Gallica, a gold mine and cultural legacy

Arnaud Beaufort,
director of Services and Networks, assistant director-general, Bibliothèque Nationale de France (BnF)

Abstract:
The digital library Gallica uses the resources of the Bibliothèque Nationale de France. With more than four hundred partners, it provides for-free access to more than six million documents. Placing it on line (in 1997) has opened contents to researchers and the general public (who now has several points of access to our cultural heritage). Gallica is striving to play a central role on the Web and in cybernauts' browsing histories. It both references documents by using the language of search engines and builds communities of users and "builders" (via whitemarking). These relays and catalysts enhance Gallica's collection thanks to human intelligence. This emerging public service provides fertile grounds for the formation of a digital legacy, both with regard to the contents and in terms exploration and appropriation.

Inaugurated on line in 1997, Gallica, the digital library of the Bibliothèque Nationale de France (BnF) and its partners, became a library of the *honnête homme* in the middle of the first decade of this new millennium. It groups the most significant works of the intellect since Ancient Times into a universal, encyclopedic, vast and multifaceted collection destined to never stop growing: books as well as engravings, photographs, scores, videos, puppets, medals, etc. This collection amounted to more than a million "documents" in 2010, five million in 2019 and six million at the start of 2020. This mass, referenced by major search engines, has attracted the interest of both academics and the general public. In fact, a study of how Gallica was being used in 2016 brought to light the importance of personal searches alongside academic research (BnF 2017). The website hosts 50,000 visits a day — 70,000 during the period of sheltering in place. In 2019, 50% of its contents (nearly three million documents) were consulted at least once.[1]

This library's freely accessible contents form an organized collection with a constantly moving perimeter as procedures for diffusing, browsing and appropriating information evolve on the Web, where access is less intuitive, simple and clear than it seems to be. In effect, the search for information and the capability of identifying pertinent sources depend more on skills and qualifications than on intuition; and the choices made about the "transmission" of information are as important as the contents transmitted.

The aggrandizement of this library has come out of several years of work for digitizing the BnF's legacy of documents, along with deposits from other libraries. Gallica is a collective, digital library with more than four hundred institutions in its wake.

Given all this, Gallica can be described as an evolving multidimensional organization. Completing the website gallica.bnf.fr are, among other things, applications (Android and iOS), an intramural version with copyrighted material and white-label products. For Gallica's development, the BnF has adopted three approaches: continually improve its referencing tools, stake out a position as a trusted party, and "capitalize" for the purpose of forming a 21st century digital heritage.

[1] This article, including quotations from French sources, has been translated from French by Noal Mellott (Omaha Beach, France). The translation into English has, with the editor's approval, completed a few bibliographical references. All websites were consulted in December 2020.

Referencing six million documents so that they can be found

Gallica, a web on the Web: Multiple entrance points

Like the Web in general, Gallica requires contents (available and directly accessible), servers (capable of handling multiple connections), a search engine (for processing complicated queries in real time), and interfaces (familiar and uninterrupted).

INTERACTIONS WITH SEARCH ENGINES. Fewer than one out of five users enter our digital library via its home page, and 39% of visits are made from search engines. Gallica is a reservoir, its center potentially anywhere, and its circumference, nowhere… If these engines naturally reference the long tail, which corresponds to a principle of scarcity, the situation is not the same for the contents that are commonly, frequently studied, or have many homonyms. For ten years now, the BnF has, been translating and structuring its data in formats familiar to search engines, and adopting the principles of Web semantics. This translation underlies data.bnf.fr; 84% of the visitors to this capstone on the variety of the library's digital contents arrive via search engines. This backbone of the system serves to pivot visitors toward other BnF websites. It has hoisted our contents up into the first ranks of the results in response to queries.

APIS: Each atom of Gallica can be found not just in its conventional form (inserted in a page on the website) but also directly via a line of code (URL). The development of IIIF has given a major boost to the interoperability of contents and their distribution. This API can be used to call and manipulate iconographic contents from a website so as to distribute them on another site without forcing the cybernaut to switch interfaces.[2] Among Gallica's users who have chosen this API, I might mention as an example a site on thoroughbred pedigrees, the French Polar Archives, or a website that sells customizable smartphone cases (or shells).[3]

RELAYS: Apart from the progress of technology, these uses depend on "relays" that involve an active presence on the social networks and the support of a community, the Gallicanauts. Owing to its form and involvement, this group of supporters reminds us of communities at the start of the Web. Its members interact with the BnF and share what they have discovered.

An increasing number of visits are referrals from Wikipedia. The visits to Gallica from this website rose more than 30% between 2018 and 2019.

Furthermore, white labels enable institutions to benefit from Gallica's infrastructure while making their digital collections accessible to their public. They thus enrich the general collection with new documents. This program has proven ever more popular. This pooling around Gallica's infrastructure has set off a virtuous circle of cooperation in response to the issues of "digital sobriety" and "smart information". Ten white labels are on line; and ten others, in the works.[4]

Browsing, exploring

A professional community active in the diffusion of contents on the Web is working on the problem of how to explore these contents. Exploration entails both an ongoing dialog with researchers and advances in technology, in particular the evolution of Gallica's search engine (Cloud View from Dassault Systèmes).

The tools under development for conducting searches among all the images on Gallica are advancing with the help of artificial intelligence. Experiments are already being conducted on certain collections. Meanwhile, the tools for searching in texts are being used to explore and index our collections; "search reports" process the results from search engines (GIOUX 2017). These tools

[2] International Image Interoperability Framework: See: https://iiif.io .

[3] Respectively: https://www.pedigreequery.com/, https://www.archives-polaires.fr/ and https://cover.boutique.

[4] https://www.bnf.fr/fr/cooperation-autour-de-gallica#bnf-gallica-en-marque-blanche

provide novel procedures for exploring documents, *e.g.*, searches by similarity or proximity (ÉQUIPE GALLICA 2019) or analytics performed on a vast corpus (LANGLAIS 2019).

Gallica helps create new jobs, but it is also revolutionizing existing jobs and activities owing to the wealth of contents to which it opens access and to the time saved for doctoral students, researchers, journalists, authors and draftsmen, like Pierre Lemaître, Daniel Schneidermann, Alain Rey, Benoît Peeters, Maylis de Kerangal and others.

However, the multiplication of entrance points, websites and applications does not suffice, not any more than the ongoing dialog with users, to make browsing through these contents a fluid, tranquil experience. Bearings are needed.

Gallica, a source of culture and a trusted party

Editing, counseling

How to immediately find the right version of Pascal's *Pensées*? Where to locate the first reference works on artificial intelligence or climate-related problems? Gallica provides not just information as such but the sources of information. Its assistance to the curious offers features for targeting certain groups, such as students in the building trades with the website *Passerelle(s)* or children with Gallicadabra.[5] The work of reading and analyzing contents done by Gallicanauts is valued and promoted. The capillarity between data.bnf.fr, the white labels and social networks yields a granular level of advice, selections, blogs, etc. "Gallica's counsel" heads the lists of the site's results.[6] All this helps the library play its role as a reference source — algorithmic processes completed with human intelligence.

This work as a mediator is concentrated on highlighting treasures (from the speech pronounced in 1981 by Robert Badinter against the death sentence to back issues of journals)[7], promoting the editions of classics (in literature, law, etc.), and drawing attention to old documents that vibrate in response to current events.

Ethics of the search engine and website

As a source and public service, the Gallica galaxy upholds the principles of security, neutrality, openness, legality and stability.

Neither the contents consulted during previous sessions nor our visitors' browsing histories are used to modify the order of search findings, or to confine users within an environment set by their preferences. When a researcher has patiently conducted queries and managed to detect a convergence between several titles, his work is not disclosed in a list of findings suggested to other cybernauts.

Involved in national and international debates on legal questions, the BnF has managed to strike a balance between, on the one hand, the access to its resources based on the principles of open data and open science and, on the other hand, the rules used in the business world. It makes a distinction between data (entirely free) and digitized contents (subject to various conditions). The open data policy (pursued since 2014) fits in with the state's policy favoring the emergence of new uses by citizens and new economic opportunities. As for the reuse of digitized contents freely accessible on line, the BnF supports the principle that they are for free in the for-free educational and academic world while setting fees for commercial players.

[5] Respectively: http://passerelles.bnf.fr/ & https://c.bnf.fr/Hav.

[6] https://gallica.bnf.fr/conseils

[7] https://c.bnf.fr/Hap

Whoever cites a document on Gallica can be sure not only that the link will remain valid but that the contents referenced will stay the same. The BnF even proposes its own application (c.bnf.fr) for shortening URLs and guaranteeing that short links remain valid. According to a study of uses in 2016, 66% of respondents said that they "often" or "always" consulted documents very attentively, as compared with 31% in 2011 (BnF 2017); this makes it useless to download backups.

Gallica, a fertile breeding grounds for the digital heritage of the 21st century

The ground explored by the BnF for the development of Gallica has implications for the heritage now being formed. As a source, Gallica stimulates new creations, and makes it possible to extend beyond the Web the features and uses that it has stimulated.

Openness and coconstruction

Besides having the advantage of shifting the assessment of results toward qualitative rather than quantitative criteria, the concentration on uses (the theme of this special issue) helps us to show how much a global approach has to gain by taking account not just of individual uses but also of their network implications — not so much a sum of juxtaposed uses by persons as an ongoing series of inspirational interactions. The BnF proposes services that respond to the need for customization: those already mentioned (Gallica's counsel and search reports), digitization on demand, "Adopt a book", etc. However customization cannot be imagined without sharing, without a collective movement.

The Web leans toward a conception of uses with which it has been closely associated, namely reading and writing. The BnF hackathon in 2016 gave birth to Gallicarte, now a part of Gallica.[8] A month after this feature for geolocating documents was launched, 5000 additional points were benefitting from it thanks to cybernauts. Other uses might be to make annotations, correct an OCR, or transcribe a manuscript. As explained by Henri Verdier and Nicolas Colin (2015, p. 58), "*There will nearly always be more intelligence, more data, more imagination and creativeness outside than inside an organization.*"

For this reason, the BnF has to offer, maintain and foster a wide range of possible uses, starting with the most basic (printing, downloading, etc.) without overlooking physical uses. "*Often described as 'distant' users, Gallicanauts remind us that physical and digital services nurture each other: 38% said they had already physically visited the BnF's facilities*" (CHEVALLIER 2017). The library's strength is that it does not separate these forms of materiality. It is probably one of the most suggested places for guaranteeing their coexistence.

Exploring legal electronic deposits

"Gallica intra-muros" opens access to a million additional documents.[9] Gallica's tools are useful for digital documents protected by intellectual property rights and for document searches in general. Thanks to the legal deposit of electronic documents, intramural Gallica will soon be bigger than Gallica itself. Researchers will be able to undertake searches in this deposit on location, in BnF's Datalab. The circle that started with Gallica is virtuous; and this tool will be all the more valuable since the BnF is France's legal depository for copyrighted material. At stake is informational superiority.

[8] https://gallica.bnf.fr/blog/21032018/gallicarte-arrive-dans-gallica?mode=desktop

[9] https://gallica.bnf.fr/conseils/content/gallica-intra-muros

Conclusion

This digital library's revolutionary dimension allows for shades of opinion, since it is a reflection of preexisting multidisciplinary and documentary practices. However the visibility of this heritage on line and the expansion of uses is related to three aspects: training for artificial intelligence on documents that are not contemporary (in particular, on the stock of images), the massive digitization of documents in French, and the light brought to current debates. To be on par, the BnF relies on the professionalism of its personnel, the close attention paid to its public, the technology (for searches improved by inquiries in natural language) and cooperation so that this online heritage be truly shared.

References

BNF [Bibliothèque nationale de France] (2017) *Enquête auprès des usagers de la bibliothèque numérique Gallica* (Rennes: TMO Regions), 119p., available via https://multimedia-ext.bnf.fr/pdf/mettre_en_ligne_patrimoine_enquete.pdf.

CHEVALLIER P. (2017) "Résultats de l'enquête au'pèsdes usagers de Gallica", blog du 10 May on https://gallica.bnf.fr/blog/10052017/resultats-de-lenquete-2016-aupres-des-usagers-de-gallica?mode=desktop.

ÉQUIPE GALLICA (2019) "La recherche par proximité: une nouvelle fonctionnalité dans Gallica" available at https://gallica.bnf.fr/blog/26112019/la-recherche-par-proximite-une-nouvelle-fonctionnalite-dans-gallica?mode=desktop.

GIOUX M. (2017) "Nouvelle fonction pour Gallica : Le rapport de recherche", blog of 9 January available on https://c.bnf.fr/G97.

LANGLAIS P.C. (2019) "Reconstituer les genres romanesque sur Gallica: essai de classification automatisée de 1500 romans (1815-1850)" available at https://scoms.hypotheses.org/986.

VERDIER H. & COLIN N. (2015) *L âge de la multitude. Entreprendre et gouverner après la révolution numérique* (Paris: Armand Colin).