RBEE IE Intern Project

Rules Based Extraction Engine

Phi Henry Nguyen

amazon

Agenda

- Introduction to RBEE
- Existing Services & Architecture
- Project Motivations & Goal
- Solution Overview
- Project Demo
- Impact & Future Work
- Learnings & Reflections
- Q/A

amazon

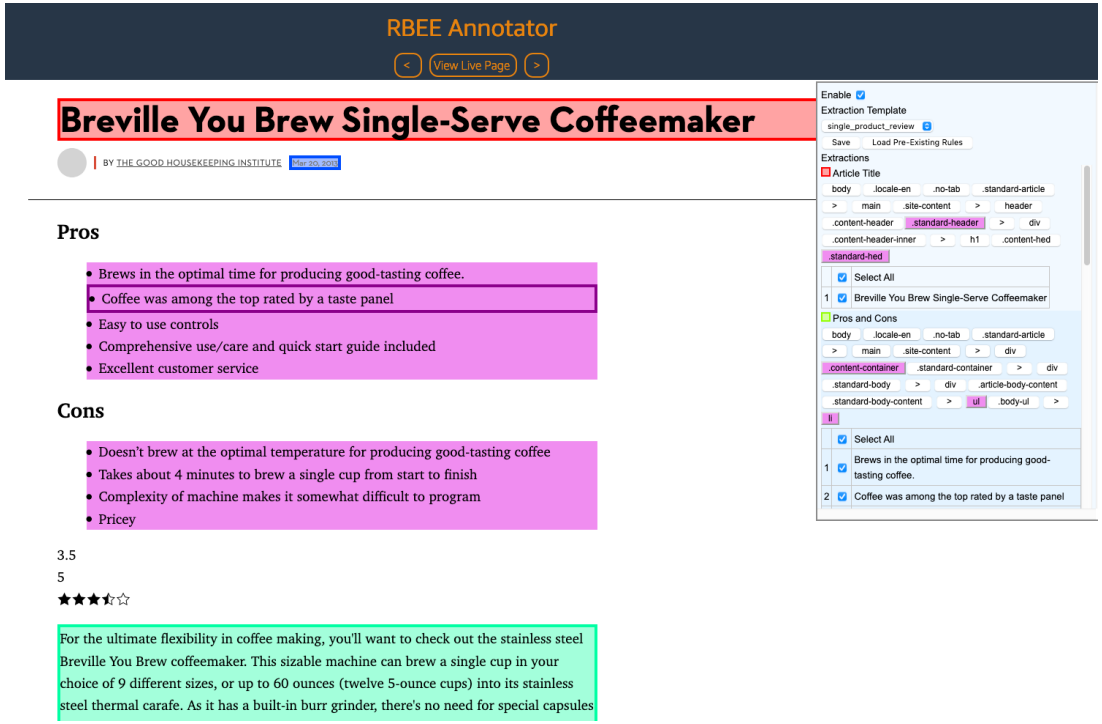
Introduction to RBEE

- RBEE – Rule Based Extraction Engine
  - RBEE is a framework of tools, rules, heuristics, and ML models for scraping information of HTML documents in an Information Extraction document processing pipeline
  - Templated scraping rules will form the basis for scraping, extraction, distant supervision of ML models, and wrapper induction
  - We will cover enhancements to the RBEE rule annotation tooling today to aid in scraping rule development

amazon

Introduction to RBEE Annotation Tooling

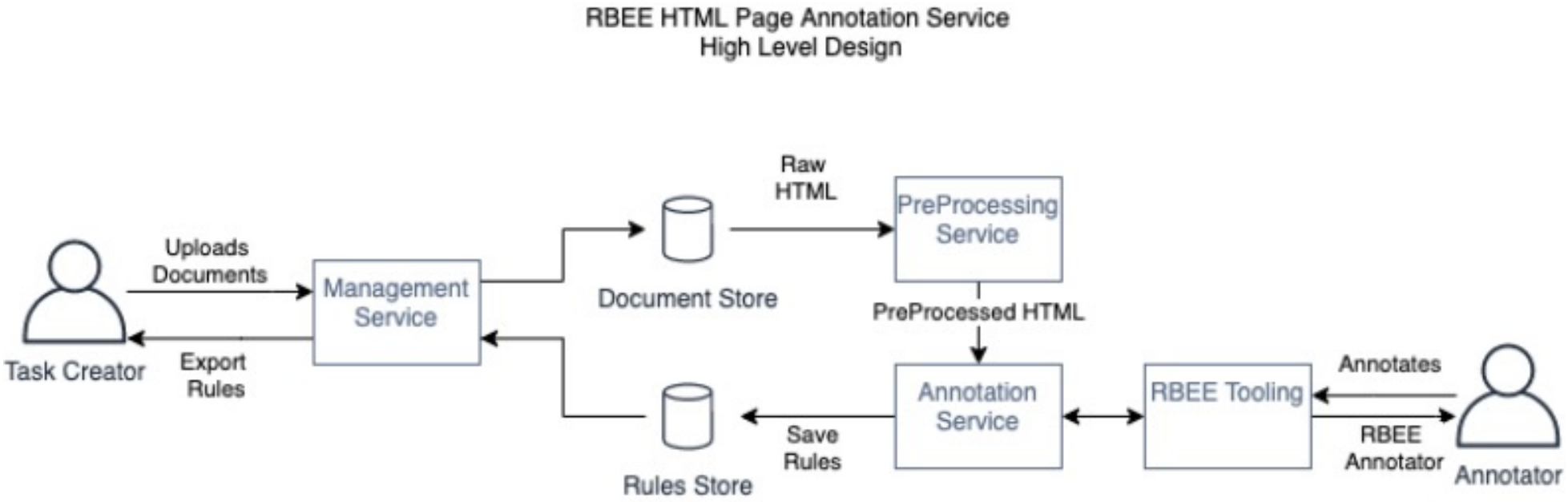
- Provides an effective way to quickly annotate web page attributes with CSS selector path rules
- Used to generate a large collection of web-sourced documents and data for building models for the Alexa Assisted Search team



Existing RBEE Annotation Services

- **Management Service**: where users can create templates, upload batches of documents, and create tasks
- **PreProcessing Service**: used to populate an HTML document with RBEE tooling for annotations
- **Annotation Service**: where users fulfill tasks by annotating web documents with RBEE tooling

Existing RBEE Annotation Tooling Architecture



Rule Annotation vs Scraping/Extraction

- **Rule Annotation**: Manually choosing elements on a page and associating its CSS path with a templated extraction type (ie “Article Title” or “Product Name”)
- **Scraping/Extraction [new]**: Applying the manually created annotations/rulesets to new documents and extracting its corresponding elements

amazon

Project Motivation

- Logical next step for RBEE Annotation Tooling
- Users need a way to apply rules across a domain of web pages to extract features
- Allows users to quickly generate annotated data for development purposes

Design Process and Project Requirements

Met with RBEE users to finalize project requirements:

- Users can upload a batch of documents for information extraction
- Users can choose a previous annotation/ruleset for use in an extraction
- Each document extraction is saved to a database
- Users can manually edit extractions
- Users have a method to view and download results
- Note: intended for development, not large scale document processing

amazon

Project Goal

To build a system that allows users to upload a batch of documents for extraction, choose a ruleset/annotation, and view the extracted documents.

Two main focuses of the project:

1. Integrate within existing RBEE infrastructure to provide users a single application to annotate documents and extract information
2. Decouple the task of extraction for development from annotations

Solution Overview

Integrated system within the RBEE Management Service

- Enables new information extraction features alongside existing functionality of the existing RBEE product

Decoupled data storage between Annotations and Extractions

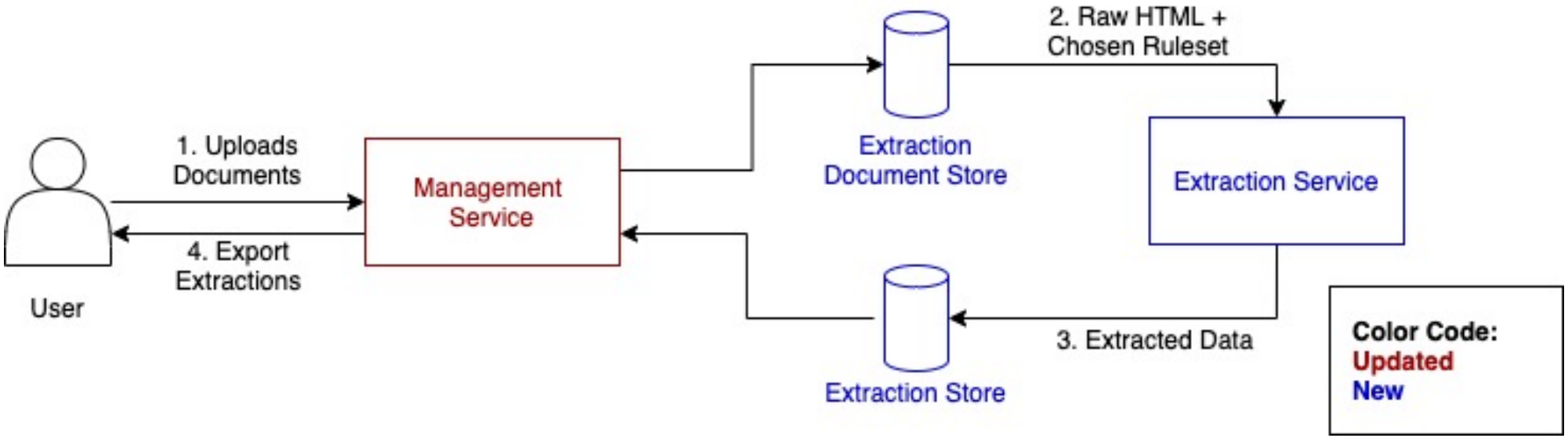
- New tables created for documents, tasks, and extractions as reusing tables would create unnecessary confusion and coupling

RBEE Extraction built out as a Lambda service

- Best suited as a lambda since it is event driven and only called when a user uploads documents to be extracted

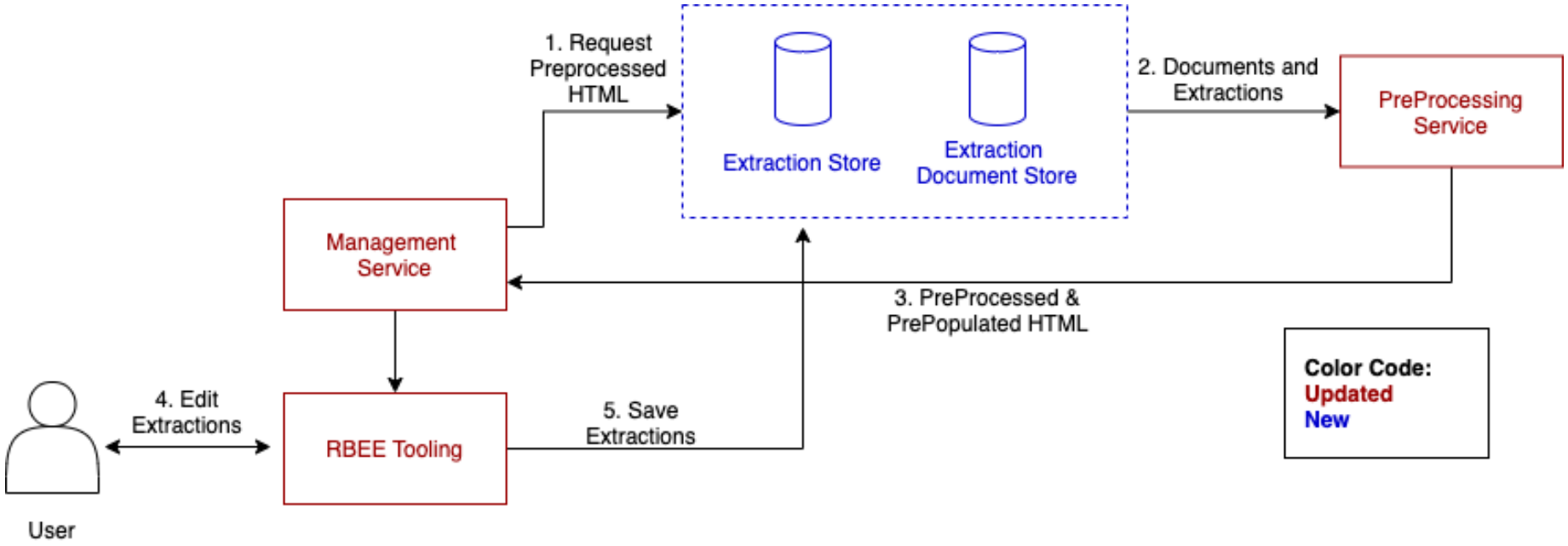
RBEE IE Architecture

Extraction Creation



RBEE IE Architecture

Extraction Editing



Project Demo

User Feedback

"RBEE is very helpful for annotating rules for different extractions from raw HTML. It renders the page and allows one to select the tags associated with a particular extraction. This is essential to create distantly-labeled data for such extractions. I look forward to the next version of RBEE that supports more complex rules (e.g. using logical operators) and that is easy for non expert annotators to use."

amazon

Performance

Tested on AWS console

Scales linearly

- Batch of 100 documents: ~51 seconds
- Batch of 20 documents: ~9.6 seconds

amazon

Impact & Future Work

This project will:

- Allow the Ambient Explorer team to utilize RBEE to generate large collections of annotated data from the internet
- Allows exploration of web scraped data through an easy-to-use web interface

Future work:

1. Rule Reconciliation- Allow users to choose multiple annotations/rulesets
2. UI Refresh – Update front end to use AWS UI
3. Functionality and usability improvements for the Annotation Service
4. Schema packaging and built-in verifications

amazon

Learnings & Reflections

- Learned to use various AWS technologies and CI/CD tools such as Brazil, CloudFormation, Lambda, DynamoDB, S3, and many more
- Experienced the whole project and system creation process through research and design, developing and addressing feedback, and deploying
- Learned about code ownership and responsibility, since I was the only person working in RBEE Annotation Tooling full-time after Michael left

amazon

Acknowledgements

Thanks!

Any Questions?

amazon