Pamela Vizner Oyarce
Digital Preservation
Professor Kara van Malssen
October 1, 2013

Digital Preservation Network (DPN)

Introduction

Institutions, as individual entities, have tried to establish systems to make digital preservation of their collections achievable. But there is so much you can do by yourself. While individual initiatives can meet the goal, implementing them can be very expensive and daunting. Community efforts present themselves as a more effective way to achieve long-term preservation, especially when dealing with large amounts of data. Collective initiatives are not new[1], but each of them present a solution to one or more issues encountered by their specific target communities[2]. This paper will present the project Digital Preservation Network, its scope, general functions and architecture, stakeholders and status of development.

DPN

The Digital Preservation Network (DPN, pronounced "deepen") is a network of interoperable, heterogeneous and trustworthy preservation repositories which goal is to preserve academic and scholar information for future generations. To do so, it establishes technical, service and legal frameworks. While the sources consulted do not mention a specific date of creation, the information about the project - especially related to its technical requirements - has been circulating online since early 2012. DPN was conceived as a preservation system "for the academy and by the academy".[3]

The institutions leading the initiative are University of Virginia, University of Michigan, University of Texas, Evans Telecommunications Company, University of Illinois, Columbia University, University of California, Stanford University, Johns Hopkins University, HathiTrust, Massachusetts Institute of Technology, UC San Diego, DuraSpace, University of Maryland, Academic Preservation Trust, Association of Research Libraries, Ohio State University, Purdue University, Cornell University and Duke University. These institutions have representatives in one or more of the three teams created for the implementation of DPN: Leadership Team (or board), Tech Team and Succession Rights Team. Five of those institutions not only participate in management and planning, but also have an active role in the architecture and functions of DPN, acting as information "nodes", as explained later in this document.

People in the three teams come from varied fields, most of them from library sciences, economy studies, information technology and copyright/policies studies.

Inside DPN, there is another level of participation: membership. Members (also known as depositors) are also academic digital repositories in the US, who not only economically support the project, but who will provide the data to DPN for its preservation. Currently, the project has more than 50 members.

So what does DPN do? DPN provides a system to preserve information in case of failure, data corruption, organization dissolution, disaster, or any situation that would put in danger its long-term preservation. This project is based on the fact that no matter how strong or

well managed an institution is, it is still an individual entity susceptible to economic, administrative, political or technological failure.
In general terms, DPN stores and replicates across the network "dark copies" of the data provided by the members. This means that the repositories in the network keep copies only for preservation, i.e. not available for end-users. However, DPN gives the possibility of restoring it in case of data loss - process called "brightening" content - providing the required technical and legal framework. It is worth noting that brightening not always means that the content will be made publicly available.
What does not do? DPN does not provide access in any form to final users: *"DPN is about preservation only"*[4]. Access or any other service can eventually be provided as determined by the contracts and agreements held between members and nodes, but these are not required to participate in the project.

DPN: Architecture[5]
The DPN architecture consists of the following basic elements: members, nodes (first and replicating), a message system and a registry.
The basic workflow is as follows: the members provide their data to one of the five nodes, which is called first node. First nodes and members determine the terms and conditions of the transfer and other services individually. The content then is replicated to the other nodes; the replicating nodes. Nodes can be considered simultaneously first or replicating nodes, depending on its relationship with the contributors and the data stored. Each node also contains a copy of the system registry, i.e. basic information of every asset in the system, which is kept in order to provide content tracking.
The communication and information transferring between nodes is done according to the general requirements of the system. DPN, rather than establishing technology solutions or strict standards for the transfer of information, determines a general framework. Thus, each first node must decide how to implement the systems and technical infrastructure to comply with the network requirements.
In order to demonstrate the trustworthiness of the network, DPN also provides frameworks for fixity auditing, process auditing, and reporting. Once again, nodes must implement solutions to fulfill these requirements.

Technical Requirements
On the technical side, DPN has established some basic requirements for nodes. The network uses Advanced Message Queuing Protocol (AMQP)[6] for messaging, BagIt[7] bags for wrapping content, and HTTPS[8] and rsync[9] for the transmission of content. It is worth to reiterate that these technical requirements only apply inside the network and that the transfer of content from members to the first nodes could or could not be done using the same standards.
DPN also establishes basic metadata requirements for the transmission of information. This metadata will be available over the network on the registry. The registry allows every node to access the following fields for each element inside the network:

- ObjectID: recording both DPN UUID and First Node provided ID, which allows each node to avoid the use of a mapping to keep track of both IDs.
- Source information: provided by the First Node
- Fixity

- Location of Copies: indicating the nodes where the data is replicated.
- Versioning Info

Encryption during transfer is not required, but it is only allowed to encrypt data; metadata would still be readable for every node.
DPN is currently working on the requirements for fixity auditing, process auditing and reporting, as established by the project timeline.

Reliability
DPN has established four different stages for data transfer. Each of them represents every possible scenario in which data travels inside the network. These four stages explain how the system works in case of failure.

- Stage 1, Ingest and Replication: the depositor transfers data to the first node. The first node replicates information to other nodes (one or more).
- Stage 2, Restoration of Content: In case of identifying data corruption after an audit process, replication nodes provide the first node a copy of the data, in order to be able to retrieve it to the depositor (if established in the agreements)
- Stage 3, First Node Cessation: in this case, replicating nodes initiate the process of brightening in order to be able to retrieve content to the depositor.
- Stage 4, Successioning: if both depositor and first node cease to exist or are no longer available to provide the required services, replicating nodes start the process of brightening in order to eventually make the information available, according to the agreements previously signed.

Project Timeline
As mentioned before, there is no public record of when the project started. However, the discussion and planning of technical requirements started in February of 2012, divided in four phases.
During Phase 1, the technical team worked on the parameters and various possible solutions and implementations for the theoretical model. Testing of the first technical features started with Phase 2, in January of 2013. Until May the team tested messaging and data transfer protocols as well as discussing the implementation of data encryption and fixity.
In Phase 3 - until August 2013- the project started with the installation of hardware at each node, the implementation of the registry and testing of transfer of larger packages of test data. Fixity checking and audit reporting was also tested.
The project is currently in Phase 4, which will last until December 2013. During this period the team will start working with data providers (members) in terms of contracts and agreements to determine how data will be ingested, what services the node will provide and what are the copyright status of each collection in order to determine the legal framework for eventual brightening processes. The technical team will also perform some test for brightening and reporting.
With all this considerations, the DPN team expects to have the network up and running at the beginning of 2014.

Succession Rights
The Succession Rights Team has been working on the legal framework to ensure the use of content inside the network in case of the dissolution of a member institution or node. The first report was released on September 2013, which includes an explanation of the case study performed and a draft of the preliminary legal structure.[10] In general terms - according to this first publish document - DPN will have the non-exclusive, perpetual and irrevocable license to: document and describe assets; store, copy and reformat content; re-disseminate copies in case of non-existing intellectual rights or in case that permissions are granted.
According to this document, DPN also requires the depositors to clear any rights issues and to sign a special agreement with donors to allow the content to be transferred to the DPN network.

DPN Features[11]
DPN, because of its node-based architecture, data transfer system and legal framework provides:

- Resilience: in case of any type of failure - whether originated at a node or depositor – the system has the capability of returning to its original state, i.e. recovering the data from one of the nodes (first or replicating depending on each case).
- Succession: the system provides the intellectual and technical information to "brighten" content in the future.
- Economies of scale: the cooperative system provided by DPN allows academic institutions to preserve large amounts of data at a much lower cost, which allows them to save money in operations.
- Efficiency: by establishing requirements applicable inside the network to facilitate communication and data transfer.
- Extensibility: to be able to accommodate to eventual changes the system might require over time, such as addition or failure of nodes and/or depositors.
- Security: in all its operations, allowing audit and report actions.

DPN is a unique project for long-term digital preservation because it provides a unified organization that is not technically nor organizationally centralized proving diversity in both areas. The use of several repositories, different geographic areas, different data storage protocols and systems makes it more likely to achieve preservation. However, if the system becomes bigger, this very same diversity can also be a problem in the future, especially when considering the brightening of large amount of content. Furthermore, while the system provides the intellectual information and technical requirements for brightening content, the highly restrictive agreements between nodes and depositors, especially regarding copyrighted material can be a huge problem; brightening, if specified by an institution, could take years. In terms of copyright, DPN presents itself as an extension of the depositors' repository. Looking it from the perspective of a Library, DPN is a friendly environment, because it provides technical infrastructure for long-term preservation including highly complex intellectual rights issues. After all "DPN is about preservation only".

Footnotes

[1] There are other collective projects to preserve digital content, such as HathiTrust (which is also part of DPN) and CLOCKSS (Controlled Lots of Copies Keep Stuff Safe).

[2] Some institutions have decided to take part of more than one initiative.

[3] DPN Wiki, available at https://wiki.duraspace.org/display/DPNC/Benefits

[4] *Digital Preservation: Saving the Scholarly Record Together*, Robin Ruggaber at the 7th International Conference on Open Repositories, liveblog available at http://or2012.ed.ac.uk/

[5] Architecture diagram on appendix.

[6] Advanced Message Queuing Protocol (AMQP) "*is an open standard for passing business messages between applications or organizations*". It has the ability to work with different platforms, available in different points in time and it can be operable remotely. More information available here http://www.amqp.org/about/what

[7] BagIt is a file packaging system created by the Library of Congress and the California Digital Library. It was designed to support the transfer of digital content within a network and also for disk-based storage. The "bag" only contains a "payload" and descriptive information known as "tags" but it does not require for it to know the structure of the payload. More information here http://www.digitalpreservation.gov/multimedia/videos/bagit0609.html

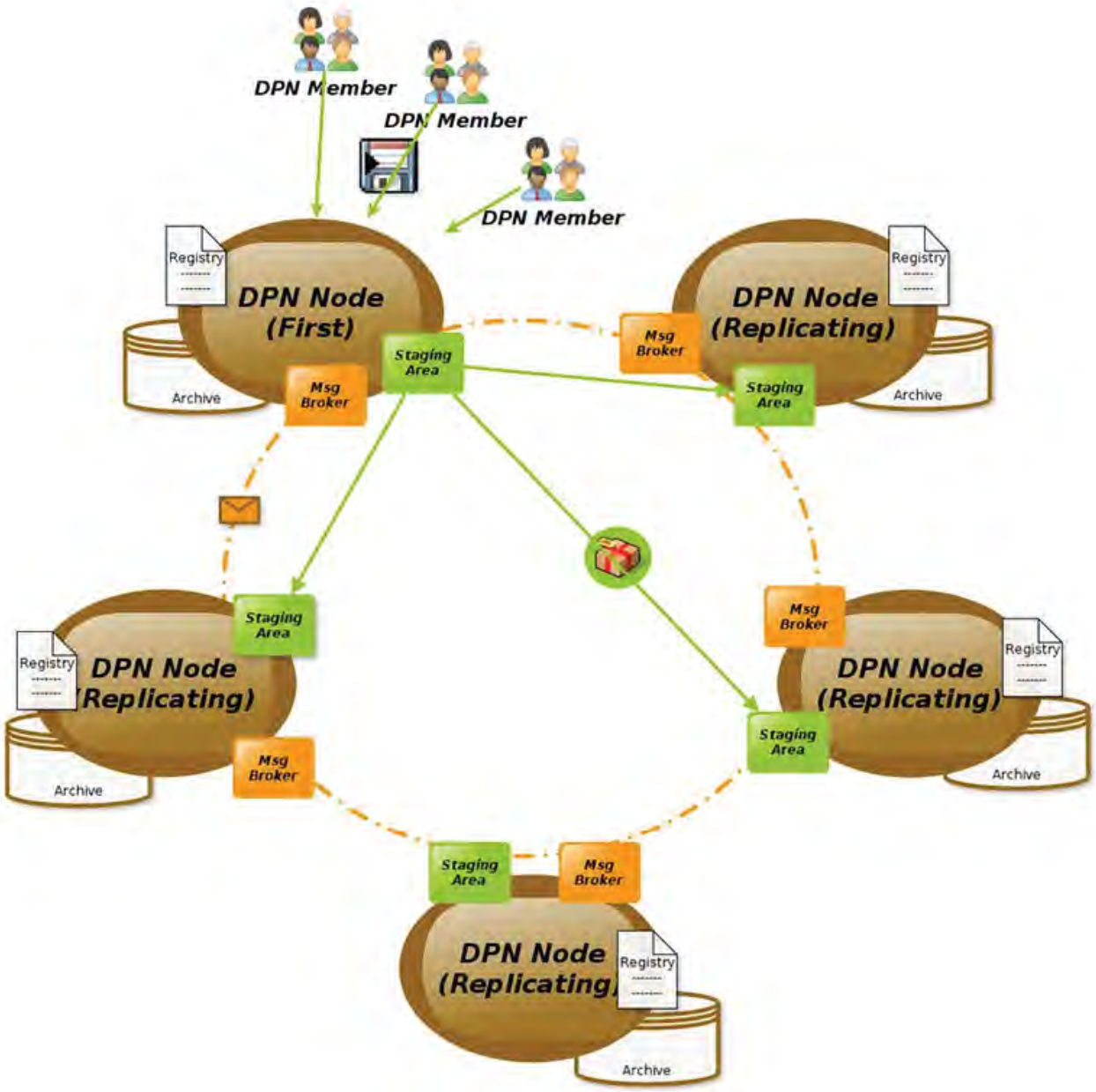
[8] Hypertext Transfer Protocol Secure (HTTPS) is a communication protocol derived from HTTP and SSL/TLS, to provide secure transfer over a network.

[9] rsync is a network and software protocol for Unix systems to synchronize files and directories. http://rsync.samba.org/

[10] *Succesion Rights and IP Working Group: Initial Report* available here https://wiki.duraspace.org/download/attachments/34639787/DPN+SRWG+Inital+Report.pdf?version=1&modificationDate=1379902384111

[11] Steven Morales, *DPN Member Technical Architecture Update*, March 20, 2013.

Appendix
DPN Architecture Diagram



Bibliography/webography

All websites last accessed on October 1, 2013.

DPN Neewsletter, May 2013
http://www.dpn.org/wp-content/uploads/2013/06/DPN_MAY_2013.pdf

Hilton, James, Cramer, Tom, Korner, Sebastien, Minor, David *"The Case for Building a Digital Preservation Network",* August 5, 2013.
http://www.educause.edu/ero/article/case-building-digital-preservation-network

Advanced Message Queuing Protocol Website
http://www.amqp.org/

Cramer, Tom, Simon, James, Video Presentation: "*The Digital Preservation Network: A Report and Discussion on DPN´s Emerging Architecture, System Protocol & Service Model"* at the Coalition for Networked Information
http://www.cni.org/topics/digital-preservation/the-digital-preservation-network-a-report-and-discussion-on-dpns-emerging-architecture-system-protocol-service-model/

Steven Morales, Video Presentation: *DPN Member Technical Architecture Update*, March 20, 2013.
https://meetings.webex.com/collabs/files/viewRecording?encryptData=0_F91505970EFE83957CE09F6686FFD1C56AA99B4968D7E04EAF2B2380858F54E1276110A6D4895729A6CD952AB22CDFDFFF5E986D274A79C7A2CAC197F3512EA29075156C1D80DEE16B5F1B8419C0AAB5_F4F482DDE14BD391C18A4811096FE700D2FFF6A9

HathiTrust Website
http://www.hathitrust.org/home

CLOCKSS Website
http://clockss.org/clockss/Home

LOCKSS Website
http://www.lockss.org/

Robin Ruggaber, *Digital Preservation: Saving the Scholarly Record Together*, at the 7th International Conference on Open Repositories, liveblog
http://or2012.ed.ac.uk/

Succesion Rights and IP Working Group: Initial Report
https://wiki.duraspace.org/download/attachments/34639787/DPN+SRWG+Inital+Report.pdf?version=1&modificationDate=1379902384111