

INFORMATION MANAGEMENT IN THE SERVICE OF
KNOWLEDGE AND DISCOVERY

A Dissertation

Presented to the Faculty of the Graduate School
of Cornell University

In Partial Fulfillment of the Requirements for the Degree of
Doctor of Philosophy

by

Lori Ann Lorigo

January 2006

© 2006 Lori Ann Lorigo

INFORMATION MANAGEMENT IN THE SERVICE OF KNOWLEDGE AND DISCOVERY

Lori Ann Lorigo, Ph. D.

Cornell University 2006

Information networks are prevalent in society. An understanding of the properties of such networks including their structure and inherent relationships can serve to assist knowledge and discovery.

In this work we examine ways in which information networks assist in creating and organizing knowledge. We first describe an environment for mathematical knowledge management. This open information system impacts the formal methods community by making data management central and allows for collaborative creation and sharing of mathematics. The architecture of that system lays the foundation for a prototype Formal Digital Library (FDL), which serves to aggregate mathematical resources and provide data-rich services across them. The foremost importance of the FDL in this dissertation, however is that it serves as a laboratory for examining information networks in general. Because of the rich structure inherent in formal mathematics, it is a well-suited domain for testing and evaluating network analysis techniques and their roles in knowledge discovery. We examine the structural properties of the FDL's contents and also the recent definitive proof of the four color theorem by Gonthier. Our analysis reveals a characteristic depth and breadth and uses Kleinberg's HITS algorithm to reveal (mathematical) hubs and

authorities. To show generality, we also examine non-mathematical and dynamic networks. In particular, we build institute and country based collaboration networks from over 200,000 scholarly publications in the physics community to model long distance collaboration trends over 30 years. The findings demonstrate the influence of graph-theoretic metrics and visualizations on discovery.

Finally, we expand our notion of links in a network and describe concepts and methods for linking data to published articles to support quality and authority. We describe our experimentation with a new authoring mechanism that incorporates data provenance to provide evidence for claims made in articles. Because math in articles is something that can be ambiguous and sensitive to errors, we see the mathematics domain as an ideal candidate for this work. However, our concepts also generalize to other domains where the accurateness and archival of data referred to in an article is of importance.