Botanic Garden
© MNHN - François Grandin.

"Digitalization required a secure, cost-efficient and scalable storage system: this is Active Circle!"

Henri Michiels, Head of IT Services

The herbarium before digitalization
© MNHN - Bernard Faye

active circle

CUSTOMER STORY

MUSEUM OF NATURAL HISTORY

STORING THE IMAGES OF THE GREAT HERBARIUM

The Museum of Natural History

Created by royal decree in 1635, what was at first the royal botanical gardens became the Museum of Natural History in 1793. Widely recognized for its expertise, its missions include the conservation and enrichment of its exceptional collections, research, education, and sharing knowledge among the public. Two thousand people work at the museum, of which 500 are researchers. They are assisted in their tasks by the IT services department under the leadership of Henri Michiels. The department manages more than 2,000 workstations and 120 physical or virtual servers located in 3 computer rooms at the botanical gardens. The systems and the storage units are connected by fiber optic cable.

The digitalization of the herbarium pushes the storage needs

The storage needs start to increase significantly around 2004, when the task of digitalizing the specimens in the collections begins. This generates many images with sizes varying between a few megabytes up to almost 200 MB. The museum responds to this need by implementing a SAN fiber optic network supporting the Fibre Channel protocol, which connects the three computer rooms at the botanical gardens. The SAN storage bays and the servers are connected to this network. A secure backup system protects the data.

However, the start of the renovation of the Great Herbarium in 2009 puts a strain on this architecture. The herbarium, which dates from the 17th century, is the largest in the world together with the one in London, containing more than 10 million sheets in A3 format. The renovation project includes reconditioning the sheets, integrating pending specimens, reorganizing the collections and digitalizing them. This last task is considered key for effectively sharing knowledge. The generated data volume needed to be stored is estimated at 500 terabytes!

Extending the SAN is not possible

The initial plan is to store the data on the SAN, but a cost analysis shows that this solution is not economically viable in light of budget constraints and the estimated data volumes for the herbarium. "We quickly discarded a solution based on a SAN, since the cost per gigabyte is too high and it had to be duplicated using backups, incurring additional software and storage hardware costs" affirms Henri Michiels, IT director at the Museum of Natural History.

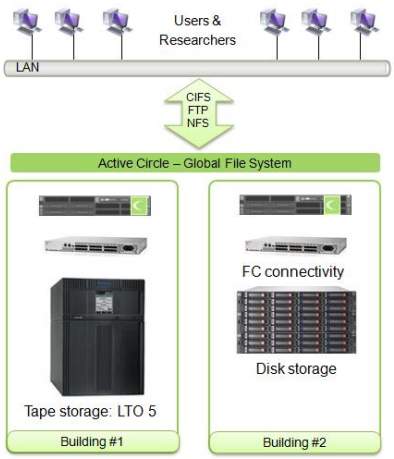
Seeking large storage space at low cost

The storage needs are plainly stated by the IT department: the data will not be accessed very frequently, but they need to be constantly available; some lag time is acceptable. The data need to be protected without having to resort to classic weekly backups; as such systems are not suitable for static data. Still, the security must not be compromised, since any loss of the original data is unacceptable when no copies exist! The infrastructure of the system must ensure long time preservation of the data at minimal purchase and operating cost, implying low energy consumption. In short, the need is for a durable, open, secure and inexpensive storage system, something which presents an equation that is difficult to resolve.

Herbarium's Gallery
© MNHN - Laurent Bessol

"One of the advantages of the solution is to secure the data without having to do backups. All new data are replicated and stored in two different physical locations. This automatic duplication in real time is the best solution"

Henri Pham – system administrator



The storage architecture

Rousseau's Herbarium
© MNHN - Bernard Faye

A merciless selection

After publishing the system specification in 2011, the museum receives various responses, but not all respond to their need. The solutions of the manufacturers are too costly and too proprietary, and they lack openness and scalability. The solution proposed by Active Circle stands out as open and non-proprietary, and it resolves the equation of storage capacity at lower cost, while guaranteeing data protection.

The Active Circle solution: An open and distributed HSM system

Active Circle proposes a solution at half the cost of the other solutions. It consists of software running on standard hardware, and provides a hierarchical file storage environment of the HSM type on disk and LTO tape. It provides continuous data protection, and it guarantees the longevity of the data due to the use of the TAR format. The solution is offered by the certified Active Circle partner CEFI as integrator, who has strong experience in deploying Active Circle solutions. CEFI has chosen hardware from Dell, which consists of an LTO-5 tape library and a disk bay. From these two storage elements, Active Circle creates a unified storage space. Each element is placed in separate buildings in order to protect the data against a disaster

Deployment and digitalization in parallel

The deployment of the solution and the digitalization is being done in parallel. The digitalization begins on a massive scale in 2010. The sheets are to be converted to digital in TIFF format, creating files of around 50 MB each. The files are compressed to 5 MB in JPEG format, which is the format used for preservation and consultation. The files in TIFF format are temporary, and are stored on external hard drives. The digitalization phase lasts two years and is completed in December 2012.

The deployment is carried out in 2012 with the installation of the hardware and software, followed by a test stage during the spring. The solution is found to comply with the specifications and goes into production. At the end of the summer, the first digitalized data are received by the Active Circle system.

The data are accessible

The general public and the researchers access the herbarium data using a web application which uses Active Circle's global file system. This provides the application with an extendable and secure file system. The access is transparent and takes only seconds for data on disk and minutes for data on LTO tapes. Data accessed on tape are copied to disk, making subsequent reads instantaneous. The access times are perfectly acceptable for reading in such a huge volume of data.

The data are secured in real time

"One of the advantages of the solution is to secure the data without having to do backups. All new data are replicated and stored in two different physical locations. This automatic duplication in real time is the best solution" emphasizes Henri Pham, the system administrator. Active Circle is in fact managing two physical storage locations as a unified storage space. For each logical file, there is a physical copy in each of the two storage locations. This is completely transparent to the users.

Storing for sharing the knowledge

Now that the Great Herbarium is digitalized, new needs are being expressed: "With this storage system, we look confidently at the future. We can respond when new requirements arise, like digitalizing in 3D using tomographic section methods and extending the digitalization to other collections. With digitalization, we make access easier for the general public and the researchers, and we fulfill our mission of sharing our knowledge. To this end, we need a storage system that is secure, cost-efficient and scalable: this is Active Circle!" affirms Henri Michiels.

Created in 2002, Active Circle develops software for organizations that manage large volumes of data: video content, images, scientific or technical data, or user information. The Active Circle solution optimizes data lifecycle management while at the same time simplifying storage administration and reducing total cost of ownership.

active circle