The Computerworld Honors Program

CASE STUDY

ORGANIZATION:

NZZ Neue Zürcher Zeitung AG

PROJECT NAME:

Archive 1780

LOCATION:
Zurich, Switzerland

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Media, Arts and Entertainment

NOMINATING COMPANY:
EMC

Summary

In 2005, Neue Zürcher Zeitung (NZZ), Switzerland's leading newspaper, created Archive 1780, a complete online archive of every page of every issue over the publication's 225-year history. The Archive 1780 team overcame major technological challenges in building the easily searchable archive of more than two million scanned newspaper pages, including text, illustrations, photographs and advertisements.

Available on NZZ's web site (www.nzz.ch), Archive 1780 gives the general public, including historians, journalists, students and other researchers, direct access to this unique and invaluable record. It provides a detailed, comprehensive insight into the social, political, economic and cultural life of Switzerland, Europe, and the wider world, not just in words but in images that convey information on major events as well as the "look and feel" of daily life over the centuries.

Introductory Overview

Zurich-based Neue Zürcher Zeitung is one of very few newspapers that can look back on 225 years of continuous publication. Of those, it may be unique in having a complete collection of every issue published. Prior to 2005, this extremely rare archive physically resided on more than 1,500 microfilms containing images of about two million newspaper pages. This film-based storage format made NZZ's extraordinary historical archive inaccessible to all but a few people and vulnerable to physical deterioration, loss or mishandling.

For the newspaper's 225th anniversary, NZZ decided to treat itself to a special present: digitizing the entire archive and making it available online with full-text search capability. "We had dreamed of doing this for a long time," said Rolf Brun, NZZ's Chief Information Officer. "Taking into account the time span covered and the technical procedure, we believe the Archive 1780 project is unique in the world."

From the outset of this massive effort, NZZ decided the traditional, manual-based approach of digitization was not acceptable and instead would automate as much of the multi-step process as possible. The first step was digitizing the microfilm images of all two million printed pages. It immediately became clear that manually scanning, "cleaning" and indexing all these images

ORGANIZATION:
NZZ Neue Zürcher Zeitung AG

PROJECT NAME:
Archiv 1780

LOCATION:
Zurich, Switzerland

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Media, Arts and Entertainment

NOMINATING COMPANY:
EMC

would be exorbitantly expensive and take 66,000 hours or dozens of man-years to complete.

In collaboration with two specialized technology and service providers, NZZ tackled the fully automated digitization of all 1,500 microfilms. While scanning the 35-mm microfilms was relatively straightforward, a number of major technical challenges arose. First, the digitization initially created very large, 80-megabyte data files per newspaper page. The size of these files had to be reduced for the project to be feasible. So the Archive 1780 team employed compression techniques to reduce them to a more manageable two to four megabytes per page. While much more compact, the resulting Archive 1780 database is a sizable 10 terabytes of content-addressed image data.

Another major technical hurdle at this stage related to some unique readability issues. The microfilm images were typically photos of pages bound into books. The text was often wavy, with the book spine clearly visible in the middle. Similarly, there was a wide variation in paper and print quality over the years. For example, after World War II, the newsprint was sometimes so thin that the reverse side of the page was visible on the microfilm. The NZZ team was able to use equalization and focusing without manual intervention to correct for these distortions on 95 percent of its microfilm library. The remaining 5 percent of pages were in such poor condition that they had to be manually corrected for digitization.

However, having cleanly scanned data was only half of the NZZ team's digitization job. They had to convert all the text into a machine-readable format in order to extract index data to make the archive fully searchable. And therein lay another unique challenge handed down from the past.

Prior to 1946, NZZ used the traditional German Gothic type; subsequently, the newspaper converted to the Antiqua font used today. However, most optical character recognition (OCR) programs do not "read" Gothic type. After an extensive search, the Archive 1780 team found that Abbyy FineReader XIX software did the job. Even then, making the text machine-readable took two minutes per page. In order for the process not to take years, NZZ and its technology collaborators implemented a Windows cluster of 20 PCs to handle the processing.

Simultaneous with the digitization process, NZZ and its partners designed, programmed and implemented a Web application for Archive 1780 access. Consistent with NZZ's IT policy, this application is based on open standards and open-source software. In addition, the archive is fully integrated with NZZ's enterprise storage infrastructure.

At the end of the conversion process, each newspaper page had been converted to a PDF with XML metadata for searchability. These PDFs comprised the 10-terabyte database of content-addressed data. Since Archive 1780 would consist of fixed, read-only data, a high-end storage system that supported read-write access was not necessary. Tapes and other removable storage media would also not be suitable, as online research requires fast, read-only access to the whole data pool.

The NZZ team met this challenge with EMC Centera, which provides for cost-effective archival storage, high availability, and fast access to large, fixed-content databases up to the petabyte range. This ensures that the publicly accessible archive will be able to deliver reliable, speedy access to its content at all times. In addition, the archive will have "room to grow" as both the size of the database and volume of traffic continue to expand.

Since its launch in September 2005, NZZ's Archive 1780 has become one of the most popular

THE COMPUTERWORLD HONORS PROGRAM

CASE STUDY

ORGANIZATION:
NZZ Neue Zürcher Zeitung AG

PROJECT NAME:
Archiv 1780

LOCATION:
Zurich, Switzerland

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Media, Arts and Entertainment

NOMINATING COMPANY:
EMC

features of NZZ Online with the number of users growing steadily each month. Most important, it gives the public an unprecedented window into recent events and the distant past. Unrivalled in its scope, Archive 1780 gives users fresh insight into major events, as well as the daily lives of people in Switzerland, Europe and the wider world for centuries past.

Benefits

There are two primary beneficiaries of Archive 1780: the people of Switzerland and the Neue Zürcher Zeitung newspaper itself, along with its parent company, NZZ Group.

First of all, the innovative and technological advanced solutions NZZ applied to the creation of the database dramatically sped up what could have been an exhaustive, manual-based process lasting decades. In fact, the technological and labor costs of the traditional approach of manually scanning two millions newspaper pages into a digital format would have been a barrier to the Archive 1780 ever becoming a reality.

Quickly and easily available via NZZ's web site, Archive 1780 gives anyone with Internet access the ability to see – not just read – every page of every issue of Switzerland's leading newspaper. Covering 225 years of continuous history, it provides a detailed, comprehensive insight into the social, political, economic and cultural life of Switzerland, Europe and the wider world, not just in words but in images that convey information on major events as well as the "look and feel" of daily life over the centuries.

Archive 1780 is an especially valuable resource for historians, students, journalists and other researchers, giving them a depth of detailed information and insight into local, national and international politics, issues and events. Previously, when the archived newspaper images resided only on microfilm, they were only available to NZZ staff and to other users upon request and approval.

At the same time, Archive 1780 was a vital part of Neue Zürcher Zeitung's expansion into online publishing and a range of new business ventures. With its high-profile launch in September 2005, the archive has been instrumental in driving online traffic to NZZ Web site.

This has been part of NZZ's ongoing growth into new media and communication ventures. These include the publication of "NZZ on Sunday," the monthly magazine "NZZ Folio," and the television program "NZZ Format." NZZ Group now includes regional newspapers as well as commercial printers and publishers across Switzerland.

Storing Archive 1780 on the EMC Centera system also has benefited both NZZ and users of the newspaper database. Because Centera utilizes breakthrough C-Clip technology, information is stored so it is unchangeable and secure, ensuring users of the authenticity of the information they retrieve. Centera provides users with sub-second response to their requests to search through hundreds of thousands of pages, making research and other projects more efficient. In addition, users enjoy reliable, 24x7 access to the Archive 1780 because Centera's self-healing architecture continuously monitors the integrity of stored objects and automatically detects and repairs software errors.

Because of all of these capabilities, the NZZ IT team has found Centera extremely easy to manage and scale to larger capacities as the Archive 1780 continues to grow with each publication day. In fact, the more information added to the Centera, price per megabyte actually decreases,

ORGANIZATION:
NZZ Neue Zürcher Zeitung AG

PROJECT NAME:
Archiv 1780

LOCATION:
Zurich, Switzerland

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Media, Arts and Entertainment

NOMINATING COMPANY:
EMC

generating even greater cost efficiencies for NZZ in the future.

The Importance of Technology

From the outset, NZZ had four primary technology goals for Archive 1780:

•Cost-effective development of online access

•Full-text search capability spanning the entire archive

•Centralized management of content-addressed database

•Integration of the archive with NZZ's existing storage infrastructure.

NZZ achieved these goals through the development of an automated, digitization process and selection of key information technologies ideally suited to its truly unique needs.

NZZ recognized the traditional process of manually scanning its two-million page microfilm library into a digital format would take dozens of man-years or 66,000 hours to complete. Committed to finding a fully automated way to digitize its vast microfilm library, NZZ turned to Germany's Fraunhofer Society, a vast network of research organizations, and its Institute for Media Communication (IMK). With expertise gained from having created a multimedia Beethoven archive, IMK offered NZZ its support in developing a comprehensive and automated digitization process for the Archive 1780 project.

In 2004, NZZ asked IMK's NetMedia competency center and the specialized ScanPlex service provider to undertake the automated digitization of its 1,500 microfilms. They applied compression techniques that reduced the size of each scanned newspaper page from 80 megabytes to two to four megabytes. Without this, implementing Archive 1780 would not have been feasible. Similarly, the technicians used equalization and focusing to convert 95 percent of NZZ's microfilm library without manual intervention. As a result, what would have taken tens of thousands of hours was accomplished in approximately two months.

Once it had created the digital image files, the NZZ team employed Optical Character Recognition (OCR) to convert the text into machine-readable format. This was necessary in order to index the archive and make it searchable. That's when the team discovered that most OCR programs do not read Gothic type, the standard font used by NZZ for about 160 of its 225 years. Ultimately, they found the Abbyy FineReader XIX program, with which they overcame this problem.

In keeping with NZZ's IT policy, the Archive 1780 developers focused on implementing open standards and open-source technology. That included selecting MaxDB, a robust, open-source database, as the software platform for Archive 1780 that integrated with NZZ's SAP application environment. In building the Web application for accessing the archive, the team made sure it integrated with the Apache Lenya Java/XML open-source content management system used by NZZ.

A key component in making Archive 1780 feasible from both a technology and business perspective is EMC Centera, a data storage system designed for archiving fixed, content-addressed data. Further, NZZ recognized it did not need read-write access and frequent backup capabilities provided by higher-end storage systems for essentially static, read-only data characteristic of Archive 1780. Therefore, it chose a more cost-effective, object-oriented storage system that

ORGANIZATION:
NZZ Neue Zürcher Zeitung AG

PROJECT NAME:
Archiv 1780

LOCATION:
Zurich, Switzerland

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Media, Arts and Entertainment

NOMINATING COMPANY:
EMC

also provided centralized data management and ensured maximum integrity of archives over the long term.

Originality

Archive 1780 and the development process that created it epitomize innovation and originality on many different levels. The nature of the online archive itself is unique: there is no other comprehensive, Web-accessible newspaper database of all issues published over the span of 225 years. After considerable research, NZZ knows of no comparable online archive. Rolf Brun, NZZ's Chief Information Officer, noted, "Regarding the time span covered and the technical procedure, we believe this project is unique in the world,"

Creating a searchable online archive of the size and scope of Archive 1780 was extremely ambitious and innovative. NZZ and its technology collaborators truly "pushed the technology envelope" in overcoming numerous technology challenges in making Archive 1780 a reality. These include:

•Automating the digitization of more than two million microfilm images,

•Employing compression, equalization and focusing techniques to overcome issues related to the size of data files and the legibility of extremely old, often distorted text and images,

•Overcoming challenges to machine-readability based on differences in typefaces over the years,

•Indexing the entire Archive 1780 database, enabling users to easily search for, identify, and access specific topics and publications by date and topic,

•Developing the open-source Web application to make this accessibility possible,

•Designing, developing and implementing an archival, cost-effective content-addressed storage system for read-only information and access.

Success

Thanks to the collaborative efforts of NZZ's IT team and its collaborators, the Archive 1780 was successfully launched in September 2005. The archive has achieved all the goals defined by NZZ at the outset of the project. It delivers fast, easy access for any Internet user to Switzerland's political, social, economic and cultural life over the past 225 years. In a sense, it enables archive users a chance to "eavesdrop" on the views and values of people over the centuries.

People who have successfully used Archive 1780 to find information and expand their cultural awareness range from casual users to professional researchers, academics and others. The archive makes it easy for NZZ journalists to read the newspaper's previous coverage on topics to gain insight and background information. In addition, Archive 1780 lets general public users do "recreational" searches such as finding the NZZ issue that was published on their birth date.

From a business perspective, Archive 1780 has firmly established NZZ as a technology leader among print publications that have successfully expanded into the online arena. Seizing the opportunity presented by its 225th anniversary, NZZ pursued the Archive 1780 initiative to raise its profile among its readership and the general public. It gave NZZ a new, revenue-generating profit center out of what had previously been an incredibly valuable but hidden, under-utilized

ORGANIZATION:
NZZ Neue Zürcher Zeitung AG

PROJECT NAME:
Archiv 1780

LOCATION:
Zurich, Switzerland

YEAR:
2006

STATUS:
Laureate

CATEGORY:
Media, Arts and Entertainment

NOMINATING COMPANY:
EMC

resource. By attracting more visitors to the NZZ Online Web site, Archive 1780 has proved instrumental in boosting the profitability of its online presence.

The Archive 1780 project successfully employed a wide range of technologies, including:

•Fully automated digitization of its vast microfilm library,

•Processing of image data without -- or with a minimum of -- manual assistance,

•Implementation of OCR and full-text indexing of microfilms,

•Development of the client-specific Web application,

•Integration of the Archive 1780 interface into the archival storage system.

In the end, the Archive 1780 project is an outstanding example of how a content provider and outside service providers can work together to turn a technology vision into reality.

Difficulty

With enthusiastic, supportive senior management at NZZ, the challenges faced by the Archive 1780 team were exclusively technical ones. One of the biggest hurdles was due to the sheer volume of two million microfilm images that had to be digitized. Manually scanning, "cleaning" and indexing all these images would have been exorbitantly expensive and taken dozens of man-years to complete. The team addressed this by automating the digitizing process, which reduced the time involved from what would have been more than 66,000 hours to about two months.

The automation process itself involved a number of technological difficulties, which NZZ and its development partners, resolved. The first was discovering that the digitization initially created very large, 80-megabyte data files per newspaper page. The size of these files had to be reduced for the project to be feasible. So the Archive 1780 team employed compression techniques to reduce them to a more manageable two to four megabytes per page. While much more compact, the resulting Archive 1780 database is a sizable 10 terabytes of content-addressed image data.

Another major technical hurdle at this stage related to some unique readability issues. The microfilm images were typically photos of pages bound into books. The text was often wavy, with the book spine clearly visible in the middle. Similarly, there was a wide variation in paper and print quality over the years. For example, after World War II, the newsprint was sometimes so thin that the reverse side of the page was visible on the microfilm. The NZZ team was able to use equalization and focusing without manual intervention to correct for these distortions on 95 percent of its microfilm library. The remaining five percent of pages were in such poor condition that they had to be manually corrected for digitization.

Once this data was cleanly scanned, the Archive 1780 team faced a new challenge in enabling OCR technology to "read" the content-addressed data. Prior to 1946, the newspaper used the traditional German Gothic type; subsequently, they converted to the Antiqua font used today. The team found that most optical character recognition (OCR) programs do not "read" Gothic type. After an extensive search, they found Abbyy FineReader XIX software, which did the job. Even then, making the text machine-readable took two minutes per page. In order for the process not to take years, NZZ and its technology collaborators implemented a Windows cluster of 20 PCs to handle the processing.