Explorations in the Digital History of Ideas

What would the history of ideas look like if we were able to read the entire archive of printed material of a historical period? Would our 'great men (usually)' story of how ideas are formed and change over time begin to look very different? This book explores these questions through case studies on ideas such as 'liberty', 'republicanism' or 'government' using digital humanities approaches to large-scale text datasets. It sets out the methodologies and tools created by the Cambridge Concept Lab as exemplifications of how new digital methods can open up the history of ideas to heretofore unseen avenues of inquiry and evidence. By applying text mining techniques to intellectual history or the history of concepts, this book explains how computational approaches to text mining can substantially increase the power of our understanding of ideas in history.

Peter de Bolla is Professor of Cultural History and Aesthetics at the University of Cambridge. His publications include *The Architecture of Concepts: The Historical Formation of Human Rights* (2013), which won the Robert Lowry Patten Award in 2015. He is the author or editor of nine books, including *The Discourse of the Sublime: Readings in History, Aesthetics and the Subject* (1989), *Art Matters* (2001) and *The Education of the Eye: Painting, Landscape and Architecture in Eighteenth Century Britain* (2003). He directed the Cambridge Concept Lab between 2013 and 2017, a £1.5 m funded project on the structure of concepts. He is an International Honorary Member of the American Academy of Arts and Sciences.

Explorations in the Digital History of Ideas

New Methods and Computational Approaches

Edited by
Peter de Bolla
University of Cambridge

CAMBRIDGE
UNIVERSITY PRESS

Shaftesbury Road, Cambridge CB2 8EA, United Kingdom

One Liberty Plaza, 20th Floor, New York, NY 10006, USA

477 Williamstown Road, Port Melbourne, VIC 3207, Australia

314–321, 3rd Floor, Plot 3, Splendor Forum, Jasola District Centre, New Delhi – 110025, India

103 Penang Road, #05–06/07, Visioncrest Commercial, Singapore 238467

Cambridge University Press is part of Cambridge University Press & Assessment, a department of the University of Cambridge.

We share the University's mission to contribute to society through the pursuit of education, learning and research at the highest international levels of excellence.

www.cambridge.org
Information on this title: www.cambridge.org/9781009263603

DOI: 10.1017/9781009263610

First published 2024
First paperback edition 2025

A catalogue record for this publication is available from the British Library

ISBN 978-1-009-26358-0 Hardback
ISBN 978-1-009-26360-3 Paperback

Additional resources for this publication at www.cambridge.org/deBolla

Contents

List of Figures *page* vii
List of Tables xi
List of Contributors xv
Acknowledgements xvi

Part I Computational Methodologies for the History of Ideas 1

1 Introduction 3
PETER DE BOLLA

2 Distributional Concept Analysis and the Digital History of Ideas 12
PETER DE BOLLA, EWAN JONES, PAUL NULTY, GABRIEL RECCHIA AND JOHN REGAN

3 Operationalising Conceptual Structure 54
PAUL NULTY

Part II Case Studies in the Digital History of Ideas 77

4 The Idea of Liberty, 1600–1800 79
PETER DE BOLLA, EWAN JONES, PAUL NULTY, GABRIEL RECCHIA AND JOHN REGAN

5 The Idea of Government in the British Eighteenth Century 98
PETER DE BOLLA, EWAN JONES, PAUL NULTY, GABRIEL RECCHIA AND JOHN REGAN

6 Republicanism in the Founding of America 122
PETER DE BOLLA

7 Enlightenment Entanglements of Improvement and Growth
PETER DE BOLLA, RYAN HEUSER AND MARK ALGEE-HEWITT 140

8 The Idea of Commercial Society: Changing Contexts and Scales
JOHN REGAN 163

9 The Age of Irritability
EWAN JONES AND NATALIE ROXBURGH 184

10 On Bubbles and Bubbling: The Idea of 'The South Sea Bubble'
CLAIRE WILKINSON 206

11 Embedded Ideas: Revolutionary Theory and Political Science in the Eighteenth Century
MARK ALGEE-HEWITT 225

12 Computing Koselleck: Modelling Semantic Revolutions, 1720–1960
RYAN HEUSER 256

Index 286

Additional appendix resources at www.cambridge.org/deBolla

Figures

2.1 Number of tokens per year. Source: ECCO. *page* 44
2.2 Correlation of word frequency with *dpf*. Source: ECCO. 45
2.3 Average (i.e. mean) number of terms on the *dpf* list after the bend in the curve is correlated with window size. Source: ECCO. 46
2.4 'Violin' plot showing density, median and interquartile range of degree for all tokens at each distance. Source: ECCO. 47
2.5 Average number of terms on *dpf* correlated with year. Source: ECCO. 47
2.6 Relationship between clustering coefficients and term frequency. Source: ECCO. 48
2.7 Relationship between clustering coefficients and distance. Source: ECCO. 49
2.8 Relationship between clustering coefficients and time. Source: ECCO. 49
3.1 Top ten *log-dpf* association scores with the focal term liberty. Date spread 1790–1795, containing 3,000 documents, 2,198,379 words. Source: ECCO. 68
3.2 Word meanings represented in a two-dimensional space. 68
3.3 Top twenty words scored by *log-dpf* association with the focal term liberty. Date spread 1790–5, containing 3,000 documents, 2,198,379 words. Source: ECCO. 70
3.4 Top twenty words scored by *log-dpf* association with the focal term rights. Date spread 1790–5, containing 3,000 documents, 2,198,379 words. Source: ECCO. 70
3.5 A network visualisation with rights and liberty as focal terms, score threshold 0.0 and rank threshold 20. Shading is assigned by a community detection algorithm. Date spread 1790–5, containing 3,000 documents, 2,198,379 words. Source: ECCO. 71

3.6 A network visualisation with rights and liberty as focal terms, score threshold 0.4 and rank threshold 30. Shading is assigned by a community detection algorithm. Date spread 1790–5, containing 3,000 documents, 2,198,379 words. Source: ECCO. 73
4.1 Number of terms in the persistent bound-lexical company derived from the common set of terms at distance 10, 50, 100. Counts presented decade by decade 1700–1800. Source: ECCO. 94
4.2 A network visualisation with liberty and republican as the focal terms; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1780–1800. Shading is assigned by a community detection algorithm. Source: ECCO. 96
5.1 A network visualisation with government as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1720–30. Shading is assigned by a community detection algorithm. Source: ECCO. 106
5.2 A network visualisation with government as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1750–60. Shading is assigned by a community detection algorithm. Source: ECCO. 108
5.3 A network visualisation with government as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1790–1800. Shading is assigned by a community detection algorithm. Source: ECCO. 109
5.4 A network visualisation with despotism as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1790–1800. Shading is assigned by a community detection algorithm. Source: ECCO. 119
5.5 A network visualisation with liberty as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1790–1800. Shading is assigned by a community detection algorithm. Source: ECCO. 120
6.1 Number of documents per year containing the word republicanism, 1700–1800. Total number of documents: 202,191. Source: ECCO graphing tool. 124

7.1 A network visualisation with improvement as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1700–10. Shading is assigned by a community detection algorithm. Source: ECCO. 150
7.2 A network visualisation with improvement as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1720–30. Shading is assigned by a community detection algorithm. Source: ECCO. 151
7.3 A network visualisation with improvement as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1740–50. Shading is assigned by a community detection algorithm. Source: ECCO. 152
7.4 A network visualisation with growth as the focal term; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1790–1800. Shading is assigned by a community detection algorithm. Source: ECCO. 153
7.5 A network visualisation with improvement, growth and increase as the focal terms; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1720–30. Shading is assigned by a community detection algorithm. Source: ECCO. 155
7.6 A network visualisation with improvement, growth and increase as the focal terms; distance 100, score expressed as *log-dpf* with threshold of 2.6 and rank threshold 20. Date spread 1740–50. Shading is assigned by a community detection algorithm. Source: ECCO. 156
9.1 Vector subtraction: *dpf* profiles of habit at distance 10, 1751–1800, minus *dpf* profiles of habit at distance 10, 1700–51. 202
9.2 Vector subtraction: *dpf* profiles of system at distance 10, 1751–1800, minus *dpf* profiles of system at distance 10, 1700–51. 204
10.1 South-sea and bubble, distance 5, 1701–1800. Source: ECCO. 213
10.2 South-sea and bubbles, distance 5, 1701–1800. Source: ECCO. 217

11.1 Graph of the semantic difference between political and revolution within the embedding spaces over time. Source: ECCO. 232
11.2 Graph of the semantic space of political and science within the embedding spaces over time. Source: ECCO. 242
11.3 Graph of the semantic space of revolution and science within the embedding spaces over time, 1700–1800. Source: ECCO. 250
12.1 Most similar words to culture, 1720–1960. Source: BPO. 267
12.2 Most similar words to station, 1720–1960. Source: BPO. 270
12.3 Historical-semantic distance matrix for culture, 1720–1900. Source: BPO. 274
12.4 Novelty scores for keywords culture, station, liberal, liberty, 1750–1930. Source: BPO. 277
12.5 Novelty scores for keywords, 1750–1940. Source: BPO. 278
12.6 Number of words whose meanings pivot in a given historical period. Source: BPO. 280

Tables

2.1 The top twelve terms having the highest *dpf* values with despotism at a distance of 50 among all documents in ECCO published from 1780 to 1800 (73,104 documents; 3.8 billion tokens; 19,474 instances of despotism). *page* 33
2.2 The top ten terms having the highest *dpf* values with focal token aristocracy at distances of 10 and 100 among all documents in ECCO published from 1760 to 1800 (118,166 documents; 6.3 billion tokens; 13,532 instances of aristocracy). 42
2.3 The top ten terms having the highest *dpf* values with focal token liberty at distances of 10 and 100 among all documents in ECCO published from 1760 to 1800 (118,166 documents; 6.3 billion tokens; 1,006,661 instances of liberty). 43
3.1 The most frequent co-occurrences (after removing the most common stopwords) with the word liberty in a random subset of 3,000 documents (2,198,379 words). Date spread 1790–5. Source: ECCO. 60
3.2 Co-occurrence frequencies for phlogiston, experiments and gas. 61
3.3 Words with highest PMI score with liberty. Date spread 1790–5, containing 3,000 documents, 2,198,379 words. Source: ECCO. 63
3.4 Words with the highest *log-dpf* (smoothed PMI) score with liberty. Date spread 1790–5, containing 3,000 documents, 2,198,379 words. Source: ECCO. 64
4.1 Raw frequency of liberty, liberty from/to and freedome, freedome from/to. Date spread 1600–40, containing 7,230 documents, 24.9 million tokens. Source: EEBO. 83
4.2 Raw frequency of liberty, liberty from/to and freedom, freedom from/to. Date spread 1600–1800. Source: EEBO and ECCO. 83

4.3 Core terms for liberty and freedom. Date spread 1680–1700. Source: EEBO. 84
4.4 Number of terms on *dpf* list for liberty and freedom, 1600–40, at spans of 5, 10, 50 and 100; and number of terms shared by each *dpf* list; number of terms on *dpf* list for liberty and freedom, 1660–70, at spans of 5, 10, 50 and 100; and number of terms shared by each *dpf* list. Source: EEBO. 85
4.5 Number of terms on *dpf* list for liberty and freedom, 1700–40, at spans of 5, 10, 50 and 100; and number of terms shared by each *dpf* list; number of terms on *dpf* list for liberty and freedom, 1760–1800, at spans of 5, 10, 50 and 100; and number of terms shared by each *dpf* list. Source: ECCO. 86
4.6 Number of shared terms on *dpf* lists for liberty(ie)/benevolence and freedom(e)/benevolence and so on at distance 10 in fifty-year segments. Total number of terms in each list indicated in brackets. Date spread 1600–1800. Source: EEBO and ECCO. 90
4.7 Number of shared terms (n) on *dpf* lists for liberty/libertie and rights at span of 10 and 100; percentage of terms shared between the two lists. Data source: EEBO and ECCO. 92
4.8 Common terms in the persistent bound-lexical company of liberty and republican in each decade from 1750 until 1800. Terms in descending value of *dpf*. Final row indicates the percentage of the terms within the persistent bound-lexical company for liberty represented by the column. Data source: ECCO. 95
5.1 The top twenty-one content words having the highest *dpf* values with focal token government at distance 100 among all documents in ECCO published from 1751 to 1800 (118,166 documents; 6.3 billion tokens). 103
5.2 The top six terms having the highest *dpf* values with focal tokens democracy, aristocracy and monarchy at distance 100 among all documents in ECCO published from 1751 to 1800 (118,166 documents; 6.3 billion tokens). 105
5.3 Pearson correlations and cosine similarities between the full *dpf* lists of despotism and selected terms at distance 100 among all documents in ECCO published from 1720 to 1740 (29,332 documents; 1.6 billion tokens; 55 instances of despotism). 110
5.4 Pearson correlations and cosine similarities between the full *dpf* lists of tyranny and selected terms at distance 100 among all documents in ECCO published from 1720 to 1740 (29,332 documents; 1.6 billion tokens; 26,446 instances of tyranny). 111

5.5 Pearson correlations and cosine similarities between the full *dpf* lists of government and selected terms at distance 100 among all documents in ECCO published from 1720 to 1740 (29,332 documents; 1.6 billion tokens; 244,535 instances of government). 111
5.6 Pearson correlations (*r* values) and cosine similarities between the full *dpf* lists of government and selected terms at distance 100 among all documents in ECCO published from 1750 to 1770 (41,829 documents; 2.2 billion tokens; 346,387 instances of government) and among those published from 1780 to 1800 (73,104 documents; 3.8 billion tokens; 794,588 instances of government). Tokens are sorted by the difference in their Pearson correlation in 1750–70 versus 1780–1800. 112
6.1 Pearson correlation between terms, 1701–1800. *Dpf* calculated at distance 10. Source: ECCO. 131
6.2 Vector subtraction of *dpf* lists for republican, 1701–20, minus 1780–1800. *Dpf* calculated at distance 5. Source: ECCO. 132
6.3 Vector subtraction of *dpf* lists for republican, 1780–1800, minus 1701–20. *Dpf* calculated at distance 5. Source: ECCO. 133
6.4 Common set derived from the top eight terms by Pearson correlation value in each time segment in Table 6.1. Source: ECCO. 134
7.1 Terms in the core for improvement, 1720–40, derived from lexis common to all *dpf* lists at all distances, 10–100. Terms in decreasing rank order of *dpf*. Source: ECCO. 147
7.2 Terms in the core for improvement, 1740–60, derived from lexis common to all *dpf* lists at all distances, 10–100. Terms in decreasing rank order of *dpf*. Source: ECCO. 148
7.3 Terms in the core for improvement, 1760–80, derived from lexis common to all *dpf* lists at all distances, 10–100. Terms in decreasing rank order of *dpf*. Source: ECCO. 149
7.4 Terms appearing in more than one core list for improvement in each twenty-year segment, 1700–1800. Cores derived from lexis common to all *dpf* lists at all distances, 10–100. Terms in decreasing rank order of *dpf*. Source: ECCO. 150
7.5 The top twelve terms having the highest *dpf* values with growth at a distance of 10 generated by a vector subtraction of *dpf* list 1780–1800 minus *dpf* list 1701–20. Source: ECCO. 154
8.1 Nouns following commercial in *The Theory of Moral Sentiments* and *The Wealth of Nations*. 169
8.2 Instances of phrase, 1750–1800. Source: ECCO. 174

9.1 The top ten terms having the highest *dpf* values for stimulus and nervous at a distance of 10 among all documents from 1780 to 1800. Total number of documents: 73,104. Source: ECCO. 188
9.2 The top ten terms on *dpf* lists for fibre and sensibility at a distance of 10 among all documents from 1780 to 1800. Total number of documents: 73,104. Source: ECCO. 189
9.3 Vector subtraction: *dpf* profiles of irritability at distance 10, 1751–1800, minus *dpf* profiles of irritability at distance 10, 1700–51. 202
10.1 Top twenty terms after vector subtraction. Source: ECCO. 222
11.1 Binding and opposing terms of political and revolution at five-year intervals between 1700 and 1795. Source: ECCO. 235
11.2 Binding and opposing terms of political and science at five-year intervals between 1700 and 1795. Source: ECCO. 244
11.3 Binding and opposing terms of science and revolution at five-year intervals between 1700 and 1795. Source: ECCO. 252

Contributors

MARK ALGEE-HEWITT is an associate professor of Digital Humanities and English at Stanford University.

PETER DE BOLLA is professor of Cultural History and Aesthetics at the University of Cambridge.

RYAN HEUSER is a research software engineer in Princeton's Center for Digital Humanities.

EWAN JONES is an associate professor for English at the University of Cambridge.

PAUL NULTY is a lecturer in the Department of Computer Science and Information Systems at Birkbeck, University of London.

GABRIEL RECCHIA is a cognitive scientist at Modulo Research Ltd, Cambridge.

JOHN REGAN is a lecturer in Literature and the Digital and the Creative Industries at Royal Holloway, University of London.

NATALIE ROXBURGH is a lecturer in English at the University of Hamburg.

CLAIRE WILKINSON is an assistant professor in Eighteenth Century English at Robinson College in the University of Cambridge.

Acknowledgements

Academic work in the Digital Humanities is fundamentally collaborative, and this in part contributes to its potential game-changing power. Not all collaborations, of course, are happy, but in this case it gives me great pleasure to acknowledge the individuals who have in the main been co-creators of the work here presented. The deep origins of the project lie in a Mellon Foundation seminar/workshop that I convened with Cliff Siskin at Cambridge, where we were hosted by the Centre for Research in Social Sciences and Humanities in 2012. We called that workshop 'The Experimental Concept Lab' and I thank all the participants in those meetings for their creative input and their willingness to withhold scepticism, at least for a while. Two of the members of that group, Ewan Jones and John Regan, would follow me into the next phase of the project. This was facilitated by the generous grant from the Foundation for the Future to Cambridge University for establishing work in the Digital Humanities. I was lucky enough to win a substantial portion of the grant, which enabled me to found the Cambridge Concept Lab, which I directed from 2014 to 2018. I thank the Foundation and its director, Andrew Thompson for his support and belief in the potential transformative power of collaborative work using computational methods in the humanities. As I also thank the then Director of CRASSH, Simon Goldhill, who also provided invaluable support and encouragement.

The Cambridge Concept Lab comprised four then post-docs, with myself as Director, and our project was to ascertain if it might be possible to discern what we thought of as 'digital signatures' for concepts. All of our work was intensely collaborative, and all four of my collaborators were equal partners in our endeavours. I thank them here – Ewan Jones, John Regan, Paul Nulty and Gabriel Recchia – for the amazingly generous spirit of their contributions. In many ways the four years of the project were some of the most intense, creative and productive moments of my research career. They were also (for the most part) fun.

A parallel Lab operated in the United State, under the direction of Cliff Siskin, and although it was not fortunate, as we in Cambridge were, to be

funded by an external grant, and therefore operated on a very different timetable supported only by good-will and interest in the project, it nevertheless provided a different axis to the work we did. I thank the participants in that Lab: Mark Algee-Hewitt, Bill Blake, Ryan Heuser, Yohei Igarashi, and Bill Warner, as I also thank its Director, Cliff Siskin, for all of their contributions to what was fundamentally a speculative project.

The Cambridge and American labs eventually took slightly different directions, but not before we had jointly presented our work on two occasions on the West Coast of the United States. The Stanford Literary Lab was our host for two days in May, 2016 and I thank the members of that Lab, and its then Director, Franco Moretti, for the invitation and interrogation we were pleased to be subjected to. We then moved south to the University of Santa Barbara, where Alan Liu hosted us, and I also thank him for his incredibly generous support for the team and the project. Bill Warner was also extremely generous while we were in Santa Barbara, and I thank him on behalf of the entire team.

As the funding for the project began to reach its terminus the Lab's thoughts turned to issues of legacy. The most pressing task in this respect was the establishment of a digital repository for the tools we had created. We were fortunate to have locally in Cambridge some very forward-looking and helpful colleagues in the University Library, and in particular I thank Huw Jones for his practical and organisational skills. The Lab's tools are now available through the Library portal at https://concept-lab.lib.cam.ac.uk/. The second issue we turned to was the development of a new area of inquiry, represented by the chapters in this book, which we thought of as the Digital History of Ideas, and we began to collect our thoughts and resources so as to bring this notion to the wider academic community. Some time later, following some early published papers, I proposed to my erstwhile collaborators that we might attempt to launch a new initiative in the history of ideas, one which explored the affordances of computational methodologies. The book you have in your hands is the result. I am pleased to note that my initial inquiry with a publisher, Cambridge University Press, and the editor of its Intellectual History list, Liz Friend-Smith was a very productive one. Liz immediately saw the potential for the project, as did the two anonymous reviewers, whom I also thank. I am also very happy to thank the production team working for the Press, Geethanjali Rangaraj and Alexander Mcleod, who oversaw the final stages of publication with extraordinary care, great skill and punctilious observation of the schedule.

My final thanks are both more conceptual and personal. Without the willingness to suspend ego, disbelief, aggressive investments in personal

advancement, truly collaborative work cannot even get started. All of the contributors to this volume left these undesirables at the door and it has not only been a pleasure working with them, it has also been conceptually formative – perhaps, dare one say, the instantiation of the notion that concepts are, to the very core, emergent forms.

Cavalaire-sur-Mer, July, 2023