DOI:10.1145/1859204.1859218

Peter J. Denning and Robert E. Kahn

The Profession of IT

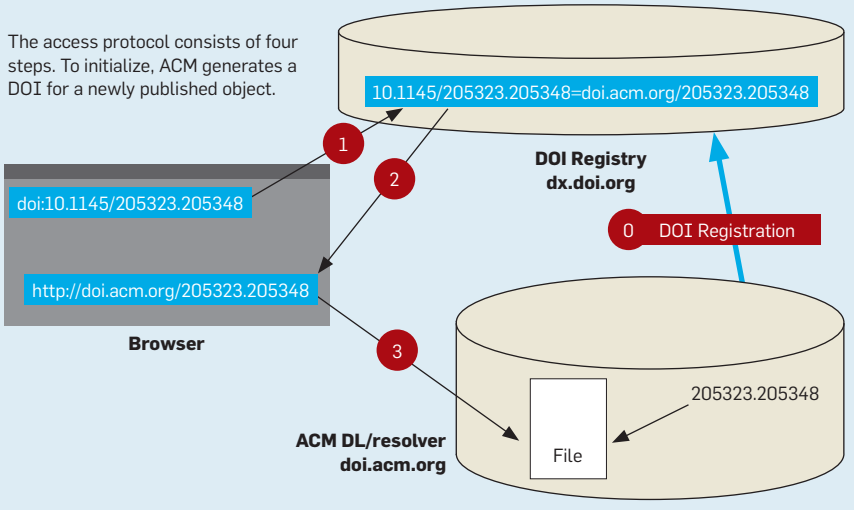
The Long Quest for Universal Information Access

Digital object repositories are on the cusp of resolving the long-standing problem of universal information access in the Internet.

Information sharing is an age-old objective. Always a challenge in the world of print documents and books, it has become truly daunting as music, movies, images, reports, designs, and other information entities are represented digitally and offered on the Internet. Numerous issues contribute to the complexity, such as file creation, formats, identifier systems (local and global), access controls, privacy controls, interoperability, searching, and rights management. The complexity is multiplied in the global Internet by the sheer amount of information available for access, the potential number of connections and reference chains, and jurisdictional issues. Reducing the complexity of information management in that universe has been a very long quest.

For the past 15 years a set of infrastructure technologies for "digital objects" has been gradually evolving in the Internet. They are now mature. We believe these technologies offer some real solutions to the conundrums of information sharing and access. We will summarize them and call attention to the significant user communities already employing them. We advocate that every IT professional become knowledgeable about these technologies and use them in the design of new systems and services.

Resolving a DOI to a URL or file name in the ACM Digital Library.



The DOI consists of ACM's unique number (10.1145) followed by a unique string chosen by ACM. ACM registers the DOI with the DOI registry (step 0). Thereafter a user can take the DOI from a citation and ask the registry to resolve it (step 1). The registry returns the URL of the object in the ACM Digital Library (step 2) . The ACM Digital Library resolver directs the access to the object specified by the DOI (step 3).

Early Attempts at Universal Information Access

Vannevar Bush is credited with the first visionary speculation about universal access to documents in 1945 ("As we may think," *Atlantic Monthly*). He proposed a hypothetical machine called Memex that stored documents on microfilm and allowed annotations and cross links. Many subsequent designers credit Bush with the inspirations for their work in computer networks.

The first among these designers was Ted Nelson, who, as early as 1960, proposed a networked system called Xanadu, which aimed at information sharing with a simple user interface. Nelson introduced topics such as hypertext, hyperlinks, automatic version management, automatic inclusion of referenced items, and small payments to authors for use of their materials.

In the middle 1960s, Doug Engelbart started the Augmentation Research Lab at SRI with the dream of amplifying collective intelligence through computer networks. Taking some inspirations from Bush and Nel-

son, he and his team developed NLS, the first working hypertext system with graphical user interface, mouse, and collaboration tools.

In the late 1980s, Tim Berners-Lee created the World Wide Web infrastructure to facilitate information access and as a potential means to implement many of Bush's, Nelson's, and Engelbart's ideas in the established worldwide Internet.

While these systems were major advances in the quest for universal information access, they left unsolved many of the difficult issues of information management noted earlier. Our objective here is to show that the Digital Object Architecture, which has been under development since the late 1980s at the Corporation for National Research Initiatives (CNRI) and is now reaching a tipping point, can provide the missing pieces of infrastructure and help us to attain universal information access.

Toward a Universal Address Space

A universal address space is the most fundamental element of a system of universal information access. A collection of ARPA-sponsored projects in the 1960s developed the first methods for doing this efficiently.

In a famous 1960 article "Man-computer symbiosis," J.C.R. Licklider expounded on the virtues of "intergalactic networks" and man-machine communications. Partly at his instigation, MIT's Project MAC undertook the construction of an operating system, Multics, which would be a "computer utility" that could dispense computing power and information access widely and cheaply. Other research organizations, such as IBM, UC Berkeley (Genie), and BBN (Tenex), were early developers of time-shared operating systems that could be networked.

Within the Multics system, information sharing was achieved by making virtual memory a common space accessible by every user on the system. The directory structure was an overlay that let users assign their own symbolic names to files; files themselves were addressed internally by their virtual addresses. Users never had to open or close files, or copy them from secondary storage to their workspaces; they simply referenced them as segments of address space.

A universal address space is the most fundamental element of a system of universal information access.

Jack Dennis, who helped design the Multics virtual memory, saw how to generalize it to allow dynamic sharing of a more general class of objects, not just files, and how to protect these objects from access not permitted by their owners. His "capability architecture" became the blueprint for object-oriented runtime systems.[2] That architecture inspired two commercial computing systems—Plessey 250 and IBM System 38—and two research projects—Cambridge CAP and CMU Hydra—that contributed much implementation knowledge. These projects all demonstrated that a large widely accessible address space would not only facilitate sharing, but it could be implemented efficiently on a single machine with a large shared file system. The capability architecture also provided a clean way of managing access and controlling rights by channeling all object references through a reference monitor protocol.[3]

The capability architecture easily generalized to homogeneous networks all running the same operating system. Unfortunately, it did not generalize well to the heterogeneous systems making up the Internet.

In 1970, the ARPANET expanded the universe of connected systems beyond a single machine to hundreds of time-shared computers. In the U.S. and around the world, other packet networks were developed in parallel with the ARPANET, including, in particular, a packet satellite network, a ground radio packet network, and various local area networks, such as token rings and Ethernet.

The Internet, introduced in 1973 by Bob Kahn and Vint Cerf, is a global information system composed of many networks and computational resources all embedded within a single large address space based on the Internet Protocol (IP) addresses. The domain name system, introduced in 1983, maps domain names to IP addresses, making it much easier for users to address computers.

Information sharing was accomplished in the Internet in various ways. In the early 1970s, the primary means were text-based email and file transfer, followed later by email attachments. A few research communities pioneered with collections of reports and software collections hosted on specific machines. Professional associations such as ACM built digital libraries to share their entire published literature. Services to back up files and provide drop points for transferring files became common.

In 1989, the Web became available as an architecture overlaid on the Internet to facilitate information sharing. The Web protocols automatically invoked a file transfer protocol with a single mouse click on a hyperlink. Hyperlinks used Uniform Resource Locators (URLs) to name objects; a URL consisted of a host name concatenated with the pathname of a file on the host file system.

The Web was an important step toward universal information sharing but it is far from perfect. Among other things, URLs are not persistent, the same URL can be reused for different purposes, and users frequently encounter broken links because files have been moved to new locations. Moreover, there are no rights-management protocols, and the only access controls are those for file access on target machines.

The Digital Object Architecture

Universal information access was not achieved in the Internet because none of the protocols was designed relative to information-sharing architecture principles. In the 1980s, prior to the development of the Web, as part of its work on digital libraries, CNRI started designing a system for enabling mobile programs (called "Knowbots") to carry out information-access tasks in a network environment. This led to the later formulation by CNRI of the Digital Object Architecture (DOA),[1,4,5] which culled out and unified four key

principles from the past projects on information access:

▸ Any unit of information represented in digital form may be structured as a digital object (DO) for access (with suitable controls) within the Internet. DOs may include digitized versions of text files, sounds, images, contracts and photos, as well as information embedded in RFID devices, chip designs, simulations, or genome codes. The structure of a DO, including its metadata, is machine and platform independent.

▸ Every DO has a unique persistent identifier, called a "handle," or generically, a "digital object identifier," that can distinguish a DO (or separately identified parts of it) from every other object, present, past, or future. Handles consist of a unique prefix allotted to an entity (such as a publisher or individual) followed by a string of symbols chosen by the entity. The "resolution" system maps handles to state information that includes location, authentication, rights specifications, allowed operations, and object attributes.

▸ DOs can be stored in DO Repositories, which are searchable systems that offer continuous access to objects over long time intervals that span technology generations.

▸ Accesses to an instance of DO Repository are made via a standard DO protocol that restricts actions to those consistent with an object's state information.

The primary components of CNRI's DOA design are the Handle System, DO Repositories, and DO Registries:

▸ The Handle System allots prefixes to registered administrators of local handle services and provides resolution services for their digital object identifiers. This system has been available as a service on the Internet since the middle 1990s and is highly reliable. It makes use of existing Internet protocols, which do not need redesign. Handle services also support domain name resolution for backward compatibility.

▸ The DO repositories use standard storage systems. They provide digital object management services with a standard protocol called digital object protocol (DOP). The DO repositories also support HTTP and DOP-over-TLS, a secure socket layer (SSL) service.

▸ The DO registries allow users to reference, federate, and otherwise manage collections across multiple repositories and allow for protected access to such information including completely private information, sharing within designated groups, and full public access.

The Web was an important step toward universal information sharing but it is far from perfect.

Examples of DOA in Use

The figure on the first page of this column shows the essence of a resolution of a DOI to a URL or a file name, using the ACM Digital Library as an example.

The International DOI Foundation (IDF) is a non-profit organization that administers DOIs for a variety of organizations, mostly publishers. The IDF has trademarked the DOI; and it uses the Handle System technology for resolution. One of the IDF Registration Agents (RAs), DataCite, manages large scientific data sets using DOIs. The largest of the IDF RAs, CrossRef, manages metadata on behalf of a large segment of the publishing industry.

The U.S. Library of Congress uses the Handle System to identify large parts of its collections. The U.S. Department of Defense (http://www.adlnet.gov) relies on the Handle System and DOI Registry and Repository to manage distributed learning material. The European Persistent Identifier Consortium (EPIC) (http://www.pidconsortium.eu) provides identifier services to the European research community. People in the legal community are implementing the DOA for wills, deeds to real property, bills of lading, and other legal instruments.

Conclusion

The quest for universal information access in networks began around 1960 and over the years yielded a set of principles to fully support universal information access. These principles include unique, persistent identifiers, protocols that map identifiers to objects (including their metadata), protocols for enforcing access and use rights, and repositories that hold objects for indefinite periods spanning technology generations. These principles offer a possible solution to universal information access and an infrastructure for more general information management.

The Digital Object Architecture was designed to exploit these principles without changing existing Internet protocols. The architecture is now widely used by publishers, digital libraries, government agencies, and many others to manage their collections and offer them for access over the Internet.

The Digital Object Architecture offers computing professionals an emerging infrastructure for managing information that goes a long way toward universal information access as well as system interoperability in the Internet. We advocate that every computing professional become familiar with these technologies and use them for new systems and applications. The Internet sites doregistry.org, dorepository.org, and handle.net contain tutorial materials and are a good place to start learning how these technologies work. C

References

1. Corporation for National Research Initiatives. *A Brief Overview of the Digital Object Architecture and its Application to Identification, Discovery, Resolution and Access to Information in Digital Form* (June 2010); http://www.cnri.reston.va.us/papers/Digital_Object_Architecture_Brief_Overview.pdf
2. Dennis, J.B.and Van Horn, E. Programming semantics for multiprogrammed computations. *Commun. ACM 9*, 3 (Mar. 1966), 143–155.
3. Graham, G.S. and Denning, P. Protection: Principles and practice. *AFIPS SJCC Conference* (May 1972), 417–429. DOI: 10.1145/1478873.1478928.
4. Kahn, R.E. and Lyons, P. Representing value as digital objects: A discussion of transferability and anonymity. *Journal of Telecommunications and High Technology Law 5* (2006).
5. Kahn, R.E. and Wilensky, R. A framework for distributed digital object services. *International Journal on Digital Libraries 6*, 2 (2006). DOI: 10.1007/s00799-005-0128-x. (First made available on the Internet in 1995 and reprinted in 2006 as part of a collection of seminal papers on digital libraries).

Peter J. Denning (pjd@nps.edu) is Distinguished Professor of Computer Science and Director of the Cebrowski Institute for Innovation and Information Superiority at the Naval Postgraduate School in Monterey, CA and is a past president of ACM.

Robert E. Kahn (rkahn@cnri.reston.va.us) is President and CEO of the Corporation for National Research Initiatives and a recipient of the ACM A.M.Turing Award.