Repolytics: Identifying Measurable Insights for Digital Repositories

Vangelis Nomikos

Intelibility Innovative Data Engineering LLC,
Atlanta, GA, USA
nomikos@intelibility.com

Abstract. This paper presents specific aspects of the Repolytics platform: a data analytics platform for digital repositories. An overview of the platform is presented alongside an example of how one can employ it's services to analyze a digital repository's data and identify quality related issues.

Keywords: metadata quality, data quality, data analytics, quality metrics, digital repositories.

1 Introduction

Data is captured, analyzed and used to drive all aspects of our lives in a data driven world. Recently, we have seen a rapid growth of digital repositories and open data catalogues being made available to the public. In the case of digital repositories that target the libraries-archives, scientific-research data domains and open data portals the market is dominated by open source solutions such as DSpace, Omeka, DKAN, CKAN due to their simplicity and low-no cost. Most solutions enable ingest / cataloguing of information either through automated means (SWORD, REST APIs) or through simple and untuitive forms. Quality assurance in most cases comprises of a number of mandatory metadata the user has to enter. The reality however is far more complex and this has a profound effect in the quality of the the ingested data.

Repolytics [1] is a platform that aims at filling this gap through intelligent data analytics. The data loaded into the platform are analyzed and specific quality metrics are presented alongside a more thorough analysis per metadata element. The metrics include metadata completeness, accuracy and consistency. Furthermore, similar metrics are calculated for the data as well.

2 Related Work

Metadata quality is an important issue for the Digital Library domain and has attracted several researcher groups to deal with it. One of the main demands is the establishment of a conceptual framework consisted of a set of well-defined quality assessment criteria such as completeness, validity, consistency, timeliness, appropriateness and accuracy constituents [2, 3]. In such a framework any assessment effort would be based on reliable indications about metadata quality. The first attempt to define a framework established a narrow set of criteria such as accuracy, completeness and

serviceability [4]. Some researches expanded the criteria set from which metadata quality is approached [6, 8], while some other narrowed the perspective focusing only on the completeness criterion [5].

An important evolution of this scientific field was the introduction of the context-dependent metadata quality approach [3, 7]. According to this approach metadata quality issues follow four major concepts: mappings, changes to the information entity, changes to the underlying entity and context changes. For these concepts a taxonomy of 22 information quality criteria was developed. The criteria were clustered to three categories: intrinsic, relation and reputational and are measured via 41 metrics [7]. Recent research suggests that a metadata quality framework doesn't have to "invent new dimensions in order to accommodate the needs of diverse communities of practice" [3] but to give the flexibility to each evaluator to assess the results within a specific context. This paper follows the context-dependent approach and assumes that metadata quality strongly depends on the viewpoint of the evaluator and should be aligned with the application domain for which the metadata were produced and used. Therefore weighting functions for these factors (evaluator viewpoint, application domain, metadata usage) should be defined and used to weight the values of the metadata quality metrics.

3 The Repolytics approach

The core principle of the Repolytics platform is the provisioning of an expandable set of middleware services that operates both on metadata and data level of a digital repository. The most fundamental services of the platform involve:

- (meta)-data integration services
- (meta)-data profiling services
- (meta)-data quality services

The Repolytics platform enables the user to load data from different data sources either directly (e.g. through a file archive) or by using one of the supported data providers such as DKAN API and OAI-PMH protocol. Each data source provides one or more metadata (e.g. OAI-DC, MODS) and data (e.g. CSV, excel) representations. Once the data has been loaded, each digital object is analyzed and the main workflow seen in Figure 1 below is executed.

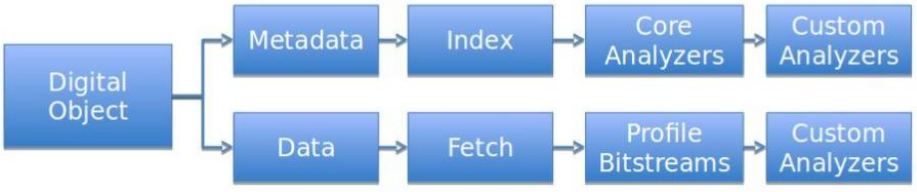

Figure 1. Overal workflow per digital object.

For each digital object, if metadata is provided, the format is identified, every metadata element is indexed and a set of core analyzers are invoked for each element. For specific elements depending on their type, a set of custom analyzers is invoked. If the

data is provided (usually throuh some kind of URL) the actual data is fetched and verified followed by a profile of each bitstream. Again, a set of custom analyzers is then invoked depending on the file type.

Some of the metrics that are calculated include:

- metadata completeness
- distinct values for each metadata element
- accuracy for specific element types (e.g. dates, actors) where all values are classified according to their class
- itemset frequencies when applicable (e.g. in the case of subject terms and keywords)

One of the primary challenges include the efficeint visualization of the results to the end user. For that reason a series of bar charts, radar charts, tables, gauge meters etc are employed per case. For example, as shown in Figure 2 a radar chart is employed to fingerprint an entire repository according to it's completeness.

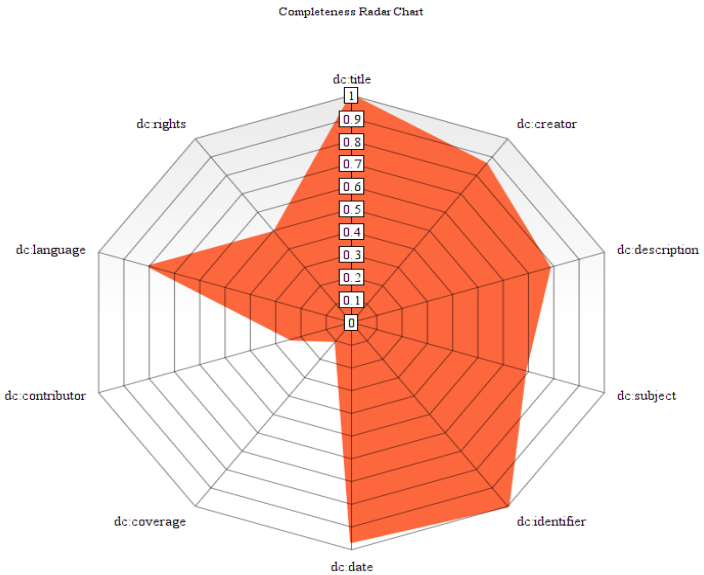

Figure 2. Radar chart showing the overall completeness for a data soure.

Similary, an accuracy detector identifies and classifies all metadata element values according to their class and provides an insight on the accuracy level of each class (low, medium, high).

Class Type	Example	Accuracy
YYYY-MM-DD	2015-12-20	Medium
YYYY	1998	Low
YYYY-MM-DD HH:MM:SS	1996-02-23 14:30:11	High
UNKNOWN	Sp.1996/23	-

4 Conclusions & Future Work

In this paper, Repolytics, a platform for repository data analytics is presented. The platform enables the use to easily load data from supported data sources and analyzes this data focusing primarily on completeness and accuracy whereas in specific cases, more specific metrics (such as itemset frequencies) are employed to help gain insight on highly subjective metrics such as consistency. The platform can also access and profile data as well.

References

1. Repolytics Platform, http://www.repolytics.com
2. Herzog, T., Scheuren, F., Winkler, W.: Data Quality and Record Linkage Techniques. Springer-Verlag, NY (2007)
3. Tani, A., Candela, L., Castelli, D.: Dealing with metadata quality: The legacy of digital library efforts. Inf. Process. Manag. 49 (2013) 1194–1205
4. Moen, W.E., Stewart, E.L., McClure, C.R.: The Role of Content Analysis in Evaluating Metadata for the U.S. Government Information Locator Service (GILS): Results from an Exploratory Study. http://digital.library.unt.edu/ark:/67531/metadc36312/citation/ (1997)
5. Margaritopoulos, M., Margaritopoulos, T., Mavridis, I., Manitsaris, A.: Quantifying and measuring metadata completeness. J. Am. Soc. Inf. Sci. Tech. 63 (2012) 724–737
6. Ochoa,X., Duval,E.: Quality Metrics for Learning Object Metadata. In: WorldConference on Educational Multimedia, Hypermedia and Telecommunications. (2006) 1004–1011
7. Stvilia, B., Gasser, L., Twidale, M.B., Smith, L.C.:A framework for information quality assessment. J. Am. Soc. Inf. Sci. Tech. 58 (2007) 1720–1733
8. Chen,Y.-N.,Wen,C.-Y.,Chen,H.-P.,Lin,Y.-H.,Sum,H.-C.:Metrics for metadata quality assurance and their implications for digital libraries. In: Xing, C., Crestani, F., and Rauber, A. (eds.) 13th International Conference on Asia-Pacific Digital Libraries, Beijing, China. Springer-Verlag, Berlin (2011)138–147