# Data profiling for data quality improvement with OpenRefine

**Conference Paper** · October 2016

DOI: 10.1109/ICITSI.2016.7858197

Tien Kusumasari
Telkom University
**17** PUBLICATIONS **27** CITATIONS

2016 International Conference on Information Technology Systems and Innovation (ICITSI)
Bandung – Bali, October 24 – 27, 2016
ISBN: 978-1-5090-2449-0

# Data Profiling for Data Quality Improvement with Openrefine

Tien Fabrianti Kusumasari
Information System Department
School of Industrial Engineering, Telkom University
Bandung, Indonesia
tienkusumasari@telkomuniversity.ac.id

Fitria
Informatics Department
School of Electrical Engineering and Informatics, ITB
Bandung, Indonesia
fiitriiaa@students.itb.ac.id

***Abstract*—Data profiling is an information analysis technique on data stored inside database. Data profiling purpose is to ensure data quality by detecting whether the data in the data source compiles with the established business rules. Profiling could be performed using multiple analysis techniques depending on the data element to be analyzed. The analysis process also influenced by the data profiling tool being used. This paper describes tehniques of profiling analysis using open-source tool OpenRefine. The method used in this paper is case study method, using data retrieved from BPOM Agency website for checking commodity traditional medicine permits. Data attributes that became the main concern of this paper is Nomor Ijin Edar (NIE / distribution permit number) and registrar company name. The result of this research were suggestions to improve data quality on NIE and company name, which consists of data cleansing and improvement to business process and applications.**

***Keywords—data profiling; Openrefine; data quality improvement; analysis single column; analysis multiple columns***

## I. Introduction

Data is an important asset to an organization or a company [1]. Organizing data can improve quality of data and be added value for the organization. One of the techniques that could be applied to ensure data quality is data profiling [2]. Data profiling is a process of examining the data available in a data source and collecting statistics and information of that data [3]. Data profiling is defined as the application of data analysis techniques to existing data stores for the purpose of determining the actual content, structure, and quality of the data [2]. Data profiling is the set of activities and processes to determine the metadata about a given dataset [5].

The purpose of data profiling and data cleansing implementation is to ensure data quality, to measure data accuracy and consistency, and to detect data duplications [2] in order to obtain data with the correct value that could be used in decision-making purposes. Data profiling could be used as an approach to assess data quality [2]. Data profiling also used to create knowledge about the data itself [2]. Result of the data profiling process must be maintained incrementally to ensure existing data is the most current and could be used for data warehousing purposes and data mining [2]. The data profiling process are also used to prepare for data cleansing, formulate query optimizations, data indexing, scientific data management and database reverse engineering [3]. Data with better quality would correspond to positive impacts to the organization. Good data quality in customer relationship management systems will save costs on new system developments, and minimizing data problems in supply chain management systems would improve production quantities [2].

Data profiling process are also used to prepare for data cleansing process [3]. Data cleansing is the next step after data profiling in order to have better quality. The result of data profiling would reveal groups of data with similar characteristics (clusters). Data cleansing not only removes existing data but also improve data by updating data that does not comply to the rules. Data removal by deletion is only done when necessary. Data cleansing are performed by special query or algorithms [5].

Ziawasch Abedjan and partners built a tool for data profiling dan cleansing called ProLOD++ [6]. ProLOD++ is a tool to analyze and improve open RDF (Resource Description Framework) data. ProLOD++ are composed of normal data profiling methods that are adapted to RDF data model. In addition to these methods, specific features for open data are included, for example : determining user generated attributes schema and association rule discovery [6].

In 2016 a research group in Qatar researched the relation between metabolomics data and diabetic disease using data profiling techniques [7]. In this research, the data technique being used are binning, smoothing, filtering and data clustering. The result of such data clustering shows that after thorough examination of data from patients with diabetic disease and control patients, some metabolites correlates with the diabetic patients.

Several application tools existed to help with data profiling process, among them are : IBM Infosphere Information Analyzer, Oracle Enterprise Data Quality, Talend Data Quality SAP Business Objects Data Insight & SAP Business Objects Information Steward, Informatica Data Explore, and OpenRefine.

Most of the tools are paid tools with relatively high budget requirements, making them difficult to implement in small to medium enterprises in developing countries. In the opposite spectrum we find open-source tools that requires much smaller budget. The focus of this paper is to provide data profiling and cleansing with OpenRefine, which is one of the open-source tools. OpenRefine, formerly Google Refine, is a data quality application with capabilities of analyzing messy data, data cleansing, and also transforming data from one format to another.

## II. Literature Review

### A. Data Profiling Analysis

In order to obtain detailed profiling results, several types of analysis are available. They are column property analysis, structure analysis, simple data rule analysis and complex data rule analysis [2].

Column property analysis (Fig. 1) is an analysis that are done by checking each values in a column and determining whether the values are valid or not [2]. The validity check process requires rules that the values need to conform to. Such rules usually found in the metadata, which are called column specification or domain definition.

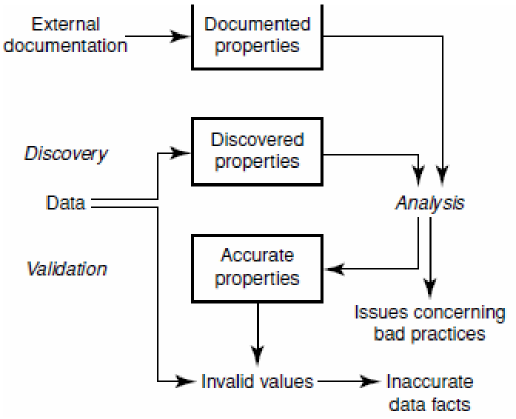


Fig. 1. Column Property Analysis Process [2]

Structure analysis is the next step after column property analysis. Two basic points that are done in structure analysis is to find data that does not comply to the rules, which shows inaccurate data, and to document the structure metadata rules. The documentation will be useful for other data structure mapping process, moving data to other system and merging data with other data elements [2].

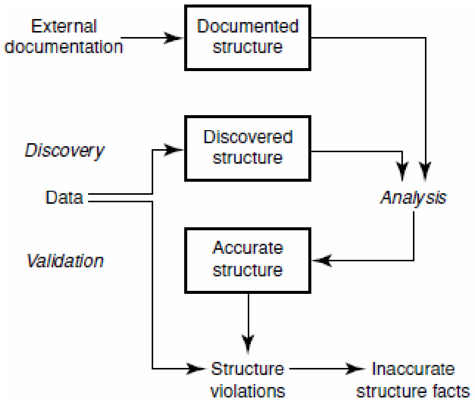


Fig. 2. Proses of Structure Analysis [2]

After column property analysis and structure analysis, data rules should be established to avoid same data mistakes. Simple rule analysis is done by applying data rule to determine whether the data is accurate or not. The rule could involve one column or several columns in the same table, or different column in different tables [2], but for simple rule it only involves one business object. Process of structure analysis can be shown in Fig. 2.

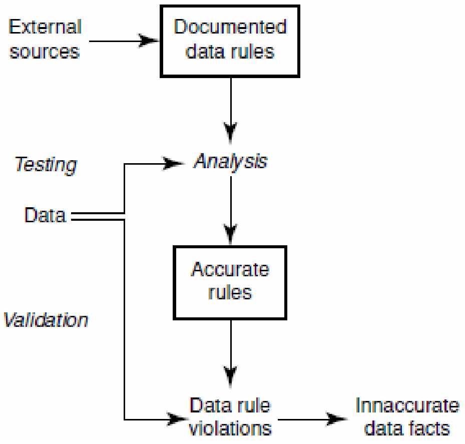


Fig. 3. Proses of Simple Rule Analysis and Complex Data Rule Analysis [2]

Complex data rule analysis is similar to simple data rule analysis, where we try to identify as many rules as possible to identify inaccurate data, but in this case a set of business objects are involved [2]. Process of simple rule analysis and complex data rule analysis show in Fig. 3.

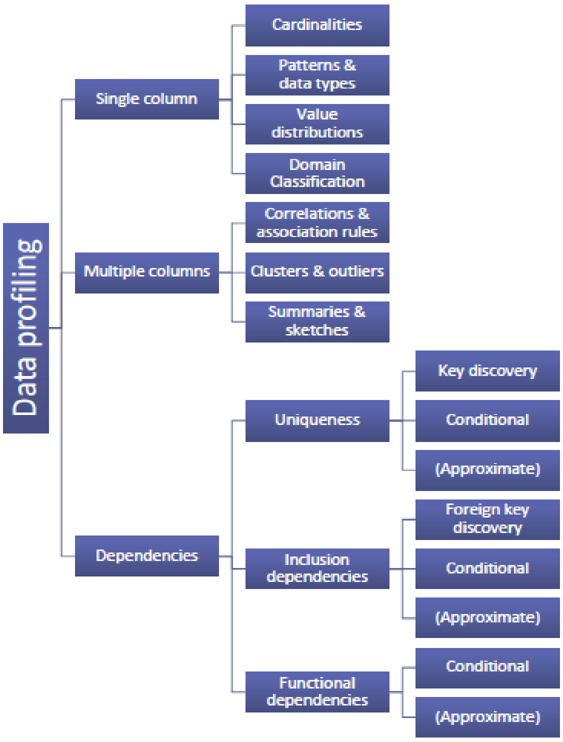


Fig. 4.Data Profiling Method Classification [4]

Data profiling could be done with several guiding processes, they are column profiling, cross-column profiling, cross-table profiling and data rule validation [9]. Classification of data profiling method can be shown in Fig. 4.

Column profiling will result in one column data pattern statistics. Cross column profiling analyzes dependency between attributes in the same table. Cross table profiling evaluates relationship and intersection between tables. Data rule validation would verify that the data matches to established data rules.

Methods of analysis for the data profiling with traditional methods which are single column and multiple columns [9]. Single column technique analyzes values, patterns, and relations from one column of data, and there also techniques that analyze dependencies of one column to another [9].

Single column method is the most basic form of data profiling [4], by assuming all data values within the column is the same kind and have the same general properties. Single column techniques consists of cardinality measurement, data type and patterns, value distribution and domain classifications. Cardinalities is useful to categorize attributes and relevancy among the attributes. Pattern and data types are used in decision-making so it is need to be defined. Data types come in several forms such as string, numbers, and characters. Each data type could have different size in the database such as varchar (12) and varchar(13). Thus data pattern and types need to be defined so it could be used as guidance in decision making.

Multiple column method analyzes not only data in one column but also multiple columns, so it would result in relationship between data elements, schema matching and clustering summary for the same data stored in different columns. A cluster is a group of data having the same actual value but different representations. The result of multiple column methods will show consistency of certain columns.

To analyze data relationship with other column, dependency methods are available. Dependency methods consists of uniqueness, inclusion dependencies and functional dependencies. Uniqueness analysis determines whether the rows have unique value in one or more key columns, and whether there are data duplication that occurred. Inclusion dependencies is used to determine whether in a data attribute that has a foreign key have the same data in the primary key. Functional dependencies are used to describe relations, constraints, and dependency between data attributes in a relation.

### B. *Data Profiling Process*

In order to get a proper data profiling result in an effective manner, a methodology that fits the requirements is needed. Data profiling uses a bottom-up approach, from the smallest level to higher levels [2]. Figure 5 shows the steps of such data profiling.

In practice, data profiling have three different phases. During initial profiling, initial basic profiling are performed and data were evaluated. In second phase, integration and automate phase, all kinds of profiling are integrated and automated to monitor changes in data profile. For the third phase, the result are reported to business users, data architects, and developers so they could take action on the data profile conditions of data.

For dynamically changing data, an approach to data profiling is needed. One of such approach is SWAN, a data profiling approach to dynamic data that consists of algorithms on insert and delete [8]. The basic flow of SWAN is to determine whether a group of data is unique or not, then a delete algorithm will be performed to delete duplicate data. For each activity being performed on the data, repository will be updated and change history is recorded.

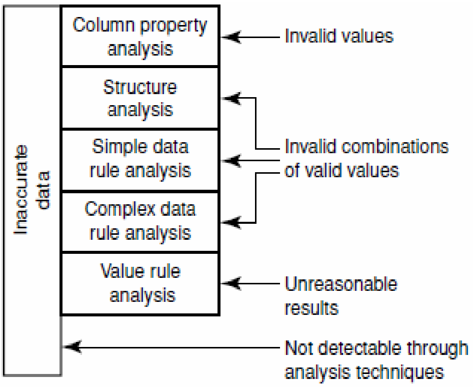

Inaccurate data
Column property analysis
Structure analysis
Simple data rule analysis
Complex data rule analysis
Value rule analysis
Invalid values
Invalid combinations of valid values
Unreasonable results
Not detectable through analysis techniques

Fig. 5. Data Profiling Processes [2]

Data cleansing could be performed using data clustering results of the profiling[5]. Each data are grouped based on their characteristic clusters, then for each group a decision making must be done to determine which data is not relevant. Data which is not relevant will be processed in data cleansing. Data cleansing process consist of permanent data deletion and fixing the data. Such process could be implemented using query or algorithms on data deletion and corrective updates [5].

## III. Method

This paper describes techniques of data profiling using OpenRefine tool. The methodology being used is by collecting types of analysis for data profiling. The types were previously referred as data profile classification. After defining the data profile classification, implementation techniques using OpenRefine tool will be described.

In this research, data profile classification that being used consists of single column analysis and multi column analysis. Single column analysis using data completeness analysis consists of null value, text pattern, text duplication, and text duplication. Multi column analysis consists of correlation and association rule.

Profiling techniques using OpenRefine tool will be performed on data received from website of the Indonesian National Agency of Drug and Food Control / Badan Pengawas Obat dan Makanan (BPOM) which are published using url http://cekbpom.pom.go.id. 5000 records of data were imported to OpenRefine by copy-pasting data from the web into OpenRefine import page. Attributes that were analysed is Nomor Ijin Edar (NIE) and registrar company name field. Nomor Ijin Edar is a number that identifies the permit to distribute food.

After data sample retrieval, the next step is to establish metadata rules on the data. The metadata rules consists business metadata and technical metadata.

Data profiling techniques using OpenRefine is determined according to data profile classification that were selected in this research. Each types of classification is specific to metadata rules. In this research, metadata rules were determined by the authors of this paper. Data profiling were performed using built-in features of OpenRefine, custom regular expressions, or combination of both. A regular expression (regex) is a pattern that the regular expression engine attempts to match in input text [11]. A pattern consists of one or more character literals, operators, or constructs [11].

## IV. Profiling Technique with OpenRefine

Analysis using data profiling in this paper only uses one data source in one OpenRefine project. 5000 row of data sample were retrieved from traditional drugs data in website (http://cekbpom.pom.go.id/). Data profiling were performed on NIE column and registrar company name, which were chosen because these two field are important data master fields that are being used for many purposes in multiple business processes across the BPOM agency.

### A. Business Rule

First step that must be done after data retrieval is defining business rule. The business rule would be used as the basis of business metadata creation. In this research, business rule are established by the authors based on interview results to selected agency staff members. Interviewing method and business rule validation would not be discussed in this paper.

The NIE column has business rule as follows : must not be empty, must be unique for each entity, and have similar alphanumeric pattern. The NIE column must not be empty, if it was empty means a product have no permit, thus violating the laws of traditional drug distribution in Indonesia. On the other hand, NIE must not be the same for different entity, because one NIE shows one certain product in a certain packaging. NIE have standard pattern rules that are established in a certain time period, showing standardization that applies nationally for the multi-region agency.

Analysis types determined by business rules that were established on NIE field and registrar company name field. NIE field profiling were done using blank analysis, duplication, and pattern. On the other hand uniqueness analysis requires multi column analysis, by combining product name field and package type field, and analyzing uniqueness of NIE in relation to these two fields.

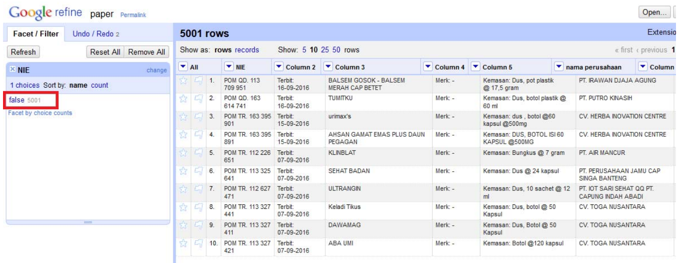
Registrar company name field were profiled using blank analysis, duplication analysis, and cluster analysis, all of which are single column analysis. Multi column analysis also performed in the registrar company name field combined with company address. As a rule the registrar company are allowed to have duplicate rows having different address. So the company name are analyzed together with the address field.

### B. Single Column Analysis

Single column analysis were performed on NIE field and registrar company name field. In OpenRefine tool, available analysis on 1 column consists of blank facet, duplicate facet, cluster (text facet) and pattern. Blank facet is a feature to analyze how many rows are blank on the column.

On the NIE field the analysis performed are : blank facet, duplicate facet, and pattern.

Blank facet analysis on OpenRefine is done by using Facet by blank feature which is found in menu 'Facet > customized facets > facet by blank' while selecting the column to be analyzed. The result of analysis is two Booleans, true and false, and row count for each of these two. "True" represents null/blank column value and "false" represents non-empty column value. Blank facet analysis results are shown in Fig. 6.



5001 rows

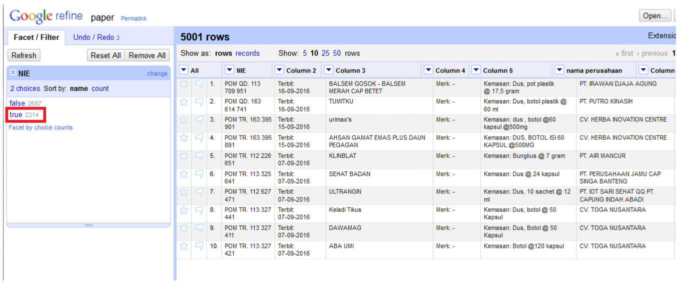
Show as: rows records  Show: 5 10 25 50 rows  « first ‹ previous 1

| All | NIE | Column 2 | Column 3 | Column 4 | Column 5 | nama perusahaan | Column |
|---|---|---|---|---|---|---|---|
| 1. | POM QD. 113 709 951 | Terbit: 16-09-2016 | BALSEM GOSOK - BALSEM MERAH CAP BETET | Merk: - | Kemasan: Dus, pot plastik @ 17,5 gram | PT. IRAWAN DJAJA AGUNG | |
| 2. | POM QD. 163 614 741 | Terbit: 16-09-2016 | TUMITKU | Merk: - | Kemasan: Dus, botol plastik @ 60 ml | PT. PUTRO KINASIH | |
| 3. | POM TR. 163 395 901 | Terbit: 15-09-2016 | urimax's | Merk: - | Kemasan: dus , botol @60 kapsul @500mg | CV. HERBA INOVATION CENTRE | |
| 4. | POM TR. 163 395 891 | Terbit: 15-09-2016 | AHSAN GAMAT EMAS PLUS DAUN PEGAGAN | Merk: - | Kemasan: DUS, BOTOL ISI 60 KAPSUL @500MG | CV. HERBA INOVATION CENTRE | |
| 5. | POM TR. 112 226 651 | Terbit: 07-09-2016 | KLINBLAT | Merk: - | Kemasan: Bungkus @ 7 gram | PT. AIR MANCUR | |
| 6. | POM TR. 113 325 641 | Terbit: 07-09-2016 | SEHAT BADAN | Merk: - | Kemasan: Dus @ 24 kapsul | PT. PERUSAHAAN JAMU CAP SINGA BANTENG | |
| 7. | POM TR. 112 627 471 | Terbit: 07-09-2016 | ULTRANGIN | Merk: - | Kemasan: Dus, 10 sachet @ 12 ml | PT. IOT SARI SEHAT QQ PT. CAPUNG INDAH ABADI | |
| 8. | POM TR. 113 327 441 | Terbit: 07-09-2016 | Keladi Tikus | Merk: - | Kemasan: Dus, botol @ 50 Kapsul | CV. TOGA NUSANTARA | |
| 9. | POM TR. 113 327 411 | Terbit: 07-09-2016 | DAWAMAG | Merk: - | Kemasan: Dus, Botol @ 50 Kapsul | CV. TOGA NUSANTARA | |
| 10. | POM TR. 113 327 421 | Terbit: 07-09-2016 | ABA UMI | Merk: - | Kemasan: Botol @120 kapsul | CV. TOGA NUSANTARA | |

Fig. 6 OpenRefine Blank Facet Screenshot

Duplicates facet analysis is a profiling technique that could be performed to identify field values that have 100% match with the values in other row(s).

Duplicate facet analysis are performed using “Duplicate facet” feature. To separate blank value rows, before doing duplicate analysis first the blank facet analysis must be performed on the same column.


Google refine paper Permalink

Facet / Filter | Undo / Redo 2

Refresh | Reset All | Remove All

NIE change

2 choices Sort by: name count

false 2687
true 2314

Facet by choice counts

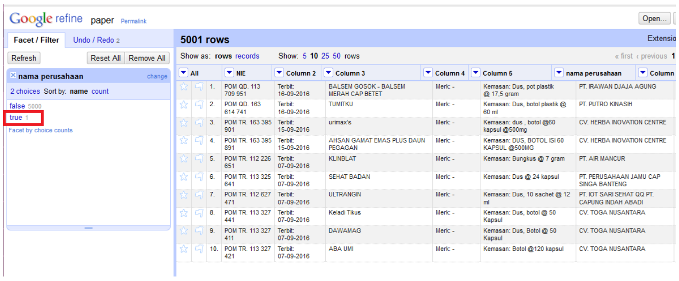
5001 rows

Show as: rows records Show: 5 10 25 50 rows « first ‹ previous

| All | NIE | Column 2 | Column 3 | Column 4 | Column 5 | nama perusahaan | Column |
|---|---|---|---|---|---|---|---|
| 1. | POM QD. 113 709 951 | Terbit: 16-09-2016 | BALSEM GOSOK - BALSEM MERAH CAP BETET | Merk: - | Kemasan: Dus, pot plastik @ 17,5 gram | PT. IRAWAN DJAJA AGUNG | |
| 2. | POM QD. 163 614 741 | Terbit: 16-09-2016 | TUMITKU | Merk: - | Kemasan: Dus, botol plastik @ 60 ml | PT. PUTRO KINASIH | |
| 3. | POM TR. 163 395 901 | Terbit: 15-09-2016 | urimax's | Merk: - | Kemasan: dus , botol @60 kapsul @500mg | CV. HERBA INOVATION CENTRE | |
| 4. | POM TR. 163 395 891 | Terbit: 15-09-2016 | AHSAN GAMAT EMAS PLUS DAUN PEGAGAN | Merk: - | Kemasan: DUS, BOTOL ISI 60 KAPSUL @500MG | CV. HERBA INOVATION CENTRE | |
| 5. | POM TR. 112 226 651 | Terbit: 07-09-2016 | KLINBLAT | Merk: - | Kemasan: Bungkus @ 7 gram | PT. AIR MANCUR | |
| 6. | POM TR. 113 325 641 | Terbit: 07-09-2016 | SEHAT BADAN | Merk: - | Kemasan: Dus @ 24 kapsul | PT. PERUSAHAAN JAMU CAP SINGA BANTENG | |
| 7. | POM TR. 112 627 471 | Terbit: 07-09-2016 | ULTRANGIN | Merk: - | Kemasan: Dus, 10 sachet @ 12 ml | PT. IOT SARI SEHAT QQ PT. CAPUNG INDAH ABADI | |
| 8. | POM TR. 113 327 441 | Terbit: 07-09-2016 | Keladi Tikus | Merk: - | Kemasan: Dus, botol @ 50 Kapsul | CV. TOGA NUSANTARA | |
| 9. | POM TR. 113 327 411 | Terbit: 07-09-2016 | DAWAMAG | Merk: - | Kemasan: Dus, Botol @ 50 Kapsul | CV. TOGA NUSANTARA | |
| 10. | POM TR. 113 327 421 | Terbit: 07-09-2016 | ABA UMI | Merk: - | Kemasan: Botol @120 kapsul | CV. TOGA NUSANTARA | |

Fig. 7. Screenshoot of Openrefine Result for Duplicate Facet  of NIE

OpenRefine is capable to profile one column using multiple analysis types. Blank facet analysis were meant to separate rows with blank column value with non-empty column value, in order to do more specific analysis to the non-empty rows. “True” analysis results are separated using exclude feature of OpenRefine.  After excluding blank values, duplicate analysis are performed using menu “facet > Customized Facets > Duplicates Facets” on the column. Duplicate facet analysis results are two Boolean values. True represents rows which have duplicate value of the column, false represents rows with unique value of the column. The result of duplicate facet analysis of OpenRefine are shown in Fig. 7 and Fig. 8.


Google refine paper Permalink

Facet / Filter | Undo / Redo 2

Refresh | Reset All | Remove All

nama perusahaan change

2 choices Sort by: name count

false 5000
true 1

Facet by choice counts

5001 rows

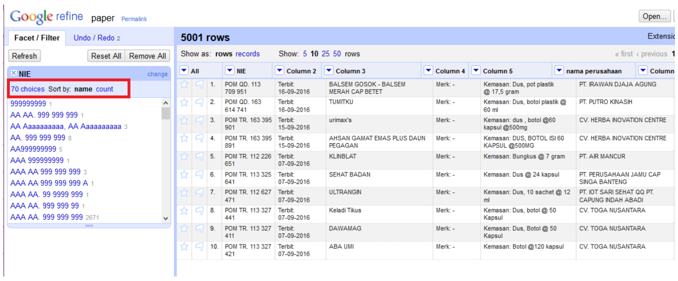
Show as: rows records Show: 5 10 25 50 rows « first ‹ previous

| All | NIE | Column 2 | Column 3 | Column 4 | Column 5 | nama perusahaan | Column |
|---|---|---|---|---|---|---|---|
| 1. | POM QD. 113 709 951 | Terbit: 16-09-2016 | BALSEM GOSOK - BALSEM MERAH CAP BETET | Merk: - | Kemasan: Dus, pot plastik @ 17,5 gram | PT. IRAWAN DJAJA AGUNG | |
| 2. | POM QD. 163 614 741 | Terbit: 16-09-2016 | TUMITKU | Merk: - | Kemasan: Dus, botol plastik @ 60 ml | PT. PUTRO KINASIH | |
| 3. | POM TR. 163 395 901 | Terbit: 15-09-2016 | urimax's | Merk: - | Kemasan: dus , botol @60 kapsul @500mg | CV. HERBA INOVATION CENTRE | |
| 4. | POM TR. 163 395 891 | Terbit: 15-09-2016 | AHSAN GAMAT EMAS PLUS DAUN PEGAGAN | Merk: - | Kemasan: DUS, BOTOL ISI 60 KAPSUL @500MG | CV. HERBA INOVATION CENTRE | |
| 5. | POM TR. 112 226 651 | Terbit: 07-09-2016 | KLINBLAT | Merk: - | Kemasan: Bungkus @ 7 gram | PT. AIR MANCUR | |
| 6. | POM TR. 113 325 641 | Terbit: 07-09-2016 | SEHAT BADAN | Merk: - | Kemasan: Dus @ 24 kapsul | PT. PERUSAHAAN JAMU CAP SINGA BANTENG | |
| 7. | POM TR. 112 627 471 | Terbit: 07-09-2016 | ULTRANGIN | Merk: - | Kemasan: Dus, 10 sachet @ 12 ml | PT. IOT SARI SEHAT QQ PT. CAPUNG INDAH ABADI | |
| 8. | POM TR. 113 327 441 | Terbit: 07-09-2016 | Keladi Tikus | Merk: - | Kemasan: Dus, botol @ 50 Kapsul | CV. TOGA NUSANTARA | |
| 9. | POM TR. 113 327 411 | Terbit: 07-09-2016 | DAWAMAG | Merk: - | Kemasan: Dus, Botol @ 50 Kapsul | CV. TOGA NUSANTARA | |
| 10. | POM TR. 113 327 421 | Terbit: 07-09-2016 | ABA UMI | Merk: - | Kemasan: Botol @120 kapsul | CV. TOGA NUSANTARA | |

Fig. 8 Screenshoot of the Openrefine Result for Duplicate Facet of Company Name

Alphanumeric pattern profiling is an analysis to determine alphanumeric pattern of a certain column and the row count for each pattern. In this case, pattern analysis could only be performed on NIE column, because registrar company name column have no such pattern.  Pattern profiling analysis in OpenRefine were performed using custom text facet. The feature could customize profiling analysis logic by using expressions, in this case we use regular expressions. The regular expression syntax to do Alphanumeric pattern analysis is as follows: Fig. 9 is screenshoot for pattern analysis from Openrefine.

```
value.replace(/[A-Z]/,'A').replace(/[0-
9]/,'9').replace(/[a-z]/,'a')
```


Google refine paper Permalink

Facet / Filter | Undo / Redo 2

Refresh | Reset All | Remove All

NIE change

70 choices Sort by: name count

| | |
|---|---|
| 999999999 | 1 |
| AA AA. 999 999 999 | 1 |
| AA Aaaaaaaaaa, AA Aaaaaaaaaa | 3 |
| AA. 999 999 999 | 8 |
| AA999999999 | 5 |
| AAA 999999999 | 1 |
| AAA AA 999 999 999 | 3 |
| AAA AA 999 999 999 A | 1 |
| AAA AA. 99 9999 999 | 1 |
| AAA AA. 999 999 99 | 1 |
| AAA AA. 999 999 999 | 2671 |

5001 rows

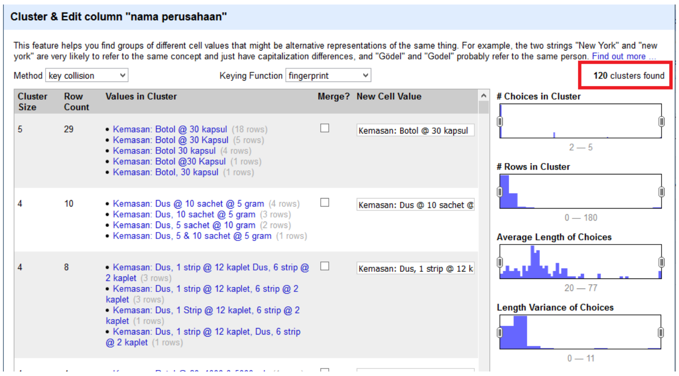
Show as: rows records Show: 5 10 25 50 rows « first ‹ previous

| All | NIE | Column 2 | Column 3 | Column 4 | Column 5 | nama perusahaan | Column |
|---|---|---|---|---|---|---|---|
| 1. | POM QD. 113 709 951 | Terbit: 16-09-2016 | BALSEM GOSOK - BALSEM MERAH CAP BETET | Merk: - | Kemasan: Dus, pot plastik @ 17,5 gram | PT. IRAWAN DJAJA AGUNG | |
| 2. | POM QD. 163 614 741 | Terbit: 16-09-2016 | TUMITKU | Merk: - | Kemasan: Dus, botol plastik @ 60 ml | PT. PUTRO KINASIH | |
| 3. | POM TR. 163 395 901 | Terbit: 15-09-2016 | urimax's | Merk: - | Kemasan: dus , botol @60 kapsul @500mg | CV. HERBA INOVATION CENTRE | |
| 4. | POM TR. 163 395 891 | Terbit: 15-09-2016 | AHSAN GAMAT EMAS PLUS DAUN PEGAGAN | Merk: - | Kemasan: DUS, BOTOL ISI 60 KAPSUL @500MG | CV. HERBA INOVATION CENTRE | |
| 5. | POM TR. 112 226 651 | Terbit: 07-09-2016 | KLINBLAT | Merk: - | Kemasan: Bungkus @ 7 gram | PT. AIR MANCUR | |
| 6. | POM TR. 113 325 641 | Terbit: 07-09-2016 | SEHAT BADAN | Merk: - | Kemasan: Dus @ 24 kapsul | PT. PERUSAHAAN JAMU CAP SINGA BANTENG | |
| 7. | POM TR. 112 627 471 | Terbit: 07-09-2016 | ULTRANGIN | Merk: - | Kemasan: Dus, 10 sachet @ 12 ml | PT. IOT SARI SEHAT QQ PT. CAPUNG INDAH ABADI | |
| 8. | POM TR. 113 327 441 | Terbit: 07-09-2016 | Keladi Tikus | Merk: - | Kemasan: Dus, botol @ 50 Kapsul | CV. TOGA NUSANTARA | |
| 9. | POM TR. 113 327 411 | Terbit: 07-09-2016 | DAWAMAG | Merk: - | Kemasan: Dus, Botol @ 50 Kapsul | CV. TOGA NUSANTARA | |
| 10. | POM TR. 113 327 421 | Terbit: 07-09-2016 | ABA UMI | Merk: - | Kemasan: Botol @120 kapsul | CV. TOGA NUSANTARA | |

Fig. 9 Screenshoot of Openrefine Result for Pattern analysis

Registrar company name profiling are performed using blank facet analysis, duplicate facet analysis, and text cluster analysis. Blank facet and duplicate facet are performed using the same manner as the NIE field. Text cluster analysis for registrar company name field are performed using combinations of blank facet, text facet, and cluster feature. Cluster feature helps you find groups of different cell values that might be alternative representations of the same thing.  Cluster analysis are performed on registrar company name field which are previously excluded from blanks.

Text facet analysis describes variations of the text in the same column. Cluster analysis results in groups of text with some degree of similarity within each group. OpenRefine tool classifies text cluster using two basic methods, key collision method and nearest neighbor method. Key collision methods are based on the idea of creating an alternative representation of a value (a "key") that contains only the most valuable or meaningful part of the string and 'bucket's (or 'bin' as it's described inside OpenRefine's code) together different strings based on the fact that their key is the same (hence the name "key collision") [12]. The Nearest Neighbor methods (also known as kNN), on the other hand, provide a parameter (the radius, or k) which represents a distance threshold: any pair of strings that is closer than a certain value will be binned together. [12].  For text cluster analysis on registrar company name field, key collision method are selected with fingerprint keying function. The fingerprinting method is fast and simple yet works relatively well in a variety of contexts and it's the least likely to produce false positives, which is why it is the default method [12]. The result of cluster analysis method on registrar company name field are shown in Fig. 10.


Cluster & Edit column "nama perusahaan"

This feature helps you find groups of different cell values that might be alternative representations of the same thing. For example, the two strings "New York" and "new york" are very likely to refer to the same concept and just have capitalization differences, and "Gödel" and "Godel" probably refer to the same person. Find out more ...

Method: key collision | Keying Function: fingerprint | 120 clusters found

| Cluster Size | Row Count | Values in Cluster | Merge? | New Cell Value |
|---|---|---|---|---|
| 5 | 29 | • Kemasan: Botol @ 30 kapsul (18 rows)<br>• Kemasan: Botol @ 30 Kapsul (5 rows)<br>• Kemasan: Botol 30 kapsul (4 rows)<br>• Kemasan: Botol @30 Kapsul (1 rows)<br>• Kemasan: Botol, 30 kapsul (1 rows) | ☐ | Kemasan: Botol @ 30 kapsul |
| 4 | 10 | • Kemasan: Dus @ 10 sachet @ 5 gram (4 rows)<br>• Kemasan: Dus, 10 sachet @ 5 gram (3 rows)<br>• Kemasan: Dus, 5 sachet @ 10 gram (2 rows)<br>• Kemasan: Dus, 5 & 10 sachet @ 5 gram (1 rows) | ☐ | Kemasan: Dus @ 10 sachet @ |
| 4 | 8 | • Kemasan: Dus, 1 strip @ 12 kaplet Dus, 6 strip @ 2 kaplet (3 rows)<br>• Kemasan: Dus, 1 strip @ 12 kaplet, 6 strip @ 2 kaplet (3 rows)<br>• Kemasan: Dus, 1 Strip @ 12 kaplet, 6 strip @ 2 kaplet (1 rows)<br>• Kemasan: Dus, 1 strip @ 12 kaplet, Dus, 6 strip @ 2 kaplet (1 rows) | ☐ | Kemasan: Dus, 1 strip @ 12 k |

Fig. 10. Screenshoot of Openrefine Result for Cluster Analysis

## V. Discussion

The profiling analysis result on 5000 sample data retrieved from BPOM Agency website is shown in Table 1. Data attributes that were analyzed are NIE and Company Name. Profiling analysis that were done on each attributes consists of duplicate analysis, blank facet analysis, cluster, and pattern analysis.

Table 1. Result of Data Profiling from 5001 Data of Tradisional medicine on BPOM using OpenRefine

| Column Name | Duplicates (%) | Blank(%) | Cluster | Pattern |
|---|---|---|---|---|
| NIE | 46 | 0 | 2 | 70 |
| Company name | 79 | 1 | 120 | - |

NIE data element is a data element that must not be empty, unique, and have one alphanumeric pattern for same commodity. According to profiling analysis results, no NIE data element were empty, but the data have many alphanumeric patterns. There are 70 patterns detected on the 5000 row data. Duplication analysis need to be combined with other elements because a single product with a single permit number could be duplicated in multiple rows if the factory location is different and also if volume or package weight differs.

Recommendation for NIE data element pattern is to create uniform alphanumeric pattern that applies to all commodity being monitored by BPOM agency, and also to have the NIE generated by an application instead of manually entered by a person. Generating NIE using application system according to established numbering rules will result in only one alphanumeric pattern for all NIE data element. Recommendation to improve existing data is to perform data cleansing by updating current data according to established pattern rule that are agreed by the entire BPOM agency.

The registrar company name data element have rules as follows, data must not be empty, and company names must match perfectly if the company is the same. One row of data have no company name, the data row should be cross-checked with original paper trail in order to determine the missing name. There are 120 cluster of data found in the company name data element, which shows that most of the data came from about 120 companies, but the fact shows these company names are not written in a consistent manner so that there are small differences among the names resulting in clusters. If the company name were consistent, the cluster analysis tool would not found cluster at all. The ideal condition is to have consistent company names so the cluster analysis would find 0 clusters.

Improvement on company name data element is to update the company name field to adopt one consistent naming for companies that are determined to be the same. The company name cleansing process are done using two steps. First step is to standardize naming automatically using string functions and the second step uses human judgement to determine which clusters is accurate. The first step consists of removing double or triple spaces, trimming spaces before and after the name, and also capitalize the company name. The second step requires a person to choose which company name follows the rule so it could be adopted for each found cluster.

Recommendation to prevent company name differences is by changing business process to obtain NIE for a product. The suggested business process change is that the registrar register the company first and verified by BPOM Agency staff, then followed by registering a product to obtain permit, referring to previously registered company. On the registrar application, features could be added to help prevent registrar to register their company more than once.

## Acknowledgment

This research supported by Information System Department, School of Industrial Engineering, Telkom University.

## References

[1] B. Dorr and P. Herbert, "Data Profiling: Designing the Blueprint for Improved Data Quality," in *SAS User Group International 30*, Philadelphia, 2005.

[2] J. E.Olson, Data Quality The Accuracy Dimension, USA, 2013.

[3] F. Naumann, "Data Profiling Revisited," 2013.

[4] J. E. Olson, Data Quality : The Accuracy Dimension, San Fransisco: Morgan Kaufmann Publishers, 2003.

[5] L. Golab, F. Naumann and A. Ziawasch, "Data Profiling," pp. 1-4, 2016.

[6] P. Berka, "Data Cleansing Using Clustering," 2016.

[7] A. Ziawasch, T. Gruetze, A. Jentzsch and F. Naumann, "Profiling and Mining RDF Data with ProLOD++," p. 1, 2014.

[8] R. Mall, L. Berti-Equille and H. Bensmail, "Metabolomic Data Profiling," 2016.

[9] Z. Abedjan, L. Golab and F. Naumann, "Data Profiling," pp. 1-4, 2016.

[10] *. A. Ziawasch, J.-A. Q. e-Ruiz and F. Naumann, "Detecting Unique Column Combinations," 2014.*

[11] microsoft, "Microsoft Developer Network," Microsoft, [Online]. Available: https://msdn.microsoft.com/en-us/library/az24scfc(v=vs.110).aspx. [Accessed 17 09 2016].

[12] OpenRefine, "Github OpenRefine," OpenRefine, 22 01 2013. [Online]. Available: https://github.com/OpenRefine/OpenRefine/wiki/Clustering-In-Depth. [Accessed 21 09 2016].

[13] N. Felix, A. Ziawasch and L. Golap, "Profiling relational data : a survey," pp. 557-581, 2015.