# MarkLogic Data Workbench Vision and Scope

Team Flash
*Computer Science Department*
*California Polytechnic State University*
*San Luis Obispo, CA USA*

October 2, 2018

# Contents

# Credits

| Name | Date | Role | Version |
|---|---|---|---|
| Joseph Buelow | October 2, 2018 | Scope and Limitations | 1.0 |
| Diana Chiu | October 2, 2018 | Competition Analyst | 1.0 |
| Bonita Galvan | October 2, 2018 | User Descriptions and Vision of the Solution 3.3 and 3.4 | 1.0 |
| Victoria Law | October 2, 2018 | Business Context | 1.0 |
| Kent Tran | October 2, 2018 | Business Requirements | 1.0 |
| Tanner Villarete | October 2, 2018 | Vision of the Solution 3.1 and 3.2 | 1.0 |

# Revision History

| Name | Date | Reason for Changes | Version |
|---|---|---|---|
| Kent Tran | October 2, 2018 | Business Requirements | 1.0 |
| Diana Chiu | October 6, 2018 | Competitive Analysis | 1.0 |
| Bonita Galvan | October 6, 2018 | User Descriptions | 1.0 |
| Victoria Law | October 7, 2018 | Business Content | 1.0 |
| Diana Chiu | October 7, 2018 | Formatting and Editing | 1.0 |
| Bonita Galvan | October 7, 2018 | User Description Editing | 1.0 |
| Victoria Law | October 7, 2018 | Business Content Editing | 1.0 |
| Jospeh Buelow | October 7, 2018 | Scope and Limitations Editing | 1.0 |
| Bonita Galvan | October 7, 2018 | Vision of the Solution 3.3 and 3.4 | 1.0 |

# 1 Business Requirements

## 1.1 Background

The proper use of data has never been as important as it is today in all fields, including traditionally nontechnical ones such as Journalism. In the past, data could be protected, organized, and standardized by limiting access to small groups of highly trained technical individuals. Now that most reporting tools are simple and self-service, there are exponentially more data sources available in widely different formats. Coupled with the evolution of privacy regulation, it has become increasingly more difficult for organizations to effectively manage their data.

Transparency, reproducibility, and verifiability are key requirements to establishing trust in published articles, research, and even legal findings. Due to the ever-expanding collection of available data-sets, it is growing more difficult to find, analyze, and share results. Organizations cannot fathom exactly what and how much data they store. Digging through all this data is extremely difficult as the privacy of sensitive information must be respected during research, the authoring process, and the final results.

These requirements put a number of burdens on data users:

1. They must be able to find data that is relevant to their topic area from amongst vast amounts of data.

2. They must carefully track the lineage of source data and the manipulations performed on that data. This metadata must be available, alongside the final published content, to journalists, academics, and other information consumers.

3. They must methodically redact or anonymize data to maintain privacy.

4. They must be able to update their results if the input data changes or analysis methods change. This includes the fact that some source data is temporal in its accuracy.

## 1.2 Business Opportunity

Our data classifying system will benefit both MarkLogic and their customers by allowing data to be more intelligently sifted, sorted, and searched. It also provides our team the opportunity to develop a system from start to finish, all the while working with MarkLogic as the customer. Over the course of the year, we will learn what it is like to develop iteratively as we will hold regularly scheduled meetings with our client to ensure the dynamic requirements are met.

## 1.3 Business Objectives and Success Criteria

The objectives for this system by which success will be measured is through the following five deliverables:

1. An Element Category GUI that allows users to see predefined and learned categories of data elements that will be recognized by the Data Classifier. It must also allow users to label data element categories as sensitive.

2. The ability to create machine learning models that will be used to identify categories of data from input data sets.

3. The ability to apply models to input data sets to do the actual categorization of the data elements.

4. The ability to process learned data by generating a machine-readable description of an input data set based on the categorization process.

5. A Data Classifier GUI that allows users to browse, edit, and add to the classifications that were performed automatically.

## 1.4 Customer or Market Needs

The need for this system by the customer, MarkLogic, is to improve upon the company's storing and searching of data. There is no pressing deadlines other than those given to ourselves in respect to the time line of the Capstone course.

### 1.5 Business Risks

There is little risk associated with this project as MarkLogic currently does not have any similar systems in place.

## 2 User Description

### 2.1 User/Market Demographics

The main demographic for MarkLogic's Data Workbench is data analysts interested in combining diverse data sets into a single classified Data Catalog. These users may have varying levels of technical skills, from highly computer-literate data analysts with coding and scripting abilities, to health care business analysts whose infrequent software use is limited to specialized systems.

### 2.2 User Personas

1. User: **Barry Allen** is 52 year old male ***Data Analyst***. He has a Masters of Information and Data Science from UC Berkeley. He has worked in the data analytic industry for over 25 years and has experience coding and scripting. His job involves using multiple database software and languages including, SQL, SAS, and BigQuery. He is familiar with statistical modeling, machine learning, and data mining. His goals are to be able to apply advanced statistics and mathematical techniques to huge data sets. he finds frustration in the time it takes to normalize various data sources and formats into one data catalog to work with. He would like our program to be able to take large data volumes with new and evolving information and categorize the data into one cohesive catalog for him to use to uncover new insights.

2. User: **Jesse Quick** is a female 25 year old ***Business Analyst*** who is fairly new to the industry. She has her Bachelors in Business Administration from Cal Poly. She is familiar with only a few business software programs and has not done much work with large amounts of data. While in college she took one class on data systems and has basic knowledge of SQL. Her job is to look for trends, interpret and evaluate data in order to gather critical information to hand over to

various stakeholders and produce useful reports. She needs to find a clean and easy way to navigate large sets of data. She wants the ability to categorize data and flag categories important to her research.

3. User: **Wally West** is a 30 year old male ***Health Care Business Analyst*** with a Masters in Health Administration from Cornell University. He has no coding experience. He has always used the same software program to visualize his data, but it can only take in one data set at a time. His job includes sifting through information to determine what opportunities to pursue that would relate to the needs of his clients and his company. He needs to be able to combine the data from multiple clinical trials, each with their own format, into one to determine which approaches are viable for the company. His frustrations include using a new program and trusting that it will combine information correctly. He will need the program to be easy to understand, create categories and determine which ones should be marked as sensitive, and hide those sensitive categories from people who receive his reports and should not have access to them.

## 2.3 User Environment

Due to the nature of the users, the following restraints are in place:

1. The system must work on popular Linux distributions, such as Ubuntu, Debian, and Fedora, as well as Mac OS.

2. There must be a GUI that is visible on modern web browsers such as Chrome, Firefox, and Edge.

3. The application must run in real-time.

## 2.4 Key User Needs

The user must be able to:

1. Upload multiple, varying data sets into the system

2. See system-defined classifications and descriptions of the data

3. Edit, add, and remove classifications

4. See and manage the *sensitive* data classifications
5. Track the lineage of data
6. Obtain a machine-readable Data Catalog created by the system.

# 3 Vision of the Solution

## 3.1 Vision Statement

1. Our goal is to organize the world's information and make it universally accessible and useful. We intend to do this by designing an all-encompassing data classifier system that includes an artificially intelligent classification system as well as a graphical user interface that the customer can utilize to effectively interact with their data.

## 3.2 Solution Overview

1. The final product will be an all-encompassing data classifier system that includes an artificially intelligent classification system as well as a graphical user interface that a customer can interact with.

## 3.3 Major Features

Major features of system will include:

1. **Element Category GUI**: Allows users to see predefined and learned categories of data elements that will be recognized by our data classifier system. Users will be presented with both category name and description. Users will be able to label data element categories as *sensitive*.
2. **Machine Learning- *Create Models***: Will use process training sets to create models to be used to identify categories from varying input data sets.
3. **Machine Learning- *Apply Models***: Apply models to user given data sets to do preform the system categorization of the data elements.

4. **Process Learned Data**: The system will be able to generate a machine readable description of the input data set based on our systems data categorization process. The output will be in the form of an XML or JSON file.

5. **Data Classifier GUI**: Allow users to browse, edit, remove and add to the classifications and tags created by our classifier system.

### 3.4 Assumptions and Dependencies

1. The user will provide the data sets.

2. The data sets will all be of the same file formats such as CSV or JSON.

## 4 Scope and Limitations

### 4.1 Scope of Initial and Subsequent Releases

Initial Release targets the end of CSC 405, March 2019. One or two subsequent releases will occur in CSC 406, April-June 2019. Detail on the contents of these releases will come closer to those dates.

### 4.2 Limitations and Exclusions

A major limitation will be the amount of hardware dedicated to this system. The system will accept multiple data sets however they must have the same structure. The system will run on Linux.

## 5 Business Context

### 5.1 Stakeholder Profiles

| **Stakeholder** | **Value** | **Interests** | **Constraints** |
|---|---|---|---|
| MarkLogic customers | Major Functionality | Ease of use | None |
| Data Scientists | Convenience | Ease of use | None |
| Bruno Da Silva | Growth | Quality of Project | None |
| Development Team | Functionality | Project Completion | None |

| Dimensions | Drivers | Constraints | Degree of Freedom |
|---|---|---|---|
| Schedule | First iteration released by the end of March, 2019; final release by the end of June 2019. | | |
| Features | | Difficulty, scope, and feature | Create a machine learning model to identify classes of information contained in the data sets. New categories may be learned by the Data Classifier. |
| Quality | | Time | |
| Staff | Bruno Da Silva | Project team consists of 6 student developers. We will meet with representatives from MarkLogic on a weekly basis. | |
| Cost | | Each student developer must spend a maximum 12 hours a week with little variation. | Each student developer can vary the amount of hours they spend on the project. |

### 5.3 Operating Environment

| | |
|---|---|
| OE - 1 | The back-end should run on a popular Linux distribution (Ubuntu, Debian, Fedora) or Mac OS. |
| OE - 2 | The user interface should be viewable in a modern web browser (Chrome, Firefox, Edge). |
| OE - 3 | The application should support MarkLogic Database Version 9 |
| OE - 4 | The application should be compatible with NoSQL databases. |

## 6 Competitive Analysis

### 6.1 Overview

MarkLogic has many competitors for both data storing and cataloging. The leading data catalog competitors are summarized below. Like MarkLogic, they all follow the Software As A Service model.

### 6.2 Waterline Data

Waterline Data aims to make large collections of varied data more accessible by making it easily searchable and providing lineage, uses, and value to the organization. Waterline Data primarily focuses on data cataloguing. They break up their product into the values it provides- data "fingerprinting", curation, search, access control, and lineage. In addition to the features their platform provides, Waterline Data also advertises easy integration with various platforms, storage solutions, and analytic tools.

Customers interested in Waterline Data can contact Waterline Data to schedule a demo and work out onboarding and pricing. They can also download a demo version from the Waterline Data website.

### 6.3 Alation

Atlation provides an automated data catalog that emphasises easy collaboration and easy-to-use browser applications. Alation's advertising targets novice users without much technical knowledge. The metadata they tag data

with is automatically determined based on context, and the cataloger learns the value of data over time by monitoring user queries. The product has a way to see a data lineage tree that includes consumers of the data being viewed. Atlation also allows users to tag data to indicate endorsement, warnings, and deprecation.

Atlation also does not put the price of their product on their website. Interested customers can schedule a demo for more information.

## 6.4 Unifi

Unifi promotes their AI and how it makes their data catalog easy to use. They also emphasise the importance of collaboratively curating data by providing features for tagging users, and discussing and sharing data. Unifi has the power to limit access to certain users, and automatically protect new data.

A demo of Unifi can be requested from their website, but they also do not list prices.

Unifi has been named a market leader in self-service data prep by Ovum in 2018. In the Forrester Wave: for Machine Learning Data Catalogs, Q2 2018, Unifi was named a leader for their intuitive UI, a single search field backed with natural language processing that can answer sophisticated questions such as "What was the revenue trend from the past three years?".