



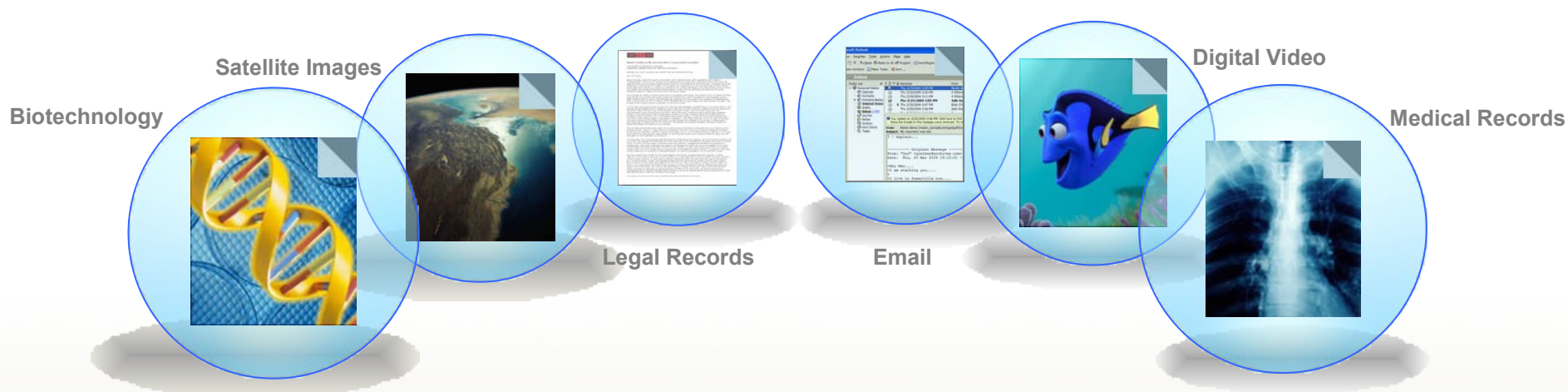
The Open Archive

Enabling Discovery

Agenda

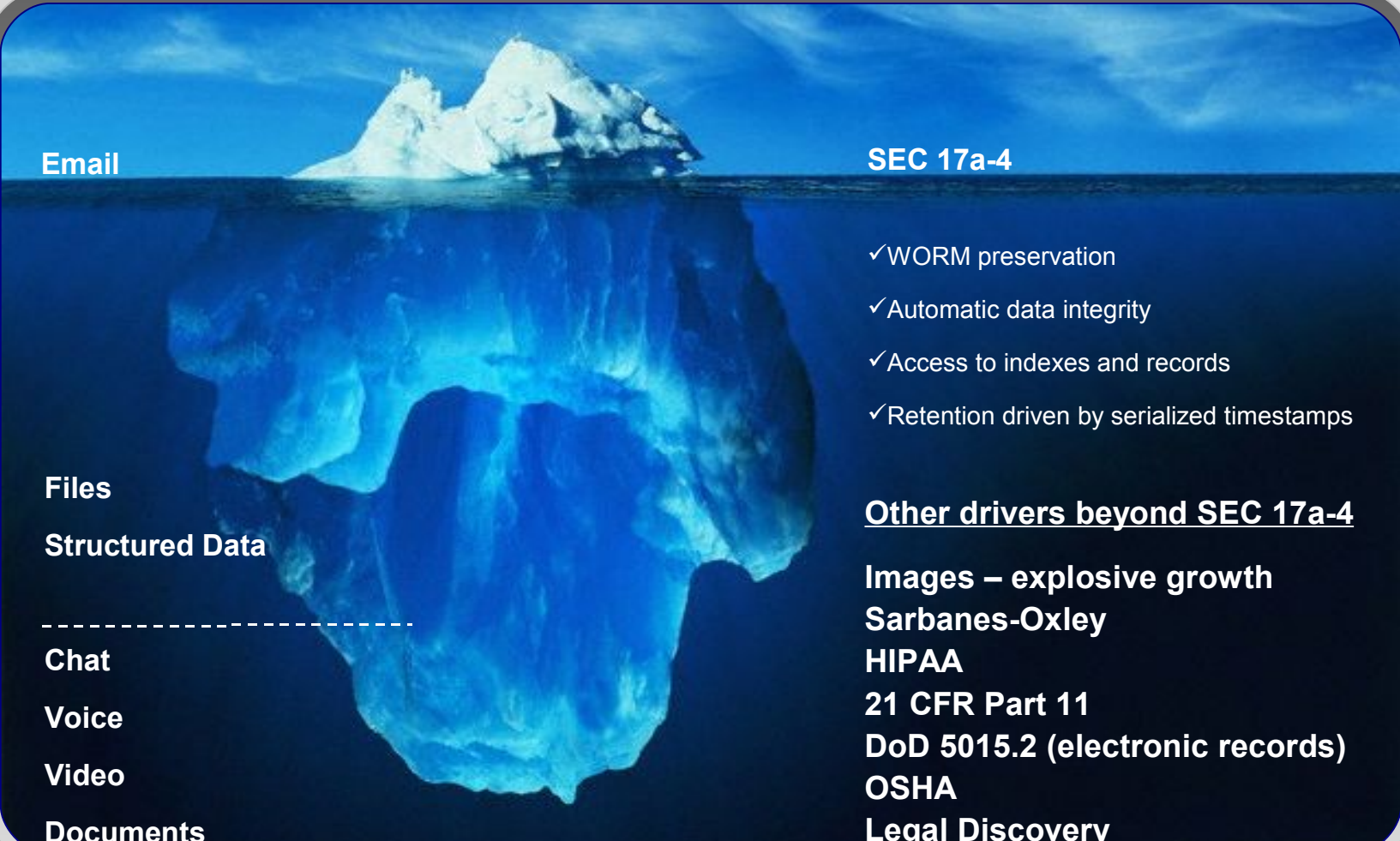
- Archives Matter
- The Open Archive
- Discovery

Archives Matter



- ❑ Fixed content data repositories storing data objects that have long-term value, that do not change over time, and are easily accessible

Email compliance: The tip of the iceberg



The image shows a large iceberg floating in the ocean. The tip of the iceberg, which is above the water line, is labeled 'Email'. The much larger part of the iceberg, which is submerged below the water line, is labeled with various data types: 'Files', 'Structured Data', 'Chat', 'Voice', 'Video', and 'Documents'. A dashed horizontal line separates the 'Email' tip from the submerged data. To the right of the iceberg, the text 'SEC 17a-4' is listed with four checkmarks and their corresponding requirements. Below that, 'Other drivers beyond SEC 17a-4' is listed with several regulatory and legal drivers.

Email

SEC 17a-4

- ✓ WORM preservation
- ✓ Automatic data integrity
- ✓ Access to indexes and records
- ✓ Retention driven by serialized timestamps

Other drivers beyond SEC 17a-4

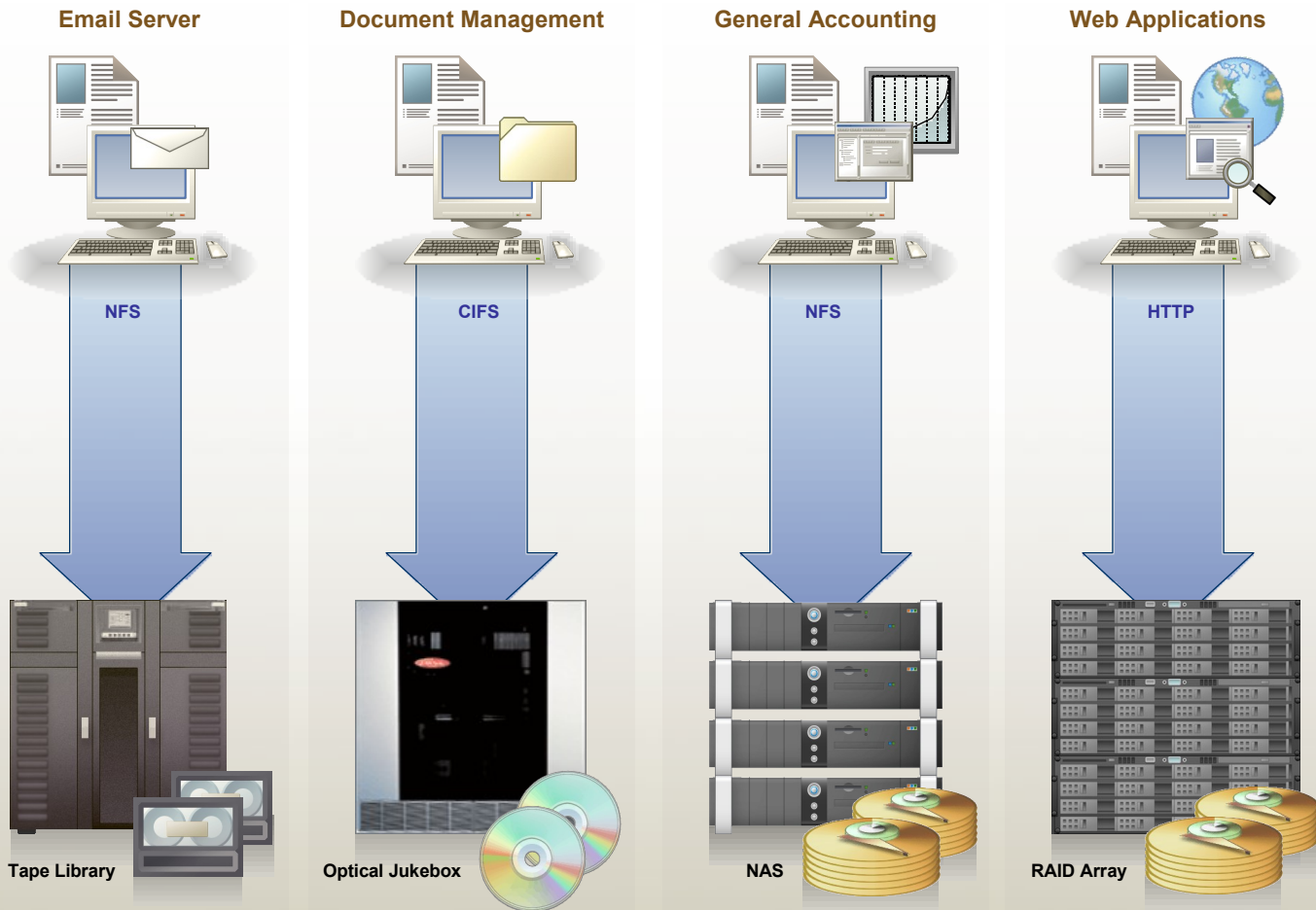
Images – explosive growth
Sarbanes-Oxley
HIPAA
21 CFR Part 11
DoD 5015.2 (electronic records)
OSHA
Legal Discovery

Files
Structured Data

Chat
Voice
Video
Documents

What is Long Term?

Beyond the applications



Beyond the hardware

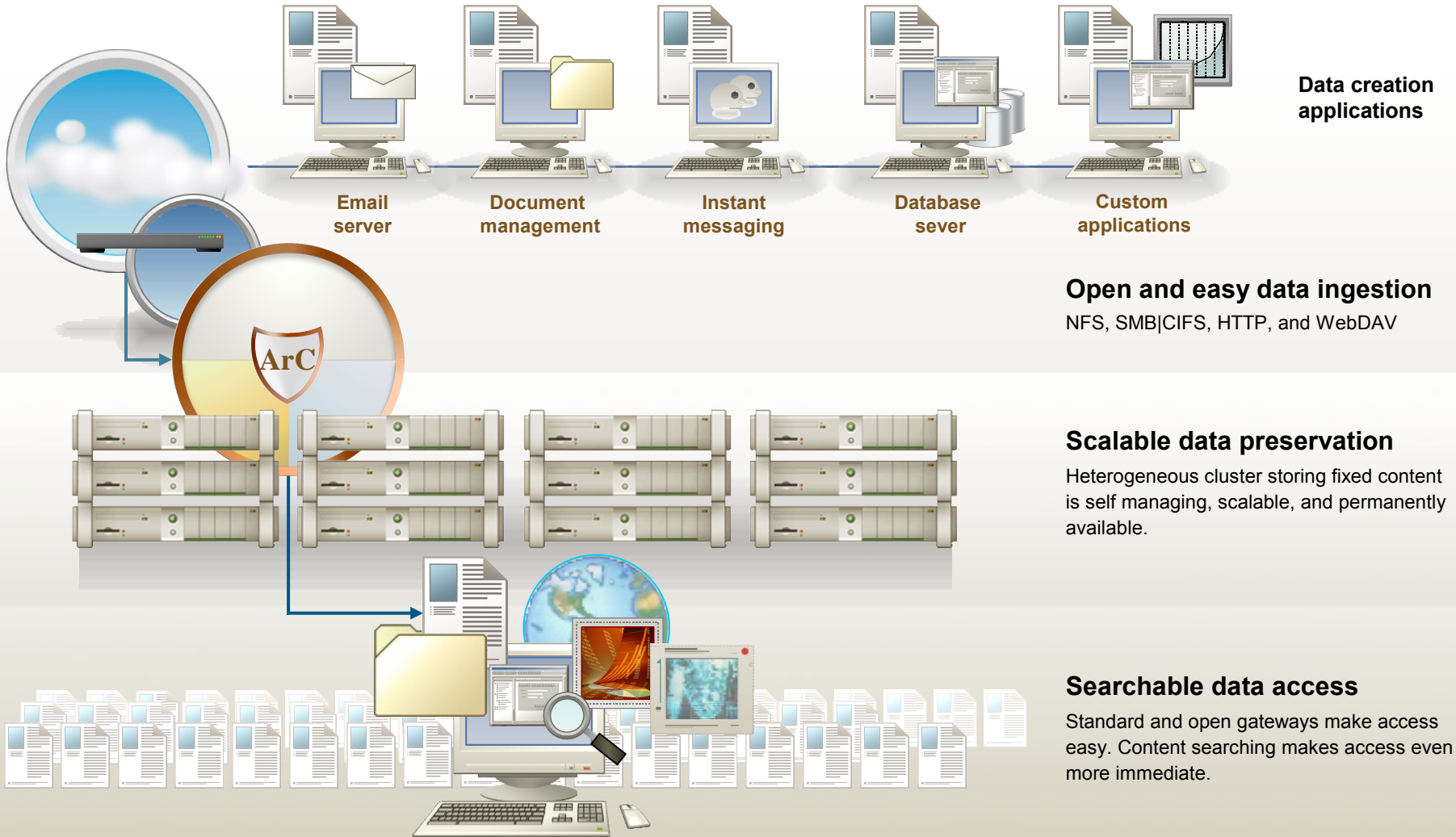
Need for a Common Archiving Platform that answers ...

How is data added to the archive?

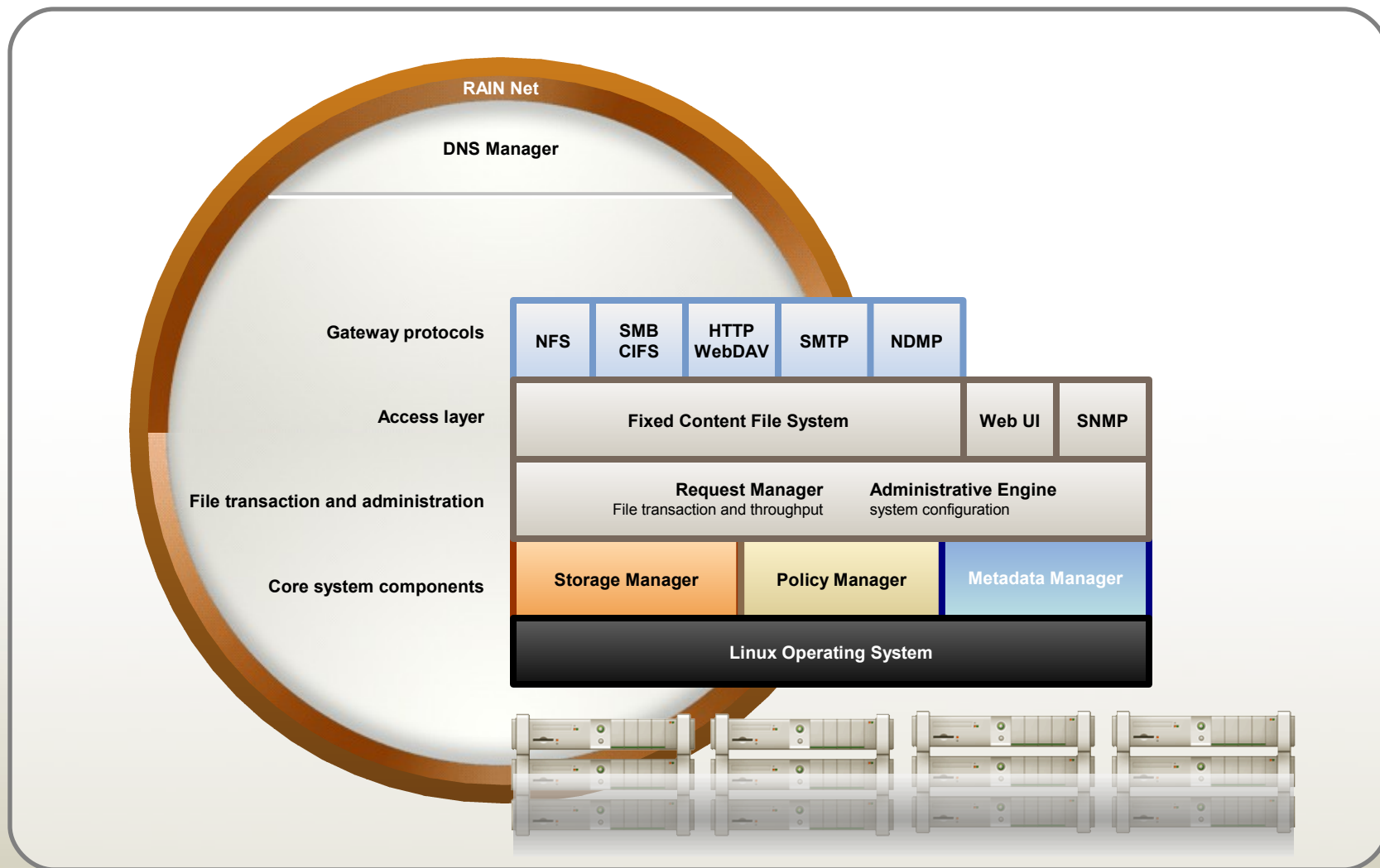
How is data stored, protected and managed over time?

How is data retrieved when needed?

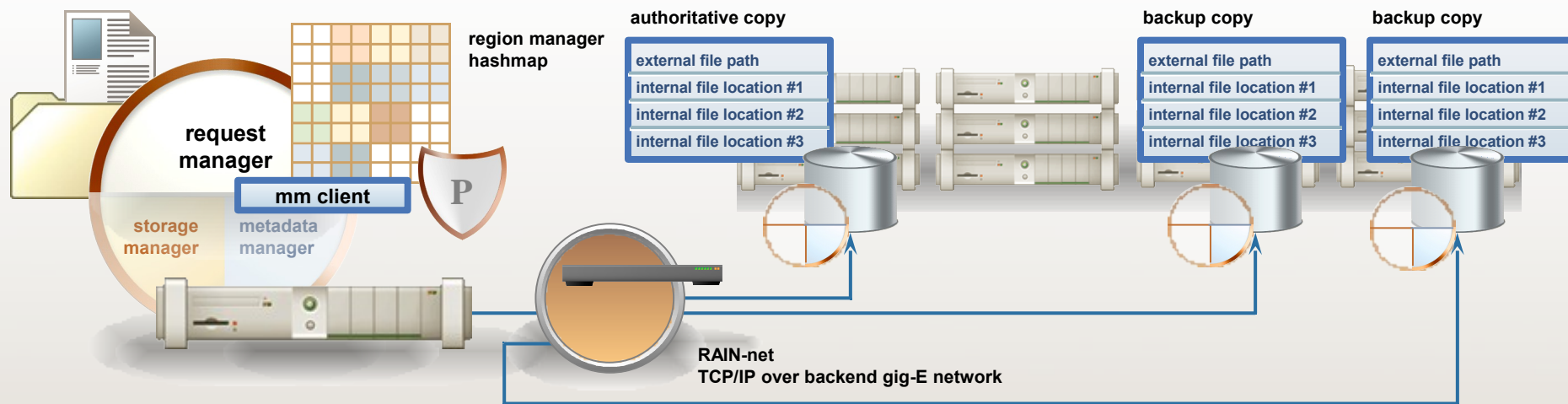
The Open Archive



Archive Node Detail



Metadata Manager: Inserting a file into ArC

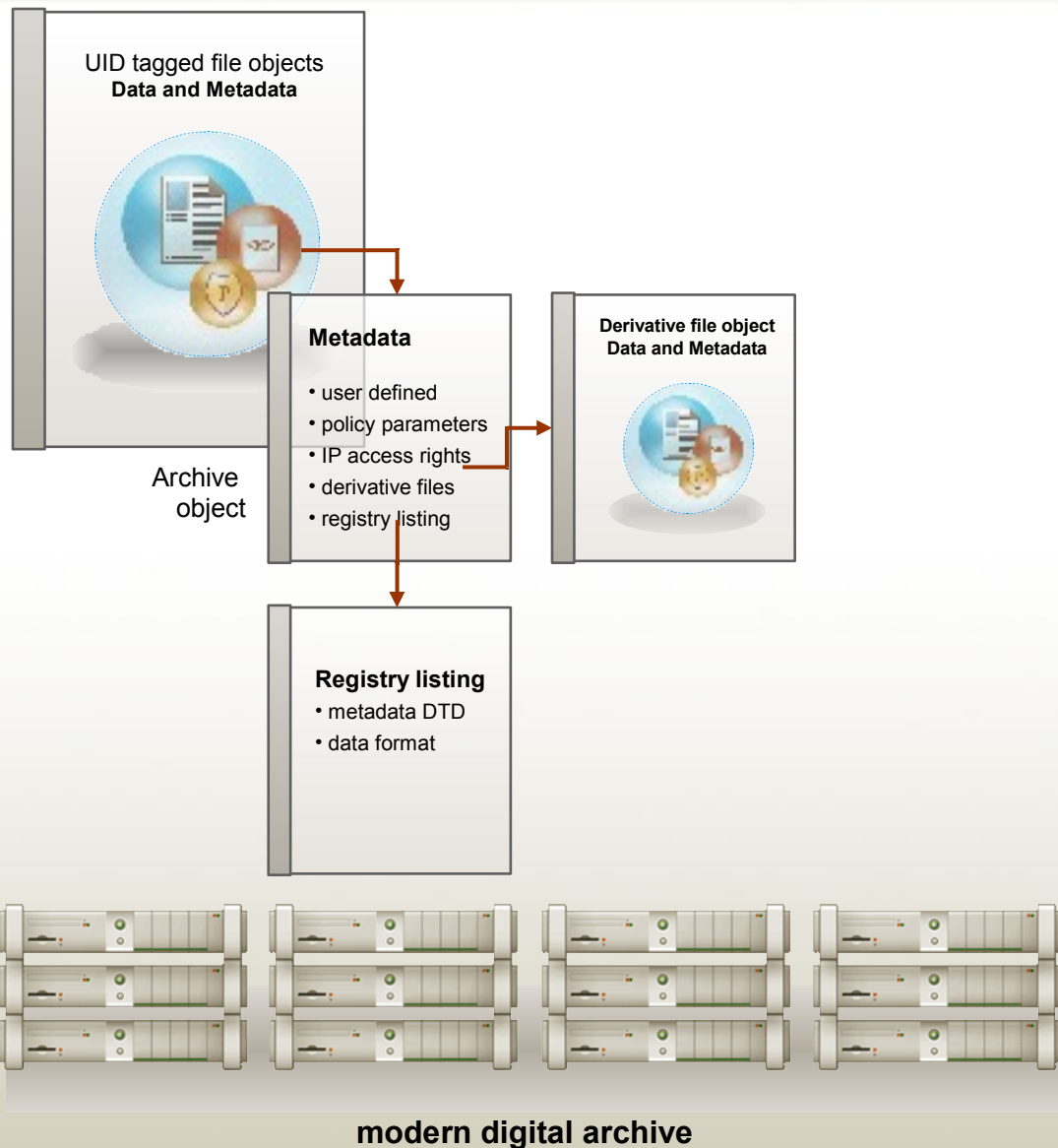


The Open Archive

Data ingestion

Data preservation

- ▶ Data stored as objects
- ▶ Maintain data integrity
- ▶ Data is location independent
- ▶ Policies are enforced



modern digital archive

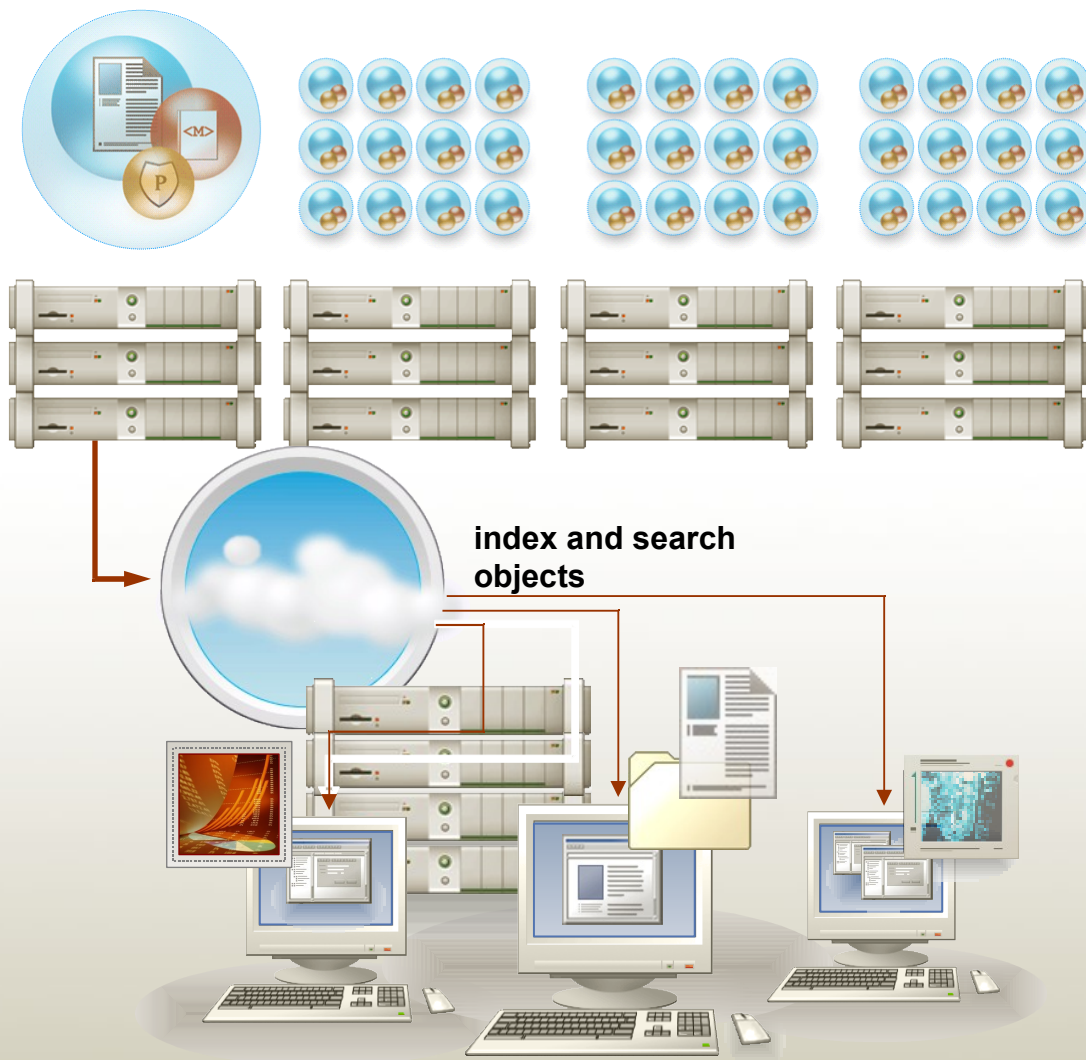
The Open Archive

Data ingestion

Data preservation

Data access

- ▶ Searchable metadata and data
- ▶ Standard file systems interface
- ▶ Open APIs: NFS, HTTP, WebDAV, SMB | CIFS
- ▶ Performance limited only by network speed



The Open Archive

- ❑ Decouple archive objects from the ingestion tools
 - Support Open data standards
 - Support Open network access protocols
- ❑ Decouple archive objects from hardware storage layer
 - Run portable archiving software
 - Assure interoperability at the network layer - not device specific
- ❑ Decouple archive objects from archiving software
 - Provide network export protocol (NDMP)
 - Support an Open transport format for archive objects

The Open Archive

A good preservation architecture

...

enables discovery

- Application independence
- Hardware independence
- Archival system independence



- Search integration
- High throughput
- High availability
- Optimize storage and indexing

Discovery: What Can Be Indexed?

- Extractable text from 370 file formats
- All extractable metadata tags (e.g., Title, Subject, Author, Category, Keywords, etc.) from the same
- File system metadata
- ArC metadata
 - Retention date
 - Authentication hash
 - Custom metadata (future)
- No fixed index schema makes the system powerful and very flexible
- Support for 77 languages makes it a global solution

Simple Search Mode

Simple Search	Structured Search	Advanced Search	Saved Queries	Change Password
----------------------	-------------------	-----------------	---------------	-----------------

Search: Simple

- All of these words
- Any of these words
- This exact phrase

© is reserved.

Structured Search Mode

Simple Search	Structured Search	Advanced Search	Saved Queries	Change Password
-------------------------------	--	---------------------------------	-------------------------------	---------------------------------

Search: Structured

Find files that match of the following...

File Type	is	Any	<input type="button" value="+"/>	<input type="button" value="-"/>
Archived Time	after	2006-01-01	<input type="button" value="+"/>	<input type="button" value="-"/>
Archived Time	before	2006-03-31	<input type="button" value="+"/>	<input type="button" value="-"/>
File Contents	contains any of	fraud	<input type="button" value="+"/>	<input type="button" value="-"/>
Author	contains	Jeff Spotts	<input type="button" value="+"/>	<input type="button" value="-"/>

[Show as advanced...](#)

© 2006 Archivas, Inc. All rights reserved.

Advanced Search Mode

Simple Search	Structured Search	Advanced Search	Saved Queries	Change Password
-------------------------------	-----------------------------------	--	-------------------------------	---------------------------------

Search: Advanced

Advanced queries syntax guide

For additional help in constructing advanced queries, please see the [list of searchable fields](#), as well as the [InStream Query Parameters Guide](#).

© 2006 Archivas, Inc. All rights reserved.

Handling Result Sets

Hold, release or delete results

Result filters

Key Terms

- [enron](#)
- [microsoft word](#)
- [mark guzman](#)
- [company](#)
- [www](#)
- [hotmail](#)
- [new york](#)
- [dow jones](#)
- [yahoo](#)
- [e-mail](#)
- [west desk](#)
- [desk x-folder](#)
- [ceo](#)
- [développement](#)
- [net](#)

page 1 of 285
go to page
results/page 10
Save as
save

Control operations
Export results
Sort results

Control operations

- Place results on hold
- Release hold on results
- Delete results

1 FW: If you haven't gotten already...start at bottom and read up

<5962632.1075845228447.JavaMail.evans@thyme>
From: martin.cuilla@enron.com **Sent date:** Tue May 29 10:39:43 EDT 2001
To: s..shively@enron.com, kevin.ruscitti@enron.com, h.lewis@enron.com, geoff.storey@enron.com

...3C465965-938A4154-90CCD37B-806912C9 X-Server-Uid: b0fe6c76-9e59-11d1-b373-00805fa7c2de X-WSS-ID: 1713474923332-01-06 X-MimeOLE: Produced By Microsoft MimeOLE V5.00.2314.1300 start from the bottom and then check out the article: http://www.washingtonpost.com/wp-dyn/articles...

Archived: Fri Apr 28 22:34:53 EDT 2006 **Retention:** Expired **Size:** 25890 bytes [Show details](#)

2 Microsoft

<15835119.1075847799107.JavaMail.evans@thyme>
From: maureen.mcvicker@enron.com **Sent date:** Thu Jan 04 04:46:00 EST 2001
To: steven.kean@enron.com

...maureen.mcvicker@enron.com steven.kean@enron.com **Microsoft** Mime-Version: 1.0 Content-Type: text/plain; charset...Maureen/McVicker/NA/Enron@Enron cc:
Subject: Microsoft Bias Suit Against Microsoft Aims at 'Flat' Workplace Hierarchies By Yochi...

Archived: Fri Apr 28 21:41:11 EDT 2006 **Retention:** Expired **Size:** 9809 bytes [Show details](#)

3 Volume Management Opening

<2412559.1075841862723.JavaMail.evans@thyme>
From: amy.fitzpatrick@enron.com **Sent date:** Tue Mar 06 05:04:00 EST 2001
To: portland.desk@enron.com

...strong organizational capabilities. Must possess excellent oral, written, and interpersonal skills. PC proficiency, including Microsoft Work, Advanced Excel, and Access.
Special Job Characteristics: Must be highly motivated. Self-starter with ability to recognize...

Archived: Sat Apr 29 02:05:11 EDT 2006 **Retention:** Expired **Size:** 2601 bytes [Show details](#)

4 Volume Management Opening

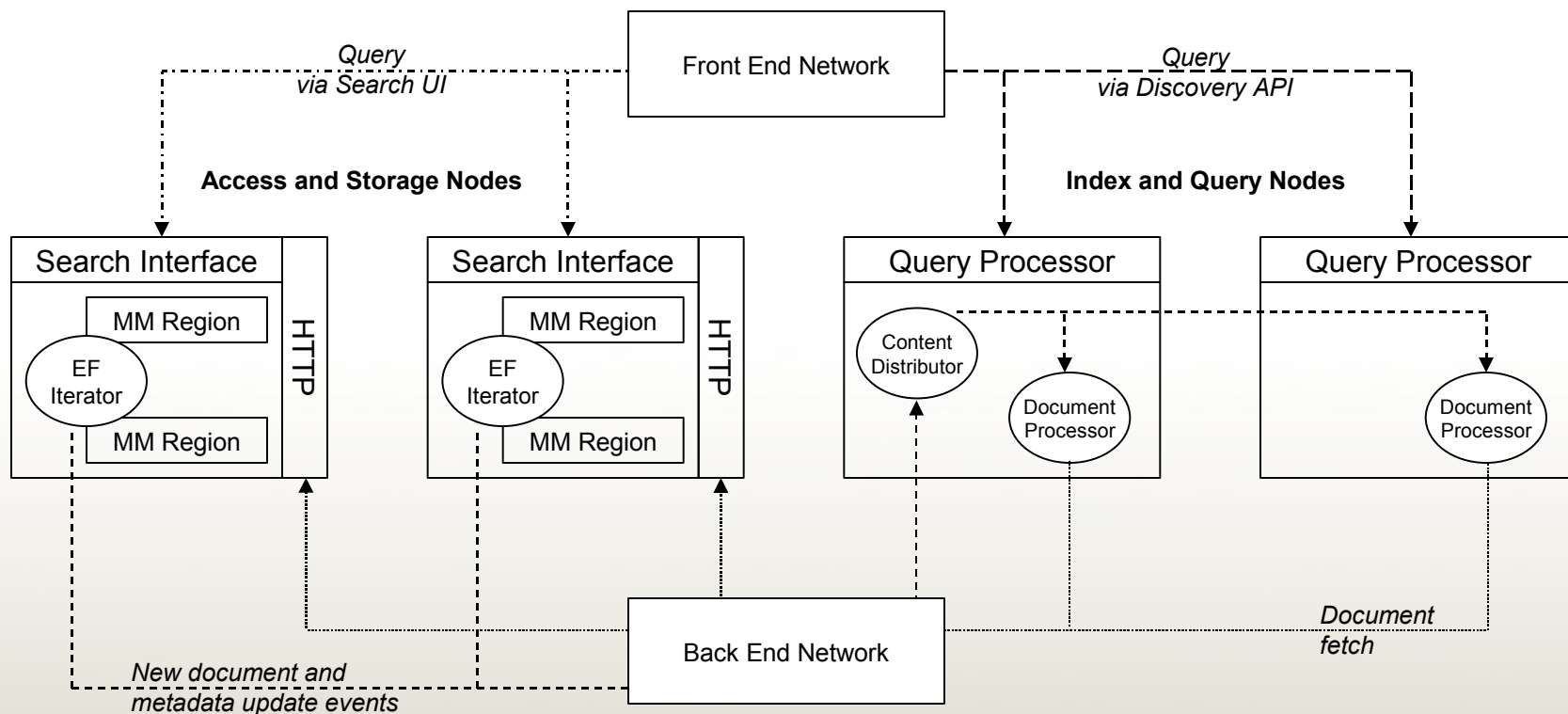
Document Format	Count
Internet Mail Message	1140
Adobe Acrobat (PDF)	936
Internet HTML	641
Plain Text	51
Microsoft PowerPoint 97-2004	22
Microsoft Word 97-98	16
CompuServe GIF	10
Microsoft Word 2002	7
.ZIP File	5
Microsoft Word 2000	5
Microsoft Excel 2000	3
WordPerfect 6.1 - 12.0	2
Microsoft Excel 5.0/7.0	2
Internet Message	2
Microsoft Excel 4.0	1

Retention	Count
Initial Unspecified	0
Never Deletable	0
Expired	2849
Not Expired	0

Other Search Functionality

- ❑ Export Result Sets for post-processing
- ❑ Save/Run/Edit Queries
- ❑ Discovery API
 - Exposes the cluster-internal query service
 - For specialized e-discovery applications
 - For federated searches across multiple archives
 - For the archive to be one of multiple targets of a federated search

Index/Search Integration Model



Query API

Programmatic access to queries and results

Highlights

- API access with factory interfaces for building queries and iterating over query results
- Control query features – Drill-down, categorization, clustering, result re-sorting, spell-check, lemmatization, find similar
- Control scope of query – Collections, filtering
- Automatic query rewrite (e.g., spell check), rewrite suggestions
- Easy to integrate with server side scripting – JSP and ASP

Interfaces

- Java, C++, .Net
- HTTP (for results returned in xml, text)

Query timeline

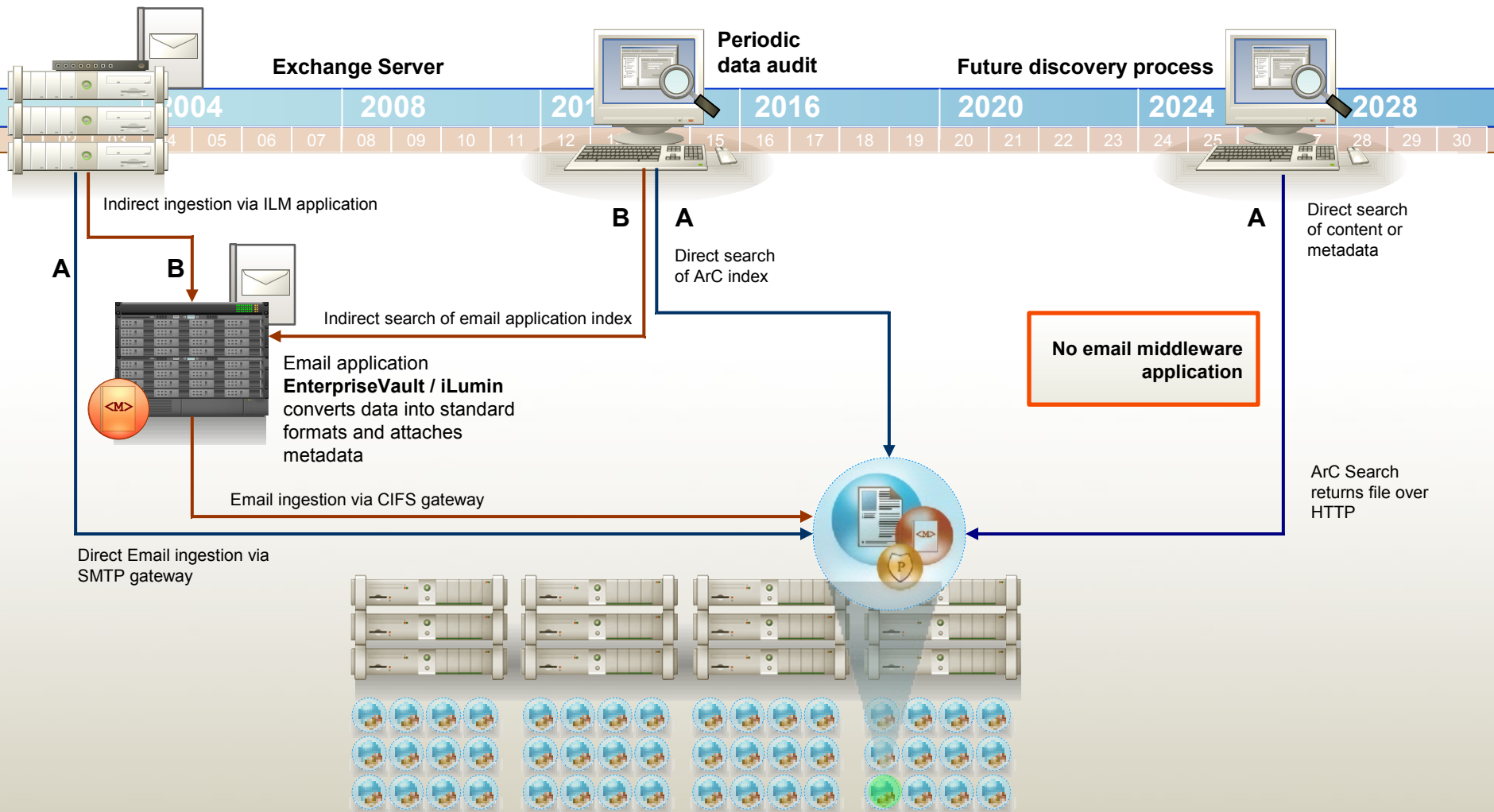
- Obtain query string
- Define query parameters
- Query transformations

The Open Archive Summary

ingestion

preservation

access





200 West Street, First Floor
Waltham, MA 021451-1121
United States
www.archivas.com

Andres Rodriguez
CTO

Phone: 617-571-1012
Email: andres@archivas.com

———— **Thank you** ————