DATA SCIENCE AND ANALYTICS IN LIBRARIES

by José Luis Preza

Abstract: *Libraries have the virtue of managing vast amounts of information. Data Science and Analytics techniques and methodologies allow libraries to fully exploit the content they hold with the goal of providing better information to their users: better search, recommendations, etc.*

Keywords: *Data Science; Analytics; Libraries; Machine Learning; e-Infrastructures Austria; Metadata*

DATA SCIENCE UND DATENANALYTIK IN BIBLIOTHEKEN

Zusammenfassung: *Bibliotheken sind in einer priviligierten Situation: Sie verwalten riesige Mengen von Daten und Informationen. Data Science und Analytics-Methoden ermöglichen es Bibliotheken, den Inhalt, den sie verwalten, voll auszunutzen, um den Nutzern bessere Informationen, Suche und Empfehlungen zu bieten.*

Schlüsselwörter: *Data Science; Datenanalytik; Bibliotheken; Machine Learning; e-Infrastructures Austria; Metadaten*

This document was prepared as part of the Deliverables of Cluster I of e-Infrastructures Austria, a nationwide project regarding the design and management of digital infrastructures for research data (www.e-infrastructures.at). See also the preprint version: José Luis Preza. (2017, March 8). Data Science and Analytics in Libraries. Zenodo. DOI: http://doi.org/10.5281/zenodo.375809.

Contents

1. What is Data Science?

Data Science is simply a discipline that combines data with programming languages, algorithms, statistics, machine learning, artificial intelligence, reporting, and data visualization, all to make sense out of data. Data Science is a very important part of Cognitive Computing that enables Artificial Intelligence.

Public, School and University Libraries are in a very advantageous position: they sit on a lot of data. The data stored in such libraries are very diverse. There are books, documents, charts, datasets, experiments, software, tables, images, videos, audio, dissertations, magazines, newspapers, processes, usage, user data, financial data, to mention a few. The challenge for libraries is not only to digitize all their content (digital objects), but also to classify, organize, link and publish all digital objects.

Up until now, most content (digital objects) has been organized and classified manually. However, manual processes are not sustainable, certainly not when you have to process millions of digital objects in a short period of time and with high accuracy. Here is where Data Science comes to the rescue.

2. Data Science application in libraries

The techniques and methods used in Data Science allow libraries to lighten the workload and get results faster than with manual processes. Concrete areas where Data Science can assist libraries include:

- Digital Object Classification/Semantics/Search: Automatic classification of digital objects (keywords, entities, concepts.[1]
- Picture Recognition and Classification: automatic classification and tagging of pictures (also extracted from video).
- Content Clustering and Segmentation: Automatic clustering and segmentation of digital objects based on content.

- **Reporting**: Generate reports from your content.
- **Predictive Analytics**: Answer the question of who is going to read/use what?
- **Machine Translation**: Automatic translation of digital objects, including e.g. Braille.
- **Speech to Text**: Extraction of audio speech to convert it into text.
- **Text to Speech**: Convert text into audible speech.
- **Plagiarism Detection**: Machine learning enables the development of advanced techniques to prevent plagiarism.
- **Analytical Platform for Institutional Repository**: Advanced reporting and analysis of digital content in institutional repositories.

3. Use Case for Analytics in Libraries: Analytical platform for an institutional repository

Most repository software applications lack a module to analyze in detail – and visually – the usage, storage and other key indicators within the repository. This is true for any open source package[2]. Commercial repository applications like Mendeley might have an analytical module. At best, system administrators create e.g. PERL scripts to extract particular information out of their systems. This information, while useful, is limited and is sysadmin oriented.

Some repositories show how often a particular digital object has been downloaded. This is normally done to show the end user how popular a particular digital object is. Naturally, having the information of how many times a digital object has been downloaded might be a useful indicator, but is not really sufficient for a Repository Manager. This information should be aggregated; displaying it next to the digital object might not be ideal to analyze downloads a as whole.

To have a good idea of what is going on in the repository, the Repository Manager requires a good overview of the content and activities within a repository.

3.1. What information should be analyzed?

1. **Searches**: it would be important to know what the users are searching for within the repository, what the needs are, what the top searches are. Are users searching for inappropriate content?
2. **Bandwidth**: another good indicator to keep track of.

3. **Storage**: self-explanatory. Storage and bandwidth data can assist the Repository Manager and the Institution to justify additional funds, plan growth and usage, estimate costs.
4. **Users and usage**: what are users doing in the repository? Who owns what?
5. **Traffic**: logging all web traffic is always a good idea. Applications like Piwik might be a good option for those repositories that do not include a web traffic management module. Things to track include visits, duration of visits, referring URLs, events (upload, download, etc.), browsers used, etc.
6. **Digital Objects**: what is inside your repository?
7. **Classifications**: usually, when a digital object is uploaded to a repository, the system will ask for a "tag" or "classification" of the object. An object can have more than one tag or classification.
8. **Audio, Video, Text, Image analysis**: what is inside the digital objects? This task can be done automatically using cognitive services.
9. **Top Ten**: the top ten biggest files, top ten most downloaded files, top ten uploaders, top ten searches, etc.
10. **Additional information** such as: files that have never been downloaded or seen, users who have never logged in, etc.

All these data should be aggregated on the fly by user, object, year, month, day, content model, etc. The analytical application should be multisite, multitenant, multiuser, web-based, and easy to use.

3.2. Phaidra Statistics

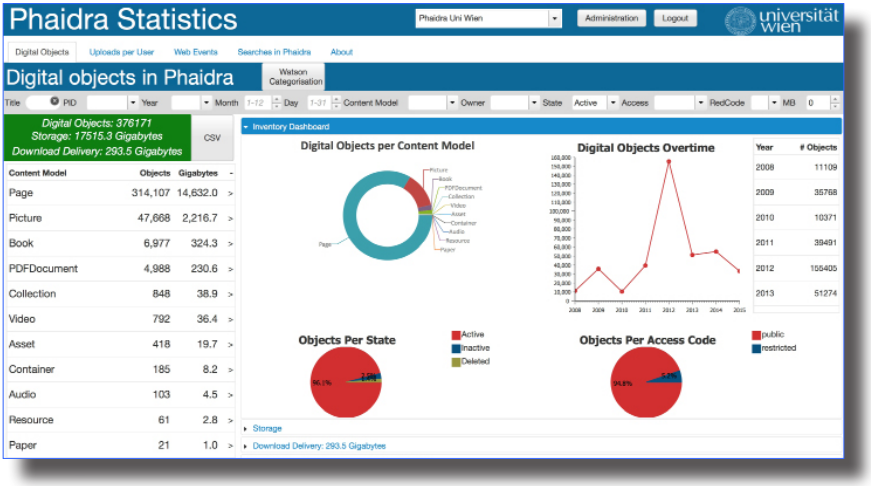
I developed such an analytical platform for Phaidra, the repository at the University of Vienna. This platform manages large amounts of metadata belonging to all the digital objects stored in Phaidra.

I also integrated IBM Watson within it to perform automatic classification of digital objects[3].

Phaidra's backend uses Fedora Commons. The frontend has been developed by the University of Vienna and uses Piwik to log traffic.

Phaidra University of Vienna: https://phaidraservice.univie.ac.at

Phaidra Statistics GitBook: https://www.gitbook.com/book/jluni/phaidra-statistics/details



José Luis Preza
E-Mail: jl@preza.org
Website: http://www.preza.org

Links

1. Wikipedia definition of Recommender System: https://en.wikipedia.org/wiki/Recommender_system
2. Analytical Platform for an Institutional Repository: https://www.linkedin.com/pulse/analytical-platform-institutional-repository-jos%C3%A9-luis-preza-d%C3%ADaz

Notes

1 See also: José Luis Preza. (2016, October 31). Automated Information Enrichment for a Better Search. Zenodo. DOI: http://doi.org/10.5281/zenodo.163933

2 See: José Luis Preza. (2016, September 25). The Best Repository for (Research) Data. LinkedIn. https://www.linkedin.com/pulse/best-repository-research-data-jos%C3%A9-luis-preza-d%C3%ADaz?trk=pulse_spock-articles

3 See: José Luis Preza. (2016, April 2). IBM Watson Analytics for Information Enrichment and a Better Search. https://www.linkedin.com/pulse/ibm-watson-analytics-information-enrichment-better-preza-d%C3%ADaz