In the last several decades there has been tremendous growth in the amount of digital content created by libraries, publishers, cultural institutions, and the general public in an effort to make content broadly accessible and useful. There are great benefits to having content available in digital form. However, unlike print objects—which, when they have been printed on acid-free paper and are held in reasonable conditions, can last for many decades with only minimal attention—digital objects may be extremely short-lived without proper attention to preservation. If you are reading this article on some sort of digital device using current software, what are the odds that twenty years from now you will be able to find it and read it with whatever device and software you will be using then? What will be the cost to locate and reproduce the original files in a format that is usable in twenty years?

Publishers, librarians, and scholars have begun to understand that their substantial and growing investment in digital object creation requires a commitment to protect the content for the long term. Although producers of digital content recognize that they need to be thinking about preservation, they may be asking themselves what steps they should be taking:

- What can we do to make a digital collection “safe enough”?
- If the IT department backs up the server, is that sufficient?
- If the high-resolution files are on an external drive in a staff member’s office, are we OK?
- If we make a tape backup every few months, are we covered?
- Do we need to work with a preservation service to ensure that our content is truly safe?

Answers to these questions depend on the type and needs of the content or collection, the content owners, the users, and the organization. However, these questions provide a useful starting point for considering various preservation options, which can be placed along a continuum that moves from near-term to mid-term to long-term preservation.

Near-Term Protection: Backup

When content is copied and stored in multiple locations to create readily available data replacements in case of equipment failure or another catastrophe, it is protected for near-term access. The proper backup of electronic assets is imperative for business continuity and is necessary to ensure that access to content will not be interrupted in the near term. A well-managed backup system can quickly resolve problems with content needed this week or next month, but it won’t help over the long term. Backup is typically implemented with commercial software, and often, content may be retrieved only via the software with which it was originally backed up. If special software or hardware is required to access the content, the long-term accessibility and authenticity of the content—key goals of digital preservation—cannot be ensured.

Mid-Term Protection: Byte Replication

In byte replication, multiple identical copies of files are created. The copies may be written to other online computers or to offline media. These replicas are typically held in diverse geographic locations, and specialized software is not needed to access the content. This diversity in copies and location, together with the lack of reliance on special software, means that byte replicas will provide content that is authentic and accessible for as long as the file formats remain usable. However, simple byte replication includes no provision for ensuring that the content is usable when the file formats are no longer current. Nor is there any inherent provision for guaranteeing that the content remains discoverable. For example, if a series of book files are byte-replicated without accessible bibliographic information describing the intellectual content of the replica, there is no assurance that a reader in the future will be able to find the specific content he or she needs. Further, if those replicated bytes are the encrypted version of the file created with digital rights management tools, or if they are in a format for which, in the future, there is no available reading device, those replicated items might not be accessible over the long term. In the case of scientific and other data sets, the lack of descriptive metadata would similarly make them less amenable to reuse (for duplication or verification of results) over the long term.

Long-Term Protection: Managed Digital Preservation

Managed digital preservation is defined as the establishment of management policies and activities that will ensure the
endurance of content over the very long term. Four goals are key to successful managed digital preservation:

- **Usability**: The intellectual content of the item must remain usable via the delivery mechanism of current technology.
- **Discoverability**: The content must have logical bibliographic metadata so that the content can be found by end users over time.
- **Authenticity**: The provenance of the content must be proven, and the content must be an authentic replica of the original as deposited.
- **Accessibility**: The content must be available for use by the appropriate community.

To successfully perform managed digital preservation, as defined here, an organization must have the following:

- A preservation mission that provides an environment conducive to the specialized planning and infrastructure needed to support digital preservation
- A sustainable economic model to support the preservation activities over the required time period
- Clear legal rights to preserve the content
- A relationship with the content provider and/or copyright owner
- Relationships with the users of the content, to ensure that their needs are met
- A preservation strategy and policies consistent with best practices, and a technological infrastructure that is able to support the selected preservation strategy
- Transparency with regard to its preservation services, strategies, customers, and content

It is important to note that backup (short-term protection) and byte replication (mid-term protection) are required elements of long-term preservation and are appropriate first steps in protecting content through preservation.

**What Is the Right Choice?**

An organization that is just beginning to contemplate and plan for long-term digital preservation may be able to take an incremental approach. The most important initial actions include (1) locating all content, (2) initiating regular backups, (3) conducting test retrieval from backups, and (4) developing a long-term preservation plan. Because of recently developed policies regarding the capturing, documenting, sharing, preserving, and disposing of research data resulting from publicly funded research, there is guidance on best practices for the preservation of data. These practices are applicable to the preservation of all digital artifacts and are directed at not only the researchers producing digital content but also their institutions. Such practices include creating persistent identifiers for all artifacts, both for identification and for citation; using well-understood formats; and clearly documenting data semantics, context (e.g., software tools used in the creation or rendering of artifacts), provenance at the time of creation, procedures for legal use and reuse of content (e.g., respect for privacy obligations), and policies for deaccessioning artifacts over time.

Organizations and individuals may develop the ability to perform long-term preservation themselves, they may develop this ability collectively, or they may partner with a third-party preservation service. An important starting point is to understand the key issues, what is at stake, and the options for moving forward with an effective preservation plan.

**Future Challenges**

Even though preserving current scholarly content is urgent, there are far more challenges ahead, driven by ongoing and rapid changes in scholarly communications. Content that comprises the scholarly record has become both more dynamic and less “bounded.” Formerly, even a digital artifact of the scholarly record consisted of a more or less discrete object, such as a journal article or book, often encapsulated in a single file or package. Increasingly, an artifact is likely to be a distributed, complex scholarly object. Its various components (e.g., article text, supporting data sets, and automated workflows from which the data was produced) can reside in more than one repository, in more than one version. For such objects to compose a scholarly record, and for that record to be preserved, we will need to create, capture, and maintain even more information about context and relationships than is provided in “classic” bibliographic metadata.

Looking ahead, we see at least two key technology challenges that will require active research in the preservation community. The first is the determination as to whether machine-learning and text-mining tools can be used to automate the collection of “good enough” bibliographic metadata (thus facilitating the automation of at least some of the work of preserving the “long tail” of small scholarly journals). The second challenge is the development of taxonomy of complex scholarly objects. This will allow an understanding of both the content and the processes/interactions surrounding that content, so that appropriate preservation actions (e.g., migration to one or more acceptable formats or establishment of an emulation infrastructure) can be taken. Ultimately, it is the responsibility of those who produce and care for valuable content to understand preservation options and take action to ensure that the scholarly record remains secure for future generations.