LEGAL INFORMATICS

This groundbreaking work offers a first-of-its-kind overview of legal informatics, the academic discipline underlying the technological transformation and economics of the legal industry. Edited by Daniel Martin Katz, Ron Dolin, and Michael J. Bommarito, and featuring contributions from more than two dozen academic and industry experts, chapters cover the history and principles of legal informatics and background technical concepts – including natural language processing and distributed ledger technology. The volume also presents real-world case studies that offer important insights into document review, due diligence, compliance, case prediction, billing, negotiation and settlement, contracting, patent management, legal research, and online dispute resolution. Written for both technical and non-technical readers, Legal Informatics is the ideal resource for anyone interested in identifying, understanding, and executing opportunities in this exciting field.

Daniel Martin Katz is Professor of Law, Illinois Tech – Chicago Kent College of Law where he directs The Law Lab. He also serves as an external affiliated faculty at CodeX, the Stanford Center for Legal Informatics and the Academic Director of the Bucerius Center for Legal Technology and Data Science. A scientist and technologist, Professor Katz applies an innovative, polytechnic approach to teaching law to help create lawyers for today’s biggest societal challenges. Both his teaching and scholarship integrate science, technology, engineering, and mathematics.

Ron Dolin is a Senior Research Fellow and Lecturer on Law at Harvard Law School, focusing on the impact of technology on the practice and nature of law. He received a B.A. in math and physics, a Ph.D. in computer science, and a J.D. He has worked at JPL, CERN, and Google, and is a licensed attorney in California. Dolin’s research includes developing legal quality metrics, examining the impact of standardized benchmarks on the legal system, and analyzing the legal market from the perspective of The Innovator’s Dilemma.

Michael J. Bommarito is an entrepreneur, educator, and investor in the legal and technology industries. His experience spans R&D, technology, business, and operations – ranging from top Am Law firms and S5+ AUM investment firms to idea-stage startups. He is affiliated with the University of Michigan, Stanford University, Michigan State College of Law, and the Illinois Tech – Chicago Kent College of Law. His research has been published in Science, Physica A, Artificial Intelligence and Law, and Quantitative Finance.
Legal Informatics

Edited by

DANIEL MARTIN KATZ
Illinois Tech – Chicago Kent Law
Bucerius Law School
Stanford CodeX

RON DOLIN
Harvard Law School

MICHAEL J. BOMMARITO
Stanford CodeX
Contents

List of Figures viii
List of Tables xi
List of Contributors xii

PART I INTRODUCTION TO LEGAL INFORMATICS 1

1.1 Motivation and Rationale for this Book 3
Michael J. Bommarito II, Daniel Martin Katz, and Ron Dolin

1.2 Technology Issues in Legal Philosophy 5
Ron Dolin

1.3 The Origins and History of Legal Informatics 24
Michael J. Bommarito II

PART II LEGAL INFORMATICS: BUILDING BLOCKS AND CORE CONCEPTS 31

A. INFORMATION REPRESENTATION, PREPROCESSING, AND DOCUMENT ASSEMBLY 33

2.1 Representation of Legal Information 35
Katie Atkinson

2.2 Information Intermediation 41
Ron Dolin

2.3 Preprocessing Data 55
Michael J. Bommarito II

2.4 XML in Law: The Role of Standards in Legal Informatics 61
Ron Dolin

2.5 Document Automation 69
Marc Lauritsen
### Contents

**B. ARTIFICIAL INTELLIGENCE, MACHINE LEARNING, NATURAL LANGUAGE PROCESSING, AND BLOCKCHAIN**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.6</td>
<td>AI + Law: An Overview</td>
<td>85</td>
</tr>
<tr>
<td></td>
<td>Daniel Martin Katz</td>
<td></td>
</tr>
<tr>
<td>2.7</td>
<td>Machine Learning and Law</td>
<td>94</td>
</tr>
<tr>
<td></td>
<td>Daniel Martin Katz and John J. Nay</td>
<td></td>
</tr>
<tr>
<td>2.8</td>
<td>Natural Language Processing for Legal Texts</td>
<td>99</td>
</tr>
<tr>
<td></td>
<td>John J. Nay</td>
<td></td>
</tr>
<tr>
<td>2.9</td>
<td>Introduction to Blockchain and Cryptography</td>
<td>114</td>
</tr>
<tr>
<td></td>
<td>Nelson M. Rosario</td>
<td></td>
</tr>
</tbody>
</table>

**C. PROCESS IMPROVEMENT, GAMIFICATION, AND DESIGN THINKING**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.10</td>
<td>Legal Informatics-Based Technology in Broader Workflows</td>
<td>121</td>
</tr>
<tr>
<td></td>
<td>Kenneth A. Grady</td>
<td></td>
</tr>
<tr>
<td>2.11</td>
<td>Gamification of Work and Feedback Systems</td>
<td>136</td>
</tr>
<tr>
<td></td>
<td>Stephanie Kimbro</td>
<td></td>
</tr>
<tr>
<td>2.12</td>
<td>Introduction to Design Thinking for Law</td>
<td>155</td>
</tr>
<tr>
<td></td>
<td>Margaret Hagan</td>
<td></td>
</tr>
</tbody>
</table>

**D. EVALUATION**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.13</td>
<td>Measuring Legal Quality</td>
<td>177</td>
</tr>
<tr>
<td></td>
<td>Ron Dolin</td>
<td></td>
</tr>
</tbody>
</table>

**PART III USE CASES IN LEGAL INFORMATICS**

**A. CONTRACTS AND PATENTS**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.1</td>
<td>Contract Analytics</td>
<td>201</td>
</tr>
<tr>
<td></td>
<td>Noah Waisberg</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Contracts as Interfaces: Visual Representation Patterns in Contract Design</td>
<td>205</td>
</tr>
<tr>
<td></td>
<td>Helena Haapio and Stefania Passera</td>
<td></td>
</tr>
<tr>
<td>3.3</td>
<td>Distributed Ledgers, Cryptography, and Smart Contracts</td>
<td>213</td>
</tr>
<tr>
<td></td>
<td>Nina Gunther Kilbride</td>
<td></td>
</tr>
<tr>
<td>3.4</td>
<td>Patent Analytics: Information from Innovation</td>
<td>239</td>
</tr>
<tr>
<td></td>
<td>Andrew W. Torrance and Jevin D. West</td>
<td></td>
</tr>
</tbody>
</table>

**B. LITIGATION AND E-DISCOVERY**

<table>
<thead>
<tr>
<th>Section</th>
<th>Title</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.5</td>
<td>The Core Concepts of E-discovery</td>
<td>257</td>
</tr>
<tr>
<td></td>
<td>Jonathan Kerry-Tyerman and A. J. Shankar</td>
<td></td>
</tr>
</tbody>
</table>
### Contents

3.6 Predictive Coding in E-discovery and the NexLP Story Engine
   Irina Matveeva
   p. 315

3.7 Examining Public Court Data to Understand and Predict Bankruptcy Case Results
   Warren E. Agin
   p. 335

C. Legal Research, Government Data, and Access to Legal Information
   p. 355

3.8 Fastcase, and the Visual Understanding of Judicial Precedents
   Ed Walters and Jeff Asjes
   p. 357

3.9 Mining Information from Statutory Texts in a Public Health Domain
   Kevin D. Ashley
   p. 371

3.10 Gov2Vec: A Case Study in Text Model Application to Government Data
    John J. Nay
    p. 393

3.11 Representation and Automation of Legal Information
    Katie Atkinson
    p. 397

D. Dispute Resolution and Access to Justice
   p. 407

3.12 Online Dispute Resolution
    Dave Orr and Colin Rule
    p. 409

3.13 Access to Justice and Technology: Reaching a Greater Future for Legal Aid
    Ronald W. Staudt and Alexander F. A. Rabanal
    p. 416

3.14 Designing Legal Experiences: Online Communication and Resolution in Courts
    Maximilian A. Bulinski and J. J. Prescott
    p. 430

 PART IV Legal Informatics in the Industrial Context

A. Challenges Facing Innovation in Law
   p. 451

4.1 Adaptive Innovation: The Innovator’s Dilemma in Big Law
    Ron Dolin and Thomas Buley
    p. 453

4.2 Legal Data Access
    Christine Bannan
    p. 467

B. Large Firm and Corporate Legal Informatics Case Studies
   p. 481

4.3 A History of Knowledge Management at Littler Mendelson
    Scott Rechtschaffen
    p. 483

4.4 Legal Operations at Google
    Mary O’Carroll and Stephanie Kimbro
    p. 501
Figures

1.3.1 This nilometer on Elephantine Island in Aswan, Egypt is one of the first applications of informatics in the law

2.1.1 CATO abstract factor hierarchy

2.2.1 Term vector spaces: (a) simple query; (b) complex query

2.2.2 Corpus Juris Civilis, republished in 1627 with commentary by Johann Fehe

2.5.1 Functions of document automation

2.5.2 Realms of knowledge management

6.1 Subfields within artificial intelligence

8.1 Example of a network of similar bills under consideration by Congress

8.2 Number of presidential proclamations or determinations, trivial proclamations, memoranda, and executive orders, from 1928 through 2015

8.3 Positive correlations between topics represented as lines

8.4 Part-of-speech tagging for a sentence in a state bill

8.5 Named entity recognition for the last part of the bill sentence

8.6 Syntactic dependencies

8.7 Co-reference resolution

8.8 Word2Vec algorithm

10.1 Anna’s process

10.2 Chart of email steps

10.3 Anna’s email process

10.4 Example workflow

10.5 Process stick for Anna’s activities

12.1 Six levels of legal design work that can improve users’ experience with the legal system, from least ambitious to most

12.1.1 88.4% of what? Failed to settle on 11.6% of losing cases?

12.1.2 Additive scoring (a) vs. (unweighted) multiplicative scoring (b)

12.1.3 Weighted multiplicative scoring

12.1.4 Geographical relevance of 0, 0.05, and 1

12.2 The emerging view of the purpose and functions of contracts

12.2.1 Linear timelines

12.2.2 Circular timeline

12.2.3 Processes with different paths

12.2.4 Variables over time

12.2.5 Synchronous progress of multiple processes or events

12.2.6 Flowchart

12.2.8 Table with bulleted lists
List of Figures

3.2.9 Table with color-coding and icons to show key differences
3.2.10 Swimlane
3.2.11 Icons in a tenancy agreement
3.2.12 Delivery diagram
3.2.13 Two different timelines, two different understandings
3.2.14 Decision tree evaluating average expected value of alternative decisions
3.3.1 A cryptographic hash
3.3.2 Linked timestamping storage in hash trees (Merkle trees)
3.3.3 Linked timestamping of case law
3.4.1 US Patent No. 6,164,870
3.4.2 A CIPO patent landscape of shale oil and gas technologies
3.4.3 Federal judicial districts by mean importance of patents litigated there
3.4.4 An example citation network
3.4.5 Automatic categorization
3.5.1 The EDRM
3.6.1 Visualization of the machine learning process
3.6.2 Mathematical objects in SVM with decision boundaries
3.6.3 A news article containing last names, nicknames, abbreviations, and metonymy
3.6.4 LSI-created word associations
3.6.5 Subdivisions within a cluster
3.6.6 Nested cluster diagram
3.6.7 Dendrogram
3.6.8 Entities extracted from an article about airline fare increases
3.6.9 NexLP Enron email communication network and content summary for emails discussing Enron
3.6.10 Example of emails under investigation
3.6.11 Failure of simple keyword search for “William Brown”
3.6.12 Story Engine summarizes data and discovers relevant details
3.7.1 This chart shows the number of cases in the data set for each disposition code
3.7.2 Case volume and success rate by judicial district, grouped vertically by federal circuit
3.7.3 Example of a decision tree
3.7.4 A small section of the final decision tree
3.7.5 Importance scores for model features
3.7.6 Confusion matrix for the three test sets, combined
3.7.7 Probability output to percentage conversion graph
3.8.1 A problem search
3.8.2 Fastcase search filters
3.8.3 Boolean search
3.8.4 Image from Fastcase’s Interactive Timeline
3.8.5 Y-axis changed to show court level
3.8.6 “Same-sex marriage” Interactive Timeline, with arrowing indicating Lawrence v. Texas
3.8.7 Close-up of Figure 3.8.6, “Same-sex marriage” case relevance
3.9.1 LENA statutory network comparing Texas and Pennsylvania statutes on epidemic emergencies with infectious diseases
3.9.2 Heat map of agent strength in statutory networks
3.9.3 Schematic view of phases of coding text by the SPH team
3.9.4 Sample public health emergency statutory provision
3.9.5 Results of initial machine learning
<table>
<thead>
<tr>
<th>Section</th>
<th>Figure Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>3.9.6</td>
<td>Results of successively applying more states’ classification models</td>
</tr>
<tr>
<td>3.9.7</td>
<td>Sample comparison of statutory provision texts</td>
</tr>
<tr>
<td>3.9.8</td>
<td>Interactive ML tool screen</td>
</tr>
<tr>
<td>3.9.9</td>
<td>Results of evaluation of the interactive ML approach</td>
</tr>
<tr>
<td>3.10.1</td>
<td>Sentence probabilities across bills for first data (a) and last data (b)</td>
</tr>
<tr>
<td>3.11.1</td>
<td>CATO abstract factor hierarchy</td>
</tr>
<tr>
<td>3.11.2</td>
<td>IBP logical model</td>
</tr>
<tr>
<td>3.11.3</td>
<td>Example abstract argumentation framework</td>
</tr>
<tr>
<td>3.11.4</td>
<td>Visualization of automobile exception domain addressing the question “Can an automobile be searched without a warrant?”</td>
</tr>
<tr>
<td>3.12.1</td>
<td>Conditional logic to determine judicial district in IL</td>
</tr>
<tr>
<td>3.12.2</td>
<td>Conditional logic to determine the end-user’s full name</td>
</tr>
<tr>
<td>3.12.3</td>
<td>Example of map feature (i.e., decision tree)</td>
</tr>
<tr>
<td>3.12.4</td>
<td>Report feature – readability score</td>
</tr>
<tr>
<td>3.13.5</td>
<td>A2J Guided Interview progress bar in Spanish</td>
</tr>
<tr>
<td>3.13.6</td>
<td>Example of a question screen in A2J Author</td>
</tr>
<tr>
<td>3.14.1</td>
<td>Matterhorn litigant access</td>
</tr>
<tr>
<td>3.14.2</td>
<td>Eligibility for review</td>
</tr>
<tr>
<td>3.14.3</td>
<td>Judge’s reviews of a litigant’s request</td>
</tr>
</tbody>
</table>
## Tables

<table>
<thead>
<tr>
<th>Section</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.5.1</td>
<td>Document automation range of functions</td>
<td>71</td>
</tr>
<tr>
<td>2.5.2</td>
<td>Document assembly products, past and present</td>
<td>77</td>
</tr>
<tr>
<td>2.5.3</td>
<td>Worksheet for document automation goals</td>
<td>82</td>
</tr>
<tr>
<td>2.10.1</td>
<td>Process study sheet</td>
<td>125</td>
</tr>
<tr>
<td>3.2.1</td>
<td>Examples of pattern structures</td>
<td>222</td>
</tr>
<tr>
<td>3.2.2</td>
<td>Categorization of contract visualization patterns</td>
<td>223</td>
</tr>
<tr>
<td>3.5.1</td>
<td>Hash values for the Gettysburg Address</td>
<td>303</td>
</tr>
<tr>
<td>3.6.1</td>
<td>LSI-computed areas of interest in Enron emails</td>
<td>327</td>
</tr>
<tr>
<td>3.6.2</td>
<td>Main named entity types</td>
<td>329</td>
</tr>
<tr>
<td>3.7.1</td>
<td>Prediction accuracy of model for different groupings of probability numbers</td>
<td>348</td>
</tr>
<tr>
<td>3.9.1</td>
<td>Comparing similarly purposed provisions from Pennsylvania (left) and Florida (right)</td>
<td>385</td>
</tr>
<tr>
<td>3.9.2</td>
<td>Results of models trained on Florida (FL), Pennsylvania (PA), and combined (FL + PA) data sets applied to PA (left) and FL (right) test sets</td>
<td>386</td>
</tr>
<tr>
<td>3.10.1</td>
<td>Synthetic summary for three topics, “Enacted” or “Failed,” in the House or the Senate</td>
<td>395</td>
</tr>
<tr>
<td>3.11.1</td>
<td>Base-level factors in CATO</td>
<td>399</td>
</tr>
<tr>
<td>3.11.2</td>
<td>CATO as ADF</td>
<td>402</td>
</tr>
<tr>
<td>3.11.3</td>
<td>IBP logical model as an ADF</td>
<td>402</td>
</tr>
</tbody>
</table>
Contributors

Warren E. Agin  Adjunct Professor, Boston College Law School; Managing Director for Digital Strategy and Solutions, Elevate Services

Kevin D. Ashley  Professor of Law, University of Pittsburgh

Jeff Asjes  Product Manager, Fastcase

Katie Atkinson  Professor of Computer Science, University of Liverpool

Christine Bannan  Policy Counsel, Open Technology Institute, New America

Michael J. Bommarito II  Fellow, CodeX – Stanford Center for Legal Informatics

Thomas Buley  Vice-President of Product, Guideline

Maximilian A. Bulinski  Research Fellow, University of Michigan Law School

Ron Dolin  Lecturer on Law, Harvard Law School; Senior Research Fellow, Center on the Legal Profession, Harvard Law School

Kenneth A. Grady  Adjunct Professor, Michigan State University College of Law

Helena Haapio  Associate Professor of Business Law, University of Vaasa; International Contract Counsel, Lexpert Ltd.

Margaret Hagan  Director Legal Design Lab, Stanford Law School Lecturer, Stanford Institute of Design (d.school)

Daniel Martin Katz  Professor of Law, Illinois Tech – Chicago Kent College of Law Academic Director, Bucerius Center for Legal Technology & Data Science; Affiliated Faculty, CodeX – Stanford Center for Legal Informatics

Jonathan Kerry-Tyerman  Vice-President of Business Development, Everlaw

Nina Gunther Kilbride  Principal, Corpening Labs

Stephanie Kimbro  Former Research Fellow, Stanford Law School; Independent Consultant, Virtual Law Practice

Marc Lauritsen  President, Capstone Practice Systems

Irina Matveeva  Adjunct Professor, Illinois Tech; Head of Machine Learning, NexLP
List of Contributors

John J. Nay CEO & Co-Founder, Skopos Labs
Mary O’Carroll Director of Legal Operations, Google; President, CLOC
Dave Orr Senior Product Manager, Google
Stefania Passera Legal Tech Lab, Helsinki University
J. J. Prescott Professor of Law, University of Michigan Law School
Alexander F. A. Rabanal Associate Director, The Law Lab @ Illinois Tech – Chicago Kent College of Law
Scott Rechtshaffen Chief Knowledge Officer, Littler Mendelson
Nelson M. Rosario Adjunct Professor, Illinois Tech – Chicago-Kent College of Law Principal, Smolinski Rosario Law P.C.
Colin Rule Vice-President for Online Dispute Resolution, Tyler Technologies
A. J. Shankar Founder and CEO, Everlaw
Ronald W. Staudt Professor of Law Emeritus, Illinois Tech – Chicago Kent College of Law
Andrew W. Torrance Paul E. Wilson Distinguished Professor of Law, University of Kansas School of Law
Noah Waisberg CEO and Co-founder, Kira Systems
Ed Walters CEO and Co-founder, Fastcase
Jevin D. West Associate Professor, Information School, University of Washington