Quelle documentation pour l'archivage du Web?

Nicolas Delaforge, Bruno Bachimont

▶ **To cite this version:**

Nicolas Delaforge, Bruno Bachimont. Quelle documentation pour l'archivage du Web?. 18es Journées Francophones d'Ingénierie des Connaissances, Jul 2007, Grenoble, France. not specified. hal-00509858

HAL Id: hal-00509858
https://hal.science/hal-00509858

Submitted on 16 Aug 2010

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Quelle documentation pour l'archivage du Web ?

Nicolas Delaforge[1,2], Bruno Bachimont[1,2]

[1] Université de Technologie de Compiègne - Heudiasyc (UMR CNRS 6599),
{nicolas.delaforge, bruno.bachimont}@utc.fr

[2] Institut National de l'Audiovisuel - DRE,
{ndelaforge, bbachimont}@ina.fr

1 Quelle unité documentaire de référence ?

Pour les institutions dépositaires françaises récemment chargées par l'état d'étendre leurs collections aux publications en ligne, l'arrivée du support web a considérablement bouleversé les conceptions traditionnelles de la documentation et de la consultation. Ce support s'est en effet révélé particulièrement résistant à la modélisation documentaire notamment en ce qui concerne l'établissement d'une unité de base préalable à tout travail de description.

Force est de constater qu'il existe encore aujourd'hui, une différence remarquable entre la définition du document telle qu'elle est présupposée dans la recommandation HTML et celle qui est appliquée par les internautes. En effet, HTML pose implicitement la page comme unité référence. Toute utilisation du terme *document* dans la norme[1] fait référence à la page et la notion de site en est tout simplement absente.

Néanmoins, les pratiques d'écriture telles qu'elles sont relayées dans les guides de conception de sites montrent qu'il est préférable d'organiser l'information en complexe documentaire et non en document monolythique. Du point de vue documentaire, la page ne représente donc pas le document à part entière, mais plutôt comme un fragment d'une communication plus vaste : le site.

Malheureusement, cette notion de site, couramment utillisée par les internautes, ne possède aucun ancrage technique fiable. Seuls l'isotopie des textes, la cohérence de la charte graphique, la pertinence des formes sémiotiques mises en co-présence au moment de la consultation indiquent au lecteur qu'il se trouve toujours dans l'espace sémantique prédéfini par l'auteur.

Sur le web plus que partout ailleurs, le support est dissout et l'attribution d'intentionnalité documentaire à un contenu web est une question d'interprétation. Certains récits interactifs en ligne[2] exploitent d'ailleurs cette faiblesse dans la définition des frontières documentaires pour créer un effet de désorientation chez le lecteur.

[1] http ://www.w3.org/TR/html401/

[2] voir par exemple le "NON-roman" de Lucie de Boutigny

Cette opposition évidente entre la norme du W3C et la pratique des internautes a pour conséquence l'obsoléscence d'un principe très pratique pour les archivistes qui permettait d'assimiler unité sémantique et unité de manipulation. De manière schématique, nous pouvons dire que le travail d'archiviste consistait à préserver les supports pour permettre ensuite au documentaliste d'en décrire les contenus. Nous voyons qu'avec le web cette organisation des tâches n'est plus possible et que pour celui qui souhaite préserver le document dans son ensemble, un travail d'interprétation doit précéder l'archivage.

2 Appréhender la masse d'information

Il est utopique de penser avoir une archive de qualité avec une documentation complète et des modes d'accès performants tout en étant exhaustif. Les archives web sont dans l'obligation d'adopter la même philosophie que le web lui-même : *le best-effort* et de trouver un équilibre entre exhaustivité et qualité de documentation en fonction de leurs moyens.

Deux modèles principaux semblent se dégager à travers les différents projets d'archivage menés dans le monde : les corpus généralistes et les corpus thématiques.

Les archives à corpus généraliste n'ont pas à "interpréter" les contenus, au mieux ils détectent automatiquement la langue des pages, pour décider de les intégrer dans leur fonds documentaire. Néanmoins, la masse d'information à traiter oriente souvent les choix techniques vers une forte automatisation des différentes étapes de l'archivage. Il est techniquement beaucoup plus simple de manipuler des pages que des sites et les architectures automatiques de ces projets ressemblent fort à celle des moteurs de recherche.

Pour les corpus thématiques, les besoins sont assez différents. Les utilisateurs des archives thématiques attendront du fait de la "spécialisation" de l'archive des moyens de recherche performants. Ce modèle aura donc tendance à s'orienter vers une documentation plus poussée. De plus l'aspect thématique nécessite une interprétation du contenu préalable à la sélection et implique donc une appréhension globale du site en tant que document.

3 Expérimentation

L'architecture adoptée par l'INA est en accord avec le domaine média dont il a reçu la charge et repose donc sur une sélection manuelle des sites à intégrer au fonds documentaire et une captation incrémentale automatisée. Plusieurs campagnes expérimentales d'archivage ont été menées. La première était une réponse à une demande interne, à savoir l'archivage des sites de chaînes du cable et du satellite. Ce premier test sur un corpus restreint mais techniquement complexe nous a néanmoins permis de mettre au jour les difficultés engendrées en documentation par un média très fluctuant. Actuellement, un deuxième corpus sur le web électoral est en cours de constitution et devrait nous permettre d'expérimenter d'autres approches documentaires plus proches des spécificités inhérentes au web.