Exploring and Curating Data Collections with CURARE

Genoveva Vargas-Solar
Univ. Grenoble Alpes, CNRS, Grenoble INP, LIG-LAFMIA
Grenoble, France
genoveva.vargas@imag.fr

Gavin Kemp
Univ Lyon, University of Lyon 1, LIRIS UMR 5205 CNRS
Villeurbanne, France
gavin.kemp@liris.cnrs.fr

Irving Hernández-Gallegos
Universidad Autónoma de Guadalajara
Zapopan, Mexico
irving.hernandez.g@gmail.com

Javier A. Espinosa-Oviedo
Delft University of Technology
2628BL Delft, Netherlands
javier.espinosa@tudelft.nl

Catarina Ferreira Da Silva
Univ Lyon, University of Lyon 1, LIRIS UMR 5205 CNRS
Villeurbanne, France
catarina.ferreira-da-silva@liris.cnrs.fr

Parisa Ghodous
Univ Lyon, University of Lyon 1, LIRIS UMR 5205 CNRS
Villeurbanne, France
parisa.ghodous@liris.cnrs.fr

ABSTRACT

This paper demonstrates CURARE, an environment for curating *raw* data collections and assisting data scientists to explore them. CURARE implements a data curation model used to store structural and quantitative metadata such as the number of columns, de name of columns and the statistics of the values of every column. It provides associated functions for exploring these metadata. The demonstration proposed in this paper is devoted to evaluate and compare the effort invested by a data scientist when exploring data collections with and without CURARE assistance.

1 INTRODUCTION

The emergence of new platforms that produce data at different rates, is adding to the increase in the amount of data collections available online. Open data initiatives frequently release data collections as sets of records with more or less information about their structure and content. For example, different governmental, academic and private initiatives release open data collections like Grand Lyon in France[1], Wikipedia [2], Stack Overflow [3] and those accessible through Google Dataset Search [4]. Releases often specify minimum "structural" metadata as the size of the release, the access and exploitation license, the date, and eventually the structure of the records (e.g., names and types of columns). This brings an unprecedented volume and diversity of data to be explored.

Data scientists invest a big percentage of their effort processing data collections to understand records structure and content and to generate descriptions. Useful descriptions should specify the values' types, their distribution, the percentage of null, absent or default values, and dependencies across attributes. Having this knowledge is crucial to decide whether it is possible to run analytics tasks on data collections directly or whether they should be preprocessed (e.g., cleaned). Therefore, several questions should be considered by a data scientist. For example, whether one or more data collections can be used for target analytics tasks; whether they are complementary or not, or whether they can be easily integrated into one data set to be analyzed; whether certain attributes have been cleaned, computed (e.g., normalized) or estimated. Keeping track of cleansing changes is important because such operations can bias certain statistics analysis and results.

Therefore, the data scientist must go through the records and the values exploring data collections content like the records structure, values distribution, presence of outliers, etc. This is a time consuming process that should be done for every data collection. The process and operations performed by the data scientist are often not described and preserved, neither considered as metadata. Thus, the data scientist effort cannot be capitalized for other cases and by other data scientists.

Curation tasks include extracting explicit, quantitative and semantic metadata, organizing and classifying metadata and providing tools for facilitating their exploration and maintenance (adding and adjusting metadata and computing metadata for new releases) [3, 4, 6]. The problem addressed by curation initiatives is to address the exploration of data collections by increasing their usefulness and reducing the burden of exploring records manually or by implementing ad-hoc processes. Curation environments should aid the user in understanding the data collections' content and provide guidance to explore them [1, 2, 5, 8].

Our work focuses on (semi)-automatic metadata extraction and data collections exploration which are key activities of the curation process. Therefore, we propose CURARE, an environment for curating and exploring data collections.

CURARE provides tools in an integrated environment for extracting metadata using descriptive statistics measures and data mining methods. Metadata are stored by CURARE according to its data model proposed for representing curation metadata (see Figure 1). The data model organizes metadata into four main concepts:

[1] https://www.metropolis.org/member/grand-lyon
[2] https://www.wikidata.org/wiki/Wikidata
[3] https://archive.org/details/stackexchange
[4] https://toolbox.google.com/datasetsearch

Demonstration already presented and published in the conference EDBT 2019.

, ,

ACM ISBN 978-3-89318-081-3.

,,

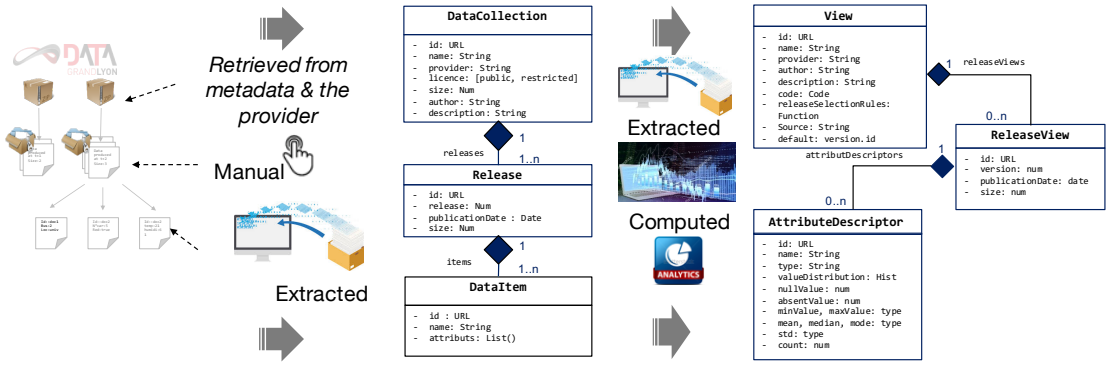

Figure 1: CURARE Model

- Data Collection: models structural metadata concerning several releases like the number of releases it groups, their aggregated size, the provider, etc.
- Release: models structural metadata about the records of a release (file) including for example the columns' name in the case of tabular data, the size of the release, the date of production, access license, etc.
- Release view: models quantitative metadata of the release records of a data collection. For example, the number of columns and the number of records in a release. A release view can store, for every column in all the records of a release, its associated descriptive statistics measures (mean, median, standard deviation, distribution, outlier values).
- View: models quantitative aggregated metadata about the releases of a data collection.

CURARE also provides functions for exploring metadata and help data analysts determining which are the collections that can be used for achieving an analytics objective (comparing several collections, projecting and filtering their content). Rather than performing series of self-contained queries (like keyword search or relational ones when possible), a scientist can stepwise explore into data collections (i.e., computing statistical descriptions, discovering patterns, cleaning values) and stop when the content reaches a satisfaction point.

The demonstration proposed in this paper is devoted to evaluate and compare the effort put by a data scientist when exploring data collections with and without CURARE assistance. Through a treasure seeking metaphor we ask users to find clues within non curated Stack Overflow releases. Then, we compare the effort for finding cues by exploring metadata extracted by CURARE. The tasks are intended to show the added value of using metadata to explore data collections for determining whether they can be used for answering target questions.

The remainder of the paper is organized as follows. Section 2 describes the general architecture of CURARE and the extraction of structural and quantitative metadata. Section 3 describes the demonstration scenario for CURARE and the tasks used for measuring user experience during the demo. Section 4 concludes the paper and discusses lessons learned from the demonstration.

2 CURARE

The curation tasks implemented by CURARE are coordinated through a general iterative workflow consisting in three steps: (i) data collections' harvesting (manually or automatically) and preservation; (ii) structural and statistical meta-data extraction; (iii) exploration.

Figure 2: CURARE data curation workflow

The workflow is iterative because data collections can be updated with new releases and therefore releases are harvested recurrently. Similarly, as a result of curated data collections exploration, new metadata can be defined and the processing step can be repeated to include new metadata and store them.

2.1 General architecture

Figure 3 shows the architecture of CURARE consisting of services that can be deployed on the cloud. The current version of CURARE are Azure services running on top of a data science virtual machine. The services of CURARE are organized into three layers that correspond to the phases of the curation workflow. They are accessible through their Application Programming Interfaces (API's). Services can be accessed vertically as a whole environment or horizontally from every layer.

- The first layer is devoted to harvesting data collections and extracting structural metadata like the size, the provenance, time and location time-stamps.
- The second layer addresses distributed storage and access of curated data and metadata.
- The third layer provides tools for extracting quantitative metadata by processing raw collections.

The structural and quantitative metadata of CURARE are organized into four types of persistent data structures implementing the CURARE data model (i.e., *data collection, release, view and release view*). According to the CURARE approach, a data collection consists of several sets of data released at a given moment. For example, Stack Overflow is a data collection consisting of packages

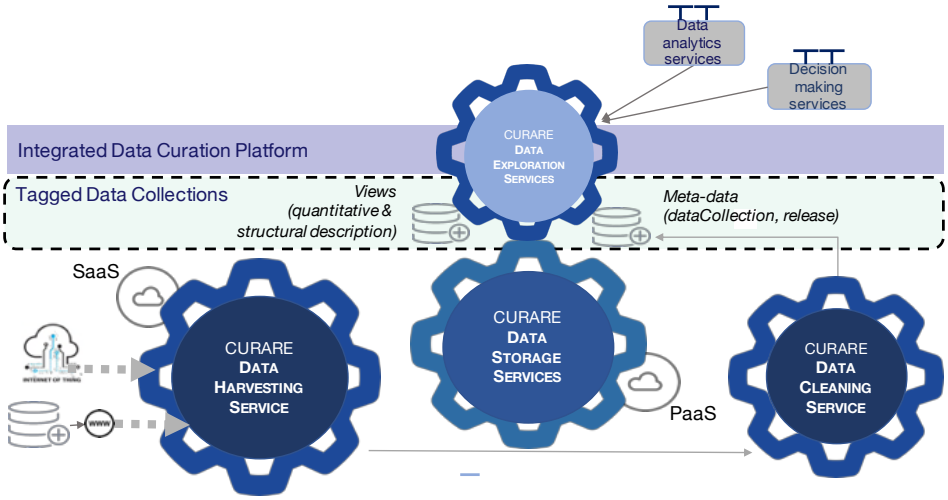

Figure 3: General functional architecture of CURARE

of records (i.e., releases) produced periodically. A Stack Overflow release provides a kind of snapshot of all the interactions happening in this social network. It is made available at some point in time. A release view, and a view, are data structures that provide an aggregated and quantitative perspectives of resp. a data collection and its several associated releases.

These data structures are then explored using ad-hoc operators within semi-declarative expressions or programs instead of directly exploring raw data collections. For example, is a release included in another? To which extent a release has new records with respect to others? Which is the percentage of missing values of the releases of a data collection?

2.2 Extracting structural metadata

The *data collection model* shown in the left hand side of Figure 1 provides concepts for representing data collections as sets of releases of raw data. Each release consists of a set of items (e.g. records). The data collections model is used by the data harvesting and cleansing services for organizing structural and context metadata related to collections (provider, objective, URL, item structure).

The demonstrated version of CURARE provides a data collection upload facility for CSV files, which is an explicit data harvesting process. During this process, the structure, the size and release date of the file are extracted automatically. Then, other metadata for example the URI of the provider and the type of license can be provided by the data scientist by filling a form.

As said before, metadata are stored by CURARE. Thus, in the exploration phase a data scientist can search for releases of a given size, with a "open source" licence, or produced by a specific provider, expressing queries or executing explicit exported methods.

2.3 Extracting quantitative metadata

The *view model* of CURARE shown in the right hand side of Figure 1 provides concepts for representing quantitative and analytic aspects of the releases of a data collection. For instance, statistics of the content including the distribution of the values of each attribute across the items of a release, missing values, null values.

Depending on the attribute type (only atomic types are considered for statistics) the strategies for computing measures can change. CURARE first defines for every attribute in a tabular file, for example, its type and it approximately discovers possible null, absent and default values. Then, it computes description statistics measures (mean, mode, median, standard deviation, distribution) of the values of every attribute of the structure of the data collection in tabular format. It interacts with the data scientist to decide what to do with those attributes with a domain containing "dirty" values. Should these values be excluded from the computation? Or how can they be adjusted if they should be considered? If values are modified, should CURARE create a new release? Should it be preserved completely tagging the modified values as modified and indicating the applied formula or procedure. This is specified by the data scientist and managed by CURARE. For computing statistics of string values, CURARE processes the strings and creates frequency matrices and inverted indexes for the terms.

CURARE processes collections with many attributes and records that can be somehow costly depending on the tasks the data is used for, and the data volume. Thus, CURARE uses a Spark environment for computing statistics measures that sits on the data science virtual machine [5]. This strategy can be related to existing approaches like the one proposed in [7].

3 DEMONSTRATION OVERVIEW

The demonstration of CURARE uses a Jupyter notebook prepared for performing the tasks intended to test the use of metadata (i.e., entities of the CURARE data model) for exploring data collections. We use the Azure Notebooks environment [6] to run the demonstration. The back-end of the demonstrated version uses the Python

[5] https://azure.microsoft.com/en-us/services/virtual-machines/data-science-virtual-machines/
[6] https://notebooks.azure.com

data science libraries, nltk [7] for dealing with the processing of string typed attributes. We use ATLAS [8], the clustered public service of MongoDB for storing metadata associated to raw data collections.

The demonstration provides 3 releases of Stack Overflow [9] made available between the 1st. and 4th. January 2018 each consisting in five files "badges", "comments", "posts", "users" and "votes". These files range between ca. 300 KB to 20 MB.

CURARE extracts structural metadata and stores them as Data Collection documents according to its data model. It extracts quantitative metadata and stores them as Views documents. For the demonstration purposes we use a notebook that shows the process and the results. The releases used in the demonstration are processed and the results are stored in Atlas. The result weights 126,5 MB in Mongo and the data collection of structural metadata 9,5 KB.

Demonstration scenario. The demonstration of CURARE shows in which situations it can be useful to have an integrated curation and exploration environment. This can be evaluated through a set of tasks proposed by the demonstration scenario with a game called *The cure challenge*. The objective is to compare the effort when performing tasks to explore Stack Overflow releases manually and with CURARE.

When the game is activated it gives access to raw data collections and to CURARE metadata. The game consists of two curation matches where a data scientist has to perform the following tasks.

(1) Data collections exploration tasks:
   - Discover the release of the Stack Overflow data collection with the best quality. That is, the one with less missing, null and default values in the posts.
   - Compare the releases size in terms of number of records.
   - Compare releases in terms of values distribution for a given attribute.
   - Compare releases with respect to the topic by comparing terms of a given attribute.

(2) Discover the attribute(s) of the Stack Overflow posts that can be used to compute the popularity of the answers and the reputation of the author (stared answers and authors).
(3) Discover the attribute(s) that can be used to identify the most trendy topics addressed in the release. Will missing, null, default values bias the observation of the trends?
(4) Choose the attributes that can be used as sharding keys to fragment the release using a hash based and an interval based strategy.

In the first match the tasks are done by exploring directly the raw data collections. The second using CURARE's exploration facilities. For both matches a dashboard shown in Figure 4 automatically measures the degree of effort given by comparing with the accuracy/precision and correctness of the answers against time used for providing an answer. These measures are completed with an explicit score for every task given by the data scientist (one star low effort, three stars high effort).

Measuring data exploration effort. The experiment is assessed by comparing the evaluation results of the tasks performed in the two matches. We evaluate:

- Whether the metadata in views provide representative information of the raw content of a release (Q_1 and Q_2)
- How easy it is to see data collection quality in terms of consistency of the structure, the degree of missing, null and absent values of the attributes (Q_2)
- Usefulness of views for exploring the data collections to determine which kind of analytics questions they can answer. (Q_3, Q_4)

4 CONCLUSION AND RESULTS

We demonstrate CURARE that implements a data curation approach integrating metadata describing the structure, content and statistics of raw data collections. Thereby raw data collections can be comfortably explored and understood for designing data centric experiments through exploration operations. The demonstration shows the usefulness of CURARE by measuring the effort of a data scientist for performing exploration tasks in predefined Stack Overflow releases, used as demonstration examples.

5 ACKNOWLEDGEMENT

This work was partially funded by the Rhône-Alpes region through the project AMBED of the ARC 7 program[10] and the LDE Centre for BOLD Cities [11]. The implementation of CURARE was partially funded by the CONACYT "beca mixta" fellowship program of the Mexican government awarded to Irving Hernández Gallegos.

REFERENCES

[1] L. Battle, M. Stonebraker, and R. Chang. 2013. Dynamic reduction of query result sets for interactive visualizaton. In *2013 IEEE International Conference on Big Data*. 1–8. https://doi.org/10.1109/BigData.2013.6691708
[2] M. Bostock, V. Ogievetsky, and J. Heer. 2011. Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics* 17, 12 (Dec 2011), 2301–2309. https://doi.org/10.1109/TVCG.2011.185
[3] André Freitas and Edward Curry. 2016. Big data curation. In *New Horizons for a Data-Driven Economy*. Springer, 87–118.
[4] Alon Halevy, Flip Korn, Natalya F Noy, Christopher Olston, Neoklis Polyzotis, Sudip Roy, and Steven Euijong Whang. 2016. Goods: Organizing google's datasets. In *Proceedings of the 2016 International Conference on Management of Data*. ACM, 795–806.
[5] Martin L Kersten, Stratos Idreos, Stefan Manegold, Erietta Liarou, et al. 2011. The researcher as guide to the data deluge: Querying a scientific database in just a few seconds. *PVLDB Challenges and Visions* 3, 3 (2011).
[6] Michael Stonebraker, Daniel Bruckner, Ihab F Ilyas, George Beskales, Mitch Cherniack, Stanley B Zdonik, Alexander Pagan, and Shan Xu. 2013. Data Curation at Scale: The Data Tamer System.. In *CIDR*.
[7] Abdul Wasay, Xinding Wei, Niv Dayan, and Stratos Idreos. 2017. Data canopy: Accelerating exploratory statistical analysis. In *Proceedings of the 2017 ACM International Conference on Management of Data*. ACM, 557–572.
[8] Sam Likun Xi, Oreoluwa Babarinsa, Manos Athanassoulis, and Stratos Idreos. 2015. Beyond the wall: Near-data processing for databases. In *Proceedings of the 11th International Workshop on Data Management on New Hardware*. ACM, 2.

[7] https://www.nltk.org
[8] https://www.mongodb.com/cloud/atlas
[9] https://data.stackexchange.com/stackoverflow/query/new
[10] http://www.arc7-territoires-mobilites.rhonealpes.fr/
[11] http://boldcities.nl/

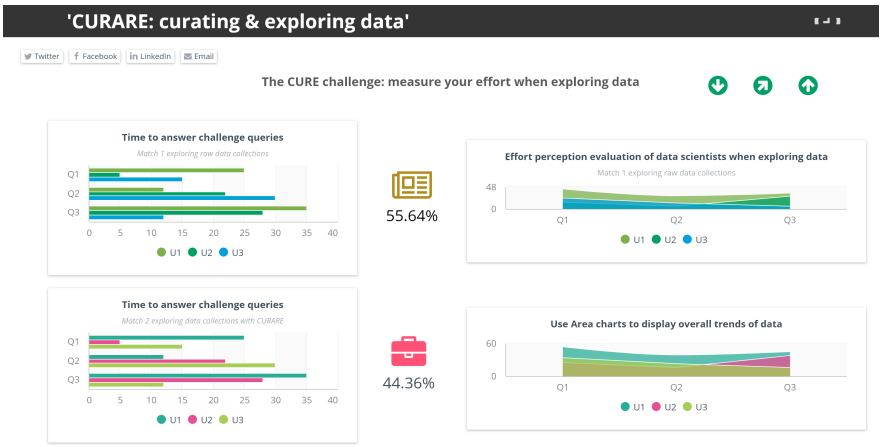

Figure 4: Assessment dashboard results of the demonstration scenario