MARKLOGIC WORLD 19

Defense Technical Information Center: Scientific & Technical Information for the DoD Community

Andrew Pedrick

Information Architect

Defense Technical Information Center

Defense Technical Information Center: Scientific & Technical Information for the DoD Community

Andrew Pedrick

Information Architect

Defense Technical Information Center

Disclaimer

All discussions of commercial products are descriptive only.

DTIC does not endorse the use of any commercial product. Specific questions as to the effectiveness of a commercial product should be addressed to the vendor.

Agenda

- DTIC
- Challenges
- Implementation

Defense Technical Information Center

DoD Field Activity under the Under Secretary of Defense for Research and Engineering.

DTIC's mission is "to aggregate and fuse science and technology data to rapidly, accurately and reliably deliver the knowledge needed to develop the next generation of technologies to support our Warfighters and help assure national security."

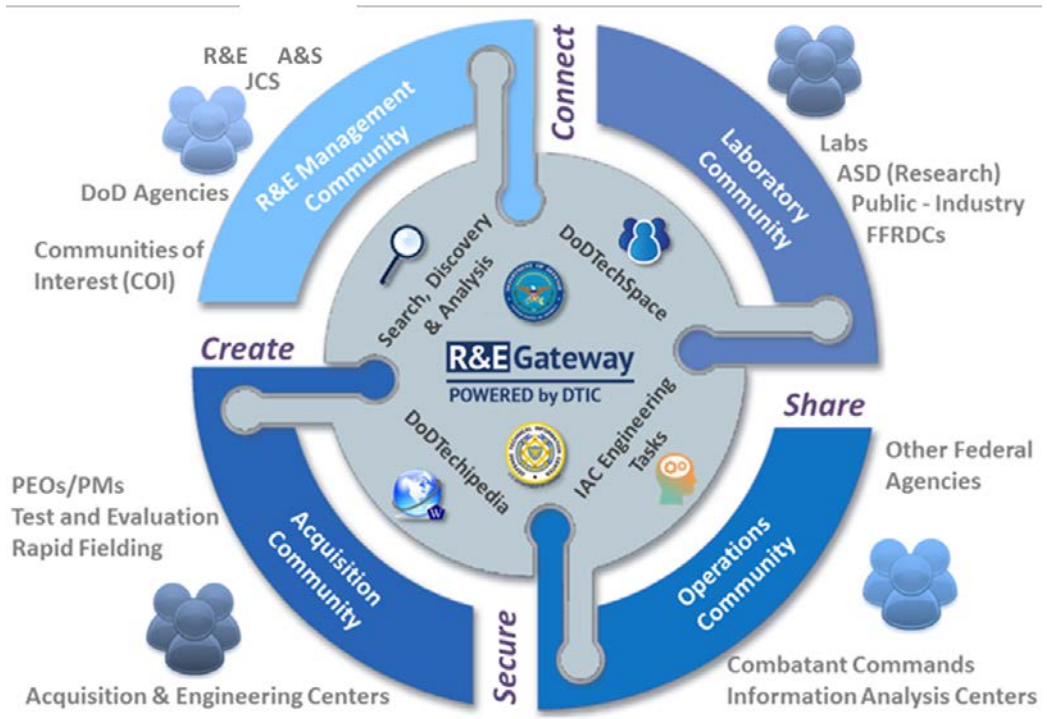
Defense Technical Information Center

- Collect and store
  - Reports, journal articles, conference papers which describe the results of DoD-funded scientific and technical research
  - Descriptions of DoD research in progress
  - Research and Development budget requests
  - Acquisition information
  - Collaborative data
- Catalog, enhance, and organize this data
- Secure the data and verify the access rights of users
- Promote the discovery, analysis, and reuse of the data through search and analytic tools

Defense Technical Information Center

DoD, other federal agencies, and industry use this data to

- Share knowledge and discoveries
- Build upon each other's efforts
- Discover expert individuals and organizations
- Prevent wasteful duplication of effort
- Determine how resources are being spent in research and development and decide how they should be allocated



Typical Use Cases

Warfighter

"Our adversaries now have laser blasters. Has anyone worked on counter-measures?"

Engineer

"I've built a light saber that needs a power source. What is the latest battery technology available?"

Engineer

"I'm starting a new project to build a light saber. Has anyone already done this?"

Scientist

"I've performed research on mentally controlling matter by leveraging 'an energy field created by all living things.' Here are my results."

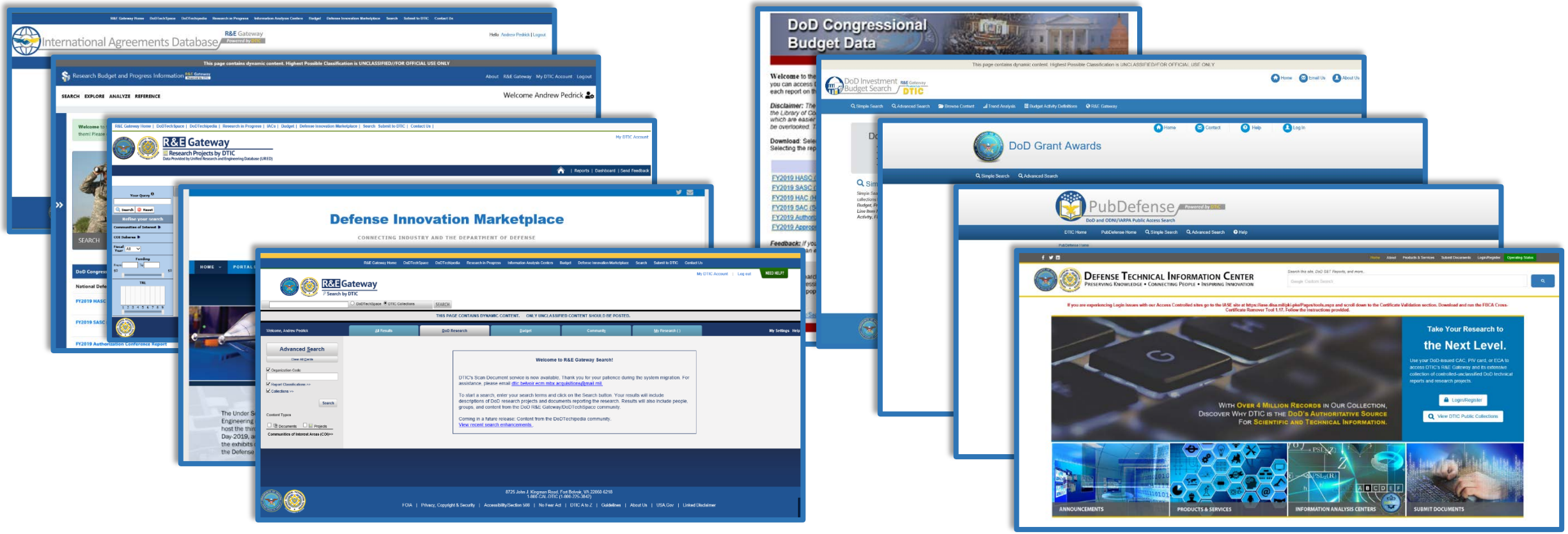
Scientist

"My field of research is telekinesis. Any new developments?"

Leadership

"How much has DoD spent on this research, what were the results, and where should we invest future dollars?"

Many Front End Products



Many Data Sets

- Technical Reports
- Information Analysis Center Technical Reports
- Public Journal Articles
- Research Projects
- Independent Research Projects
- Independent Research Projects Archive
- Primary Wiki
- Secondary Wiki
- Collaborative Tool
- Organization Taxonomy
- Subject Thesaurus
- Grants
- Research and Development Budget Requests
- Procurement Budget Requests
- House and Senate Budget Reports
- Technology Transfer Agreements
- Acquisition Information
- International Agreements
- National Defense Industry Association Conference Proceedings
- User Database
- Usage Logs
- User Bibliographies
- User Saved Searches
- User Preferences

Many Silos

Leaflet, License CC BY-SA 3.0

Diego Delso, delso.photo, License CC-BY-SA

DMahalko, Dale Mahalko, Gilman, WI, USA. License CC BY-SA 3.0

Challenges

- Storage
- Security
- Search
- Analytics

Challenges: Storage

- Data was not collocated for reuse and analysis
- Slow to introduce new data
- Slow to repurpose data
- Difficult to add elements to core collections
- Data flows at high risk
- Many data copies
- Data flows were high maintenance

Security at DTIC

- Attribute-based element-level access control
- Controlled access to documents, records, elements, and services
- Many document markings, some with several values, some negating others
  - Distribution Code
  - Distribution Reason
  - Multiple types of classification markings
  - Export Control
  - Copyright
  - For Pay
  - Status of Effort
  - Special Indicator
  - Completed Flag
  - Record Status
  - Accession number range
- Inconsistently represented in the data
- Inconsistent policies per collection

Challenges: Security

- Custom built element-level security system was:
  - Used by only flagship search product
  - Controlled only primary collections
  - Difficult to update and expand
  - Created additional copies of data
  - Out of sync with search index
- Difficult to keep security policies consistent when implemented in multiple systems

Challenges: Search

- Missing basic capabilities
- New index required complete reload of data
- Inaccurate and inconsistent result counts
- Didn’t support analytical queries
- Could not return all results
- Created additional copies of data
- Data load met only a single purpose

Semantic Use Cases

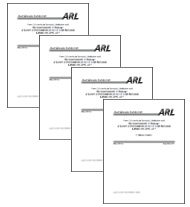
DoD agencies…

submit budget requests

that fund research projects

performed by scientists

that produce reports



viewed by other researchers

from the same or other agencies

leading to further research or procurements

Challenges: Analytics

- No platform to store and query semantic data
- No data warehouse for Business Intelligence tools
- User data, usage data, content data not collocated
- Analytic tools multiplied data flows and copies

Solution

- Consolidate data into single repository for reuse…
  - that can store unstructured and structured data with many different schemas…
  - has built-in access control…
  - has built-in full-featured search…
  - supports analytic queries and Business Intelligence tools…
  - stores and queries semantic data…
  - all in one system.

Master Data Repository (MDR)

Storage

- **Consolidate data** once into a single repository for reuse, linking, and integration
- Ingest data that is organized in a variety of ways, **reducing time to market for new data**
- **Reduce maintenance & risk** by eliminating data flows and integration points
- Support **create, update, delete** of data

Enterprise Search

- Enable **advanced search** capabilities and **accurate result counts**
- Support **analytical queries**
- Couple search with the database, **reducing time to market for new data sets**

Access control

- Build access control into the repository to ensure **centralized security and data view logs**
- **Enforce consistency** by applying policies once instead of within each application

Semantics

- Introduce platform for storing and querying **linked data**

MDR Program Components

Data

- Content data
- User data
- Metrics reporting data
- Master data
- Catalog of content data
- Taxonomy tags
- Inferred data
- External data
- Targeted data quality enhancements

People

- Train technical staff
- Train management staff on components, process, capabilities
- Socialize benefits
- Communicate progress to stakeholders
- Appoint Data Owners and Data Stewards

Process

- Establish procedures for migrating products
- Integrate in Project management
- Publish Metadata guidance
- Initiate Data Governance
- Require Data Management

Technology

- Provision hardware
- Install NoSQL database
- Refactor Interfaces
- Create services for data upload, input, query, export
- Ingest data
- Build Access Control
- Integrate with DTIC architecture

Distinctions of the DTIC Use Case

- Primarily content and usage data, not business or transactional data
- No data warehouse, but rather a series of search engines
- Existing live front-end product
- Attribute-based element-level access control

Master Data Repository Timeline

2013 – Preliminary Market Research
2014 – Pilot; Market Research; NoSQL Request For Information
2015 – Request For Proposals; Acquisition
2015-2016 – MDR Phase 1; Research Projects Unclassified
2016-2017 – MDR Phase 2; Research Projects Classified; International Agreements
2017 – Assessment; Training
2018 – New Information Analysis Center Content Ingest
2019 – New Interfaces; ML9; Cloud migration; New Technical Report Ingest
2020 – Advanced Security; Classified cloud migration; Semantics

2015-2016

Master Data Repository Phase 1

- Replaced backend system for flagship search product
- Ingested 15 Data Sets totaling over 4M records and over 1.5M PDFs. Continuous feeds for four data sets, ad hoc updates for two
- Established 20 new collections for records of user activities, co-located with content data
- Implemented Access Control for all content collections
- Created 154 indexes
- Established simple search language for complex searches. Delivered all capabilities users had lost in previous system (accurate counts, Boolean operators, truncation, field searching, etc.)
- Updated flagship search front end to use MDR with minimal disruption to users

Research Projects

- Updated secondary search front end to use Master Data Repository

Data Ingest

Method:

- Custom Java Ingester – four content collections, three run daily
- Legacy custom Java Ingester – one content collection, runs daily
- MLCP – two content collections with seven schemas, three user data collections

Envelopes

Raw Data

```
<?xml  version="1.0" encoding="UTF-8"?>
<mdr:Record Type=" TR " xmlns:mdr="http://dtic.mil/mdr/record">
 <tr:Raw xmlns:tr="http://dtic.mil/mdr/record/tr">
  <tr:TrFix>
   <tr:UnclassifiedTitle>Improved Polyurethane Storage Tank Performance</tr:UnclassifiedTitle>
   <tr:TitleClassificationCode>U</tr:TitleClassificationCode>
   <tr:AccessionNumber>A613634</tr:AccessionNumber>
   <tr:CorporateAuthor>SEAMAN CORP WOOSTER OH</tr:CorporateAuthor>
   <tr:SourceCode>648970</tr:SourceCode>
   <tr:GeoPoliticalCode>39</tr:GeoPoliticalCode>
   <tr:ReportDateText>20140630</tr:ReportDateText>
   <tr:AbstractClassificationCode>U</tr:AbstractClassificationCode>
   <tr:PaginationOrMediaCount>0198</tr:PaginationOrMediaCount>
   <tr:MonitorAcronym1>XD</tr:MonitorAcronym1>
   <tr:MonitorSeries1>DLA/FB</tr:MonitorSeries1>
   <tr:ContractNumber1>SP6066-04-D-5442-1029</tr:ContractNumber1>
   <tr:DescriptorClassificationCode>U</tr:DescriptorClassificationCode>
   <tr:IdentifierClassificationCode>U</tr:IdentifierClassificationCode>
   <tr:CitationStatusCode>A</tr:CitationStatusCode>
   <tr:DistributionStatement>Approved for public release; distribution is unlimited.</tr:DistributionStatement>
   <tr:CitationClassificationCode>U</tr:CitationClassificationCode>
   <tr:ReportClassificationCode>U</tr:ReportClassificationCode>
  </tr:TrFix>
  <tr:TrDistCode>
   <tr:Code>1</tr:Code>
   <tr:Text>APPROVED FOR PUBLIC RELEASE</tr:Text>
   <tr:Code>0</tr:Code>
   <tr:Code>0</tr:Code>
  </tr:TrDistCode>
  <tr:TrlFgs>
   <tr:Code>110900||130400||201100||210400</tr:Code>
   <tr:Text>PLASTICS||CONTAINERS AND PACKAGING||MECHANICS||FUELS</tr:Text>
  </tr:TrlFgs>
  <tr:TrlType>
   <tr:Code>4</tr:Code>
   <tr:Text>INDUSTRIAL/COMMERCIAL</tr:Text>
  </tr:TrlType>
```

Standardized Data

```
 </tr:Raw>
 <meta:Metadata xmlns:meta="http://dtic.mil/mdr/record/meta">
  <meta:UnclassifiedTitle>Improved Polyurethane Storage Tank Performance</meta:UnclassifiedTitle>
  <meta:TitleClassification>
   <meta:Code>U</meta:Code>
   <meta:Description>Unclassified</meta:Description>
  </meta:TitleClassification>
  <meta:AccessionNumber>ADA613634</meta:AccessionNumber>
  <meta:PersonalAuthors>
   <meta:PersonalAuthor Num="1">Kral, J. A.</meta:PersonalAuthor>
   <meta:PersonalAuthor Num="2">Uhler, S.</meta:PersonalAuthor>
   <meta:PersonalAuthor Num="3">Fenske, S.</meta:PersonalAuthor>
   <meta:PersonalAuthor Num="4">Bradenburg, F.</meta:PersonalAuthor>
   <meta:PersonalAuthor Num="5">Burgess, S.</meta:PersonalAuthor>
   <meta:PersonalAuthor Num="6">Russell, A.</meta:PersonalAuthor>
   <meta:PersonalAuthor Num="7">Kidney, R.</meta:PersonalAuthor>
  </meta:PersonalAuthors>
  <meta:CorporateAuthor>SEAMAN CORP WOOSTER OH</meta:CorporateAuthor>
  <meta:SourceCode>648970</meta:SourceCode>
  <meta:GeopoliticalCode>39</meta:GeopoliticalCode>
  <meta:OrganizationType>
   <meta:Code>4</meta:Code>
   <meta:Description>Industrial/Commercial</meta:Description>
  </meta:OrganizationType>
  <meta:ReportDate>
   <meta:String>2014-06-30</meta:String>
   <meta:Date>2014-06-30</meta:Date>
  </meta:ReportDate>
  <meta:Abstract>The Improved Polyurethane Storage Tank Performance program was sponsored by Defense Logistics Agency
Energy (DLA Energy) to study polyurethane coated membrane fabrication processes that are or may be used in the production
of collapsible fuel tanks (CFT). The objective of this effort was to provide technical information that will assist in the
improvement of polyurethane storage tank performance through the development of improved tank construction, fabrication
techniques and quality control procedures. This research program expands on the FY2008 published effort by integrating in
CFT fabrication and quality control recommendations and evaluating tank leaks through laboratory and actual field study.
```

Full text and binary description

```
 </meta:Metadata>
 <mdr:PDF>
  <mdr:FullTextPath>/stishare/stidataU2/ft_u2/a613634.pdf</mdr:FullTextPath>
  <mdr:FullTextBytes>10368230</mdr:FullTextBytes>
  <mdr:FullTextExists>1</mdr:FullTextExists>
  <mdr:FullTextIndexed>1</mdr:FullTextIndexed>
  <mdr:FullText>
   Improved Polyurethane Storage Tank Performance      Page 1 of 197

   FY2009 Final Technical Report
```

Envelope Approach Scenarios

	Scenario				Envelope Approach		
	State of Source Data	Display Reqs	Search Reqs		"Raw" Data	"Standardized" Data	Data Set Schemas
A.	RDBMS: had materialized view or otherwise flattened	All fields	Most fields		All fields	Nearly all fields	3
B.	RDBMS: no materialized view	All fields	Most fields		No raw form	All fields	1
C.	XML/JSON: multiple schemas	Few fields	Few fields		All fields	Few fields	7
D.	RDBMS: COTS product	Few fields	Few fields		No raw form	Few fields	1
E.	RDBMS: user data, simple tables	Few fields	One or two fields		No raw form	All fields	3

Data Documentation

Data Set 1

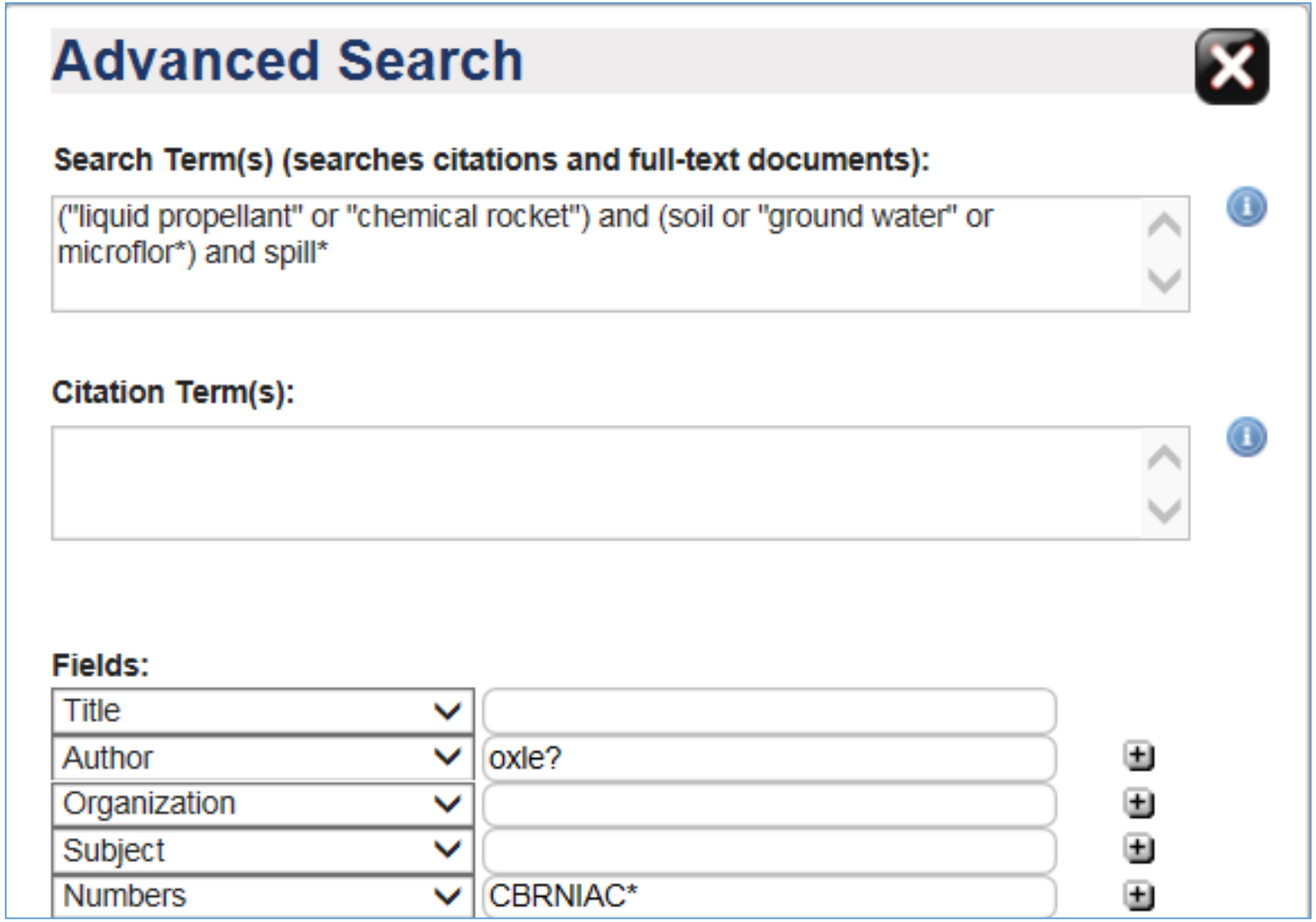
Index Constraint	Index Type	Display Label	Source Table	Source Column	Sample Data	Data Type	Transform	Target Raw XML	Target Standard XML

Data Set 2

Index Constraint	Index Type	Display Label	Source Table	Source Column	Sample Data	Data Type	Transform	Target Raw XML	Target Standard XML

Search

- Accurate result counts
- Return all results
- Boolean operators (AND, OR, NOT) and nesting
- Truncation, wildcards
- ‘Expert’ field searching



Security

- If document markings are:
  - Record Classification is unclassified
  - Document Distribution code is F
  - Document is Export control
  - Record status is Active
  - Abstract is marked Approved for public release
  - Document Classification is Confidential
  - Document is not copyrighted
  - Document is not proprietary
- And the User:
  - Is authorized for Distribution A, C, D
  - Has a Secret clearance
  - The company is approved for export control
  - Is using a secure network product
- And the document is in the TR collection

- Then:
  - Allow user to view record
  - Allow user to view abstract element
  - Do not allow document to be downloaded
  - Do not allow user to request the document be digitized
  - Allow user to request permission to view the document

2016-2017

Master Data Repository Phase 2

- Updated ingest and access control system for classified policies
- Migrated new backend to flagship search product on classified network

Research Projects

- Deployed updated secondary front end to classified network

International Agreements

- Migrated third application to MDR

2017-2018

Assessment & Training

- Health assessment of implementation, MarkLogic 9 evaluation, custom training

New Information Analysis Center Content Ingest

- Began work to ingest primary collection from new cloud-based document management system

Operations & Maintenance

- Bugs and maintenance

2019

New Search Interface & Budget Tools

- Add indexes and data transformations for two new search interfaces

Operations & Maintenance

- Remove dependencies on Oracle
- Ingest new data sets
- Conduct Advanced Security pilot

MarkLogic 9

- Migrate to new version of MarkLogic

Cloud

- Migrate secure network to cloud

Service Agencies ↑	Requested Funding
AF	$1,093,904,000 (23%)
Army	$584,571,000 (12%)
CHEM	$208,111,000 (4%)
DARPA	$894,798,000 (19%)
DHP	$961,218,000 (20%)
DTRA	$155,415,000 (3%)
Navy	$822,019,000 (17%)
OSD	$90,129,000 (2%)

2020

- Classified cloud
- Apache NiFi
- Advanced Security
- Semantics
- More data sets

Master Data Repository

Storage

- √ **Consolidate data** once into a single repository for reuse, linking, and integration
- √ Ingest data that is organized in a variety of ways, **reducing time to market for new data**
- √ **Reduce maintenance & risk** by eliminating data flows and integration points
- √ Support **create, update, delete** of data

Enterprise Search

- √ Enable **advanced search** capabilities and **accurate result counts**
- √ Support **analytical queries**
- √ Couple search with the database, **reducing time to market for new data sets**

Access control

- √ Build access control into the repository to ensure **centralized security and data view logs**
- √ **Enforce consistency** by applying policies once instead of within each application

Semantics

- √ Introduce platform for storing and querying **linked data**

MDR Program Components

Data

- √ Content data
- √ User data
- √ Metrics reporting data
- • Master data
- √ Catalog of content data
- √ Taxonomy tags
- • Inferred data
- • External data
- √ Targeted data quality enhancements

People

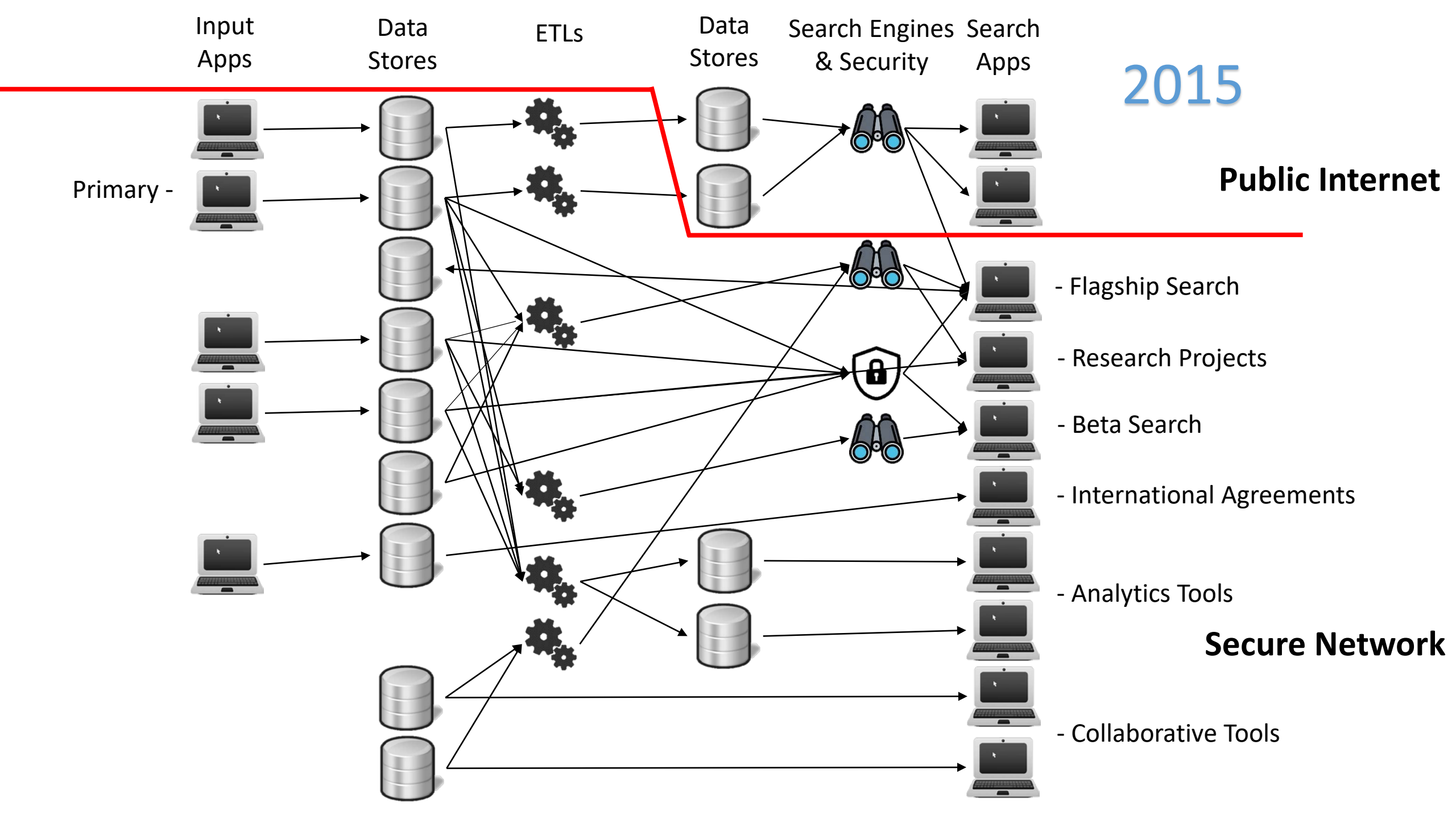
- √ Train technical staff
- √ Train management staff on components, process, capabilities
- √ Socialize benefits
- √ Communicate progress to stakeholders
- √ Appoint Data Owners and Data Stewards

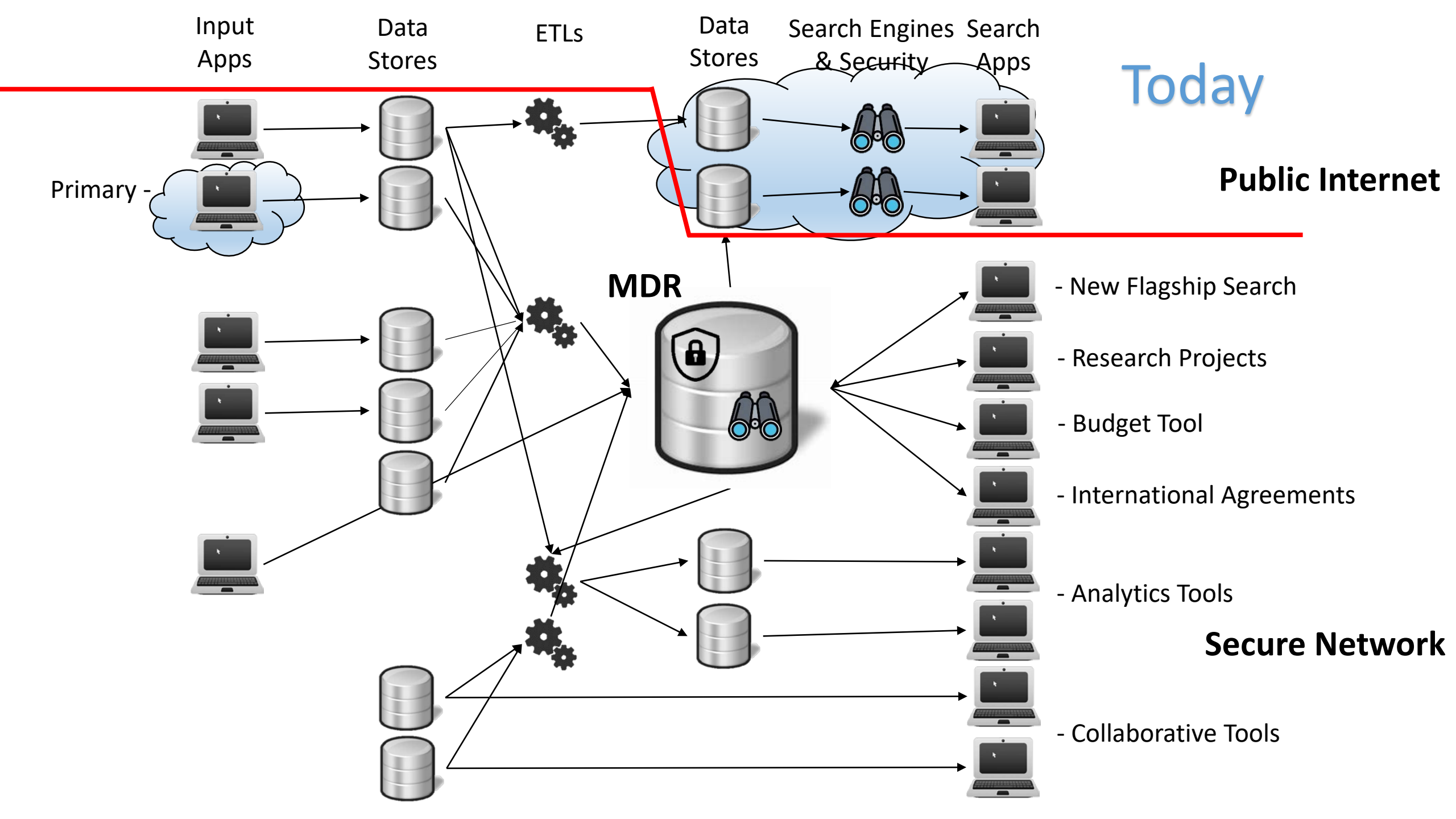
Process

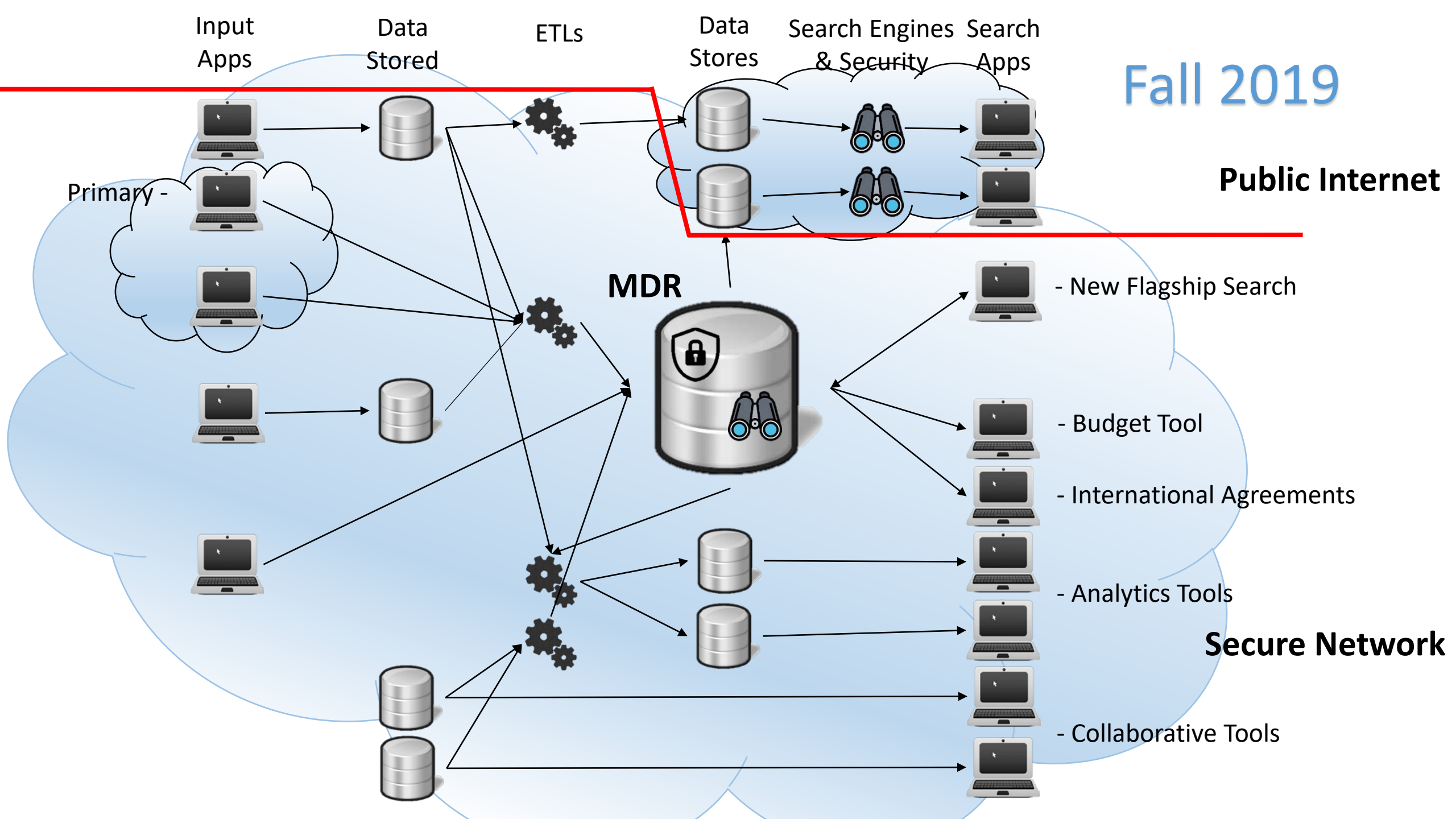
- √ Establish procedures for migrating products
- √ Integrate in Project management
- √ Publish Metadata guidance
- √ Initiate Data Governance
- √ Require Data Management

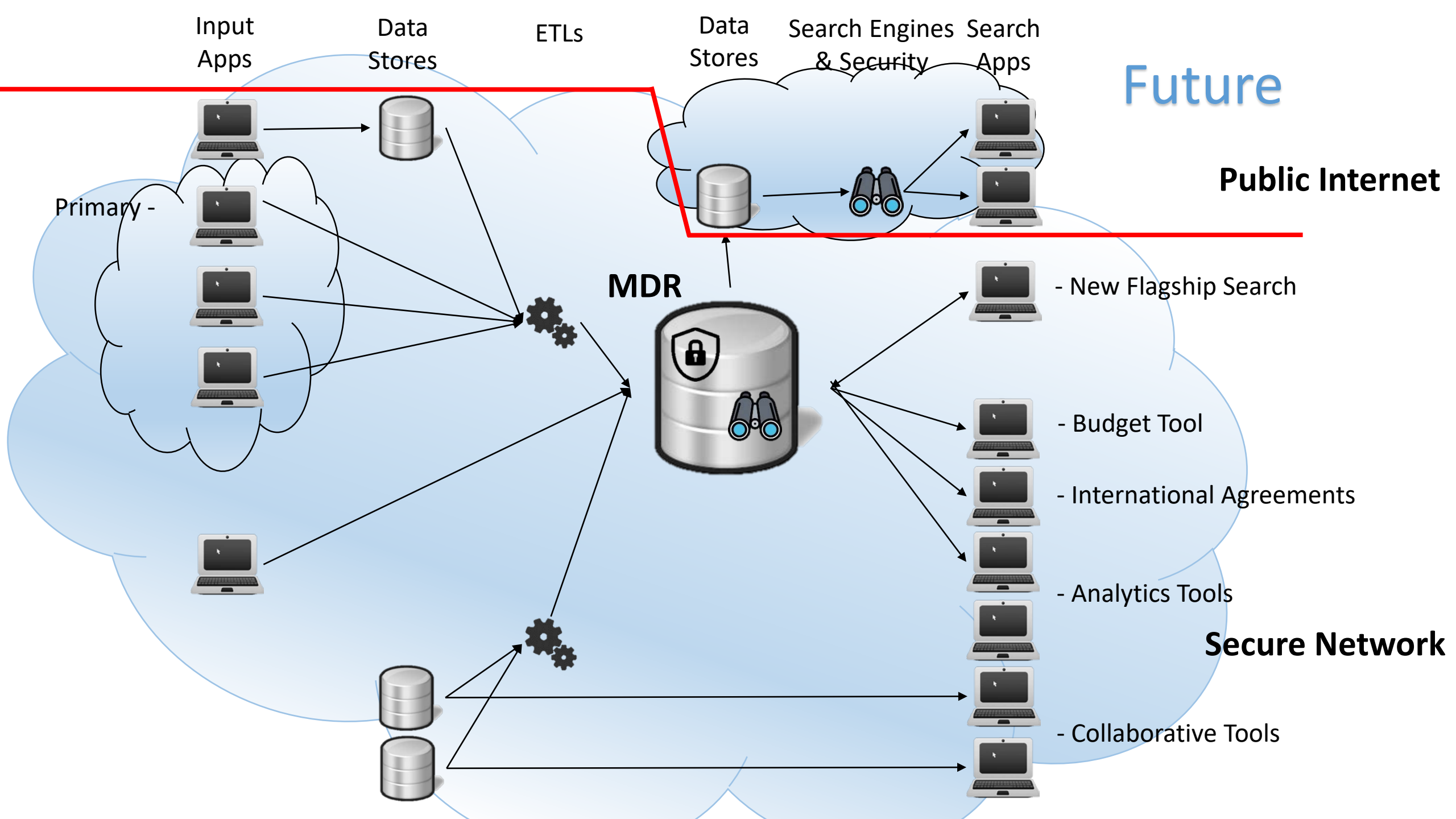
Technology

- √ Provision hardware
- √ Install NoSQL database
- √ Refactor Interfaces
- √ Create services for data upload, input, query, export
- √ Ingest data
- √ Build Access Control
- √ Integrate with DTIC architecture


Input Apps
Data Stores
ETLs
Data Stores
Search Engines & Security
Search Apps
2015
Primary -
Public Internet
- Flagship Search
- Research Projects
- Beta Search
- International Agreements
- Analytics Tools
Secure Network
- Collaborative Tools


Input Apps
Data Stores
ETLs
Data Stores
Search Engines & Security
Search Apps
Today
Primary -
Public Internet
MDR
- New Flagship Search
- Research Projects
- Budget Tool
- International Agreements
- Analytics Tools
Secure Network
- Collaborative Tools


Input Apps
Data Stored
ETLs
Data Stores
Search Engines & Security
Search Apps
Fall 2019
Public Internet
Primary -
MDR
- New Flagship Search
- Budget Tool
- International Agreements
- Analytics Tools
Secure Network
- Collaborative Tools


Input Apps
Data Stores
ETLs
Data Stores
Search Engines & Security
Search Apps
Future
Public Internet
Primary -
MDR
- New Flagship Search
- Budget Tool
- International Agreements
- Analytics Tools
Secure Network
- Collaborative Tools

Continuing Challenges

- Learning NoSQL and XQuery
- Leveraging out-of-the-box solutions
- Applying lessons learned
- Consistent resourcing
- Consistent processes and tools

Contributors to Success

- Clear and consistent goals with tangible functional and cost benefits
- Buy-in from leadership
- License expiration
- NoSQL database
- Pilot project and professional services
- Involve all directorates
- Data and customer product knowledge
- Focus on the backend
- Remove dependencies
- Follow vendor recommendations

Thank You!