A Framework for Digital Object Self-Preservation

Charles L. Cartledge
Old Dominion University
Department of Computer Science
Norfolk,VA 23529 USA
ccartled@cs.odu.edu

ABSTRACT

The author presents a plan to design, simulate, create, field and test an infrastructure that supports long term preservation of digital data based on the idea that the digital data should have an active role in its own preservation. The infrastructure traces its roots the the Kahn-Wilensky Framework for digital services. The infrastructural links between the digital data create a small world graph for efficient digital library activities and communication. A novel way to create the small world graph based on the incremental addition of nodes to the existing graph and their independent determination how to relate to other members of the graph is given.

Categories and Subject Descriptors

H.3.7 [**Digital Libraries**]:

General Terms

Algorithms, Design, Experimentation

1. INTRODUCTION

The research question that I am investigating is: can the traditional digital library model become more efficient by augmenting the digital data with the capacity for self preservation and providing an infrastructure where that can occur. Currently, the traditional approach to digital preservation is for archivists working a digital library to curate, maintain and monitor the digital data under their purview. This approach is working now and will probably continue for a while, but eventually the sheer volume of digital data will overwhelm the archivists, resulting in irretrievable loss of data.

I am interested in the long term preservation of digital data and the creation of a web based infrastructure that will support preservation built on the premise that the digital data should be able to preserve itself. I believe that looking at the preservation problem from the perspective of the digital data looking out into the digital library (versus looking at the data from the perspective of the library) and then having it attempting to make copies of itself is a novel and innovative view that has not been adequately addressed to date. In researching the goal of enabling self-preservation; applicable current and historical persistent infrastructures have been evaluated and ideas from graph theory used to define the method for linking digital data for efficient communication and ultimately safely distributing redundant copies across the underlying computer network. To support this research, I have developed a simulator to test my ideas.

I end with a discussion of the work I have done to date, where I plan to go in the immediate future and how these efforts contribute to the science of digital libraries.

2. RELATED WORK

I am investigating the creation of an autonomic system that can be used for the preservation of digital information and data. At the lowest level is an infrastructure where the digital data must live. At the highest level are the infrastructural links that define how these digital parcels are connected one to another. Between the two extremes is a self organizing mechanism that operates at the digital parcel level using data and information that the parcel has gleaned while communicating one to another and establishing the infrastructural links. These low and high views of the system point to two distinct areas of related work.

2.1 Infrastructure related work and ideas

My work is heavily influenced by Kahn and Wilensky Framework (KWF) [14] and their technical definition of a digital object. Applicable details of their work are detailed in section B. Lots Of Copies Keeps Stuff Safe (LOCKSS) is a distributed decentralized peer-to-peer infrastructure for the preservation of bits and bytes [26]. My approach is similar, but extends the desire for preservation from the archivist into the digital object. OceanStore provides a durable storage utility atop an infrastructure of untrusted servers and supports nomadic data [17]. I am focusing on the idea of trusted servers and data that is stationary, except during directed digital migrations. Intermemory [8] focused on exchanging a finite-term donation for unbounded storage rights. The collective storage capacity of all donors is viewed as a very large persistent block of contiguous RAM, where each donor has access to a small portion of that RAM for their own purposes. This could include long term persistent storage of digital information. My work compliments the ideas proposed by Hunter and Choudhury [13], whereas they put forward the idea of a semi-automated communications between repositories, mine is takes place at a lower level and is a behind the scenes approach.

2.2 Graph construction related work and ideas

The relationship between my work and existing work in

JCDL'09, June 15–19, 2009, Austin, Texas, USA.
ACM ——————/–/–.

graph construction can be best addressed using graph construction related terms, see Appendix C for an explanation of terms and ideas.

My approach contrasts with others [24] where connections to other nodes are proportional to the destination's degree count. We can use preferential attachment only to select the first node when starting to consider connecting to any nodes as compared with algorithms that starts with a graph (or lattice) and then grows a "small world" by the addition of new links [10, 12, 9, 15, 21]. Or, by connecting a node to a fixed number of vertices based on their degree [5], or even creating a small world graph from a random one [11], whereas we construct a small world one from the beginning. Some approaches grow a graph based on preferential attachment (or "fitness") [16, 4, 6]. A survey of small world graph construction techniques and analysis is given [23], but none discuss the creation of small world graphs based on locally gleaned knowledge in a manner that we demonstrate.

3. BACKGROUND

My area of investigation crosses the disciplines of Digital Libraries (DLs) and graph theory using ideas based on the KWF. A summary of DL activities as they relate to this discussion is included in Appendix A. An overview of the KWF and some the infrastructures that can be traced directly back to it are in Appendix B. Graph theoretical terms and ideas are used extensively through this discussion. Those that are germane are listed in Appendix C. Section 6 ties each of these disparate disciplines together.

These are terms and ideas that are unique to the discussion. They are used in sections 4, 5 and 7.

Friend: DOs are nodes within a graph. The nodes that any particular DO is directly connected to is considered its friend.

Family: Ultimately, I am interested in imbuing DOs with the ability to preserve themselves through replication. All DO that are replicas of each other belong to the same family.

Parent: The family member that was first inserted into the graph and is responsible for ensuring that enough family members are created to meet its preservation goals.

4. INVESTIGATIONS TO DATE

I am taking the concept of a repository as defined by the KWF, and expanding on it to include:

> ... any computer system whose primary function
> is to store digital material for use in a library.
> *William Y. Arms*[2]

Overlaying the repository concept onto a simulated web architecture and focusing on building graph structures from a node's perspective (to wit: node's perspective ⇒ small world graph ⇒ web architecture ⇒ repository concept). I have developed an algorithm [7] that creates a small world graph by incrementally adding one node at a time. Each node makes friendship connections to existing members of the graph independent of outside control or guidance. This is different from most other approaches that have some sort of omnipotent controller that has knowledge of the entire graph and makes global decisions.

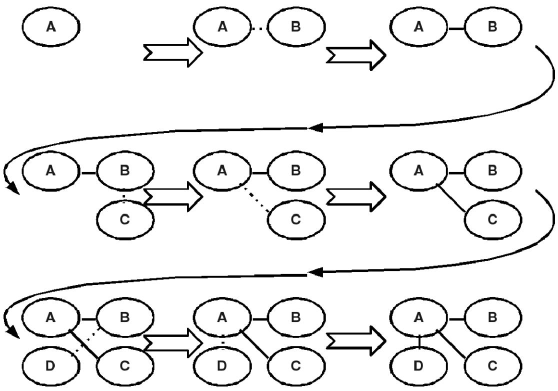


Figure 1: The unsupervised small world growth algorithm with 4 nodes. The dashed lines are communications and the solid lines are friendship links. See the text for a full explanation.

Based on a review of graph structures, their characteristics and attributes, small world graphs appear to be the most practical choice for minimizing a graph's size, communication costs and construction effort. Small world graphs also emulate natural processes and occur often in nature and human endeavors, where regular and random graphs are relatively infrequent.

Following is a description of how four nodes are processed and is shown in Figure 1. All nodes have passed the DL accession activity and have entered the storage phase. The simulation allows for different methods to select an initial friend, including: always choose the same one, select one at random from all that are known, and use preferential attachment as a selection criteria. Each new node begin to "wander" the preservation system looking for an initial friend. Node A is the first DO in the preservation system and so stops wandering immediately. Wandering node B is inserted into the system and communicates with node A. Node B makes a friendship link with Node A because it is the only other node. Wandering node C is inserted into the system, and as a result of random chance contacts node B first. Node C gets B's list of friends and then tries to form a friendship link with node B. Based on a roll of the die, C does not form a friendship with B. C then looks at its internal list of potential friends (its "to be visited" list) and contacts node A. C gets node A's list of friends and attempts to form a friendship with node A. Because node A is the last node in C's "to be visited" list and because C had already visited all of A's friends, C forms a friendship link with A. Node D now enters the system and, like node C happens to communicate first with node B. Like C, node D fails to connect to node B and then tries its luck with node A. Node D forms a friendship link with Node A, but does not form a link to A's friends B or C. In general, a "wandering" node is introduced to an existing node in the preservation system. The wanderer gets a list of friends from the non-wandering node. If the wandering node doesn't form a friendship link with this particular non-wandering node, it will select another non-wandering node that it has learned of from all its conversations with other non-wandering nodes it has encountered. When the wandering node finally makes a friendship link, it will then look back at all the nodes that it did not connect with as well as some that it was intending to communicate with and make friendship links with some of them.

There are a number of important things to take away from the above discussion. They are:

1. Links are infrastructure not HTML navigational. The links between friends and family are infrastructure links only. These links serve as a way for DOs to send messages from one to another. Messages could include things like requests or replies relating to the status or condition of a particular DO, termination directive if there are too many family members, establishment of new friendship links, or other such maintenance activities.

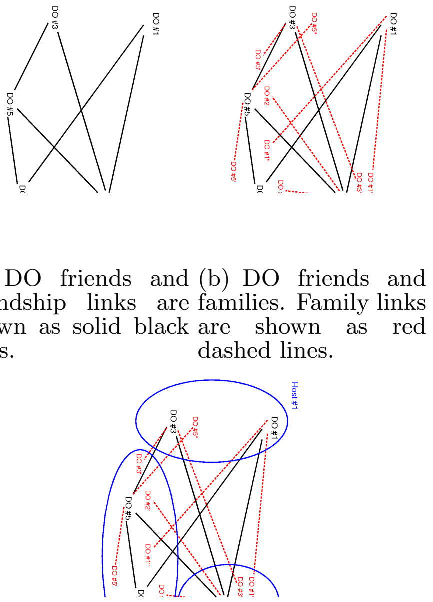
2. Links are important. From a graph theoretical point of view, a node has little or no value, it is the set of edges that give the graph structure and form. The same can be said for the links between DOs. Without the links, the DOs would not be able to communicate with anyone else. They would be inanimate objects.

3. Friendship links are used to support the replication process. Each DO has been preprogrammed to replicate itself in order to preserve itself. All DOs live on a host, and friendship links allow any particular DO to query its friends about which host they live on. If a DO needs to replicate itself and it has a friend that lives on a different host and if there is room on that host for an additional DO then the DO can replicate itself onto the new host. Replication is how a family grows from a single copy (a parent) to multiple copies (a family). The friendship links, family links and how family members are spread across various hosts is shown in Figure 4.

5. FUTURE INVESTIGATIONS

The investigations to date have been based on simulations of different graph sizes, differing numbers of hosts and different minimum and maximum copies for preservation. The results have been promising because the graphs have been predictable, the distribution of DOs and preservation copies across hosts reasonable. Figure 3 shows simulation results for a 1000 node graph spread across 1000 hosts where the DOs attempted to replicate themselves between 3 and 10 times. Not all DOs were able to replicate themselves as often as intended because some of their friends lived on the same host as they did. Having multiple copies of a DO on the same host does not increase the likelihood of survival.

In the near term there is a need to investigate the robustness of these autonomously created small world graphs in the face of accidental and targeted failures. Because of the nature of small world graphs, there are typically a relatively small number of articulation nodes (nodes that are vital to the many paths in the graph). If these nodes are lost due to accident or network distribution then the graph may become disconnected [1]. The graphs that I have created need to be evaluated and tested in these types of scenarios. If the graph becomes disconnected then the DO families will have to take some sort of action to elect a new parent and return to the business of creating replicas. If a graph that was disconnected reassembles itself (for instance a major network

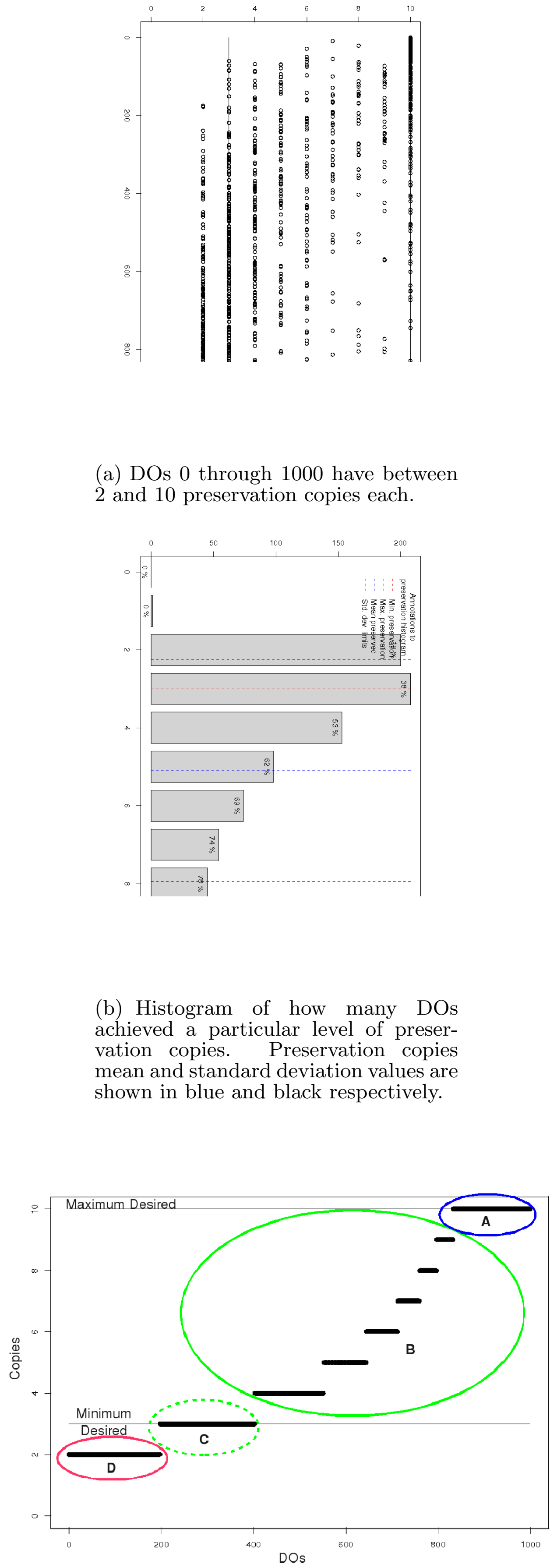


(a) DO friends and friendship links shown as solid black lines.

(b) DO friends and families. Family links are shown as red dashed lines.

(c) DO friends, families and hosts. Hosts are shown as solid blue ovals.

Figure 2: A layering of DO friends, families and hosts. A DO may have many friends. A DO may have many family members. DO friends and families live on hosts.



(a) DOs 0 through 1000 have between 2 and 10 preservation copies each.

(b) Histogram of how many DOs achieved a particular level of preservation copies. Preservation copies mean and standard deviation values are shown in blue and black respectively.

(c) DOs sorted by number of preservation copies achieved. DOs in area D (red) are in a "time of scarcity," in area A (blue) a "time of plenty," in B or C (green) are in between .

Figure 3: A case where 65% of 1,000 DOs have preserved at least their minimum number of copies.

router becomes operational again, or something else that has caused a major disruption is corrected) then the DO families will have to take corrective actions to ensure that the correct number of replications are in place.

Simulations can be useful to predict how things will operate in the real environment, but nothing replaces actually building a system and testing it in a real hardware and operational environment. I view this as at least a three stage evolution. The first is to develop a module to an Apache server that will transparently and with minimal communications impact, implement the small world creation, replication and maintenance ideas that I have put forth. This module would then be tested in a controlled test environment and its performance measured and assessed. Assuming that this phase goes well, then the module would be inserted into a semi-controlled education environment for testing on live systems.

One major type of data that will become available during this testing is the actual communications costs involved with creating and maintaining the system. In a simulation environment, there are no communication costs and nearly instantaneous. The costs and times in a real environment are expected to be considerably different. If communications are uncontrolled, grow too fast for the size of the graph, consume too much bandwidth or processor resources then the design and implementation will have to be revisited. All of these "automated" communications between and amongst the DOs will not preclude the human archivist from exercising ultimate veto power over any and all actions.

6. DIGITAL LIBRARY ACTIVITIES USAGE SCENARIOS WITH THE PROPOSED FRAMEWORK

The normal/traditional DL activities and their mappings into both the KWF and my area of investigation are shown in Table 1 and are explained in the following short list of scenarios (assuming that the small world graph of DOs already exists and that it is connected).

Change media: Assume that an old device is no longer considered safe or reliable and a replacement has been installed. All DOs that are currently on the old device, must migrate to the new. A message could be crafted telling those DOs on the old device to migrate to the new one. Sending the message to any DO will result in all members of the graph receiving it and if applicable acting on it.

Change in format, scenario #1: Assume that an old data format will no longer be supported and that all DOs that currently utilize the old format will have to be converted to the new format. A format conversion service has been developed somewhere on the network and the interface to the service is something that the DOs can utilize. Because the DO graph is connected, passing a message to any DO will result in the message being passed to all DOs. If the message is that all DOs with old style data formats are to use a particular web service to convert to the new format would result in a complete update of all DOs regardless of their physical or logical location.

Change in format, scenario #2: The primary communications in the above scenarios has been between friends

DL Activity	KWF correspondence	Area of investigation
Appraisal and selection	Outside of KWF	No change
Accession	Outside of KWF	No change
Storage	Outside of KWF	For simulation purposes, all DOs (friend and family members) are assumed to have the same storage requirements.
Access	Part of RAP	Intuition at this time points to a small world graph structure as being the most viable configuration for efficient access to all graph members,
System engineering	Outside of KWF	The intent is to imbue the DOs with an ability to maintain themselves as part of their small world.
Change media	Outside of KWF	Could fit exactly into my area of interest.
Change format	Within RAP	Could fit exactly into my area of interest.
Incorporate standards	Within RAP	Could fit exactly into my area of interest.
Build migration paths	Outside KWF	No change.

Table 1: Mapping of traditional DL activities into the Kahn-Wilensky Framework and into areas of investigation.

and a message to one DO will be received by all DOs. The DO that receives the message will be the parent of its family. If graph operations are to be executed on a "not to interfere" basis then the parent will slowly forward the message on to all members of its family. Family communication is separate and apart from friendship communication. The format migration that the parent experienced could be replicated by each family member, or the parent could simply send an updated copy to each family member and thereby reduce the load on the conversion service.

Access: Assume that a new search or information retrieval algorithm has been developed and it is desired that all DOs be subjected to this new technique. A message could be sent to any one of the DOs about the location and availability of the service. The message would make it through to all the DOs and they would comply, requiring only that the archivist inform one DO.

7. EXPECTED CONTRIBUTIONS TO THE SCIENCE OF DIGITAL LIBRARIES

The contributions to the Science of Digital Libraries will be in:

Exploration into the use of small world graphs for storing DOs: Small world graphs appear to have potential for use as a model for storing, access, communicating with and managing DOs. This effort is focused on seeing if it can be done, what lessons there are there and furthering our understanding of naturally occurring structures as applied to DL.

Application of autonomic philosophies to preservation: Autonomic principles of self configuring, self healing, self optimizing, self monitoring and self protecting are embedded in the way that the preservation graph is created, organized and maintained based on local data and information that the nodes gather. Global knowledge, guidance, control and oversight, while available is not required.

Development and fielding of tool to transparently assist archivists:

If development of the Apache module proceeds as anticipated then it will transparently assist DL archivists in many of the routine chores and tasks. By providing transparent and automated tools, the archivists will be able to more effectively and efficiently manage larger DLs.

8. SUMMARY

I feel that I have conducted enough preliminary research and analysis to believe that the concept of imbuing DOs with a preprogrammed number of preservation copies within a supportive infrastructure that supports preservation is a viable approach. My simulations indicate that it is possible to create graphs that have the small world like clustering coefficients and average path lengths based on locally gleaned data. I feel that it is time to move out from simulations and field a system that could assist digital library archivists.

9. ACKNOWLEDGEMENTS

I am thankful and grateful for the comments and recommendations from the JCDL reviewers. Their efforts, perspectives and viewpoints are much appreciated. This work supported in part by the NSF, Project 370161. Graph visualization and analysis was done using Pajek software (http://vlado.fmf.uni-lj.si/pub/networks/pajek/).

10. REFERENCES

[1] R. Albert, H. Jeong, and A.-L. Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794):378–382, July 2000.

[2] W. Y. Arms. *Digital Libraries (Digital Libraries and Electronic Publishing)*. The MIT Press, December 1999.

[3] M. Baker, K. Keeton, and S. Martin. Why traditional storage systems donâĂŹt help us save stuff forever. In *Proc. 1st IEEE Workshop on Hot Topics in System Dependability*, pages 2005–120, 2005.

[4] A.-L. Barabási, R. Albert, and H. Jeong. Scale-free characteristics of random networks: The topology of the world wide web. *Physica A: Statistical Mechanics and its Applications*, 281(1):69–77, June 2000.

[5] B. Bollobás, O. Riordan, J. Spencer, and G. Tusnády. The Degree Sequence of a Scale-Free Random Graph Process. *Random Structures and Algorithms*, 18(3):279–290, 2001.

[6] G. Caldarelli, A. Capocci, P. D. Los Rios, and M. A. Munoz. Scale-free networks without growth or preferential attachment: Good get richer. *Disordered Systems and Neural Networks*, 2002.

[7] C. L. Cartledge and M. L. Nelson. Unsupervised creation of small world networks for the preservation of digital objects. *Joint Conference on Digital Libraries*, page (submitted for publication), June 2009.

[8] Y. Chen, J. Edler, A. Goldberg, A. Gottlieb, S. Sobti, and P. Yianilos. A prototype implementation of archival intermemory. In *Proceedings of the fourth ACM conference on Digital libraries*, pages 28–37. ACM New York, NY, USA, 1999.

[9] P. Duchon, N. Hanusse, E. Lebhar, and N. Schabanel. Could any graph be turned into a small-world? *Theoretical Computer Science*, 355(1):96–103, 2006.

[10] P. Duchon, N. Hanusse, E. Lebhar, and N. Schabanel. Towards small world emergence. In *SPAA '06: Proceedings of the eighteenth annual ACM symposium on Parallelism in algorithms and architectures*, pages 225–232, New York, NY, USA, 2006. ACM.

[11] B. Gaume and F. Mathieu. From random graph to small world by wandering. Technical Report 6489, Unité de recherche INRIA Rocquencourt, Domaine de Voluceau, Rocquencourt, BP 105, 78153 Le Chesnay Cedex (France), 2008.

[12] K. I. Goh, B. Kahng, and D. Kim. Universal behavior of load distribution in scale-free networks. *Physical Review Letters*, 87(27):278701+, December 2001.

[13] J. Hunter and S. Choudhury. A semi-automated digital preservation system based on semantic web services. In *JCDL '04: Proceedings of the 4th ACM/IEEE-CS Joint Conference on Digital Libraries*, pages 269–278, 2004.

[14] R. Kahn and R. Wilensky. A framework for distributed digital object services. *International Journal on Digital Libraries*, 6(2):115–123, 2006.

[15] J. Kleinberg. The small-world phenomenon: An algorithmic perspective. In *Proceedings of the 32nd ACM Symposium on Theory of Computing*, 2000.

[16] K. Klemm and M. Eguíluz, Víctor. Growing scale-free networks with small-world behavior. *Physical Review E*, 65(5):57102, 2002.

[17] J. Kubiatowicz, D. Bindel, Y. Chen, S. Czerwinski, P. Eaton, D. Geels, R. Gummadi, S. Rhea, H. Weatherspoon, C. Wells, et al. Oceanstore: An architecture for global-scale persistent storage. *ACM SIGARCH Computer Architecture News*, 28(5):190–201, 2000.

[18] C. Lagoze and J. Davis. Dienst: An architecture for distributed document libraries. 1995.

[19] C. Lagoze, H. Van de Sompel, M. Nelson, and S. Warner. The Open Archives Initiative Protocol for Metadata Harvesting-Version 2.0. *Retrieved November*, 20:2004, 2002.

[20] C. Lagoze, H. Van de Sompel, M. Nelson, S. Warner, R. Sanderson, P. Johnston, A. Gentil-Beccot, S. Mele, A. Holtkamp, H. O'Connell, et al. Object re-use & exchange: A resource-centric approach. *Arxiv preprint arXiv:0804.2273*, 2008.

[21] N. Mathias and V. Gopal. Small-worlds: How and why. *Disordered Systems and Neural Networks*, 2000.

[22] M. L. Nelson and K. Maly. Smart Objects and Open Archives. *D-Lib Magazine*, 7(2):1082–9873, 2001.

[23] M. E. J. Newman. Models of the small world: A review. *J.STAT.PHYS.*, 101:819, 2000.

[24] V. Nguyen and C. Martel. Analyzing and characterizing small-world graphs. In *SODA '05: Proceedings of the sixteenth annual ACM-SIAM symposium on Discrete algorithms*, pages 311–320, Philadelphia, PA, USA, 2005. Society for Industrial and Applied Mathematics.

[25] S. Payette and C. Lagoze. Flexible and extensible digital object and repository architecture (FEDORA). *Lecture Notes in Computer Science*, pages 41–60, 1998.

[26] V. Reich and D. Rosenthal. LOCKSS: A permanent web publishing and access system. *D-Lib Magazine*, 7(6):1082–9873, 2001.

[27] D. Waters and J. Garrett. *Preserving Digital Information. Report of the Task Force on Archiving of Digital Information.* The Commission on Preservation and Access, 1400 16th St., NW, Suite 740, Washington, DC 20036-2217 ($15); World Wide Web: http://www. rlg. org, 1996.

[28] D. J. Watts. Networks, dynamics, and the small-world phenomenon. *The American Journal of Sociology*, 105(2):493–527, 1999.

[29] D. J. Watts and S. H. Strogatz. Collective dynamics of 'small world' networks. *Nature*, 393:440–442, June 1998.

APPENDIX

A. DIGITAL LIBRARY (DL) ACTIVITIES

Digital library archivists have a number of roles and responsibilities [27]. These include:

Appraisal and Selection: All systems have a finite (albeit large) amount of storage space, therefore it isn't possible to store DO. The archivist is responsible for the selection of those DO items that are deemed to be most valuable.

Accession: Once a DO has been selected for storage, it must be prepared for the storage in the archives.

Storage: The placement of the DO onto a media of some type. Consideration is also given to the anticipated frequency of access, number of redundant copies and some sort of hierarchical organization.

Access: Ensuring that the DL is accessible via a network with the appropriate bandwidth and protocols for delivering the DOs.

System engineering: Defining and maintaining the interlocking requirements of media and data formats, hardware and software upon which the DL depends.

DL archivists have to take a long-term view of the DOs under their purview. With this comes the realization that both DOs and the media on which they live will eventually become obsolete. In order to meet the DL's responsibility for long term preservation and to address the continuing obsolescence problem; archivists must have strategies to migrate their DOs from the old to the new. These are a few migration strategies:

Change media: Current magnetic and optical technology is subject to "bit rot" [3] and if left unattended will eventually corrupt enough to the media so that the DO will become unrecoverable. As more stable media becomes available, DOs on older and less stable media have to be copied from the old to the new.

Change format: A multiplicity of data formats can become unmanageable. A DL could decide to change, or convert a DL from its original format to a more manageable and "standard" format.

Incorporate standards: DLs, like any other user of digital data, benefits from adherence to well published and accepted standards.

Build migration paths: DL archivists can communicate and educate DO creators about better and more efficient techniques that can be used in the creation of DOs. Incorporating the ideas of digital preservation early makes the inevitable migrations later easier.

B. THE KAHN - WILENSKY FRAMEWORK (KWF)

The Kahn-Wilensky Framework (KWF) describes the fundamental aspects of an infrastructure that is large and which supports digital information services, of which a DL is one [14]. The KWF has a number of elements (*digital objects, handles, metadata, repositories, handle generators, originators, uses, global naming authorities, local naming authorities,* and a *repository access protocol* that operate in concert to support digital information services. Selected members of this list are of particular interest. They are:

Digital object: In a technical sense, a data structure whose primary components are digital material (for instance, a representation of the archived DO) and other data.

Global naming authority: A user makes a request to the GNA for and is provided with a unique identifier.

Handle: A unique identifier from a GNA that is used to access a Digital Object.

Repository: A network accessible storage system where digital objects can be stored for later retrieval.

Repository access protocol: A way for depositing and retrieving digital objects from a repository.

KWF has has directly influenced the following systems and infrastructure models:

1. Dienst's location independent identifiers [18],
2. Flexible and Extensible Digital Object and Repository Architecture (FEDORA) theoretical foundations lie partially with the KWF and influenced the design of their DigitalObject [25],
3. Buckets/Smart Objects, Dumb Archives (SODA)'s concept that DOs can have more than just digital representations of objects [22],
4. Open Archives Initiative - Protocol for Metadata Harvesting (OAI-PMH) concept that a DO can expose things about its contents and thereby assist external entities that desire to interact with it [19], and
5. Open Archives Initiative - Object Reuse and Exchange (OAI-ORE) recasts the repository-centric notion of digital objects into a bounded aggregation of Web resources through the use of resource maps [20].

C. GRAPH THEORETIC TERMS AND IDEAS

Graph theory has a vocabulary and set of tools that can be used to describe and compare the attributes of composite entities made up of elemental entities that are connected together in some manner. By using graph theoretic terms and ideas, the connections between DOs can be evaluated in terms of communications efficiency, average distance between any member of the graph, or the expected size of the graph that has some number of members. Some basic general graph theoretic terms, graph types and terms specific to this discussion are given below.

C.1 General graph theoretic terms

A casual vocabulary of terms and ideas that are applicable to this discussion:

Graph: A composite structure made of vertices and edges. The size of the graph (number of vertices) is denoted as N. In this discussion, a graph is composed of DOs that have infrastructural links. Infrastructural links are separate and distinct from HTML navigational links.

Vertex: From a graph theoretic point of view, an elemental entity. When I talk about about a vertex, it is a DO in the KWF sense.

Edge: A connection between two vertices. A edge that loops back to it originating vertex is not allowed. For the purposes of this paper, an edge is an infrastructural link.

Node: The same as a vertex.

Degree: The number of edges connect a particular vertex to other vertices.

Clustering Coefficient (CC): Is a measure of local graph's local structure. It is a value between 0 and 1 for a vertex that reflects the percentage of vertices that it and another vertex that it is directed connected to share in common.

Path: The edges from one vertex to another vertex.

Connected: All vertices can be reached from any vertex via a path of some length.

Diameter: The longest, shortest path between all vertices in the Graph.

Path length: The number of edges in a path from one vertex to another.

Average path length (L): The mean of all paths between all vertices in the graph.

Network: Classically a graph where the edges have a weight. In my context, all edges have a weight of 1, so the word network is used interchangeably with graph.

C.2 Selected types of graphs

There are a staggering number of named types of graphs. Within this discussion, the focus is on a very small number and even of those one is of primary interest.

1. *Regular*: A graph where the minimum number of edges of any vertex equals the maximum number of edges of any vertex. All vertices have the same number of edges K (i.e., the same degree).
2. *Lattice*: A graph where every node at location (a,b) has edges to nodes at locations (a-D,b), (a+D,b), (a, b-D) and (a,b+D) where D is some constant, but arbitrary offset.

3. *Random*: A graph where node A is connected to node B based on some probability. These stochastic connections are made from any node A to any node B.

4. *Social networks*: A graph where the average CC is high and the graphic structure mimics human activities.

5. *Small world*: A graph where the average CC is high and the average path length is low [29].

Small-World graph creation was introduced in Watts and Strogatz [29]. They began with a lattice, then perturbed the structure by randomly swapping links based on some probability. This creates a graph whose normalized clustering coefficient (CC) and normalized average length (L) curves have a shape similar to that shown in Figure 4. CC is a measure of local graph structure, while L is a measure of global graph structure [28] The Watts and Strogatz [29] technique for creating a small-world graph is based on setting a threshold parameter ρ between 0.0 and 1.0 and then rewiring each link in the lattice based on a random number compared to ρ. Normally, a "small-world" is defined as that region where the average path length (L) is short and the clustering coefficient (CC) is high. Both of these factors are controlled by ρ. This plot of the normalized path lengths and normalized clustering coefficients shows the "small-world" region as that area between the L knee on the left and the CC steep slope on the right. ρ is in the nominal range of 0.01 to 0.1 inclusive. A graph that exhibits L and CC curves approximating the shapes shown in Figure 4 has "small-world" properties. The region in Figure 4 to the right of the L knee is the domain of the regular graphs. While the region to the left of the steep slope in the CC curve belongs to random graphs. Figure 5 illustrates a regular, small world and random graphs each with 20 nodes. The graph has 20 (N = 20) nodes and each node is connected to 4 (K = 4) others. As a regular ring lattice graph the probability that an edge is not connected to its 2 second neighbors (2 neighbors to the left 2 to the right) is 0. As the graph evolves towards a random state, the probability of being not connected to its 2 second neighbors approaches 1. An analytic comparison of the CC and L for regular, small world and random graphs is shown in Table 2.

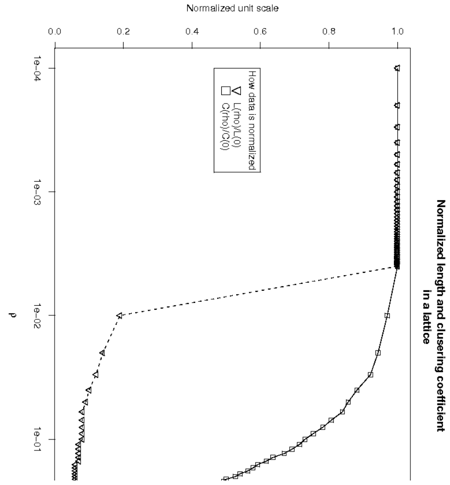
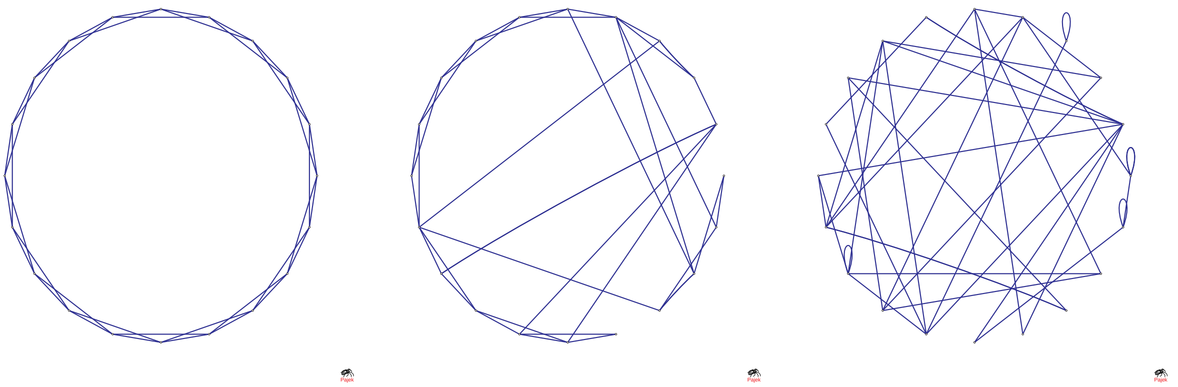


Figure 4: A "small-world" graph exists along the continuum between a regular lattice and a random graph.

Graph type	L (average path length)	CC (cluster coefficient)
Ring lattice	$\frac{N}{2*K}$	$\frac{2*(K-1)}{2*K}$
Small world	$\frac{N}{2*K} >$ L $> \frac{ln(N)}{ln(K)}$	$\frac{2*(K-1)}{2*K} >$ CC $> \frac{ln\langle\rho*K\rangle}{N}$
Random (ρ is probability of linkage)	$\frac{ln(N)}{ln(K)}$	$\frac{ln\langle\rho*K\rangle}{N}$

Table 2: Expected average path length and expected clustering coefficient for selected graph types.



(a) Regular ring lattice

(b) Small world

(c) Random

Figure 5: The metamorphosis of a graph from a regular state through a small world and finally to a random state.