Preserving websites in contemporary art : stakes and difficulties in collecting the Web at the Bibliothèque nationale de France

Introduction : history of the crawl of the Web and legal context :

From 2001, the Bibliothèque nationale de France (BnF) has been experimentating collects of the French Web, thus anticipating the extension of its mission concerning the legal deposit to Internet.
The first experimentations were almost "amateur" if one may qualify such manipulations of the Internet. We just contacted the authors of some chosen websites and asked them for a copy, available on tape at the time, on a digital media or by mail, at their convenience.

Those primary tests determined the creation of a service specifically in charge of collecting the French Internet, depending from the Legal deposit department, which concerns every type of media. This department includes today 5 librarians supported by 4 engineers from the IT department.

We proceed according to two methods :
- The first is called "broad crawl", covering all the websites included into the .fr domain.
- The other, called "focused crawl", is an intellectual selection made by specialists of all the encyclopaedic domains of the BnF collections.
All types of crawls are performed by the Heritrix robot.

The "focused crawls" aim at spotting interesting websites to be protected at all costs and that are not covered by the broad crawls. This location of sites integrates ephemeral data (events such as a festival, an evolution or a closure of a website).
Since August 1th, 2006, the DADVSI law (copyright and similar rights in the information society) makes an exception concerning copyright by allowing the on-line automatic harvest of the French Web. The BnF is appointed to this

mission of preservation. The implementation decree for the law, published quite recently last December, opens new perspectives by allowing the BnF to obtain passwords to the websites requiring registration or to subscription-based sites. Furthermore we can ask now for data obstructing the passage of the robot.

The collection today represents 250 terabytes. It was completed in 2006 by the purchase from Internet Archive of a retrospective archive from 1996 till 2003 for websites in the .fr domain.

1. A documentary policy for the Web :

The Literature and Art department at the BnF, which was one of the first department to work on this project, established ever since 2006 a documentary policy. Widely inspired by the novelty of the Internet media, guided by the existing sites, it tried to define documentary opportunities in art offered by this new media, while trying to think about the limits of such a large domain. This policy was very harsh towards the quality of the collects all along the experiments, as far as the websites in art, particularly contemporary art, showed themselves often difficult to collect.

The documentary goals :

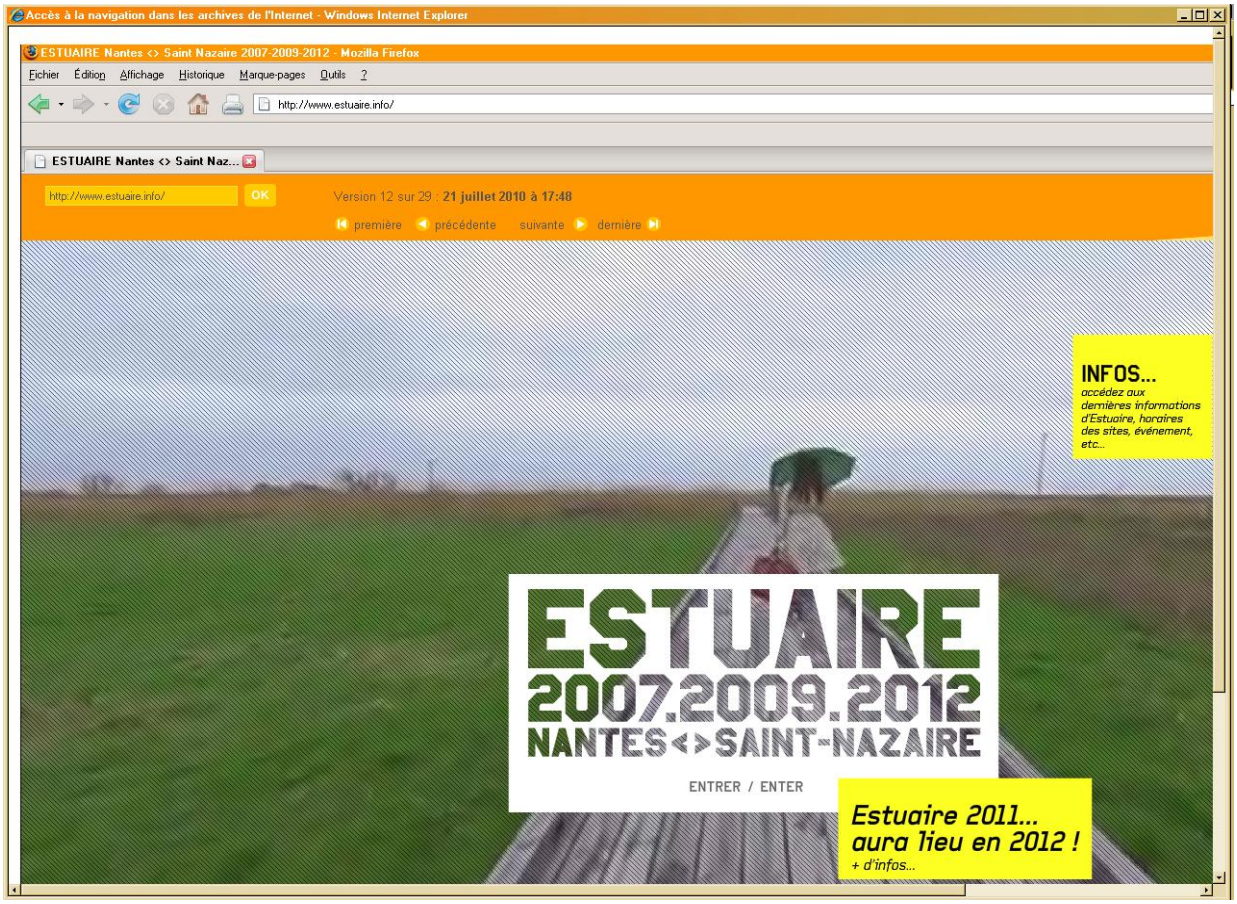
- First, we want to collect an exceptional documentary wealth for art.

Thanks to the Internet we obtain an unequalled easiness of consultation. It is possible to inventory very numerous visual objects and to update them frequently, and this for a very low cost.
Works not listed nor published on paper are at last visible. Many websites of art galleries supply biographic information on their artists.
The Web is a space of freedom for emergent or dissident artists and for forms of art on the fringe of society. It is moreover a flexible and controllable tool for spreading information.

- An other goal is to archive art current events and its short-lived demonstrations: exhibitions, festivals, on-line works, ephemeral works, websites of temporary exhibitions, programs of museums, artistic events. And also information websites, magazines, networks, blogs, any media allowing a better perception of contemporary world.

- Then we try to cover all domains of art and art history from graffiti to restoration.

- Finally we share the harvest of contemporary creation with other departments of the BnF which traditionally acquire works by media : the Audiovisual department, in charge of the multimedia and on-line creation, the Prints and Photographs department, and the Books'history department in charge of artists' books and graphic art.



http://www.estuaire.info (Archives de l'Internet, 21 juillet 2010)

Three main documentary directions :

- Contemporary creation and above all artists. Numerous artists have their personal website, but it is quite recent for those who do not use the digital media. Daniel Buren for example has already reserved a domain name of website in 2001 but opened it only in 2004. Artists often group together into networks, for aesthetic, political, technical or simply logistic motives. Net-artists, using the Web or who create interactive works, group together into research websites, collectives of artists or on-line reviews. Marginal artists like

graffiti artists or performers also find with Internet a means of distribution away from the official circuits.

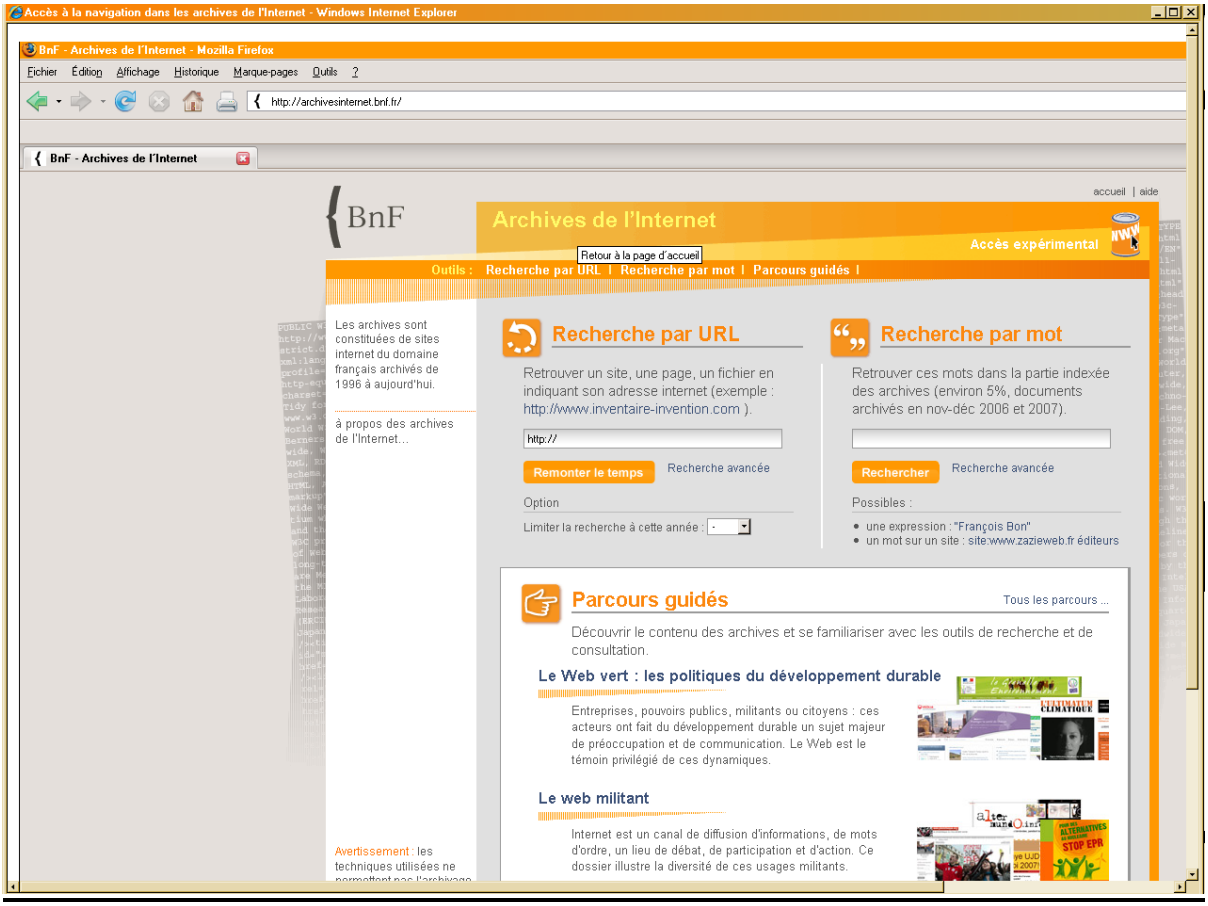
Also concerned is the contemporary artistic world : websites of galleries, precious sources of biographic information on artists, networks of historians and researchers, lists of discussion and websites concerning the training of artists such as art schools'programs.

- Current events are the second main direction : we collect websites of temporary exhibitions, virtual web exhibitions, rare but always interesting because they are on a unique media, websites of museums, events, biennials, conferences, virtual visits. Websites of information, reviews, TV, information networks, blogs, any media allowing to make creative activity perceptible.

- Finally we are interested in documentary resources on art, they deal with editorial production, websites concerning heritage, inventories, auction houses. Bibliographic databases are not harvested, as well as picture databases, since their access is through a form which is not adapted to the automatic harvest we use.

2. The user interface :

Available for consultation only in the heritage library after accreditation, in order to protect the copyright, the " Archives de l'Internet " proposes two means of access : a simple search by URL as well as thematic selections.
Art is present through a selection of blogs : some of these blogs are not on-line any more, or can now only be accessed for their most recent archives.



(Archives de l'Internet : front page)

3. The specific limits in collecting art websites :

"Focused crawls" are made with a robot which copies every site in depth. We supply it with the precise address and with frequency parameters (monthly, annual, 1 or 2 times a week). It also collects the links selected by the site, thus allowing us to capture its cultural environment.
The quality of these collections is then checked by the librarians who had established the lists. The captures are excellent when websites are constituted of texts, pictures and when their architecture is similar to that of a static page like blogs.

However several other difficulties appear as soon as websites contain:

- drop-down menus
- animations with navigation choices, since a robot cannot recognize an object to click on which evolves on the screen. Its action is very simple, it follows the

links that it meets on every page, but certainly not plastic metamorphoses or moving objects.
- panoramas : these animations are generally made with Flash technology which the robot does not integrate, the problem is well known and identified.
- animation that is too long at the opening of a site : it can prevent the robot from pursuing the harvest
- specific programs exist made by the authors of websites themselves which prevent the visit of robots and the capture of their contents.
- videos are harvested in a uneven way, it depends on the time spent on the website by the robot.

In all these cases, we obtain the indication of a missing image, and in the presence of animations, the site sometimes does not open at all, leaving the Internet user in front of a blank page.

Paradoxically pictures can be collected by other means, because referenced by other websites, in that case the link on the original site cannot be followed while the picture is well recorded in the Archives, but not linked.

Another problem is the increasing use of Flash technology, some websites which were collected well in their beginnings since they were created with very simple softwares, now modernize their presentation with Flash, and today, we are no longer able to harvest them.

Conclusion :

If in the field of art, the results of the crawls were sometimes particularly poor at their beginnings, the evolution of the quality is obvious, thanks to the gained experiences, to the patient identification of the technical difficulties, to the modification of the settings and to the refinement of harvesting strategies.
Even if the textual elements of information collected are important (legends, explanations about the works of artists, movements, galleries, institutions, etc.) it is clear that we cannot be satisfied if we don’t capture the images.

We still thus have to improve, but art websites due to their formal wealth, constitute a corpus of experiment which is situated at the forefront of our research. If a part of the Web has already disappeared, website creators, aware of this volatility, rather quickly archived the successive states of their sites.

But in ten years the crawls have largely improved, the performances of the robot have been constantly refined, so that we can hope very soon to offer the public images which will be the faithful reflection of contemporary artistic activity.

Sites shown :

http://www.documentsdartistes.org
http://www.space-invaders.com
http://www.danielburen.com
http://www.magasin-cnac.org
http://taniuchi.fr
http://www.missticinparis.com
http://www.domaine-chaumont.fr
http://www.centrepompidou-metz.fr
http://lunettesrouges.blog.lemonde.fr
http://annemalherbe.blogspot.fr

Françoise Jacquet, Librarian, Art department, Literature and art department of the Bibliothèque nationale de France. In charge of the collection of the French websites in art.