Data Quality in the NYPR Archives

Introduction

The New York Public Radio Archives (NYPR) Archives states as its mission

> preserving [NYPR's] organizational and programming legacy for future generations of producers, researchers and listeners. The Archives collects, organizes, documents, showcases and makes available work generated by, and produced in association with, all NYPR entities.

For almost 20 years, the Archives has used the PBCore schema to describe its assets. Developed by a consortium of US broadcasters, PBCore aims to provide a common language among public broadcasters.

By aiming to accommodate the diverse requirements of many diverse broadcasters, PBCore offers remarkable flexibility. However, this flexibility, coupled with years of data entry with virtually no quality control, has resulted in inconsistent entries in the Archives' PBCore-based database (nicknamed "cavafy").

To tackle this issue, in 2018 the Archives began mapping PBCore elements to a set of 18 essential descriptive elements, with one-to-one correspondence to physical and digital objects[1], and limiting the use of each descriptive element to one occurrence per object[2]. Sixteen of the essential set of 18 fields come from Microsoft's elements in the RIFF INFO chunk of WAV files (later adopted by the EXIF consortium), while two fields come from the audio file's BEXT chunk (developed by the EBU as part of the Broadcast Wave File specification). This essential set of 18 metadata elements, alongside the development of simple ingest protocols (namely, a single point of entry --i.e., a folder), allow for much easier implementation of quality assurance, and has resulted in a dramatic improvement in data consistency and speed of data entry. The data model for these elements is both simple and linear: each object has its essential metadata expressed as one row in a spreadsheet.

The eighteen fields are Title, Contributor, Date, Creator, Keywords, Description, Genre, Producing organization, Show/series, Copyright, Original medium, Comment, Engineer, Software, Metadata source, Provenance, Technician, and Coding history. This report focuses on the first nine listed, which are more related to discoverability.

Five thousand audio files have been described with these 18 essential elements, and Archives expects to use them to describe almost 200,000 more audio files over the next eight years. This document explores ways to evaluate the fields' usefulness in a more systematic way.

---

[1] PBCore uses a relation between "assets" (descriptions) and "instantiations" (physical and digital objects), roughly equivalent, respectively, to the top entity "Work" and bottom entity "Item" of the Functional Requirements for Bibliographic Records (FRBR) model.

[2] Most of PBCore's elements are "unbound", meaning that there are no restrictions, for example, to how many titles an object can have, or even how many titles of one type an object can have.

Data quality studies

Similar studies on discoverability and data quality have emerged since the world wide web allowed users to access materials directly, without the traditional library and archives "gatekeepers". The Research branch of OCLC, a global consortium of 15,000 libraries, has published several high-quality reports; a particularly relevant one may be OCLC's *The Metadata is the Interface* (2009). We have used OCLC's 2009 report, *Online Catalogs: What Users and Librarians Want,* as a model for this report. The question in that report is: What is data quality for end users and for information professionals?

The OCLC report indicates that data-quality expectations differ for librarians vs. end users. Librarians need certain fields to perform their jobs, while end users focus on access. For example, the top priority for librarians of all types is to merge duplicate records, while end users' top priorities are immediate access to a proxy item (in our case, ability to listen to the audio) and accessing more subject information. As stated, this report focuses on nine fields related to discoverability. It also attempts to get a glimpse of what kind of materials our end users seek.

For the present study, we consider NYPR Archives staff as the information professionals that are equivalent to the "librarians" group in the OCLC report –the people who perform the cataloging and data management of materials. We consider end users everyone else, although we further divide this group into "internal" NYPR users and "external", non-NYPR users. As we will see, they exhibit some significant differences.

Methodology

The OCLC report used surveys to assess data quality expectations for end users. For the present paper, we have used three sources of data:

1. 2,305 normalized entries covering four years (2017-2020) of data from the Archives' **reference request forms**. These forms are filled out either directly by requesters or by Archives director Andy Lanset afterwards if the request was entered elsewhere (e.g., via a phone call or an e-mail).
2. 35 replies to a **Content Department survey** sent to the top 10 archives reference-service users and then to the 295 NYPR Content Department members.
3. A conversation about data quality among the four NYPR Archives staff members. Informally, NYPR Archives staff report adequate satisfaction over the adoption of these 18 fields in terms of usefulness and ability to disambiguate.

The first source, the reference request forms, provide most of the data on external users. After normalizing the data in these forms, required fields in the form were used to parse out rows in the spreadsheet. To detect patterns in the natural language[3] used to create a request, we used Regular Expressions (RegEx); we then manually eliminated false positives. For example, to explore "topics" in the natural language, we detected words such as "about", "regarding", or "on", and then deleted, for example, patterns that used "on" to denote a date (e.g. "on July 2, 2009"). There was no systematic way to include false negatives.

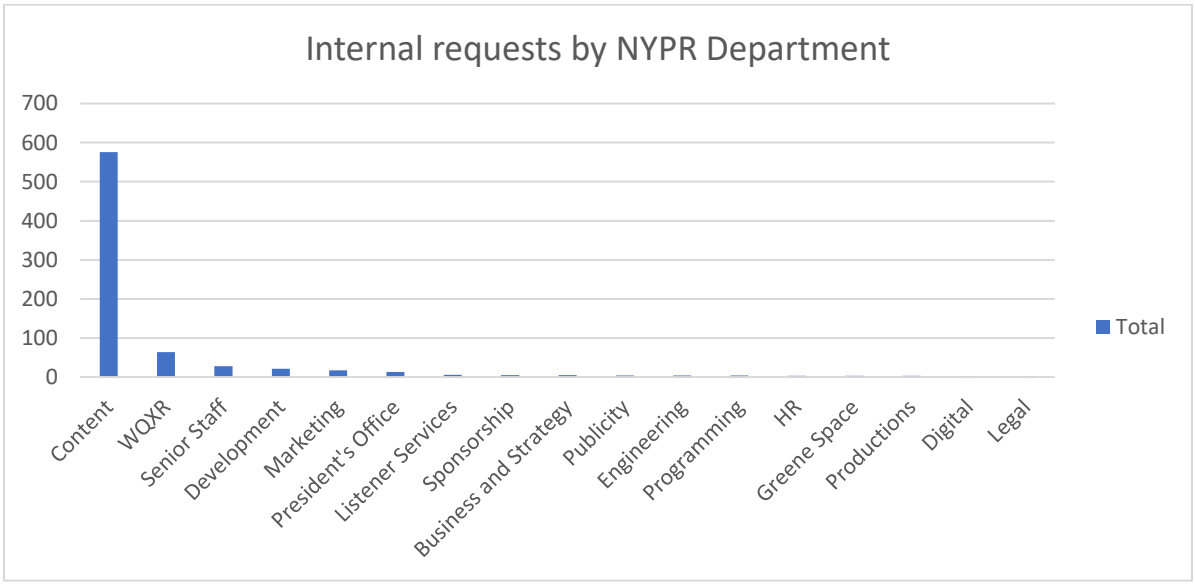
[3] When Archives staff fills out these forms, the language appears to be considerably less "natural". This is to be expected, since the form is just used as a record-keeping device, rather than a means to communicate with a human being.

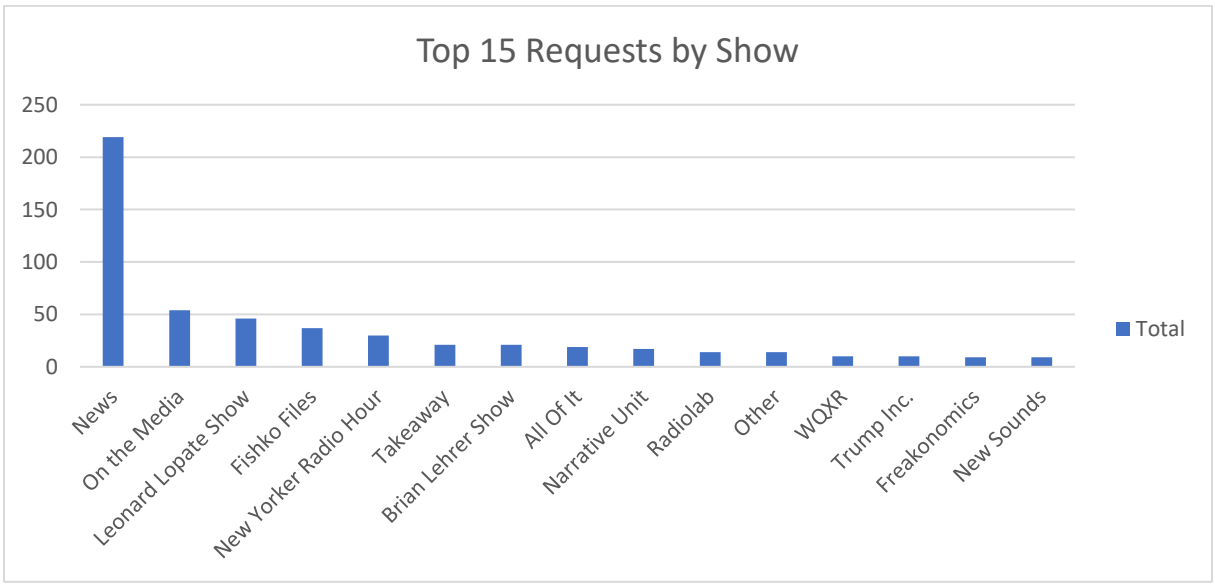
The NYPR Archives users

The two main groups of Archives users are NYPR staff and external requesters. In the years covered, one third of requests received by NYPR Archives were from NYPR staff, and two thirds were external requests.

NYPR Staff

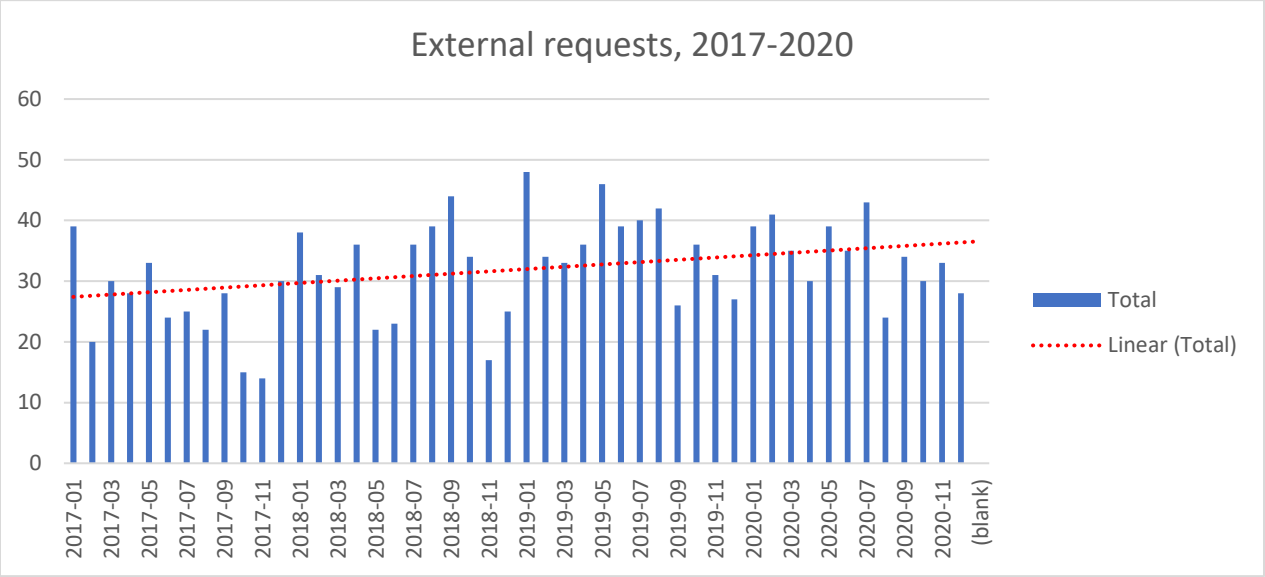
The NYPR Archives was originally established to mainly serve NYPR staff. Three-quarters of NYPR staff requests come from the "Content" department (576); of these, one third (219) are from the News Department. Other heavy users include individual shows such as On the Media (54), the Leonard Lopate Show (46), Fishko Files (37), New Yorker Radio Hour (30), the Takeaway (21), the Brian Lehrer Show (21), All of It (19), and Radiolab (14).





External requests

External requests have generally increased over 2017-2020, reaching a peak of 48 in January 2019; this could be due to increased presence in sites such as dp.la or americanarchive.org, or simply increased traffic in NYPR digital properties such as wnyc.org and wqxr.org. The proportion of external requests has greatly increased. In 2017 they were less than half; in 2020 they represented two-thirds of all requests, generating $60,000 in licensing revenues[4].

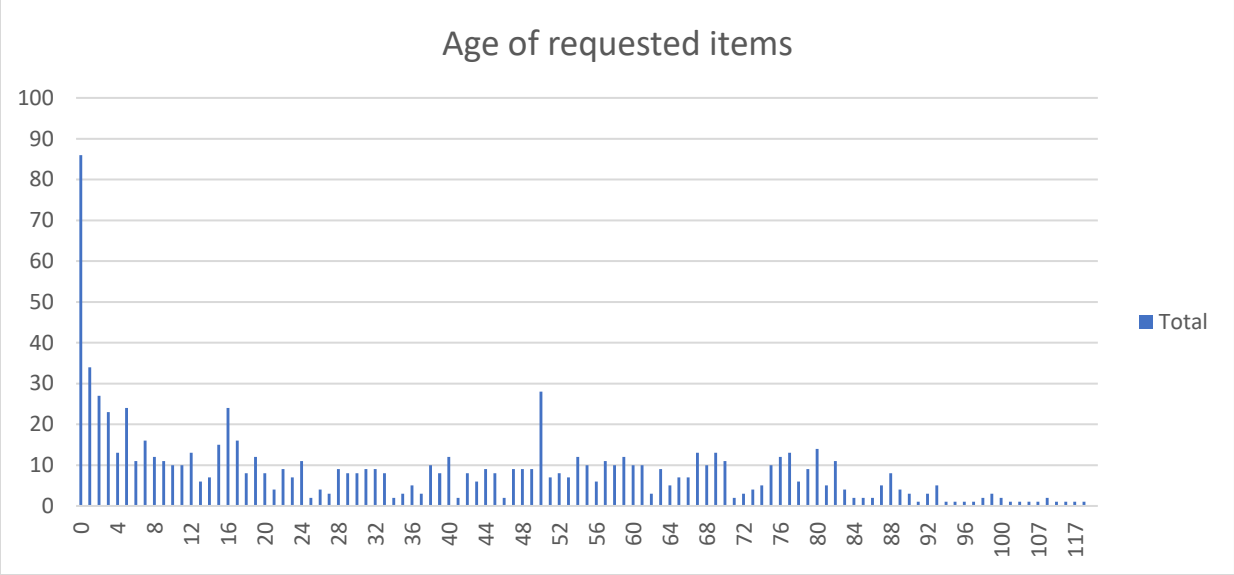


[4] NYPR Business Office

Profile of requests

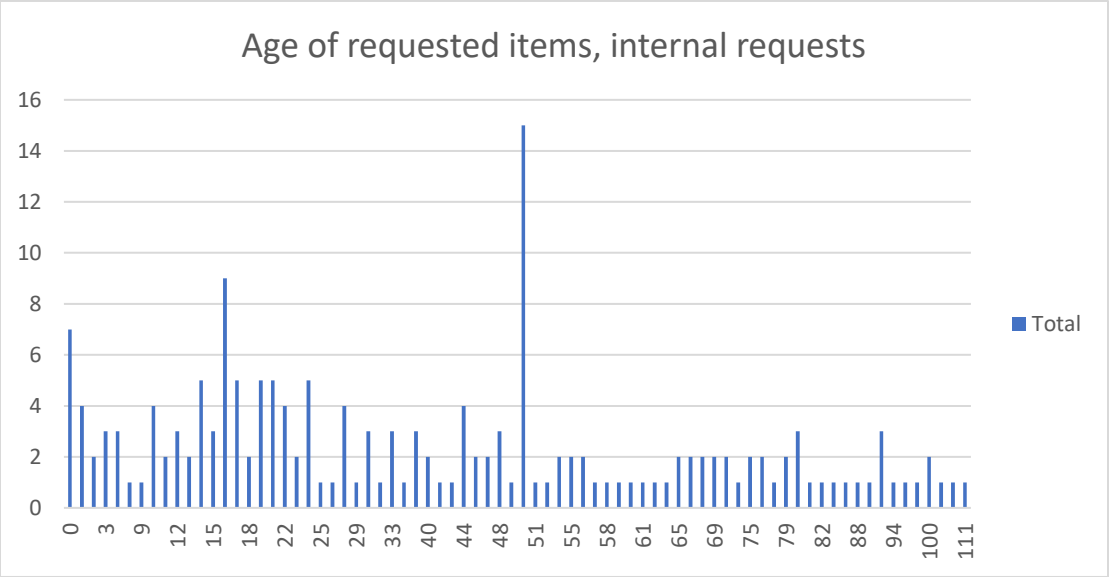
During 2017-2020, 70% of requests included audio, with more than half the requests requesting *only* audio. The top request parameters included names (mentioned in 60% of requests), dates (40% of requests include a year), topics (almost 20%), and genres (almost 16%). Shows and producing organizations (WNYC, WQXR, Greene Space) were mentioned in about 10% of the requests. (These percentages are not mutually exclusive).

Of the requests that mention a year, almost 10% request material from that same year; more than one quarter of requests were from materials 10 years old or less. The rest of years are fairly evenly distributed, with a noticeable peak at 50, probably due to anniversaries of events.



NYPR staff requests

The most significant difference in NYPR internal requests is that the age of materials requested trends considerably older, with a maximum at 50 years old.



Top request parameters for NYPR staff requests include names (almost 50%) and dates (almost 25%). Topics are mentioned in less than 8% of requests; genres, shows and producing organizations are all mentioned in less than 4% of the requests.

Some of the generally lower numbers may be due to the fact that many of the internal requests were filled out afterwards by Andy Lanset, which means the request came via other means. Naturally, the request descriptions are less wordy.

The 35 results from the internal Content Department survey generally match this profile. Four out of five respondents expressed most interest in materials older than five years; all of them included audio as their top request; persons, subjects and dates were the most important fields when searching for materials.

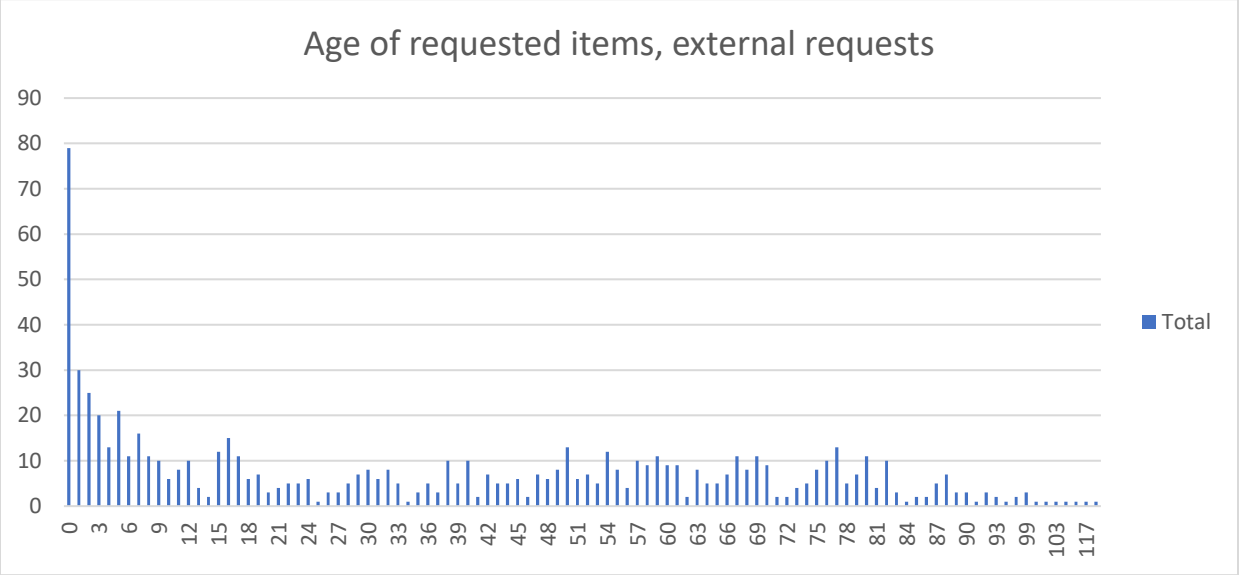
However, there were some surprising results of the internal survey with regards to discoverability. They include:

- Almost half of respondents found copyright info "least important" when looking for material
- More than half of respondents found location of recording and producer info "least important" when looking for material
- About 40% of respondents found a full transcript "least important" when looking for material

External requests

External requests drive the trends stated above for overall requests, with proper names (60%), dates (50%), topics (almost 25%) and genres (about 20%) as the main parameters. Producing organizations (about 17%) and shows (about 15%) round out the profile.

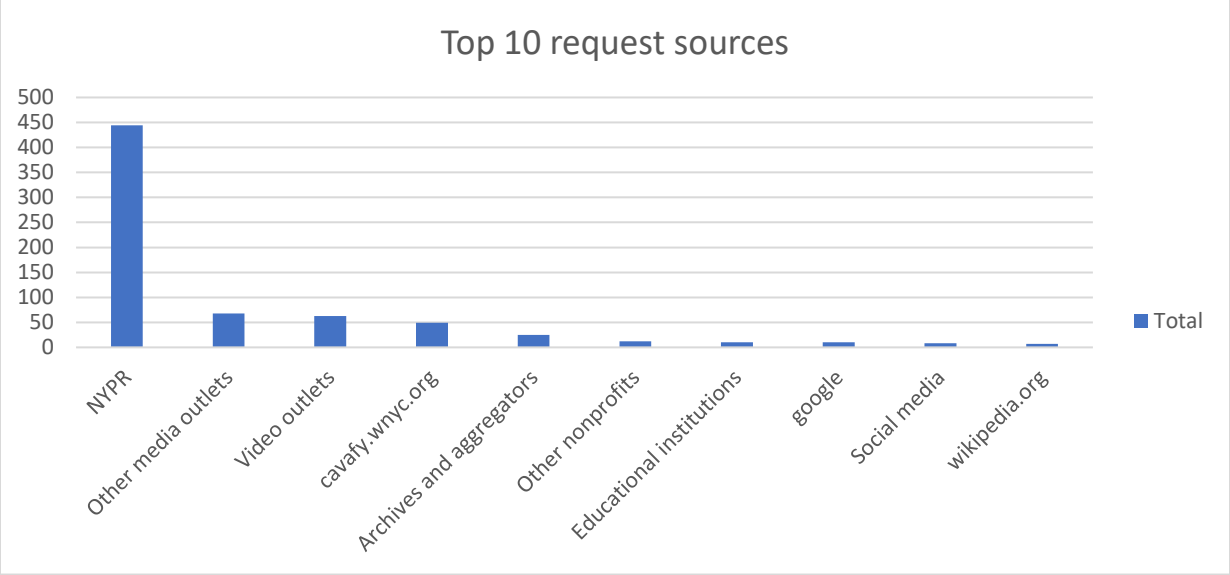
The age of materials requested trends much lower for external requests. More than 5% were requests for materials from that same year, while materials less than 10 years of age accounted for almost 10% of external requests. There is barely a peak at 50 years.



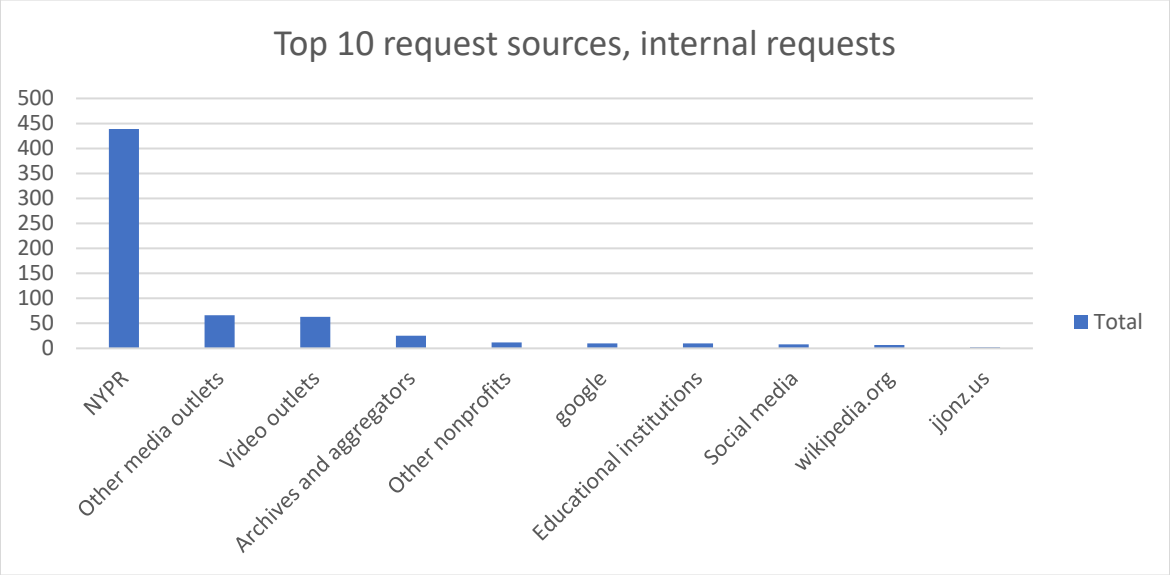
Item references

Over one-third of requests mention a reference where the requested items appear. Almost half of them reference an NYPR digital property such as wnyc.org or newsounds.org; the rest are spread among media and video outlets, archives and aggregators, and other sources.

The relatively rare occurrence of Google may have more to do with Google being a gateway to other URLs, and that Google links are awkward to share.

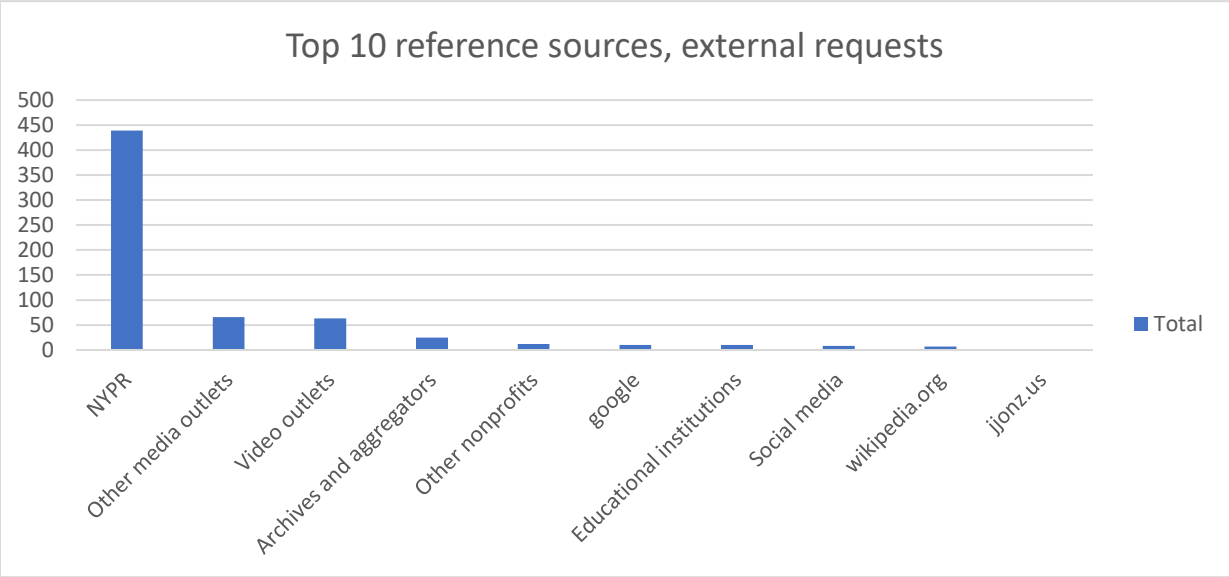


The difference between internal NYPR users and external users is dramatic. When internal users mention a reference (which is less than 10% of the time), it is most often the Archives' internal catalog, cavafy. Nevertheless, over 90% of internal users do not mention cavafy as a reference. It seems NYPR users prefer to contact Archives staff directly. One respondent said, "I didn't know about the online archives link! I hope I haven't asked too many questions that I could have answered myself."



The internal Content Department survey corroborates these findings. Virtually all respondents use Google often; two thirds of respondents use youtube. Other sources used often, albeit at considerably lower numbers, include libraries, Archives staff and DAVID. Concerningly, only three respondents mention using cavafy often.

On the other hand, when external users mention a reference, they mention one of our digital platforms more than three-quarters of the time. This means that our native digital platforms are very useful indeed to users; it also may mean that our data is not widely available on other platforms. Over 40% of requesters did not provide a reference, suggesting that discoverability may still be an issue.



Archives staff

Currently (2021), NYPR Archives is staffed by three full-time employees and one consultant. Their needs range from answering reference requests to generating manifests for external digitization vendors.

On April 5, 2021, the four NYPR Archives staff members discussed their desired data quality enhancements. In no particular order, these desired enhancements include:

- Better defined fields, e.g. Series, Collection, Genre
- Better defined entities, e.g. assets/instantiations/segments, and their interrelations (e.g. compilations)
- Full concordance between physical and digital items and their catalog entries
- Consistent data entry, e.g. dates, technical notes, location
- Consistently formatted text, e.g. from csv imports
- Consistent use of authorities or references

Archives staff hopes that the consistent implementation of the NYPR Archives Ingest Protocol, alongside the use of the station-wide Digital Asset Management System, will help solve many of these issues.

Conclusions

- Requests for NYPR materials have increased over 2017-2020. The proportion of external requests has increased considerably
- Internal requests, mostly from the Content Department (and especially News), tend to look for much older material. External requests tend heavily towards recent items
- Requests mention names, dates and topics (in this order) most often. Genres, shows and producing organizations appear less often
- Over one quarter of external requesters mention an NYPR digital property –much higher than any other source. This may mean that, although discoverability may still be an issue, our digital properties are still the best place to find recent items
- The Archives internal catalog, cavafy, remains virtually unused by NYPR staff (other than Archives staff)
- Aggregators such as the American Archive of Public Broadcasting are mentioned only in less than 2% of external requests, while Google and youtube are mentioned in more than 4%. Archives may want to consider increasing discoverability via search-engine optimization. Consistent application of data quality standards will help discoverability for both sources