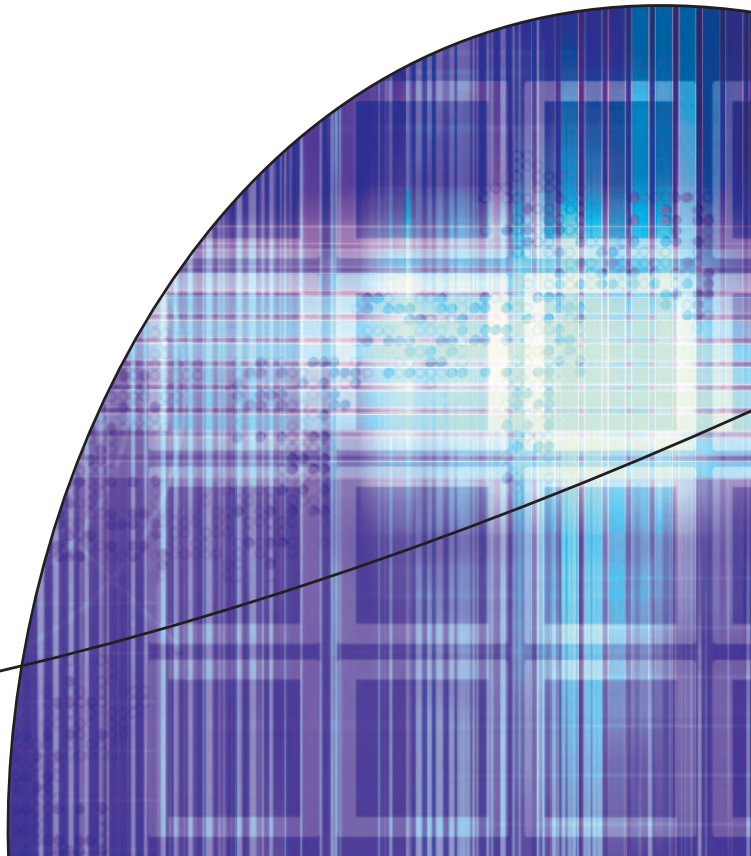
HITACHI
DATA SYSTEMS

Active Archive

Solutions Brief

A Blueprint for Long-term Preservation
of Business-critical Digital Assets

Partner Beyond Technology



Executive Summary

In 2002, the U.S. Securities and Exchange Commission (SEC) found Deutsche Bank Securities, Goldman Sachs, Morgan Stanley, Salomon Smith Barney, and U.S. Bancorp Piper Jaffray in violation of SEC Rule 17a-4. For failing to preserve e-mail communications and other business documents, the companies were fined a total of US$8.25 million.

During the trial on antitrust charges, the United States District Court in Baltimore demanded that Microsoft produce old e-mail messages to be introduced into court evidence. Although the cost and inconvenience of searching 25,000 e-mail backup tapes was substantial, Microsoft had to bear the burden, knowing that, if the e-mail messages were not produced, the judge could instruct the jury to make assumptions unfavorable to the company.

Digital documents, such as e-mail, are now receiving significant attention from litigators, government and industry regulators, and, through necessity, corporate executives. Maintaining these fixed-content documents in their original form, and providing quick retrieval when needed, demands a new approach to long-term data storage.

To satisfy evidentiary and compliance mandates, records must be protected from harm and safeguarded against unauthorized access and manipulation. This requires a combination of hardware and software that provides immutable storage, secure authentication, and audited access, which can satisfy legal and regulatory mandates for guarantees of a record's authenticity.

Long-term Data Preservation

Data is, by nature, transitory. The ones and zeroes that record the rise and fall of a stock price or the value of goods in a digital ledger are subject to constant modification by applications, and deletion at the click of a mouse. But not all data can be allowed to change or disappear so easily. Many organizations are now required to preserve digital records for much longer than originally intended.

Like other items of data, e-mail, corporate documents, medical, legal, financial records, digital audio, video, and images are hosted on conventional storage systems. But, unlike the working data in the transactional databases of the organization, these relatively new forms of digital information have fixed content.

Fixed Content

Fixed content means that an item reflects a particular real-world event that happened at a point in time—for example, an X-ray image of a broken arm, an e-mail message, a completed digital video, or a filing made to a government agency. For the item to remain valuable in the future, its content must remain fixed to accurately reflect the original state. For most organizations, fixed-content data comprises the bulk of all data storage needs. Although industry analyst estimates vary, fixed content is thought to consume

Figure 1: *Various types of digital assets require archiving.*

between 80 percent and 90 percent of all storage capacity. Fixed content is also being created faster than traditional information assets. This is driving demand for long-term archive storage capacity (see Figure 1).

Why Retain Fixed-content Assets?

Prior to the use of digital records, physical fixed-content items were indexed and filed, providing a permanent archive that could be accessed when needed. However, digital fixed-content files frequently have no physical counterpart. Maintaining a permanent record of digital fixed-content items is becoming a priority for many organizations.

The list of reasons for retaining fixed-content assets grows continually. Not only do these digital assets provide a historical record for the organization, but also regulators and government agencies now stipulate data retention and protection best practices for a wide range of industries. Organizations ignoring these requirements face stiff penalties and legal consequences.

E-mail is now considered a standard source of evidence in legal proceedings. Searches through e-mail archives are a routine first step in any legal discovery process. Organizations that do not have ready access to historical e-mail messages are handicapped during litigation, with the disruption and expense of manually searching through tape backups to satisfy the discovery process sometimes outweighing the cost of a settlement.

For industries that have digitized existing workflows, such as broadcasting and health care, the enormous amount of digital information produced daily is forcing IT to rethink conventional storage architectures.

Although many of the digital assets produced are needed immediately, their use becomes less frequent over time. Storing all of these assets on top-of-the-line storage systems is not a cost-effective approach.

The Challenge of Digital Archival

Today's IT infrastructures were not designed with long-term data archival in mind. Information on conventional storage systems is invariably considered working data, and it is only archived to tape or optical media as part of routine backup processing. Unfortunately, tape and optical media do not satisfy the requirements of a modern active archive. These media are cumbersome to access, slow to search, and have no means to guarantee preservation of fixed-content records over long periods of time.

What Would a Modern Digital Archive Look Like?

Rather than reinvent the wheel, IT planners can refer to substantial existing research on the topic of records management and archiving. In the physical world, the long-term preservation of important records and artifacts is the domain of library sciences and archiving. Practitioners in these diciplines have given much thought to the problem of data preservation and have developed several general theories and best practices.

Traditional archives provide a repository for records the organization has selected to preserve. The archive serves two fundamental goals: records must be preserved unchanged, and records must be easily accessed. These properties apply equally well to archives containing fixed-content digital assets.

Loading Information to the Archive

Fixed-content digital assets are created by a wide variety of enterprise applications. The digital archive loading process allows multiple applications to stream data into the archive simultaneously for long-term preservation and storage.

The process of loading also allocates metadata to each fixed-content record. Indexes, using the metadata, speed the search and retrieval of archived items. The loading process also associates a retention period with the content. Retention periods ensure content cannot be changed or deleted from the archive until a predetermined period of time has passed. In a traditional archive, the loading of records also represents a physical and legal handing over of custody of the content.

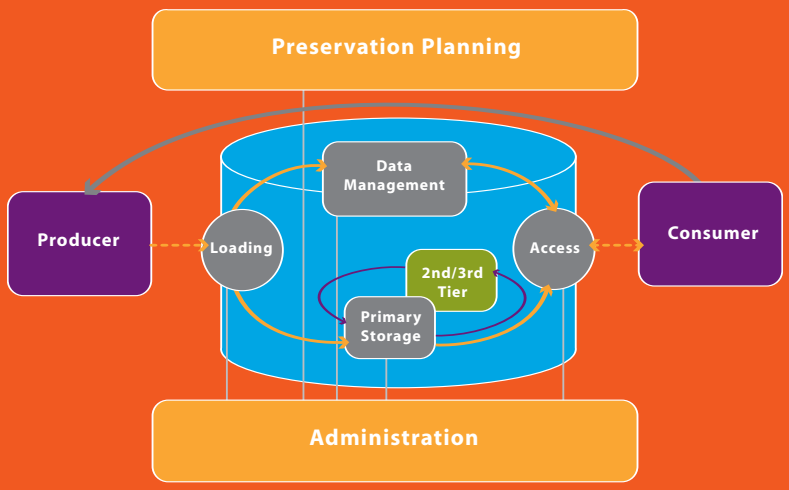
Authenticity and the Chain of Custody

The contents of a digital archive are only of value if authenticity of each record is guaranteed. In the world of library sciences, the term provenance refers to the ability to vouch for the origin and chain of custody of an archived record. An item can only be certified as authentic if the chain of custody is unbroken, which means that the item has always been securely managed since it was loaded into the archive. If an item's chain of custody cannot be proven, its reliability as evidence is significantly reduced.

Digital records face the same provenance problems as traditional archived items. For example, an auditor assessing an organization's compliance with the Sarbanes-Oxley Act is unlikely to vouch for the authenticity of archived e-mail records if it cannot be proven that the content has not been tampered with.

The OAIS Reference Model for Content Archival

The Hitachi Content Archive Platform supports a common set of archival functionality outlined in the ISO Reference Model for Open Archival Information Systems (OAIS). The OAIS model has been used as a foundation for some of the largest archives ever designed, including those managed by the U.S. National Archive and Records Administration and the gargantuan Planetary Data System Archive managed by the National Aeronautics and Space Administration (NASA).



Sidebar Figure 1: *The Hitachi Data Systems Common Model for Active Archive solutions leverages the ISO Reference Model for Open Archival Information Systems with modifications to describe an active content flow.*

The OAIS model is also extensible, allowing the specialized needs of individual implementations to be satisfied. The Hitachi Data Systems implementation of the OAIS model defines the processes for managing an active archive that supports the information lifecycle.

Access
To access archived data assets, consumers must have sufficient rights and security privileges to execute queries and make retrieval requests.

Added Value
The ability to repurpose and reuse archival content in new and innovative applications can enhance the value of existing data assets.

Administration Layer
The administration layer provides an interface for managing all components in the archive solution, including retrieval requests.

Archival Storage
The original data asset is maintained in the storage infrastructure of the archive. This can mean storing data on online disk, nearline disk, tape, or any combination of these media.

Consumer
The OAIS model defines the consumer as an end user or an application capable of searching the index and requesting data assets for retrieval.

Data Management
Metadata associated with the archived content is maintained in a data management system. The system allows users to manage policies associated with the archived content and to query and retrieve data using searchable indexes.

Loading
The loading mechanism indexes relevant metadata from the unstructured data assets and establishes authenticity by generating a unique identifier or signature, which is sometimes referred to as a hash.

Preservation Planning
During preservation planning, policies that control management of data assets within the archive environment are defined. For example, the value and policies associated with a piece of information determine whether it is maintained on primary disk and then migrated to second- or third-tier storage.

Producer
A producer of data assets is any entity that submits data to the archive. Examples include applications, e-mail servers, and medical imaging systems.

During the loading process, e-mail, file systems, and databases publish content to the digital archive. Once loaded, a preservation layer authenticates the digital records and stores them in a "write once, read many" (WORM) file system. This prevents unauthorized modification or deletion of the archived content.

Preserving Digital Content

To guarantee the authenticity and accessibility of stored digital records many years after they were originally created, every component of the archive must be capable of being upgraded as new technologies become available. For example, encryption algorithms used today will likely be useless five or ten years from now, as faster computer processors allow existing routines to be cracked.

The preservation layer of a digital archive ensures that stored content is available for access using technology that is current at the time the search is conducted. This can be a tremendous challenge when data formats and application versions are continually changing. An essential feature of the digital archive is the ability to upgrade hardware, software, data format, and encryption routines without jeopardizing the chain of custody of fixed-content assets. This also means that before-and-after copies of each data transformation must be kept as an audit trail, so that subsequent researchers can determine if any unauthorized changes took place during conversion.

Search and Retrieval

The ability to successfully search and retrieve stored records is a key function of an archive. If records cannot be located they are effectively lost, and all attempts to preserve the content are wasted. A digital archive is almost certain to contain many millions of records, and, therefore, efficient, high-performance search and retrieval of information is critical.

Whether an archive is conventional or digital, the curators and managers of the information face a common problem: it is almost impossible to anticipate how future users of the archive will want to query the information. Some users will know exactly what information they are looking for and will be able to use specific indexes to locate items. Other users will need to browse the archive looking for relevant information. Digital archives can also support data-mining technology, allowing researchers to gain insights into archived content unavailable through other search and retrieval mechanisms.

Metadata provides a source for indexes that allow the flexible search for content in a digital archive. Support for multiple access protocols also allows different applications to act as a front end to the search and retrieval process. The use of open protocols guarantees records in the archive will always be accessible, no matter how search and retrieval technology changes.

Tiered Storage and Archiving

The recent escalation in demand for enterprise storage capacity has lead many IT organizations to consolidate and centralize storage resources. Storage area network (SAN) and network attached storage (NAS) technologies have produced significant efficiency improvements, allowing administrators to manage much more capacity using a common set of storage management tools and procedures.

In these new consolidated environments, IT planners favor storage solutions that integrate seamlessly with the existing infrastructure. Solutions that require a unique management and configuration approach lower administrator productivity, reduce efficiency, and raise the overall cost of storage.

Considering a digital archive's potential to consume enormous amounts of storage capacity, it is imperative that the archive integrates seamlessly into an existing IT infrastructure. This is the only way to ensure that the archive remains cost-effective as it scales. Support for storage networking and common storage management tools will allow an archive to be treated as simply another tier in a multitiered pool of efficiently managed storage.

Hitachi Content Archive Platform

The Hitachi Content Archive Platform is a new approach to long-term fixed-content data preservation. Designed to seamlessly integrate into an existing enterprise storage infrastructure, the high-performance, high-availability, highly scalable archiving solution satisfies an organization's regulatory compliance requirement by ensuring the secure, long-term preservation and fast search and retrieval of valuable business records (see Figure 2).

A first in SAN-based digital archival solutions, the Content Archive Platform uses world-class Hitachi storage systems to provide scalability, availability, and performance, satisfying the growing demand for long-term fixed-content storage. With built-in authentication, protection, and retention capabilities, the highly available platform guarantees archived content will be continually available for access for years to come.

The Hitachi Content Archive Platform provides:

- WORM file system and time-based retention at the object level
- Content authentication with digital signatures
- User-selectable MD5, SHA1, SHA256, SHA512
- Embedded full-text index, search, and retrieval for content discovery
- Standard file system access, for browsable view of archived content
- Simple integration methods
- Standards-based interfaces—NFS, CIFS, HTTP, WebDAV
- Support for multiple applications on a common archive

Built around a mainstream computing platform, deployed with a cluster of inexpensive nodes, and backed by Hitachi storage systems, the Content Archive Platform dramatically reduces the burden of managing archival fixed-content storage. The solution leverages existing storage management and business continuity processes from Hitachi Data Systems, including seamless movement of data between the archive tiers. Using management software common to the entire storage environment lessens the training burden, cuts deployment time, streamlines information flow, and improves administrator productivity. The archive supports industry-standard documents and formats, including structured, semi-structured, and unstructured data, guaranteeing future archival and retrieval applications access to stored information.

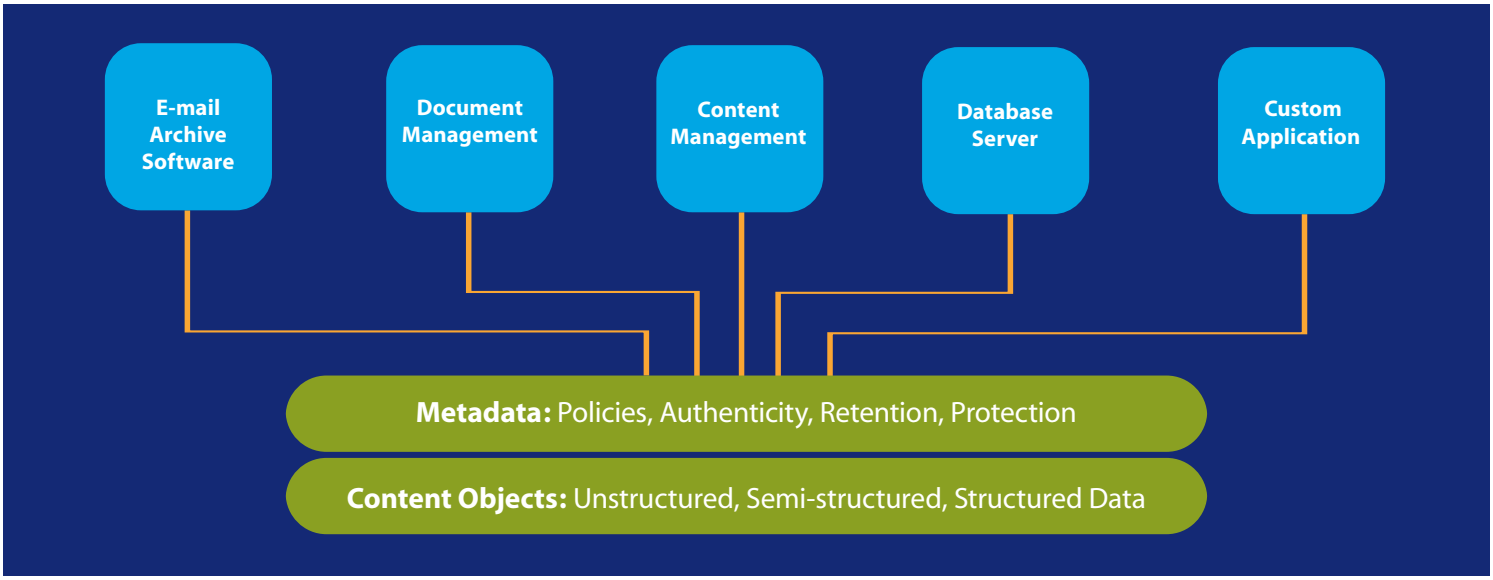


Figure 2: *The Hitachi Content Archive Platform creates an Active Archive.*

The Hitachi Content Platform delivers clear business benefits:

- Regulatory compliance
- Standards-based architectural model
- Simplified data storage and retrieval
- Integrated archive management within the storage environment

Application Optimized Storage™ Solutions: Aligning Business and IT Objectives

Application Optimized Storage™ solutions from Hitachi Data Systems provide an integrated approach to developing a storage infrastructure based on business requirements rather than technology features. The solutions are founded on a multitier, heterogeneous storage infrastructure supported by application, content, data, and storage services.

Application Optimized Storage solutions allow the storage infrastructure to respond to the specific performance, availability, functionality, and cost requirements of each application. By delivering capacity and services based on the specific needs of the business, Application Optimized Storage solutions substantially lower the overall cost of the storage infrastructure.

The Content Archive Platform takes Application Optimized Storage solutions to a new level, satisfying business application demand for high-performance, scalable, and highly available fixed-content archival storage. Providing content services layer functionality, the Content Archive Platform fully interoperates with application, data, and storage services for seamless integration into an existing infrastructure.

Building an Active Archive Business Case

This solutions brief provides a first step on the road to understanding the implications and challenges of ensuring the secure, long-term preservation of fixed-content digital assets. The information offered here illustrates not only the difficulty of providing archiving services, but also the substantial benefits that accrue from implementing a comprehensive solution based on proven storage technology. After reading this brief you will be able to build a business case for addressing the challenges of long-term data archival.

Although reading is a necessary first step in understanding the ramifications of a comprehensive archival solution, Hitachi Data Systems strongly recommends that you engage our Global Solution Services group before seriously undertaking a long-term archival strategy. Global Solution Services can help you to design and implement the optimal Content Archive Platform to meet your business and application requirements.

Implementing the Hitachi Content Archive Platform

The Implementation Service for Hitachi Content Archive Platform helps you implement an active archive in your IT environment. Global Solution Services consultants will undertake all associated project management responsibilities and install and configure one Content Archive Platform and, if requested, coordinate third-party resources for the configuration of any supported independent software vendor (ISV) products. This service provides you with a fast fixed-content active archive implementation and addresses the need of IT project leaders for solid workflow management of complex IT tasks.

To learn more about how Hitachi Data Systems can help with your digital archive storage plans and to read more about active archiving, please visit www.hds.com or call Hitachi Data Systems at 888 234 5601, ext. 950, to explore an engagement that will result in the optimal solution for your digital archival storage needs. (Hitachi TrueNorth™ Channel partners should contact their Channel managers for information.)

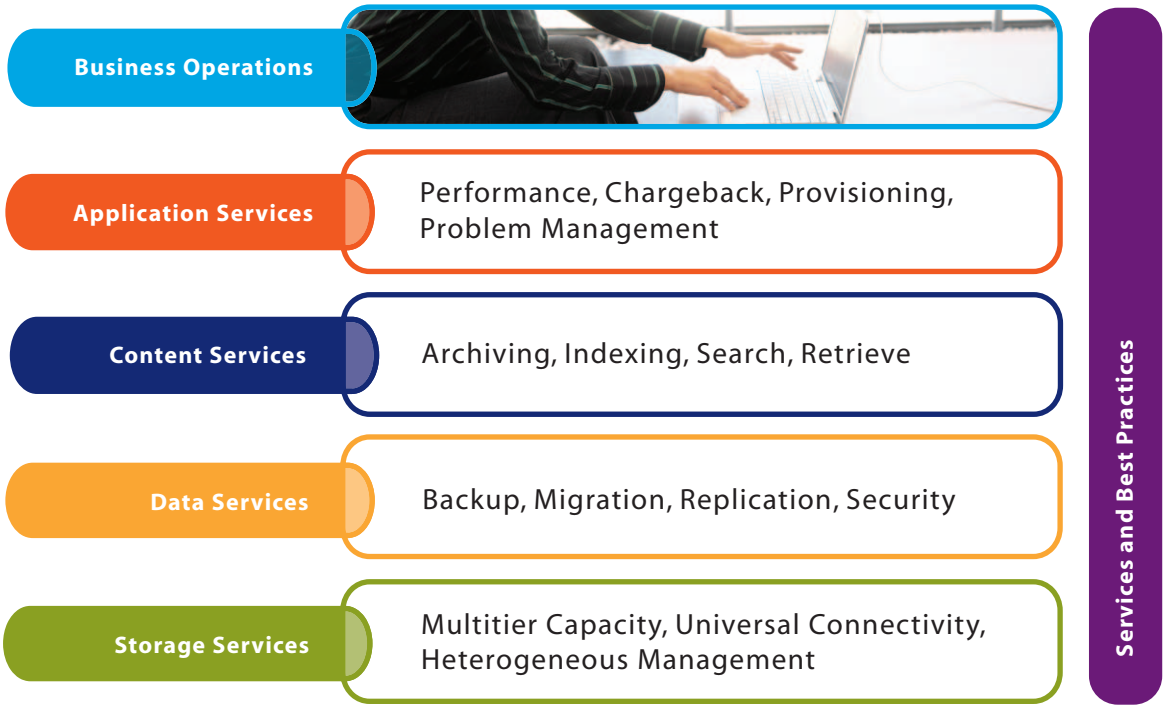


Figure 3: *Each Application Optimized Storage solution is built on a framework for aligning business and IT objectives.*

Hitachi Content Archive Platform

Software	Hitachi Content Archiver, powered by Archivas®
Server Chassis Type	Four 2U rack-optimized chassis
Server Processor	Two Intel Xeon processors at 3GHz with 800MHz frontside bus (FSB) and 2MB L2 cache
Server Memory	4GB 400MHz DDR2 ECC SDRAM (4 x 1GB modules)
Hard Drive	Two 36GB Ultra320 SCSI SCA 15,000RPM hot-swappable hard drives
RAID Card	LSI MegaRaid Ultra320-2 two-channel SCSI 64/66 PCI RAID controller with 128MB cache
Standard Disk Controller	Integrated dual-channel Ultra320 SCSI with one internal and one external connector (supports embedded RAID-0 and RAID-1)
High-availability Kit	Second 700W hot-swap power supply, four additional fans, and monitoring capability
Storage	Hitachi TagmaStore® Workgroup Modular Storage, model WMS100 (two systems) 2GB cache 30GB x 400GB, 7200RPM disk drives (12TB raw capacity) RAID-5 (6 + 1) with hot spares; RAID groups include RAID-0, RAID-1, RAID-2, RAID-3 (9.6TB raw)

Hitachi Data Systems Corporation

Corporate Headquarters
750 Central Expressway
Santa Clara, California 95050-2627
U.S.A.
Phone: 1 408 970 1000
www.hds.com
info@hds.com

Asia Pacific and Americas
750 Central Expressway
Santa Clara, California 95050-2627
U.S.A.
Phone: 1 408 970 1000
info@hds.com

Europe Headquarters
Sefton Park
Stoke Poges
Buckinghamshire SL2 4HD
United Kingdom
Phone: +44 (0) 1753 618000
info.eu@hds.com

DISK-595-00 June 2006