CIRDS: CTP Integrated Research Data System

Mark Gingrass[1], Vibha Kumar[1], Christine Wang[1], Wei Chen[2], Geetha Potluri[2], Alex Abramson[2], Nirmala uralikrishna[2], Wei Wu[3]
FDA/CTP/Division of Research Science Informatics[1], SAIC[2], MarkLogic[3]

CHALLENGES / BUSINESS NEEDS

Distributed information across both internal and external sources creates a significant challenge for the Center for Tobacco Products (CTP) personnel responsible for the Agency's tobacco regulation. Specifically, the CTP researchers and scientists need to

- search heterogeneous data sources for comprehensive studies (web URLs, different searching capabilities and configurations, import/export formats, etc.),
- access disparate systems use different technologies to express similar concepts (syntax and semantic level),
- maintain advanced search capabilities across heterogeneous applications, which leads to higher overall system costs, and
- find a User Interface (UI) and host environment for the iDAT (Industry Analysis Document Tool) data given that its Web UI will be decommissioned – this also posts technical challenges of ingestion of new data sources.

GENERAL INTRODUCTION

CTP Integrated Research and Data System (CIRDS) has been developed to address the above challenges. CIRDS employs MarkLogic's Enterprise NoSQL-based operational and transactional data integration platform to provide a Web-based UI for all CTP researchers and scientists to explore different data sources in a unified way.

CIRDS has gone through the following stages:

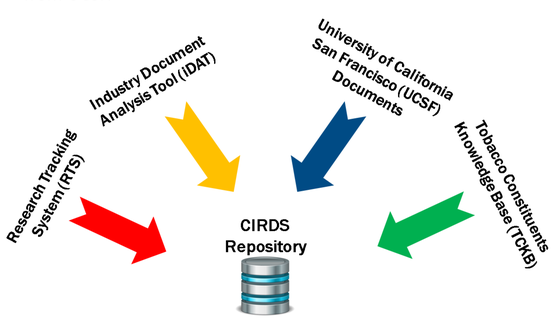
- From 2016 to 2019 CIRDS has been designed and developed as a Proof of Concept (POC) solution.
- In the past year CIRDS has been revised and improved to a state that paves a solid groundwork for advancing itself as a production system.
- Currently, CIRDS is in a transition process from POC to production.

OBJECTIVE

The objective is to introduce CTP's Integrated Research Data System (CIRDS) platform that allows users to search, explore, discover and disseminate information from multiple data sources.

So far, four disparate data sources have been ingested:

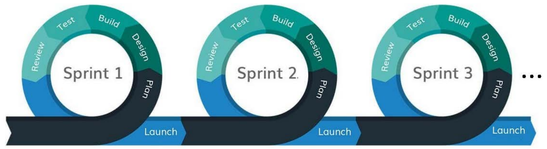
- CTP's Internal Research Tracking System (RTS),
- CTP's internal Tobacco Constituent Knowledgebase (TCKB),
- CTP's internal Industry Document Analysis Tool (iDAT),
- and the publicly available Truth Tobacco Industry Documents from UCSF.



SOLUTION APPROACH

Agile Methodology

CIRDS development and requirements gathering follow an Agile approach with small steps starting from November 2016 until September 2019. The development effort between September 2019 to September 2020 has been extremely rigorous following an Agile Scrum methodology with close communications with the COR and the CIRDS and iDAT users, streamlined tasks, adequate and dedicated QA/testing, ambitious and successful software releases, and the adaption to changing requirements.



CIRDS High Level Architecture

The CIRDS integrates multiple data sources into one repository employing multiple workflow steps:

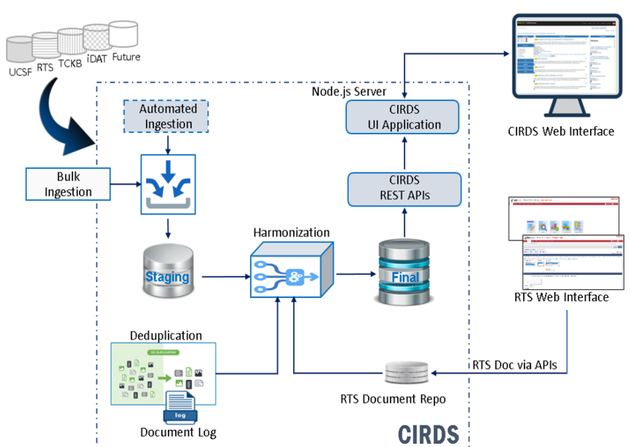
Data Harvesting- identifies the data source(s) and individual data sets within the source repository that is targeted for the system.

Data Ingestion- the data sets from the various sources are input as JSON formatted data records (a format for structuring data transmitted between a server and a web application). These records are stored in a "staging" database.

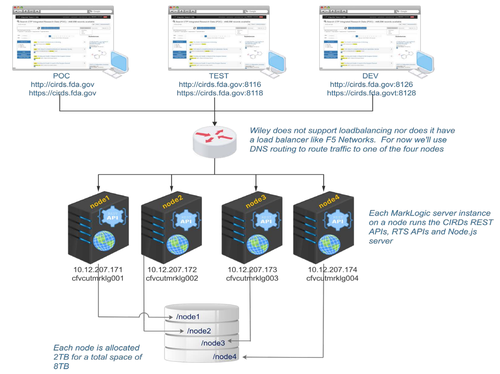
Harmonization- includes adding common fields for documents/records and identifying and indexing substances that are mentioned within each of the RTS projects and UCSF document records.

Cured Data Repository-the harmonization step produces the CIRDS Final Repository that is employed by the CIRDS Proof of Concept (POC) application. This repository is where records are stored for the query process and allow users to retrieve the records after completing a search.

CIRDS Components and Data Flow



CIRDS System Infrastructure

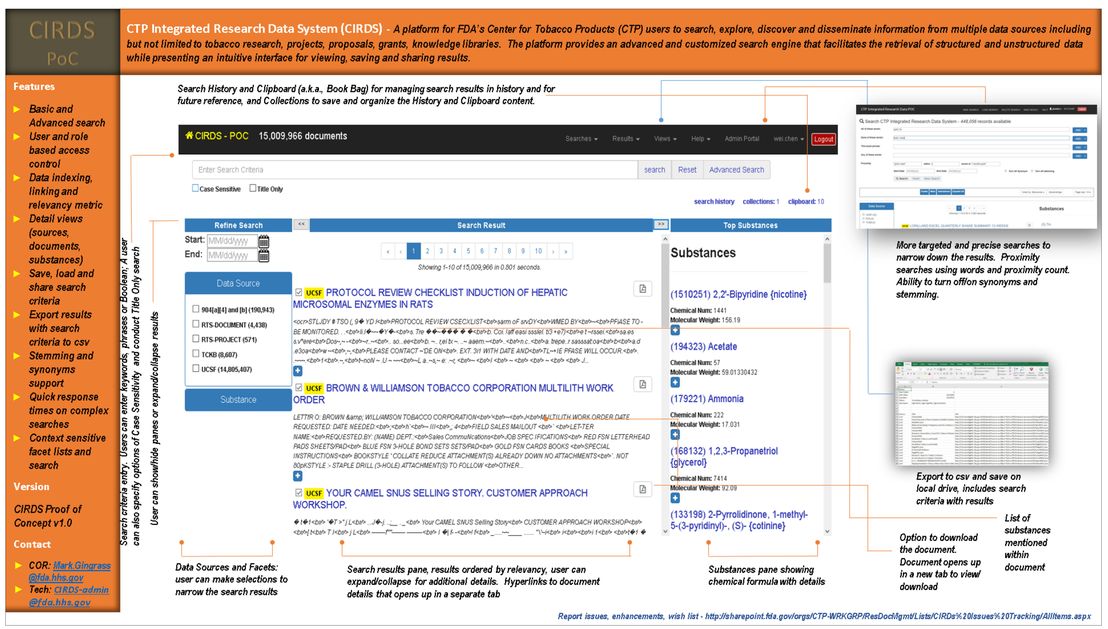


RESULTS

Features

- Unified UI displays all the data sources (pre/post search)
- Powerful search capability with customizable synonyms list and flexible options including Clipboard and Collections, etc.
- Effective system monitoring, management and security
- Highlighted search results with ability to Save and Export to MS Excel or Endnote.
- Developed a streamlined data ingestion process
- Employed MarkLogic as the integrated document search management system
- Adopted role-based user management

CIRDS Interface



Deliverable and Achievements

- Ingestion of multiple heterogenous data sources, including RTS, TCKB, iDAT (with 200K documents from its Dev/Test server), UCSF Truth Tobacco Industry Documents (about 15M documents)
- Accommodation of more than 15 million records
- Close communications and positive feedbacks from the CIRDS and iDAT userbase
- Secure data transfer channel
- The MarkLogic upgrade (v10.0-2.1) and the associated system revisions
- Robust application with the functionality to support every-day needs of its user community.
- 8 Sprints and the corresponding software releases have been completed in the past year
- A large set of documentation maintained at the CTP SharePoint site, the whole CIRDS code base stored in CTP SVN, system development records hosted in CTP Jira system.

NEXT STEPS

The enhanced Proof of Concept (POC) of CIRDS has gone through rigorous requirement analysis, system design, software development and revisions, QA/testing, and multiple rounds of feature updates and releases. CIRDS will enter the preproduction and production stage in the coming year.

In the meantime, CIRDS is going through

- a transition period until late 2020 to support the upcoming preproduction and production effort,
- minor software development and bug fixes, and
- The development of documentation and training materials.