How to Make Millions of Documents Accessible Through AI and a Digital Library? The Case of Gallica, French National Library

Antoine Henry

Abstract The chapter introduces Gallica, a project hosted by the Bibliothèque nationale de France (BnF). Since its inception in the fourteenth century, the library has been developing new missions and striving to make collections more accessible. One key tool for the BnF is Gallica, an online digital library that concentrates much of the R&D conducted by the library to ensure accessibility. Since 1997, the project has been experimenting with new features, including HTR, OCR, image, and data mining, to facilitate research within its vast collection of ten million documents. The library employs an incremental approach to scaling up proof-of-concepts into industrialised innovations that can be used by its 60,000 daily users. Meanwhile, the development of artificial intelligence contributes to cultural changes within the library. Through its central role, the national library also shares its achievements with other libraries in France and at the European level.

Keywords Handwritten Text Recognition · Optical Text Recognition · Digital Library · Data mining · Image mining · Deep learning · Bibliothèque nationale de France (BnF) · Gallica · Artificial intelligence

1 Introduction

The French National Library, known as Bibliothèque nationale de France (BnF), has its roots in the royal library established by Charles V in 1368. Today, it operates as a public institution under the supervision of the Ministry of Culture. The library is divided into two main locations in Paris: Richelieu, which serves as the historical site, and François-Mitterrand.

A. Henry (✉)
University of Lille, ULR 4073—GERiiCO—Groupe d'Études et de Recherche Interdisciplinaire en Information et Communication, Lille, France
e-mail: antoine.henry@univ-lille.fr

I. Mergel, C. Schmidt (eds.), *AI Innovations in Public Services*,
https://doi.org/10.1007/978-3-032-01344-6_9

As the legal deposit institution since 1537, the BnF is responsible for preserving and archiving all printed materials published in France. This mission includes now Web legal deposit, with over 2 petabytes of data stored online. The library is one of the largest in the world and plays a significant role in its community.

In 2023, the BnF had 2262 employees, an annual budget of €253 million and the library welcomed over 1.4 million visitors.

2 Description of the Project

The project that is studied is Gallica [1]. Gallica is an online digital library created by the Bibliothèque nationale de France (National Library of France), offering a vast collection of French and international cultural heritage [2]. This comprehensive resource provides access to a wide range of materials, including books, journals, newspapers, manuscripts, maps, photographs, and other multimedia content. Approximately 100 people contribute to Gallica, with 19 employees solely devoted to the project. These staff members are spread across various departments, such as the IT Department, Collection Department, and the Cooperation Department, which serves as the central hub for Gallica. The Cooperation Department plays a crucial role in guiding the initiative, with ten of its members hailing from the IT Department and nine from the collaboration section.

Gallica mostly uses OCR (optical character recognition) and HTR (handwritten text recognition). According to an AI expert from the library: "The first experiments took place in 2010. At the time, the BNF was already involved in a number of research projects, particularly at the European level, with major research projects in the 2010s, 2007s and 2015s focusing on the digitisation of cultural heritage on a European scale. The stakes were mass digitisation, mass OCR and handwriting recognition." Those technologies are important to ensure the accessibility of documents that have been digitised, although the concept of artificial intelligence (AI) had not been mentioned at the time. In addition, the project also uses natural language processing (NLP), data mining, and image mining. This allows users to do textual research on images or digitised documents. With the rise of deep learning (2014), for them it's also an opportunity to have a "hybridisation of approaches, types of collection and types of documents." This way, they could make a massive digitisation and pursue the processing using the same technological pipeline.

Figure 1 is a schema of how AI has been implemented in the library [3]. We are here able to see which projects were contributing to which fields.

Figure 1 illustrates the evolution of the AI-related projects in the library. We can identify the various technologies involved in Gallica according to the needs (analysis, language processing, text/image recognition). The BnF articulates an internal process for projects at the national or European level. Indeed, as pointed out through interviews, the library is very active at the European level (i.e. the working group dedicated to AI at the Conference of European National Librarians—CENL—is led by a French expert from the BnF), even if she still has capacities to do internal development.

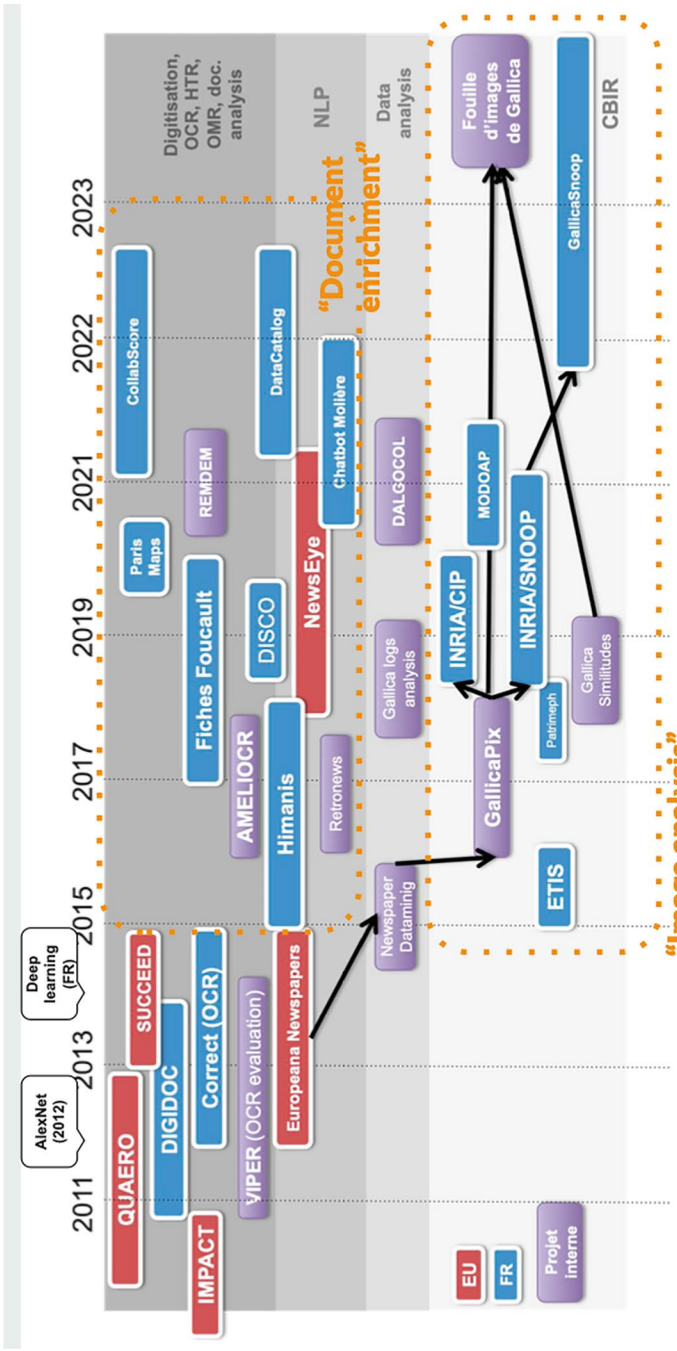

Fig. 1 Timeline of the development of AI in the national library. (Source: Jean-Philippe Moreux [4])

All these projects contribute in the end to Gallica (https://gallica.bnf.fr/) or to the global catalogue of the library https://catalogue.bnf.fr/index.do. Related to the ecosystem of Gallica, additional services from the BnF, like Mandragore [5], are also available online.

An important aspect of the library is ensuring that data is easily accessible. That's why the BnF created an API endpoint [6]. Through this API, developers can access data from Gallica in a machine-readable format. The library also has a data endpoint to request raw data: https://data.bnf.fr/

2.1 Need(s) Behind the Implementation

Gallica was established in 1997 with the aim of making materials free from intellectual property rights accessible to everyone. Currently, over ten million documents are available online via Gallica. Following the introduction of Google Books in 2005, the director of the French National Library published an article titled "Quand Google défie l'Europe," highlighting the risks associated with this project. As a result, Europeana [7] was launched, featuring content primarily from Gallica.

To enable people to access BnF's collections directly online, Gallica is engaged in digitisation efforts. With over ten million documents and 60,000 daily users, it is among the largest digital libraries worldwide. Artificial intelligence has been utilised through technologies like Optical Character Recognition and Handwritten Text Recognition to enhance document readability and streamline search processes. With such a vast collection of documents, how can the library assist users in finding what they are looking for? This is where various AI projects, such as image mining, come into play to help users discover relevant information. The head of Gallica was emphasising that: "We need functions that can be used by the general public, everywhere. These aren't very specific things, such as artificial intelligence capable of reading stock market codes or whatever. What we needed was a relatively 4×4 technology, so to speak, that could be used throughout Gallica."

Among the significant milestones were European projects for digitising press in 2008, 2012, and 2015. These initiatives contributed to improving OCR and the digitalisation of BnF's collections.

Another notable milestone is the GallicaPix project, initiated in 2016, which employed a deep-learning approach for image mining. This proof of concept proved successful, prompting the launch of scale-up efforts. These factors contribute to the reflection of AI in libraries and lay the groundwork for coordinating efforts.

The AI roadmap (2021) [8] from the BnF is related to the purpose of the library, and it's contributing to building tools to reach them (image mining, AI cataloguing, content personalisation, HTR, and OCR). The roadmap [9] encompasses four main projects that are interconnected with Gallica:

- **Image mining in Gallica** (querying images in Gallica based on similarity and generated keywords)
- **Handwritten text recognition (HTR) to be integrated into Gallica** (such a technology applies to handwritten texts but also to ancient printed works and texts written in less spoken languages)
- **Cataloguing** (daily cataloguing support, automatic mechanism expansion and improvement, implementation of LRM model, etc.)
- **Personalised content recommendation with an ethical perspective** (that is to say, respectful of diversity and data privacy, etc.)

2.2 Actors Involved

Gallica was started by the Bibliothèque nationale de France. As mentioned, the project is managed by the Cooperation Department with the support of departments like IT or the Collection Department.

Initially, the library conducted research and development. Later, they began collaborating with research teams, such as INRIA or ETIS Lab, on specific projects. The BnF also partners with European projects like Quaero [10], Europeana Newspapers [11], and NewsEye [12].

Two of their partners are the National University Library of Strasbourg (BNUS) and the National Library of History of Arts (INHA). Through Gallica, the BnF has federated a group of 276 partner organisations that use the platform. The idea of this partnership is the following: the BnF gives access to Gallica's backend in order to be used by the other library, and in exchange, the partner library gives access to its documents that could complete BnF's collections. These partners meet one or twice a year to discuss their needs, but the prioritisation and decision-making are ultimately made by the BnF: "Based on these needs, which may be those of partners or users, our own or those of partners, through surveys, feedback from the field, the emails we receive, requests, we draw up a roadmap which we consider to be a priority, which again, the choices are made on the basis of pragmatic factors, i.e. budgets, the ability to implement these new services or the improvement of existing services."

Table 1 describes the co-production phases of the project "Gallica", launched by the Bibliothèque nationale de France (BnF). The co-production types indicate the level of collaboration and involvement from various stakeholders in each phase based on the phases described by Mergel et al. [13]. Co-initiation was driven solely by BnF's willingness to start the project. In contrast, co-design involves concurrent phases with libraries collecting feedback and needs, while co-implementation features simultaneous collaboration with technical/research partners. The co-use/production moment brings together a group of libraries using Gallica as a white-label product, leveraging it in exchange for shared documents among 276 partners. Lastly,

Table 1 Synthesising the work that the national library is doing to collaborate

Phase	Co-production type	Description
Co-initiation		None, it was the willingness of the BnF to launch the project
Co-design	Concurrent co-production phase	The library is collecting feedback/needs and then chooses what to implement and how
Co-implementation	Concurrent co-production phase	With libraries like the BNUS or INHA and technical/research partners like INRIA or Mistral
Co-use/production moment	Concurrent co-production phase	With a group of libraries that are using Gallica as White-label product (30 cases). They can use Gallica in exchange for their documents. They have 276 partners that are using Gallica
Co-evaluation	Retrospective co-production phase	The evaluation is done by the BnF itself

the co-evaluation phase is retrospective, meaning that the evaluation was conducted by BnF itself.

2.2.1 Organisational Level

The project is led by the cooperation department and is under the supervision of the direction of services and networks. Figure 2 shows where the cooperation department (in charge of Gallica) is located in the organigram of the library.

There is specific support at the organisational level, as Gallica is explicitly identified in the AI roadmap. The project has a specific political sponsor, Tiphaine Vacqué, who is in direct relation with Kevin Riffault, the managing director.

2.2.2 External Actors

For the purpose of the project, they are working with partners that are mostly research teams (e.g. INRIA, a research centre in France) or technical partners (e.g. IT companies). Those actors are contributing to the implementation of a specific technology or contribute to answering a specific need. This point is difficult to handle as they are looking for actors who are able to work on technologies that are evolving at the same time. Besides, the library needs to have competencies to manage and maintain up and run the system. Those difficulties put a lot of pressure on the internal team.

We can note that for now, individual patrons and their representatives and associations are not involved in the discussion or the development of those functionalities. Even if a Gallica community did exist on Twitter (we were speaking about Gallicanautes), feedback from direct users is more informal or through investigations made by the audience studies department.

Fig. 2 Organigram of the BnF (June 2024)

2.3 Challenges

Gallica, a significant project, comes with numerous challenges. In an institution like the BnF, where AI is transforming the way people work, internal culture is shifting towards greater technological adoption. An internal expert expresses that "We have defined 4 main priorities, which are not exclusive, but there are 4 priorities. There's the image mining project, which aims to create an iconographic repository based on Gallica that will identify all the illustrations. There's handwriting

recognition to do OCR on everything, not just printed matter. Help with indexing and cataloguing is very wide-ranging. It can help with current cataloguing. Every year, 70,000 or 80,000 books are published. It can also enrich existing cataloguing, the retrospective of the general catalogue. Indexing can also mean, for example, better indexing of the web archive collection, and ensuring that it's not just 5 petaoctet of massive data sets. That was the third. The fourth is conservation aid. These are more optimisation and decision-making technologies. How to optimise storage and shop organisation? A new site is going to be built in Amiens to house the press collections. How do we take all these collections here and optimise their migration to Amiens depending on the physical space available and the demand for consultation here? This is clearly a problem of optimisation. For us, it's a bit of a novelty."

Yet, we need to acknowledge that, even if it's a priority for the library, some employees remain hesitant to digitise everything, fearing it may deter visitors from visiting the library.

The biggest challenge currently is scaling up experimentation to industrial product levels. Although experiments have shown promising results, implementing them at a production level requires substantial investment and remains difficult. Indeed, one IT expert stresses that "What already exists, 20 years–30 years of production, it's harder to make information systems that are running at full speed evolve than it is to create. It's paradoxical, but it's easier to create something from scratch. The very destructive side of deep learning, which evolves very quickly, and which has an iterative life cycle, is both an advantage and a disadvantage, because you create models from data and you can spend your time improving the models, so it's a fairly new approach for a central IT department." In the meantime, the library must ensure high-quality services, necessitating verifying that new functionalities work well and integrating them smoothly into its current IT environment.

Another pressing issue is evaluation. AI projects are costly, but assessing their impacts is currently challenging. Scaling up experimentation also raises questions about evaluation. With limited funding, the library must decide which features to develop for various user groups (citizens, companies, other libraries) and tailor its approach to meet their specific needs and practices.

3 Results

3.1 Organisational Level

With 27 years of experience thanks to Gallica, the BnF has gained significant expertise in handling projects related to artificial intelligence (AI). Initially, these projects were not labelled as AI-related, but with the emergence of deep learning and breakthroughs, the BnF established an AI team to support this evolution. Today, this

phenomenon is also part of the library's roadmap, supported by a dedicated sponsorship at the political level.

The AI team coordinates initiatives on the field and shares them to maximise their impact.

3.2 Value Created and Co-created

Gallica is a showcase for the BnF and its expertise in the AI field. By implementing this kind of technology, the library is creating public value at various levels:

- To other libraries by sharing with them the outcomes of the project (like in the context of the CENL or through associations like Liber or IFLA). Also, directly when libraries (like the BNUS) [14] are using Gallica for their own collection.
- To patrons through the implementation of this technology in their services. This could help users of Gallica or the general catalogue to access more precisely what they were looking for. This could also be a way to ease this process of finding, as their systems require important information literacy (e.g. to know how to do some research).
- To the private sector, for those who are using resources from the BnF for their activities and currently tech companies that are looking for qualitative structured data to improve their models.

AI presents various opportunities for creating value at different levels within the BnF:

- Service level: AI can contribute to achieving the library's mission by facilitating users' access to documents. For instance, OCR and HTR technologies enable digitisation of materials, image mining enhances search capabilities, and AI cataloguing tools streamline document indexation.
- Internal level: Librarians can benefit from AI tools that assist with various internal tasks, such as identifying documents requiring restoration and improving content classification.
- Process level: AI applications can also improve library processes, both internally and externally. By optimising workflows, staff can focus on higher-value tasks, ultimately leading to better services for users.

3.3 Lesson Learned

At present, there are few feedback/lessons learned as the projects are still in their experimental stages with little evaluation of performance metrics. The library has not yet established methods to assess the efficacy of its prototypes. Interviewees

highlighted the importance of gathering people from various services to be part of AI projects to make sure they will contribute to it and that they understand how AI can be useful for them (and not a threat). To achieve that, they suggest working on the improvement of AI literacy within the library.

However, AI presents various opportunities for creating value at different levels within the BnF:

- Service level: AI can contribute to achieving the library's mission by facilitating users' access to documents. For instance, OCR and HTR technologies enable digitisation of materials, image mining enhances search capabilities, and AI cataloguing tools streamline document indexation.
- Internal level: Librarians can benefit from AI tools that assist with various internal tasks, such as identifying documents requiring restoration and improving content classification.
- Process level: AI applications can also improve library processes, both internally and externally. By optimising workflows, staff can focus on higher-value tasks, ultimately leading to better services for users.

4 Conclusion

To conclude this case about Gallica, we could enlighten that Gallica "plays a part in this because everything in Gallica is in one big well of saved data for eternity. That's what we're trying to do. Once we've archived and preserved all that, it's distributed. That's where Gallica comes into its own, providing access to the collections for as many people as possible." This citation of one interviewee makes the correlation between the fundamental missions of the BnF and Gallica. As they aim to grow to 20 million documents by 4–5 years, otherwise as the head of Gallica states: "If you don't, this content won't be visible and will ultimately be useless." In this perspective, the need for technical support to ensure accessibility of collections is more prominent than ever.

As they are going deeper on their AI-driven digital transformation journey, they will face both challenges and opportunities. On the one hand, the adoption of AI requires significant investments in infrastructure, training, and personnel, which may be a barrier for some libraries, particularly those with limited resources. Additionally, there are concerns about data privacy, bias, and job displacement that must be carefully addressed. However, on the other hand, the potential benefits of AI far outweigh these challenges. By leveraging AI, National Libraries can unlock new levels of efficiency, innovation, and user engagement, ultimately enhancing their role in supporting lifelong learning, cultural development, and social progress.

Acknowledgements The work for this chapter was supported by the European Union's Horizon Europe Research and Innovation Programme under Grant Agreement ID 101061516.

Appendix

See Table 2.

Table 2 Overview of the Gallica project at the French National Library

Case and project name			
Gallica—French National Library			
Country	**Number of employees**	**Type of AI solution**	**Year and maturity level**
France	2128	OCR, HTR and image/data mining	1997, well underway/major milestones accomplished
Project description			
Gallica is the digital library of the BNF. This digital library gives access to more than ten million free documents			
Need(s) behind implementation	**Actors involved**	**Challenges**	
Information retrieval, cataloguing, ease the research inside the collections	Mostly internal (IT, Collection, and Cooperation departments), external partners like BNUS, INHA, INRIA, or technical partners (i.e. Mistral)	Moving from prototyping to implement and launch live	
Results			
Organisational level	**Value created and co-created**	**Lesson learned**	
A specific department is in charge of Gallica; this is the coordination department. The project has a sponsor at the political level and is an important project for the BnF related to AI	New services for users, new services for librarians and for partners (like other libraries), and change at the process level to ensure a smoother experience as well for users and for employees	A specific department is in charge of Gallica; this is the coordination department. The project has a sponsor at the political level and is an important project for the BnF related to AI	

References

1. https://gallica.bnf.fr/accueil/fr/content/accueil-fr?mode=desktop
2. Additional publications are being shared by the library in open access: https://bnf.hal.science/BNF_DSG/
3. To follow the activity of the BnF, you could also follow the work of AI4LAM a working group about the use of AI in Library, Archive and Museum: https://sites.google.com/view/ai4lam/working-groups-and-chapters
4. To go deeper inside the work done at the BnF, Jean-Philippe Moreux, an AI expert at the library, is publishing papers about experiments that have been done: https://www.researchgate.net/profile/Jean-Philippe-Moreux/research
5. Mandragore is a digital platform developed by the Bibliothèque nationale de France. The database is giving access to 200,000 illuminations, drawings, seals and bookbinding decorations already described, with structured information and features like image mining
6. An endpoint is a connection where data sources send their data in a structure way in order to be process: https://api.bnf.fr/
7. https://www.europeana.eu/fr
8. https://www.bnf.fr/sites/default/files/2022-01/Poster_AI%20Roadmap_BnF_202112.pdf
9. https://www.bnf.fr/en/artificial-intelligence-bnf
10. The project could be found here: https://actions-recherche.bnf.fr/BnF/anirw3.nsf/IX01/A2013000354_quaero
11. The project could be found here: https://www.europeana.eu/en/collections/topic/18-newspaper
12. The project could be found here: https://cordis.europa.eu/project/id/770299
13. I. Mergel, N. Edelmann, N. Haug, Co-production phases in the development and implementation of digital public services. Perspect. Public Manag. Govern. **8**(2), 1–13 (2025)
14. https://www.bnu.fr/fr/numistral

Antoine Henry is an assistant professor in information and communication sciences and a member of GERIICO's Axis 4, whose theme is the circulation of information and the organisation of knowledge. His research work focuses mainly on digital issues (AI, digital commons, ethics of algorithms), the transformation of organisations, and collective intelligence.

He is also a member of the board of directors of the French chapter of the scholarly association ISKO and a member of the research group of the Centre Internet et Société (UPR 2000 of the CNRS) in which he co-manages the working group AI, art, and creativity.