

Data Science and Analytics in Libraries

José Luis Preza

<http://www.preza.org>

jl@preza.org

Vienna, November 29th 2016

This document was prepared as part of the Deliverables of the Metadata Workgroup of e-Infrastructures Austria, a nation wide project regarding the design and management of digital infrastructures for research data www.e-infrastructures.at

Keywords: Data Science, Analytics, Libraries, Machine Learning, e-infrastructures Austria, Metadata

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.

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, numbers, 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.

Data Science application in Libraries:

The techniques and methods used in Data Science allow Libraries to ease 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. See a related document I wrote titled: “Automated Information enrichment for a better search”, Zenodo DOI: <https://zenodo.org/record/163933>)

- **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:** Make reports out of your contents
- **Predictive Analytics:** who is going to read/use what.
- **Machine Translation:** Automatic translation of digital objects, including Braille.
- **Speech to Text:** extraction of Audio Speech to convert it into text.
- **Text to Speech:** convert text into audible speech
- **Plagiarism:** with machine learning, advanced techniques can be developed to prevent plagiarism
- **Analytical Platform for Institutional Repository:** advanced reporting and analysis of digital content in institutional repositories.

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 (see my article [“the best repository software for research data”](https://www.linkedin.com/pulse/best-repository-research-data-jos%C3%A9-luis-preza-d%C3%ADaz?trk=pulse_spock-articles))
https://www.linkedin.com/pulse/best-repository-research-data-jos%C3%A9-luis-preza-d%C3%ADaz?trk=pulse_spock-articles

Commercial repository applications like Mendeley might have an analytical module.

At best, system administrators create perl (or whatever) scripts to extract particular information of their systems. This information although useful, is limited and is sysadmin oriented.

Some repos show how much 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 helps a bit, but is not really sufficient for a Repository Manager. This information should be aggregated, having it next to the digital object might not be optimal to analyze downloads as a whole. for a Repository Manager

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.

What information should be analyzed?

1-Searches: it would be important to know what the users are searching within the repository, what the needs are, what the top searches are. Are users searching for inappropriate content?

2-Bandwith: another good indicator to keep track of.

3-Storage: self explanatory. Storage and Bandwith data can assist the Repository Manager and the Institution to justify additional budgets, plan growth and usage, estimate costs.

4-Users and usage: what are users doing in the repo? Who owns what?

5-Traffic: logging all web traffic is always a good thing. 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 url, events (upload, download, etc), browser, 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 10 biggest files, top ten most downloaded files, top ten uploaders, top ten searches, etc...you get the idea..

10- some other information like: 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.

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 (see my article [“IBM Watson for information enrichment for a better search”](#)).

Phaidra uses Fedora Commons in the backend. 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>

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>