ENTERPRISE CONTENT AND SEARCH MANAGEMENT

FOR BUILDING DIGITAL PLATFORMS

Shailesh Kumar Shivakumar

with website

Enterprise Content and Search Management for Building Digital Platforms

Shailesh Kumar Shivakumar

IEEE
computer
society

IEEE PRESS

WILEY

For general information on our other products and services or for technical support, please contact our Customer Care Department within the United States at (800) 762-2974, outside the United States at (317) 572-3993 or fax (317) 572-4002.

Wiley also publishes its books in a variety of electronic formats. Some content that appears in print may not be available in electronic formats. For more information about Wiley products, visit our web site at www.wiley.com.

Printed in the United States of America.

10 9 8 7 6 5 4 3 2 1

To my parents, Shivakumara Setty V and Anasuya T M,
from whom I borrowed love and strength
To my wife, Chaitra Prabhudeva, and my son, Shishir,
from whom I borrowed time and support
To my in-laws, Prabhudeva T M and Krishnaveni B,
from whom I borrowed help and courage
and
To all my schoolteachers who bestowed lots of love and
knowledge upon me

Preface

Disruption in digital technologies has opened up an entirely new realm of possibilities for enterprises. Harvesting new-age digital technologies can redefine the ways business is done online and can potentially give numerous possibilities to reengage with stakeholders such as consumers, partners, resellers, and others. Digital technologies enable enterprises to provide on-demand, customer-centric, personalized, contextual, and meaningful content from anywhere, anytime, on any device. Digital-enabled business models reshape customer experiences and form the key differentiators. As a result, the digital user will be meaningfully engaged bringing in productivity, loyalty, and long-term relationship. On the B2B front, digital technologies have also opened up new realms of possibilities through process optimizations, enterprise integrations, and other developments, and a digital technology ecosystem has reshaped infrastructure and operations sides of things through hardware consolidation and cloud enablement.

Digital technologies are disrupting most of the business domains, technology ecosystems, and business processes. Due to its wide range of benefits and long-term strategic impact and competitive advantages, enterprises across domains are embracing digital revolution. In today's hyper-connected world, word-of-mouth promotion is given preference over sponsored ads, Facebook "likes" count more than expert rating, and enterprises strive hard to convert a visitor into a brand advocate using digital technologies.

DRIVERS AND MOTIVATIONS FOR THE BOOK

Modern enterprises face multiple challenges in building a robust enterprise digital ecosystem. The challenges are multifold in nature and consists, among other things, of internal challenges concerning employee productivity, process optimization, information management, content management, and big data management. Coupled with these are external challenges such as, among others, Omni-channel customer engagement, social and collaboration integrations, personalized presentation, and competitive pressures.

Based on our experience, the most effective way to address these challenges is to provide a *robust information management system* consisting *of seamless relevant information discovery*. This book tries to address these two fronts by exploring various concepts in digital content management (for information management) and enterprise search (for efficient information discovery). This book takes a differentiated view

through a combined focus on content management and enterprise search. During the process it aims to help organization build robust digital platforms using proven best practices, practical models, and time-tested techniques discussed in the book.

We can map the technology topics (CMS and Enterprise Search) discussed in this book with the modern digital platform's capability as shown in the following diagram:
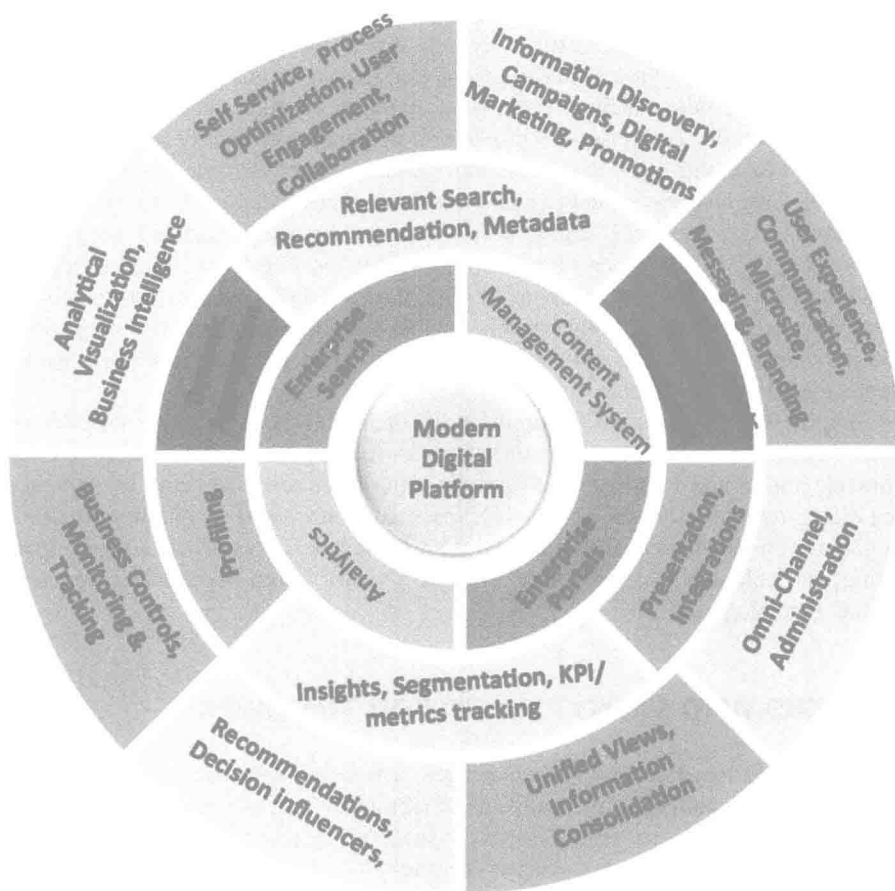
The first layer depicts the core digital technologies, namely ontent management system (CMS), enterprise search, portals, and analytics. The second layer maps the technical capabilities offered by corresponding technologies. CMS provides robust content management, workflows, documents, and asset management whereas search provides relevant search and recommendations. Both CMS and search would enable metadata-tagging capabilities. The outermost layer depicts the business capabilities

enabled by corresponding technology capabilities. Content management enables intuitive user experience, communication, messaging, branding, and micro-site. Search and CMS combined enables promotions management, campaigns, and marketing and relevant information discovery capabilities.

The diagram depicts the importance of role played by CMS and enterprise search in a digital scenario. CMS and search form the information management backbone for a digital enterprise. The book tries to cover the capabilities discussed under CMS and search umbrella and relevant analytics capabilities wherever applicable.

Key Differentiators of the Book

The key differentiators and novel aspects of this book are summarized in the following list:

- **Wide coverage of modern methodologies and techniques:** We have covered emerging technologies such as micro services architecture in content management, CMS-based customer experience platform (CXP), Big Data search, semantic search, Omni-channel content enablement, JCR and CMIS standards, content analytics, SEO, and KPIs. We have detailed trends in CMS and enterprise search we have noticed and have provided good coverage of emerging trends. CMS is explored from security, infrastructure, and performance viewpoint as well.

- **Content frameworks:** The book covers many practically proven models and techniques related to CMS evaluation framework, content migration framework, search evaluation framework, and other aspects that can be used in real-world digital engagements. Comprehensive CMS, search, and DAM evaluation templates are given in Appendixes C, D, and F, respectively.

- **Elaborate content strategy discussion:** As content strategy forms the core of content management, we engage in an in-depth discussion of it in Chapter 2 along with a detailed case study. All chapters in Parts I and II are organized to realize various elements of content strategy discussed in Chapter 2. We have also provided a content strategy template in Appendix A to complement the concepts discussed in Chapter 2.

- **Case-study-based approach:** All core topics (such as templates, workflows, content security, performance, metadata, document management, content migration, and such) have detailed in-context case studies to provide the practical flavor to the topic discussion. The Online Wiley book support material section provides content case studies to explain the best practices used in real-world engagement. Case studies are used as tools to reinforce the theory concepts and provide practical applicability for them. Online support material

also has an elaborate end-to-end digital program case study covering CMS and enterprise search for a digital e-commerce platform.

- **Sample code and configuration:** We have provided sample code while discussing JCR migration concepts to elaborate the concept in Chapter 9. We have also given the configurations that can be used to address security vulnerabilities and optimize content performance in Chapters 11 and 12, respectively.

- **Reference architectures:** The book provides reference architecture for various CMS and search-based applications. Reference architecture of CMS-based customer experience platform, knowledge management system, digital marketing platform, and e-commerce platform are elaborated.

- **Proven best practices and checklists:** We have provided elaborate practically proven best practices while discussing key topics (such as content services, content security, templates, etc.). We also provided content management checklist in Appendix B section. Architects and managers for content and search engagements could use this.

- **Content integrations:** We have dedicated Chapter 7 to integrations with CMS providing details about optimal integration techniques with CMS.

- **Synergies between enterprise content and search:** This book tries to explore the synergies between enterprise content and search systems to build a robust digital platform. Metadata, taxonomy, SEO, analytics, and digital program management are explored from this dimension.

- **Practically proven models and best practices:** We discuss various models and best practices related to content such as template design, workflow design, and Omni-channel content design that are successfully employed in various practical engagements.

- **Architecture concepts:** There is an in-depth coverage of various architecture concepts for content management and digital search. Practitioners can use this as reference architecture in digital programs.

- **Reusable templates:** We have provided CMS evaluation template, search product evaluation, and content strategy template in the appendix sections. Readers can use it for content programs.

MAIN THEMES AND FOCUS AREAS

Main themes and focus areas of this book are:

- **Digital content management and enterprise search:** The primary focus areas of this book are digital content management (primarily Web content and digital assets through Web content management [WCM] concepts) and enterprise search. Wherever necessary, the book also elaborates other supporting

systems/components such as digital asset management (DAM) systems, document management system, workflow management, and Web analytics, among others.

- **Technology and product agnostic view:** The concepts, methodologies, techniques, and best practices discussed in the book are product and technology agnostic. Wherever necessary, concrete examples are drawn from specific technologies and products to explain the concept.

- **Open source frameworks:** Many of the concrete examples are drawn from open source products. Some reference architectures are also developed using open-source components. The intention is to help readers leverage open-source technologies while creating digital systems.

- **Proven practical methodologies and best practices:** We have elaborated many proven models and best practices in areas such as content migration, CMS evaluation framework, content performance, content security, and such. This would help the content and search practitioners apply these frameworks and techniques.

- **Challenges and best practices:** While discussing core portal technologies such as integrations, content management, search, and others, we have discussed the commonly encountered challenges/pitfalls and the best practices.

CHAPTER ORGANIZATION AND TARGET AUDIENCE

The book is organized in three parts with 14 chapters. The online Wiley book support material section provides various supporting material such as content case study and end-to-end digital case study. Part I consists of six chapters that introduce reader to core concepts of content management. We look at content strategy, CMS basics, CMS architecture, templates and workflow, information architecture, taxonomy, and content metadata. Part II includes six chapters and extends the content management concepts and elaborates on topics related to integration, content standards, DAM and document management, content migration, CMS evaluation, content validation, content analytics, content security, content performance. Part III consists of two chapters and is mainly dedicated to discussing basics of enterprise search and advanced search.

We have provided six appendix sections: Appendix A provides a content strategy template, Appendix B provides a checklist for various content management activities, Appendix C is a CMS product evaluation template, Appendix D is the enterprise search product evaluation template, Appendix E provides sample Java code for adding a JCR node, and Appendix F provides an evaluation template for DAM platforms.

The following is the high-level summary of various chapters along with intended target audience:

Chapter	Main topics	Target Audience
Part I: Content Management Basics For Digital Platforms		
Chapter 1: Introduction to Digital Platforms	Enterprise digital ecosystem, enterprise content management concepts, digital strategy, content strategy, digital content management overview, enterprise search overview	Digital architects, enterprise architects, program managers, business analysts, and senior business executive
Chapter 2: Content Strategy	Content strategy overview, strategy challenges, strategy essentials, content characteristics, requirements elaboration, content strategy definition process, content strategy phases, content strategy elements, content strategy case study	Content architects, content strategists, CMS developers, and enterprise architects
Chapter 3: Basics of Content Management System	CMS drivers, CMS design principles, CMS attributes, CMS capabilities, Discussion of CMS systems (WordPress, Drupal, Joomla), and Apache Jackrabbit	Enterprise architects, content architects, and CMS developers
Chapter 4: Content Management System Architecture	CMS design and architecture, CMS architecture patterns, CMS value articulation framework, CMS solution component design, Multi-site design, content folder design, content URL design, localization design, CMS infrastructure design, content strategy realization in CMS, CMS reference architecture, customer experience platform design, knowledge management system design, and content marketing platform design	CMS architects, CMS developers and enterprise architects
Chapter 5: Templates and Workflows	CMS template design, authoring and presentation templates design and usage, template-user interface, template development case study, workflow design, workflow optimization, workflow case study	Content architects, content authors, and CMS developers
Chapter 6: Content Information Architecture, Taxonomy, and Metadata	Designing intuitive information architecture (IA), elements and goals of IA, taxonomy concepts, taxonomy and metadata, metadata types, metadata standards (Dublin Core and SKOS), metadata case study	Enterprise architects, content architects, content strategists, and information architects

Chapter	Main topics	Target Audience
Part II: Advanced Content Management		
Chapter 7: Content Integration and Content Standards	CMS integrations with security systems, TMS, search engine, portals, presentation engines, metadata systems, feeds, DAM, CMIS integrations. Content standards: HTML/XML, DITA, JSON, SCORM, Web service formats (REST and SOAP)	Content architects, Integration architects, CMS administrators, CMS developers, and enterprise architects
Chapter 8: Digital Asset Management and Document Management	DAM objects, architecting DAM system, DAM challenges, document management system capabilities, document management evolution and road map, document management case study	Content architects, CMS developers, content authors
Chapter 9: Content Migration	Migration drivers, migration design considerations, migration challenges, migration checklist, migration governance, migration automation, migration case study	Enterprise architects, CMS architects, CMS developers and program managers
Chapter 10: Content Maintenance – Validation, Analytics, KPI, SEO, and Evaluation	Content validation types, validation checklist, content analytics, content KPIs, content analytics design, content analytics case study, content SEO strategy, content SEO best practice, CMS evaluation framework	CMS architects, CMS developers, CMS QA team and program managers
Chapter 11: Content Security	Content security vulnerabilities, XSS, CSRF, denial of service, clickjacking, generic content security scenarios, SSO, penetration testing, security best practices, security testing case study	Content architects, security architects, enterprise architects
Chapter 12: Content Infrastructure and Performance Optimization	Content performance optimization, CMS-level performance optimization, infrastructure-level performance optimization, content performance KPIs, content performance testing	Content architects, enterprise architects, performance engineers, CMS developers

Chapter	Main topics	Target Audience
Part III: Enterprise Search Technologies		
Chapter 13: Introduction to Enterprise Search	Enterprise search drivers, search overview, search trends, search evolution, search capabilities, basic search features, advanced search features, Apache Solr and ElasticSearch features	Enterprise search architects, information architects, enterprise architects, and search developers
Chapter 14: Advanced Enterprise Search	Federated search, features, architecture and challenges of federated search, relevancy rank adjustment, personalized search, semantic search, semantic search process, people search, Big Data search	Search architects, enterprise architects, program managers, and search developers

DECLARATION

- Utmost care has been taken to ensure the accuracy and uniqueness of the book content. Any inaccuracies or inconsistencies are entirely my own. If you think any corrections are needed, or for any other feedback, please write to Shailesh.shivakumar@gmail.com

- In a few chapters I have used the features of popular and open-source WCM products to explain the concepts. The explanation is for educational purposes only and should not be considered as a product or technology recommendation or evaluation. The CMS plugins and modules used to illustrate examples and concepts are also for educational purposes only; they should not be interpreted as recommendations or evaluations. Comprehensive evaluation template and framework are provided in the appendix section.

- All open-source tools mentioned are in public domain as open source at the time of writing of this book.

- I acknowledge the trademarks of all products, technologies, and frameworks being used in this book.
  - WordPress, Joomla, Drupal, and dotCMS are registered trademarks and are the legal property of their respective owners.
  - Documentum is a registered trademark of EMC Corporation.
  - Oracle, Oracle Access Manager, WebCenter, WebLogic, OHS, and Java are registered trademarks of Oracle and/or its affiliates.
  - Synaptica is the registered trademark of Synaptica, LLC.
  - SiteMinder is the registered trademark of CA Technologies.
  - WebSphere, Tivoli Access Manager, IBM WCM, IHS, and DB2 are registered trademarks of IBM and/or its affiliates.
  - AEM and CQ5 are registered trademarks of Adobe and/or its affiliates.

- FAST, SharePoint, and SQLServer are registered trademarks of Microsoft and/or its affiliates.
- Liferay is a registered trademark of Liferay and/or its affiliates.
- Intel is a registered trademark of Intel Corporation.
- All other trademarks or registered trademarks are the legal property of their respective owners.

Acknowledgments

I am blessed to be surrounded by knowledgeable colleagues, subject matter experts, and friends who are happy and willing to help. I would like to acknowledge them and show my gratitude for their immense help, incredible support, and cooperation.

I would like to convey my sincere and heartfelt thanks to Elangovan Ramalingam, Arun Sugumar, Ashwin Raju, Verma V.S.S.R.K, Subramanian Narayanan, Ajay Kumar Sharma, Shikha Mahajan, Aanchal Sikka, and Nagarajan Kuppuswamy for their valuable inputs and review comments. I feel blessed in the company of these gifted colleagues.

I would also like to convey my sincere thanks to managers Shankar Bhat, Rahul Krishan who encouraged and supported me in all my initiatives.

My sincere and heartfelt thanks to Professor Dr. Viraj Kumar for his moral support and patience. He has been a continuous inspiration to me.

I really appreciate the efforts and support given by Rahul Krishan and Sreenivasa Kashyap at Infosys for ensuring timely reviews and approvals of this book.

I would also like to recognize and thank Dr. P. V. Suresh for his constant encouragement and immense support.

My special thanks to Mary Hatcher, Melissa Yanuzzi, Brady Chin, Allison McGinniss, Alex Castro, and the editors, designers, and publishing team at the IEEE and Wiley for providing all necessary and timely support in terms of review, guidance, and regular follow-ups. The team at Wiley took special care in design to make this book look beautiful.

About the Author

Shailesh Kumar Shivakumar is a Senior Technology Architect at Infosys Technologies Limited with over 15 years of industry experience. His areas of expertise include digital technologies, software engineering, Java enterprise technologies, performance engineering, and digital program management. He is a Guinness world record holder of participation for successfully developing a mobile application in coding marathon. He has four patent applications including two US patent applications in the area of Web and social technologies.

He was involved in multiple large-scale and complex digital transformation programs for Fortune 500 clients of his organization. He also provided on-demand consultancy in performance engineering for critical projects across various units in the organization. His has hands-on experience on breadth of technologies including Web technologies, digital technologies, and database technologies and has worked on multiple domain areas such as retail, manufacturing, e-commerce, and avionics, among others. He was the chief architect of an online platform that won "Best Web Support Site" award among global competitors.

He is the author of two technical books: *Architecting High Performing, Scalable and Available Enterprise Web Applications* and *A Complete Guide to Portals and User Experience Platforms*. He is a regular blogger at Infosys Thought Floor, and many of his technical white papers are published in Infosys external site. He has delivered two talks at Oracle JavaOne 2013 conference on performance optimization and project management and has presented a paper at IEEE conference. He also headed a center-of-excellence for digital practice. He led multiple thought-leadership and productivity improvement initiatives and was part of special interest groups (SIG) related to emerging Web technologies at his organization.

He holds numerous professional certifications including TOGAF 9 certification, Oracle Certified Master (OCM) Java Enterprise Edition 5, Sun Certified Java Programmer, Sun Certified Business Component Developer, IBM Certified Solution Architect – Cloud Computing, IBM Certified Solution Developer – IBM WebSphere Portal 6.1, and many others.

He has won numerous awards including prestigious Infosys Awards for Excellence 2013–14 "Multi-talented thought leader" under "Innovation – Thought leadership" category, "Brand ambassador award" for MFG unit, "Best employee award," delivery excellency award, and multiple spot awards and received honor from executive vice chairman of his organization. He is featured as "Infy star" in Infosys Hall of fame and recently led a delivery team that won the "best project team" award at his organization.

He holds an engineering degree in computer science and has done executive management program from Indian Institute of Management, Calcutta. He lives in Bangalore, India.

About the Companion Website

This book is accompanied by a companion website:

www.wiley.com/go/shivakumar/enterprisecontent

The website includes:

- Additional Book Chapters
  - Chapter 15 – Content Management Case Studies
  - Chapter 16 – Digital Case Study: Building a Modern Digital E-commerce Platform Using CMS and Search

- Appendices A-F
  - A – Content Strategy Template
  - B – Content Management Checklist
  - C – CMS Evaluation Template
  - D – Enterprise Search Evaluation Template
  - E – Java Code for Adding an Image Node to Jackrabbit Workspace
  - F – DAM Evaluation Template