Information Dynamics - An Information-centric Approach to Digital Library Interoperability[*]

Ronald L. Larsen
University of Maryland
College Park, Maryland 20742, USA
rlarsen@deans.umd.edu

KEYWORDS: digital library, interoperability**,** system design, integration, analysis

ABSTRACT

Acquisition, organization, management, retrieval, and distribution of information are fundamental purposes of digital libraries and their supporting infrastructures. Interoperable digital libraries pose particularly difficult system design issues. Interoperability research has focused largely on syntactic and semantic interoperability. In this paper, a third form of interoperability, *analytic interoperability* is proposed, with a framework in which to consider it. Since *information* is the essential commodity of interest, a comprehensive interoperability design should take into account the fundamental properties of information, including representation, composition, relationships, and dynamics. *Information Dynamics* considers how the nature of information can be used to achieve analytic interoperability.

INTRODUCTION

The growth of networked information resources, largely through the Internet and the World Wide Web, is both a result and a source of the growing interest in digital libraries. Digital libraries have emerged as the vehicle for organizing collections of digital information, much as traditional libraries have done for print and related media. They are becoming a major component of the global information infrastructure. But little standardization exists among digital libraries, and it can be argued that the international standards process is poorly suited to the rapid pace of technology development that has become familiar on the Web. Alternatively, developers of digital libraries focus more on *interoperability* among heterogeneous, or federated, systems. Andreas Paepcke [i] describes these as "cooperating systems where individual components are designed or operated autonomously." He suggests that "the ultimate goal for such a system is to have components evolve independently, yet to allow all components to call on each other efficiently and conveniently."

The rapid advancement of digital libraries throughout the 90's and the near-daily announcement of new technologies and systems all but assure that these trends will accelerate in the current decade. The hardware technology exists to create ever more complex networks of systems. However, our ability to design, implement, operate, maintain and support these increasingly complex systems lags. One reason for this is the paradigm used in system design remains largely *process-centric*, founded on economic principles inherited from many generations past of Moore's Law. Digital libraries provide the need and the opportunity to design around *information-centric* principles.

[*] This work is supported partly by DARPA / Rome Labs, Department of the Air Force, under contract F306020020578 to the Department of Computer Science at the University of Maryland. The views, opinions, and/or findings contained in this report are those of the author and should not be interpreted as representing the official policies, either expressed or implied, of the Department of the Air Force, DARPA, DOD, or the U.S. Government.

Information Dynamics[ii] is an alternate design paradigm that takes such an information-centric perspective. In this approach the role information plays is central; system design considers what information is needed and when, where the information is, and what happens to the information as it moves from one place to another. In an Information Dynamics framework, information is treated as a dynamic entity and its dynamics (e.g., location, timeliness, value) are explicitly considered. Information processing is performed through *actions* carried out as a result of explicit *choices.* Any action that carries out a transformation or related processing of information consumes resources, requiring these resources over some period of time. One can think of this in terms of actions that occupy a subspace in resource/time space. All actions take time and, therefore, have an impact on the dynamics of information. Further, information is considered to have value within a *context* and with respect to the achievement of a *goal*. To this end, the value of information typically changes over time within a given context.

The use of information for effecting time-dependent control and related decision processes is not new. Physical systems respond adaptively to linear-quadratic-Gaussian (LQG) controllers, for example, when the physics is well understood and controllers have rigidly bounded responsibilities. But decision making in network-based, distributed systems for which there is no nice physics poses a different class of problem. Early investigations into distributed decision-making led to a wealth of research in game theory and later team theory, [iii,iv,v,vi] but none of this work explicitly considered the temporal effects on the value of information, and how that affected system performance. Recent literature on autonomous agents[vii,viii] reports work in distributed, multi-agent decision-making (using, for example, state space approaches such as Markov Decision Processes) but still avoids consideration of the temporal issues underlying the usage and value of information. Interoperability research has distinguished between *syntactic* and *semantic* interoperability. While the boundary between them is blurry, syntactic interoperability typically refers to agreement on structural relationships within communication, while semantic interoperability addresses common interpretations of term usage and meaning. But this speaks to *how* information is shared among heterogeneous components. While information is used throughout these approaches, and differently in most of them, techniques that explicitly consider the role of the time/value of information (the *when/what*) and the reason for communicating (the *why*) are lacking. Analytic interoperability addresses these topics by considering intentions, or understanding what the purpose of the interaction is and using this to optimize actions.

In this paper, the dynamic nature of information is considered with regard to decision processes and the resulting implications on the design of interoperable federated digital libraries. For illustrative purposes, information dynamics is first applied to a networking problem.

WHAT IS INFORMATION?

Information is a property, characteristic, or description of something physical, logical, virtual, or conceptual, including other information. It may be a group, an action, a choice, or a relationship between any of these things. Information has value *within a context*. Relationships between information may be direct or indirect, and exist whether they are enumerated or not. Relationships may be static or dynamic. The causality principle applies. Information in the present can only affect the future; it cannot change the past. It may change the interpretation of the past, but it cannot change the past, itself. Further, delays involved in the movement of information assure that knowledge of a remote entity's state is necessarily delayed; the "present" system state as understood by any system entity will, therefore, actually reflect the collective state of system components at past time instant(s).

Every digital library consists of a large number of functional components and a much larger number of digital objects (the contents). Considering all possible pieces of relevant information and their relationships yields an arbitrarily large amount of information. In a typical system design only a small amount of relevant information is collected and used.

When multiple pieces of information are related, their relationship may be considered a higher level of information. Action is required to establish the validity of a relationship. When the *relationship* is explicit, the related information can be derived *implicitly*. In this regard, information that can be derived from other information based on known relationships is implicit information.

Information Has Value

Information has value (or *utility*) within a given context. Context can be considered the domain of the utility function relevant to the task at hand. The value of information may change with time within a given context, and its value may differ in different contexts at the same time instant. Information value may also depend on relationships with other information. Information is represented in *information variables* that include the item of information and associated metadata. A context vector defines the domain within which a relevance computation (e.g., cosine or dot product) can be made in order to establish the importance of the variable to the task at hand. Metadata associated with a variable may also include a measure of confidence (e.g., a probability distribution function) by which the significance of the variable can be assessed. The value, or utility, of the variable is a function of both context and confidence, both of which may also be a function of time. Note that while both context and confidence are formally required to understand the role of an information variable in a system, they are rarely considered explicitly in the design of systems.

Explicit information may be acquired through direct observation, communication, or inference. Only explicit information can be communicated. Implicit information inferred from known relationships by different observers or by communicating entities may still differ, unless they also share a common context and a common understanding of relationships.

Communication requires agreement on representation and protocol (the *how*), but that is only the beginning. More fundamental is the determination of *what* is to be communicated, *when*, and *why;* for every action, including communication, consumes resources and time. Knowledge of what information is required, and where and when it is needed, is a crucial part of federated system design. But for a federated digital library system, it is also the very reason for its existence.

Federated Systems

Consider the implications of information dynamics on a federated system. Such a system has a collection of entities (processing resources) capable of carrying out certain operations. A specific distributed system, designed to carry out a specific mission, uses physical resources to carry out actions and to store and move information. When such a system is interacting with an external physical system it may also have sensors and actuators.

A federated system is a collection of autonomous nodes with an interconnecting infrastructure for communication. Each node maintains a *perceived reality,* based on its prior model of the operating environment and explicit information it receives. Explicit information can be integrated into the model to update the perceived reality. Perceived reality at no node can ever be assured to be identical to *actual reality*, particularly with respect to non-local events. Transmission delays assure that information received from any remote node is, by definition, historic. Further, it is not sufficient to receive messages; they must be interpreted and integrated into the local perceived reality.

Information Dynamics incorporates an information-centric system view that explicitly considers temporal aspects of information, the value of information, and the role of implicit information. Within this framework, a system can capitalize on dynamic system behaviors that would otherwise be liabilities. Decisions are explicitly based on perceived reality, taking into account the environmental model, its dynamics, information sources, and their interactions. The value of information, conditioned on confidence levels and context vectors, plays a key role in system operation.

EXAMPLE: ROUTING

As a concrete example of the application of information dynamics to a practical problem, consider link state routing in a computer network. Shortest path routes from a source to a destination are determined according to the current known state of links. Routing algorithms typically measure their state (e.g., queue length) periodically, estimate the waiting time, and broadcast this information to neighboring nodes, which use it for best-route determination.

Consider the basic characteristics of link performance. In a typical network each link continuously transfers packets, as presented, up to the capacity of the link. Considering a link as a server, its steady state behavior can be characterized in terms of the mean, w, and the variance, v, of the waiting time. Let w(t) be the waiting time at a particular link as measured locally at time t. Assuming that the measurement is done correctly, the variance of this measurement, v(t), is zero. Given no additional information about the state of the link, the estimated waiting time $w(t_1)$ at some later time t_1 will necessarily be based on w(t) and our knowledge of w. This estimate will have a variance, $v(t_1)$, which will be nonzero. In fact, the variance will be an increasing function of the difference t_1-t, tending towards the steady state value v. Given w(t) and v(t), the actual values of $w(t_1)$ and $v(t_1)$ can be estimated with knowledge of the stochastic behavior of the link.

In this example, the basic information variable is the waiting time estimate for the link and the variance estimate is its confidence indicator. Recognizing that communicating w(t) to another node in the network takes time, any new estimates should take into account the dynamics of the situation. Depending on the characteristics of the link, the estimates $w(t_1)$ and $v(t_1)$ may come so close to the steady state values w and v that the new measurements will have negligible impact on the link information retained by another node. As a consequence, communication can be significantly reduced for link-state routing without decreasing the quality of routing decisions by considering the variance in delay estimates. Each node does need steady-state information about the links. Note that if the steady state conditions change regularly, that knowledge can also be shared. By explicitly considering the value of information in this simple algorithm, information dynamics improves the design of the routing scheme.[ix] Early results suggest a savings of at least 25% in routing control information is achievable using information dynamics approaches to link state routing.

It is important to understand that although the above example uses statistical measures, the information dynamics framework is not limited to handling quantitative information. It can be equally effective in using fuzzy or purely qualitative information.

DIGITAL LIBRARY IMPLICATIONS

A digital library has been defined as “the collection of services and information objects that support users in dealing with information objects and their organization and presentation, available directly or indirectly via electronic means.”[x] This is a sufficient working definition; digital libraries allow individuals and organizations to efficiently and effectively identify, assemble, correlate, manipulate, and disseminate information resources, regardless of the medium in which the information exists. Digital libraries provide tools to navigate and manipulate information in a multimedia, multilingual, multidisciplinary world. Task context, user values, and information provenance are critical elements in the information seeking process, but have yet to become part of the digital library infrastructure.

But how might information dynamics concepts be introduced into digital library design? Consider the information retrieval functions of digital libraries. A user formulates a query from a client entity, which sends the query to a search engine operating over some set of repositories. The repositories use the query terms to suggest materials in the local collection that may be responsive to the query, and a ranked list of responses is constructed. Either in middleware or in the client, itself, the responses from the multiple repositories are merged into a ranked list that is presented to the user, who is then responsible for perusing the list in the hope of finding relevant materials. Much effort has gone into developing high performance search engines, typically measured by *precision* and *recall* over a test corpus. But while precision and recall are used to evaluate performance in carefully constructed test scenarios, the results of these evaluations have not typically been used to *control* search engine performance. Consider, for example, the results of the TREC6 Conference shown in Figure 1, which is a typical display of state-of-the-art performance for information retrieval engines. While curves with higher precision and recall are generally superior, when these systems are placed into operation, the user has no control over where on these curves the system will perform for a particular search. Control parameters are set within the system implementation and are totally opaque to the user.

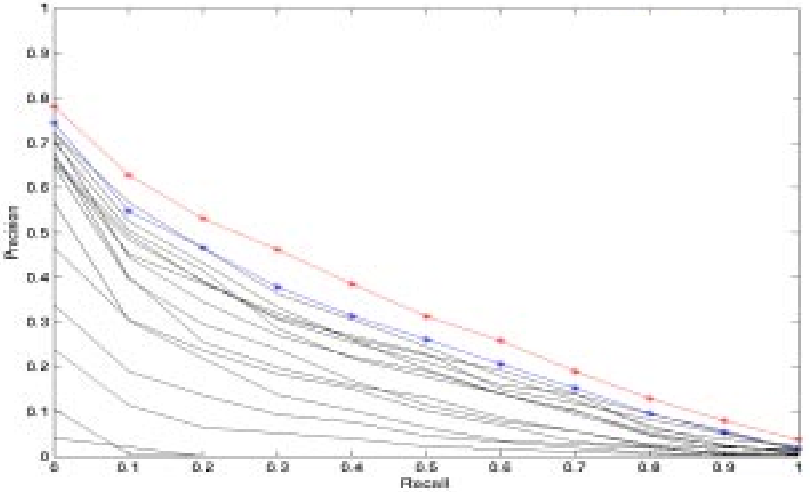

Figure 1. Precision & Recall curves from the TREC6 Conference [xi]

Information dynamics brings to the fore this kind of trade-off. It explicitly recognizes that each operation is unique, but that there is information available to tailor system performance to the specific character of the operation. For example, consider providing in the user’s query a parameter that alludes to the purpose (or goal) of the query. It may be that the user is only casually familiar with the field and needs introductory material. The search would be more effective, perhaps, by applying substantially more weight to precision than to recall. As the nature of the query moves from casual interest through tutorial (instructional), fact-finding (known-item search), research (focused inquiry), to survey (comprehensive exploration), the search engine could deliver more relevant responses by progressively shifting its weighting more toward recall and away from precision.

In like manner, information dynamics suggests strategies to incorporate *context* into queries. Query term ambiguity results in false returns from search engines. If the user incorporates multiple terms into a query, or is encouraged to do so through iterative refinement of a query by responding to irrelevant returns from poorly framed queries, the detrimental effects of term ambiguity can be reduced. But information dynamics suggests an alternate approach, capitalizing on the term disambiguation refinement work of ontologies and *entry vocabularies*.[xii] If the search engine is given sufficient information to limit the domain of search, there is reason to believe that the precision of returned results would improve.

Information retrieval operations in digital libraries are largely state-less transactions, particularly across multiple patrons. In stark contrast to the way reference librarians learn and improve their performance, later users of search engines rarely benefit from earlier searches conducted by those engines. But they could. Just as caching improves communications performance by capitalizing on temporal aspects of information demand, information retrieval can benefit from prior searches of like nature with a similar context. This represents the very early stages of analytic interoperability.

Digital library research has touched on analytic interoperability, without using the term. Several research projects have examined implicit and explicit collaborative techniques for improving an individual's success in information retrieval, for example, by capitalizing on prior search activities conducted by other individuals.[xiii] [xiv] Query languages and tools identify digital library materials across federated collections that are similar to the characteristics expressed in the query. These characteristics focus on the information artifact and are only beginning to consider non-bibliographic attributes to improve the search. Examples include identifying the types of individuals who have been reading specific material, the value they associated with it, and the paths they traversed to find it. These approaches require instrumented digital libraries, in order to build the perceived reality and set the context that will enable improved performance at the user-level. This goes beyond issues of functionally compatibility among federated systems, to a mission-oriented control structure designed to improve both the qualitative and quantitative performance *as perceived directly by the end-user*.

Digital libraries face significant technology challenges. These include real time ingest (capturing, interpreting, cataloging, and indexing high rate multimedia data flows in real time), federating distributed repositories (organizing heterogeneous distributed information sources into comprehensive discipline-oriented, user-accessible repositories), and cross-lingual interaction (automatically accessing and using information across multiple natural languages).[xv] Information dynamics holds the potential to raise the level of interoperability among users and digital libraries from a high dependence on syntax, structure, and word choice to greater exploitation of semantics, context, and concepts, thereby extending information search and filtering beyond purely bibliographic criteria to include contextual criteria related to the task, to the user, and to the time constraints of the user.

Scalability and interoperability are well-known, fundamental requirements for digital libraries. Scalable repository technology must support the federation of thousands of repositories, present to the user a coherent collection of millions of related items, and do this rigorously across many disciplines. Information dynamics holds the potential of addressing these issues with more than brute force bandwidth and capacity. As the size and complexity of information objects increases, so does the bandwidth required to use these objects. Time-critical applications requiring real-time interactivity push the bandwidth requirements even higher. *Broadband interoperability* refers to the dramatic changes in the user's work style that become feasible when the user's inputs are no longer constrained to a few keystrokes or mouse clicks. Information-centric design founded on context, utility (or information value), and temporal relationships offer the potential for real-time adaptation of scalable network and repository services

CONCLUSIONS

Whereas *information* management is the mission of complex systems, *process* management remains the dominant design and implementation approach. Since information is the essential commodity, effective design strategies should explicitly address the fundamental properties of information. The first principle is to recognize the distinction between information and its representation. Computer systems are only capable of manipulating representations and it is through the processing of representations that we attempt to carry out the processing of information. These representations are limited in that they capture only a limited portion of the generic information. Moreover, processing changes the nature of information in ways that are not necessarily intended or anticipated. *Implicit information* must also be understood and elucidated.

The second fundamental principle of Information Dynamics is that information has value in *context*. Processing affects the value of information. Movement, representation, and storage also affect information value. But the ramifications on system design are rarely considered. The third fundamental principle of Information Dynamics is that the value of information changes with time. Understanding the role time plays in the value of information impacts the applicability of information. Communication of information takes time. When the delay caused by communication becomes large, the value of the information may be reduced sufficiently that its communication may not only have been unnecessary, but may, in fact, be detrimental.

The principles of Information Dynamics presented here represent ongoing work to understand the fundamental characteristics of information within a federated system context. The objective is to develop information-centric models of system design and operation. The framework has shown the potential for bringing about a significant advancement in the way information is handled in systems.

[i] Paepcke, A., S. Chang, et al., "Interoperability for Digital Libraries Worldwide," Communications of the ACM, Volume 41, 4, April 1998.

[ii] Larsen, R., A. K. Agrawala, and D. Szajda, "Information Dynamics: An Information-Centric Approach to System Design," International Conference on Virtual Worlds and Simulation, San Diego, CA, January 2000.

[iii] Bacsar, T. and Olsder, G. J., *Dynamic Noncooperative Game Theory*, Academic Press, New York, Mathematics in Science and Engineering, v.160, 1982.

[iv] Greenwald, A., "Modern Game Theory: Deduction vs. Induction," TR 1998-756, New York University, 2/24/98.

[v] Greenwald, A., "Learning to Play Network Games," TR1998-758, New York University, 2/24/98.

[vi] Owen, G., *Game Theory*, Saunders, Philadelphia, PA, 1968.

[vii] Stone, P. and Veloso, M., "Using Decision Tree Confidence Factors for Multi-Agent Control," *Proceedings of the 2nd International Conference on Autonomous Agents*, ACM Press, NY, 1998, pp. 86-91.

[viii] Washington, R., "Markov Tracking for Agent Coordination," *Proceedings of the 2nd International Conference on Autonomous Agents*, ACM Press, NY, 1998, pp. 70-77.

[ix] Ahn, Sungjoon and A. Udaya Shankar, "An Application of the Information Dynamics Framework: Adapting to Route-demand and Mobility (ARM) in Ad hoc Network Routing," Computer Science Department, University of Maryland, March 2001.

[x] Leiner, Barry M., "The Scope of the Digital Library," Draft prepared for the DLib Working Group on Digital Library Metrics, January 16, 1998, Revised October 15, 1998, http://www.dlib.org/metrics/public/papers/dig-lib-scope.html

[xi] See the proceedings of the Sixth Text Retrieval Conference (TREC) at http://trec.nist.gov/

[xii] Buckland, M., Chen, A., Chen, H., Gey, F., Kim, Y., Lam, B., Larson, R., Norgard, B., and Purat, Y. "Mapping Entry Vocabulary to Unfamiliar Metadata Vocabularies," D-Lib Magazine, Vol.5 No.1, January 1999.

[xiii] Kantor, Paul B., Endre Boros, Benjamin Melamed, Vladimir Meñkov, "The Information Quest: A Dynamic Model of User's Information Needs," Rutgers University, http://aplab.rutgers.edu/ant/

[xiv] See http://www.packhunter.com

[xv] Larsen, Ronald L., and Jean Scholtz, "Digital Libraries and Scholarly Communication," to be published.