2019 IFLA WLIC Athens

Submitted on: 28.06.2019

Bringing Library and Museum Resources Together: How Can Artificial Intelligence Help (based on the Ivan Tsvetaev's Book Collection Project)?

Ekaterina Igoshina
Research Library, Pushkin State Museum of Fine Arts, Moscow, Russia
ekaterina.igoshina@arts-museum.ru.

Juliya Dubrovskaya
Research Library, Pushkin State Museum of Fine Arts, Moscow, Russia
julia.dubrovskaya@arts-museum.ru

Abstract:

A few years ago at the Pushkin State Museum of Fine Arts Research Library we began a long-term digitalization program of the most valuable and/or fragile parts of our book collection. For a start, we decided to divide this process into a number of smaller programs, each dedicated to a different collection. One of our first and obvious choices to digitize (and thus make visible and fully accessible) was the collection of our museum's founder and first director Ivan Tsvetaev (1847-1913), that consists of more than 850 volumes and vividly highlights his scientific interests, academic ties and hobbies. While digitizing Tsvetaev's books we discovered quite a large number of dedicatory inscriptions (nearly 250) — a very useful complementary source of information on scientific and civil circle of our founder. All the inscriptions were scanned, decoded and translated into Russian, where necessary. Each of the inscription images was also enriched with brief biographies of the inscriptions' authors (donators).

First of all (and that was the most obvious step) we joined our various digital resources into a common semantic space — a digital collection — using available technical tools that were in our disposal. By compiling descriptions of all units related to the collection in our catalog and linking them with our unified authority data system (such as subject and name indexes), we managed to create a very comfortable in use semantic fragment, that better describes Ivan Tvetaev's personality and the history of our museum's creation through the prism of his personal interests and professional contacts.

Then we started contemplating the way of possible connection of this collection to other complementary museum resources, that are kept in other departments and described in other IT-systems. One of the possible approaches to bring all these uncoordinated sources together is applying artificial intelligence technology. We found this way intriguing and rather promising.

Our paper is aimed to explain the possible way of data relationship to design a dialogue between the voice assistant and the user while searching through library, museum, and possibly external resources (based on the example of a certain digital collection).

Keywords: museum library, digital collections, data relationship, artificial intelligence technology, voice assistant.

Introduction

A few years ago we began a long-term digitalization program aimed to preserve from decay the most valuable and/or fragile parts of the book collection of the Pushkin State Museum of Fine Arts Research Library. In order to give a general idea of our Library's possessions for all those who are not acquainted with it, we would like to point out the following issues:

- our Library opened its doors to the public on May 31st, 1912 (on the same day the whole Museum was inaugurated);
- at the time of the opening and during the first two decades of its existence the Library used to serve as an important auxiliary instrument for the museum researchers and collection curators as well as the public reference art library accessible for external users (most of them were ordinary museum visitors);
- its core collection originates from the Fine Arts Cabinet of the Moscow University Philological Department where the Museum's founder Ivan Tsvetaev used to be a head before becoming in 1912 the first director of the newly founded museum;
- nowadays our collection comprises about 220 thousand storage units including books, brochures and periodicals that cover the whole history of Western art from the ancient times to our days;
- nearly 40 thousand storage units of the whole collection are the so called rare and valuable materials (either books or periodicals) that need special care and are obviously the first to be the subject for mass and accurately planned digitalization program.

For a start, following the most successful world digitalization practices, we decided to split the process into a number of smaller programs covering different collections. One of our first and mostly obvious choices to digitize (thus making it visible and contributing to a greater public access) was the collection of the above-mentioned museum's founder and its first director Ivan Tsvetaev (1847-1913). It was donated to the Emperor Alexander III Fine Arts Museum Library (future Pushkin State Museum of Fine Arts Research Library) after Tsvetaev's death in 1913.

Tsvetaev's book collection consists of more than 860 volumes and vividly highlights his scientific interests, academic ties and hobbies. Due to the owner's interest to the ancient italic languages and archeology as well as to the history of Western art from antiquity till the end of the XVI century, most of his books are dedicated to these two issues. Being a university professor, he never had enough financial resources neither to purchase really rare or well-illustrated books nor to bind his books in a splendid manner. Despite the fact that the largest part of his collection consists of academic publications in relevantly modest bindings, for us it presents a rich source of historical information on the process of planning, building and then opening of our Museum, as well as on the person, whose role in founding our Museum is recognized as crucial.

While digitizing Tsvetaev’s books we discovered quite a large number of dedicatory inscriptions (nearly 250 items) — a very useful complementary source of information on scientific and civil circle of our founder. A significant part of the inscriptions represented simple and rather formal dedications, while others contained quite long texts referring to the special sort of relations that tied him together with the authors of the inscriptions. Among those who had left their handwritings on Tsvetaev’s books were his numerous colleagues (professors of philology, archeology and history of art from Russia and all over the world), his former university students and some close family friends. All the dedicatory inscriptions that we discovered within Tsvetaev’s books were carefully scanned, decoded and translated into Russian, where necessary. Each of the inscription images was also provided with brief biographies of their authors (or donators).

First of all (it was the most obvious step) we integrated our various digital resources into a common semantic space inside the Library software environment — the digital collection — we did it with the help of technical tools that were available at our disposal. By compiling descriptions of all units related to the collection into our Catalogue and linking them up with our unified authority data system (such as the subject and name indexes), we managed to create a rather user-friendly semantic fragment which better describes Ivan Tvetaev’s personality as well as the history of our Museum’s founding through the prism of his personal interests and professional contacts.

And then we have started working on the problem of providing a more thorough information on historical liaisons which we received while decoding donative inscriptions. It is so valuable to us when studying sources as it contributes to establishing ties and cross reference between this collection and other relevant museum sources that are kept in other departments and are based in a different IT-system. But the problem is that even within one and the same institution there can be no technical solution for developing semantic cohesion of the museum items belonging to different departments, not mentioning the ties to the external sources.

For example, there are two small brochures among Tsvetaev’s books donated to him by his ex-student Nikolay Romanov at the very beginning of the XX century. Both brochures contain almost identical and very short inscriptions “To my dear and deeply respected professor Ivan Vladimirovich Tsvetaev from the devoted author”. Nikolay Romanov, one of the most brilliant Tsvetaev’s students, graduated from the Moscow University in 1890, and in 1898 he entered the Committee for the Museum of Fine Arts Foundation (thus becoming Tsvetaev’scolleague and most loyal assistant). In 1923 (which is ten years after Tsvetaev’s death) he succeeded Tsvetaev taking the position of the Museum director and holding it until 1928. Browsing our Museum catalogue (a special database for all sorts of the Museum items compiled by our collection curators from different departments) one can easily find more than 20 museum objects related to Nikolay Romanov: some of his photos, 10 graphic portraits in various techniques, scan of his periodical publication on the Museum history as well as his book-plate. By showing this example we would like to stress the idea that various associated resources which are identified separately in the Library and the Museum IT-systems actually tend to complement each other and create a thematic space dedicated to a certain person, but in practice fail to fulfill this target when searched separately.

Weighing various advanced alternatives offered by the IT-market of nowadays in the area of integrative search, we have taken interest in the cutting edge solution which made a break-through within latest two years. These are the artificial intelligence technologies.

While using this term, “artificial intelligence”, we imply the type of program applications that are able to simulate certain human intelligence processes using complex mathematical interpolation. For instance, they are able to recognize visual images, spoken words, to make decisions assisted by algorithms and to develop trends based on statistical analysis. Some program artificial intelligence (AI) solutions are becoming easily affordable at the moment. The largest search engines open their program libraries, voice recognition applications and machine learning algorithms.

The speech recognition technologies implemented in the form of the so-called voice assistants or smart assistants were well received and are now quickly winning public recognition. According to the forecasts of analytics agencies, by 2020 from 30 to 50 % of search queries on the internet will be performed by verbal asking. And there is a reasonable explanation for this. The voice assistant skill is a new type of working with databases. The user receives the required information while conducting natural conversation. It enables the user to keep his daily rhythm and manage the database in the way that corresponds to modern ergonomic requirements to technologies, as well as bringing quality and quantity ratio of the search results to the amount most suitable for human perception.

The voice query doesn’t require from the user knowledge of special syntax of search queries or metadata standards describing the resources. The question can be formulated in any form that the user finds convenient in plain human language. Even one such characteristic of voice aid applications makes them a matter for serious consideration of introducing this technology to bibliographic search.

The speech recognition technologies split the phrase into words, then the specially developed search script defines semantic dominants, and finally the variations of submitted users’ queries get transformed into the so-called key phrases which guide further search. And it is worth mentioning that the number of such recognized phrases is constantly growing — artificial intelligence “learns” from each query.

Apart from that, the important characteristic of the smart assistant search is the opportunity to interact with the user who is not left alone facing the search results, but is offered a continuation of dialogue in the form of the search conversation. So, there is a chance to let the user get more satisfying results, and more than that, there is a chance to receive feedback on the quality of the query fulfilled.

However, the above-mentioned opportunities do not cover all the possible advantages of using AI technologies while processing bibliographic databases and electronic information resources. The speech recognition technology contains powerful algorithms of the so called “machine learning”. The software libraries which carry out these algorithms are being constantly upgraded and developed by world largest developers in the area of cognitive technology and are open to public access.

Machine learning algorithms are in fact the core of artificial intelligence technology. They provide extraction of rules, functions, dependencies within the data items based on mathematic models and neural network principles. And if in the earlier days the development of the algorithm of database search was fully programmed by human, and the data required preliminary preparation, namely, formalization to this algorithm, nowadays the machine learning algorithms can analyze the content of the database as well as self-study on the samples of successful results of search queries by finding interrelation between the structure of the database referring to the user’s query and the query itself. Thus, the more number of fulfilled queries the machine learning algorithm receives, the more exact model of search structure on the possible source it will be able to carry out.

Therefore, theoretically machine learning algorithms will be able to cope with the problem of accuracy in following the standards while preparing metadata.

Regretfully, we must note that despite the developed standards of defining the items of the database, there can be various ways to interpret them, and so even the volumes of one and the same multivolume set defined in the bibliographic base by various people in various years can vary tremendously, so that the classical search algorithms will fail to find both descriptions on one and the same search query.

Integrating databases to accurate following the standards is often impossible to achieve primarily because of the amount of data subjected to rewriting.

The result of implementing AI technologies into the search through museum databases can be depicted in the following scheme.

What we do have now: separate search through various databases.

		Search in	Library automated information system (AIS)
		Search in	Museum AIS
		Search in	Archive AIS
		Search in	External AIS

What can be done using AI* applications: integrative search

Voice (Speech)	Intellectual	Search in	Library AIS Museum Archive External data set
- AI* speech recognition, transmission of articulated words into search - The script of the dialogue with the user (what to search for, where to search for, how to submit the results)	- The script of the search in bibliography using AI* machine learning libraries for text analysis with no regard to the rules of reading the format, basing only on the queries fulfilled and their estimation - The user's action logging system as a learning model (user's actions, related queries) - The database of best answers to standard queries		Integrating various resources for search into one storage in JSON format (on shedule), including exernal LOD-published and other types of datasets.
In a plain natural language	Search through text with no regard to the standard of inputting the data and its compliance		Cross search

*The points where AI technologies are applied are marked with underlining

By far we are not aware of the examples of such fully spread AI systems to bibliographic databases, but we think that implementing AI technologies to museum databases will result in the unified self-learning system which will be flexible enough and speak natural language neglecting inhomogeneity in data and delivering quite sufficient results with no need for preliminary additional processing of data.

It is worth mentioning that while dwelling on solving the problem of inaccurate following the standards of describing the content of collections, we do not at all state that standardization is not an actual issue anymore. AI technologies, from our point of view, are just a universal and flexible means of producing data which doesn't contravene the metadata standards but make the impact of inconsistencies less crucial for search results. Following metadata standards we can connect data itself but presenting it to the final user requires user friendly approach able to carry out search everywhere where possible and will deliver the result in clear terms.

Artificial intelligence able to recognize various types of datasets and apply suitable syntax for extracting it can take the place of special software which works with only one algorithm and format of the data.

And let's get back to our Ivan Tsvetaev's Book Collection Project that in our case became a model fragment of the data for developing the model of the data for voice assistant.

In March 2018 the famous Russian search engine system Yandex launched for beta testing the Yandex Dialogues platform. This platform provides free access to speech recognition software implemented in the form of the intelligent assistant Alice. Thus, every software developer is able to apply this software for his own projects. Alice is the Russian equivalent of Google Assistant, Siri, Amazon Alexa or Windows Cortana. We find it both challenging and ambitious.

Using this kind of voice assistant required preliminary writing of a small program, the so called "skill" which contained the script of conversation of with the user. In the frame of our model experiment we combined two types of data: the data of the bibliographic base and the descriptions of the items sorted by person related to the collection from the museum accounting records. Developing this "skill" we came to understanding of the fact that the most important issues of the whole thing is predicting and modeling user's queries on the matter and developing of the dialogue script between Alice and the user, and vice versa — between the user and Alice.