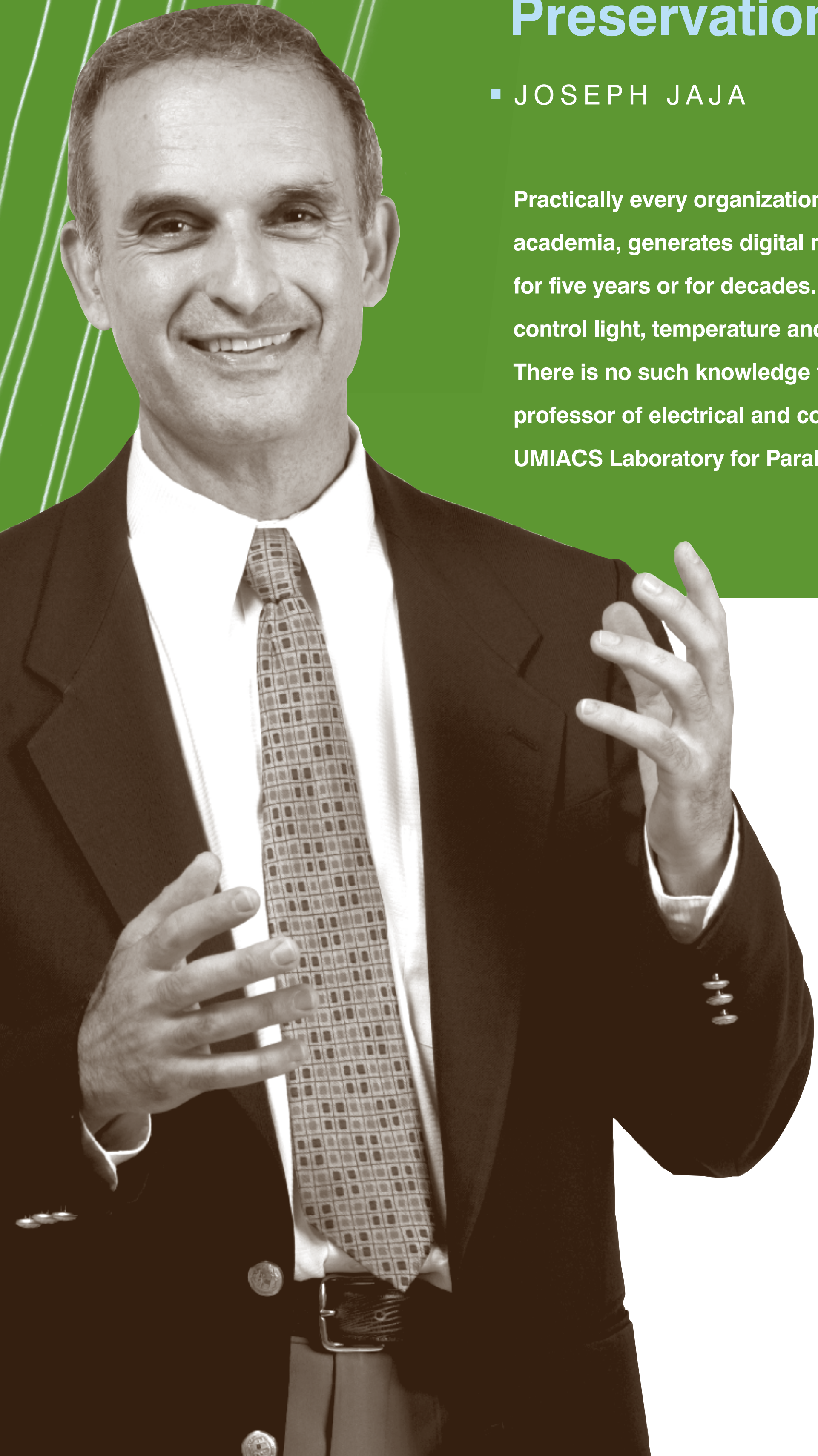
UMIACS ▪ ▪ ▪ ▪ University of Maryland Institute for Advanced Computer Studies



Enhancing Digital Preservation

▪ JOSEPH JAJA

Practically every organization, whether in government, business or academia, generates digital material that needs to be preserved, whether for five years or for decades. "Traditional archives have known how to control light, temperature and humidity to preserve physical artifacts. There is no such knowledge for electronic data," says Joseph JaJa, professor of electrical and computer engineering and a member of the UMIACS Laboratory for Parallel and Distributed Computing.

"Almost all records now are electronic. It's very easy to produce a lot of data electronically. The amount of data that needs to be preserved is growing extremely quickly," JaJa says. Since making copies of electronic data is quick and straightforward, it's easy to take its integrity for granted, but most existing digital storage systems have short lifetimes, and copies of data can easily be infected, changed or destroyed, especially when files are available on the Internet, where data is easily accessible to the public but particularly subject to security risks. Even power outages can cause data loss.

"Unfortunately, electronic materials are much more fragile than physical artifacts like paper, paintings and books," says JaJa. In addition, software is always

Enhancing Digital Preservation

▪ JOSEPH JAJA

To contact any researcher in UMIACS, go to **www.umiacs.umd.edu**.

changing, often making old files obsolete and inaccessible. The problem affects everyone from families who would like their grandchildren to see their digital photos to the National Archives, responsible for preserving the nation's historic records forever.

JaJa and his research group at UMIACS have been working on the development of technologies that will encapsulate essential characteristics of data in a way that is independent of platform and will evolve efficiently as new technologies emerge. "You need a layered architecture where you can modify one piece, instead of rewriting the whole software," JaJa explains.

Since 2000, JaJa has been working with the National Archives to develop methods for preserving and providing access to federal electronic records for the long term. In collaboration with the San Diego Supercomputing Center, the National Archives and Stanford University, JaJa's team has developed tools to ingest distributed data into an archive and automatically monitor the archive to ensure the authenticity of files and to manage format obsolescence. JaJa used his background in parallel and distributed computing to create a grid structure that prevents against data loss in the event of security breaches or operational errors. His group built hardware as well as software for the National Archives and continues to meet with colleagues there regularly.

JaJa's group also developed a new system for entering data for storage. The system allows the many different people who generate government records to submit their records easily. Many users can submit data at the same time, while providing context and annotation. While the information is temporarily stored on a "loading deck," automatic verification tools as well as human archivists can ensure that the information is entered properly. This system is in use at the National Archives. The Library of Congress is also considering using the system as its main way of ingesting data for its digital preservation program.

For his work with the National Archives creating better ways to store and access electronic records for the long term, JaJa was awarded the 2006 Internet 2 IDEA (Internet Driving Exemplary Applications) Award. Co-recipient Robert Chadduck, a computer engineer at the National Archives, calls JaJa and his UMIACS research team "more than invaluable colleagues."

JaJa is also considering better ways of storing electronic data—such as art, papers and videos—produced at the University of Maryland itself. The goal of archivists is not only to preserve the content of digital art, for example, but also its look and feel. Scientific data also needs to be safeguarded. Some data, for example about the atmosphere and the planet, will only become more valuable with time. In every case, technology for digital preservation needs to have the capacity to be scaled up, should be able to handle data in varied and changing formats and should have a system for ensuring the long-term authenticity of its records.

For expiring formats, JaJa sees the need to set up registries with conversion services that can alert archivists before a format expires. Tools should continually monitor and audit data, he says. Human archivists should be alerted whenever there is a discrepancy, for example because of a virus corrupting online data. The archivist can compare among stored copies to weed out damaged versions of files. Time stamps and other tools could be used to detect corruption by malicious users.

"There are many facets to the problem, including economics and politics," says JaJa. He leaves decisions about what should be stored and who should pay for it to others, focusing instead on creating the technology.

— Profile written by Karin Jegalian