.

My university colleagues, especially Prof. Dr. Elke Greifeneder and Prof. Dr. Vivien Pretras, have given advice and picked up tasks that I could not get done. Our Vice-President for Research, Prof. Dr. Peter Frensch, also deserves credit for the many discussions we had about research integrity issues. Many other colleagues and students who are too numerous to name individually should be thanked for conversations and idea as well.

I would like to thank Gobinda Chowdhury, Lynn Silipigni Connaway, and John Budd for their helpful comments and suggestions.

Finally I would like to thank Elsevier for funding the Humboldt Elsevier Advanced Data and Text Centre, which made part of this research possible. And I want to recognize Humboldt-Universität zu Berlin for providing an intellectual atmosphere in keeping with the traditions Hegel, Planck, Einstein, and other notable scholars.

Michael Seadle
Berlin, December 2016