Trustworthy Policies for Distributed Repositories

Synthesis Lectures on Information Concepts, Retrieval, and Services

Gary Marchionini, *University of North Carolina, 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.

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

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

Dynamic Information Retrieval Modeling No Access
Grace Hui Yang, Marc Sloan, Jun Wang

Learning from Multiple Social Networks No Access
Liqiang Nie, Xuemeng Song, Tat-Seng Chua

Scholarly Collaboration on the Academic Social Web No Access
Daqing He, Wei Jeng

Scalability Challenges in Web Search Engines No Access
B. Barla Cambazoglu, Ricardo Baeza-Yates

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

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

Information Communication
Feicheng Ma

Social Media and Library Services
Lorri Mon

Analysis and Visualization of Citation Networks
Dangzhi Zhao and Andreas Strotmann

The Taxobook: Applications, Implementation, and Integration in Search: Part 3
Marjorie M.K. Hlava

The Taxobook: Principles and Practices of Building Taxonomies: Part 2
Marjorie M.K. Hlava

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

The Taxobook: History, Theories, and Concepts of Knowledge Organization: Part 1
Marjorie M.K. Hlava

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

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

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

Information and Human Values
Kenneth R. Fleischmann

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

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

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

Information Retrieval Models: Foundations and Relationships
Thomas Roelleke

Key Issues Regarding Digital Libraries: Evaluation and Integration
Rao Shen, Marcos Andre Goncalves, and Edward A. Fox

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

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

The Answer Machine
Susan E. Feldman

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

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

Search User Interface Design
Max L. Wilson

Information Retrieval Evaluation
Donna Harman

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

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

Information Concepts: From Books to Cyberspace Identities
Gary Marchionini

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

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

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

Multimedia Information Retrieval
Stefan Rüger

Online Multiplayer Games
William Sims Bainbridge

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

Reading and Writing the Electronic Book
Catherine C. Marshall

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

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

XML Retrieval
Mounia Lalmas

Faceted Search
Daniel Tunkelang

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

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

New Concepts in Digital Reference
R. David Lankes

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

© Springer Nature Switzerland AG 2022

Reprint of original edition © Morgan & Claypool 2017

All rights reserved. No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopy, recording, or any other except for brief quotations in printed reviews, without the prior permission of the publisher.

Trustworthy Policies for Distributed Repositories

Reagan W. Moore, Hao Xu, Mike Conway, Arcot Rajasekar, Jon Crabtree, and Helen Tibbo

ISBN: 978-3-031-01175-7 print
ISBN: 978-3-031-02303-3 ebook

DOI 10.1007/978-3-031-02303-3

A Publication in the Springer series
SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #51

Series Editor: Gary Marchionini, University of North Carolina, Chapel Hill

Series ISSN 1947-945X Print 1947-9468 Electronic

Trustworthy Policies for Distributed Repositories

Reagan W. Moore, Hao Xu, Mike Conway, Arcot Rajasekar, Jon Crabtree, and Helen Tibbo

DataNet Federation Consortium, University of North Carolina at Chapel Hill

SYNTHESIS LECTURES ON INFORMATION CONCEPTS, RETRIEVAL, AND SERVICES #51

ABSTRACT

A trustworthy repository provides assurance in the form of management documents, event logs, and audit trails that digital objects are being managed correctly. The assurance includes plans for the sustainability of the repository, the accession of digital records, the management of technology evolution, and the mitigation of the risk of data loss. A detailed assessment is provided by the ISO-16363:2012 standard, "Space data and information transfer systems—Audit and certification of trustworthy digital repositories." This book examines whether the ISO specification for trustworthiness can be enforced by computer actionable policies. An implementation of the policies is provided and the policies are sorted into categories for procedures to manage externally generated documents, specify repository parameters, specify preservation metadata attributes, specify audit mechanisms for all preservation actions, specify control of preservation operations, and control preservation properties as technology evolves. An application of the resulting procedures is made to enforce trustworthiness within National Science Foundation data management plans.

KEYWORDS

trustworthiness, long-term preservation, digital repositories, data management, policy, digital preservation, ISO 16363

Contents

Preface . . . . . **xiii**

Acknowledgments . . . . . **xv**

1 Introduction **1**
1.1 Motivating Implementation Plan for a Trustworthy Repository 1
1.2 NSF Data Management Plan Example 4
1.3 Generic Infrastructure for Managing Technology Obsolescence 6

2 Trustworthy Repository Description **9**
2.1 Global Variables 9
2.2 Metadata Attributes 10
2.3 Documents 14
2.4 Policies 19
2.5 Policy Functions 20

3 ISO 16363 Organizational Infrastructure **27**
3.1 Governance and Organizational Viability 27
3.2 Organizational Structure and Staffing 31
3.3 Procedural Accountability and Preservation 34
3.4 Financial Sustainability 43
3.5 Contracts, Licenses, and Liabilities 45

4 ISO 16363 Digital Object Management **51**
4.1 Ingest: Acquisition of Content 51
4.2 Ingest: Creation of the AIP 62
4.3 Preservation Planning 85
4.4 AIP Preservation 89
4.5 Information Management 95
4.6 Access Management 98

5 ISO 16363 Infrastructure and Security Risk Management **101**
5.1 Technical Infrastructure Risk Management 101
5.2 Security Risk Management 116

6 Trustworthy Repository Implementation 119
7 Summary 127
References 129
Author Biographies 131

Preface

The definition of repository trustworthiness has been under development for more than 10 years. The original specification was created in 2005 by the National Archives and Records Administration (NARA) based on the tasks required for an Electronic Records Archive (ERA) [19]. The list included 852 tasks, of which more than 200 tasks could be automated using policy-based data management systems. In collaboration with the Research Libraries Group, the RLG/NARA Trusted Digital Repository (TDR) criteria [12, 18, 20] were then defined, which reduced the number of necessary computer actionable tasks to about 175. The list was further refined to create the TRAC standard, the Trusted Repository Audit Checklist [21], which included about 120 computer actionable tasks. The tasks were designed to manage systems that complied with the Open Archival Information Systems standard [16].

The repository audit and certification working group [1] for the Mission Operations and Information Management Services (MOIMS) of the Consultative Committee for Space Data (ISO MOIMS-rac) provided additional refinements to the trustworthiness assessment criteria, culminating in the ISO-16363:2012 standard [2, 3, 17]. Trustworthiness criteria are defined for the categories of Organizational Infrastructure, Digital Object Management, and Technical Infrastructure and Risk Management. For each category, explicit tasks are specified that need to be enforced for a repository to be trustworthy. A total of 113 tasks are specified by the ISO standard. A research question is how many of these tasks can be automatically enforced using computer actionable policies.

In this book we analyze the tasks in the ISO 16363 standard. For each category and each task, a set of policies is defined. Computer actionable rules are proposed for automating each task, persistent state attributes are defined that are needed to track the outcome of each task, and example reports are presented for implementing a trustworthy repository. The policies are available on http://github.com under the repository DICE-UNC / policy-workbook / ISO / Rules. This paper defines a minimal set of policies, state information, and operations that are needed to automate tasks required for managing a trustworthy repository.

Acknowledgments

The development of the iRODS data grid was funded by the NSF OCI-1032732 grant, "SDCI Data Improvement: Improvement and Sustainability of iRODS Data Grid Software for Multi-Disciplinary Community Driven Application," (2010–2013). The results presented in this paper were funded by the NSF Cooperative Agreement OCI-094084, "DataNet Federation Consortium," (2011–2016). The iRODS Consortium developed the pluggable architecture used by the DataNet Federation Consortium. The software systems described in this paper are available as open source software on http://github.com.