Rapport sur l'usage du Captcha patrimonial CapthAN

Projet SNI 2019

Auteur : Christopher Kermorvant (kermorvant@teklia.com)
Date : Août 2023
Version : 1.0

Introduction

Avec l'augmentation continue de l'usage du web, les sites internet sont devenus des cibles privilégiées pour de multiples activités malveillantes, allant du spam aux attaques automatisées. Dans ce contexte, garantir la sécurité et l'intégrité des informations en ligne, tout en préservant l'expérience utilisateur, est devenu une préoccupation majeure pour les administrateurs de sites web. C'est dans ce cadre que les CAPTCHA (Completely Automated Public Turing test to tell Computers and Humans Apart) ont été introduits.

Les CAPTCHA sont des tests conçus pour distinguer les utilisateurs humains des robots lors d'une interaction avec un site web. Typiquement, un CAPTCHA présente à l'utilisateur un défi sous forme d'image ou de texte, qui requiert des capacités cognitives que seuls les humains sont censés pouvoir mobiliser pour les résoudre. Ces défis peuvent aller de la reconnaissance d'images ou de séquences de caractères déformées à l'identification d'objets dans des photos.

L'usage des CAPTCHA est devenu courant sur de nombreux sites web pour différentes raisons. Premièrement, ils empêchent les robots d'envoyer des formulaires à répétition, protégeant ainsi les sites contre le spam ou les tentatives d'inscription automatisées. Deuxièmement, ils servent de barrière de sécurité contre certaines attaques informatiques, telles que les attaques par force brute. Enfin, dans des contextes où la préservation et la sécurisation des informations sont primordiales, les CAPTCHA peuvent garantir que l'accès et les interactions avec le contenu sont réalisés par des humains.

Dans ce rapport, nous étudierons l'utilisation d'un service de captcha patrimonial développé dans le cadre du programme Services numériques innovants (SNI) du ministère de la Culture. Ce projet, appelé CaptchAN, a été lauréat de l'appel à projet SNI en 2019.

Description du Projet CaptchAN

Le projet CaptchAN, né lors du hackathon des Archives nationales en 2018 et récompensé alors pour son caractère novateur, vise à transformer le monde de la sécurité des formulaires web en intégrant des éléments patrimoniaux au mécanisme traditionnel des captchas.

Nature du Service

CaptchAN est une interface de sécurité destinée aux formulaires web, couramment désignée sous le terme captcha. Contrairement aux captchas traditionnels, CaptchAN exploite des données patrimoniales, offrant ainsi une expérience nouvelle aux utilisateurs tout en garantissant que le formulaire est bien complété par un humain et non par un robot.

Objectifs Principaux

Le service CaptchAN vise plusieurs objectifs :

1. *Sécurité Informatique*: les données patrimoniales, comme les documents manuscrits ou des œuvres d'art numérisées, offrent une opportunité pour exploiter la faiblesse des algorithmes d'intelligence artificielle sur ce type de données. En effet, ces données n'ayant pas été utilisées pour entraîner les modèles d'IA actuels, elles résistent à une analyse automatisée. Cette difficulté offre une opportunité pour renforcer la sécurité par l'utilisation de données patrimoniales dans les captchas.
2. *Promotion du Patrimoine Culturel*: en utilisant des contenus patrimoniaux numérisés, comme ceux de la BNF, des Archives Nationales ou du musée de Bretagne, le projet vise à mettre en avant la richesse culturelle française face à la dominance des contenus commerciaux habituellement utilisés dans les captchas. Cela renforce la présence et la valorisation de la culture et du patrimoine dans le paysage numérique.
3. *Exploitation Collaborative*: CaptchAN cherche à bénéficier des contributions des utilisateurs pour aider à valider et vérifier la qualité de la numérisation de corpus patrimoniaux. En sollicitant une participation des internautes, le système aide les institutions culturelles partenaires à assurer la qualité des métadonnées associées à leurs corpus.

Partenariats et Corpus Exploités

Les corpus susceptibles d'être utilisés par le service CaptchAN sont :

- Des ouvrages patrimoniaux illustrés numérisés
- Des corpus manuscrits issus de projets de reconnaissance automatique ou de crowdsourcing
- Des collections d'images ou photographies, soit indexées automatiquement ou de manière collaborative, soit non indexées

Lors du projet CaptchAN, une collection d'ouvrages illustrés de zoologie de la BNF et une base de noms manuscrits issues de la base de donnée Leonore des Archives Nationales ont été utilisées.

Structure d'un projet CaptchAN

Un projet CaptchAN fait intervenir 3 partenaires comme présenté sur le schéma ci-dessous :

1. *Un site partenaire* qui implémente le service de captcha sur certaines de ses pages, généralement les pages nécessitant l'accès aux données personnelles des utilisateurs comme la connexion à leur compte.

2. *Une institution culturelle* qui fournit le corpus numérisé pour lequel elle souhaite obtenir une indexation
3. *TEKLIA* qui configure, opère et supervise le service de captcha patrimonial

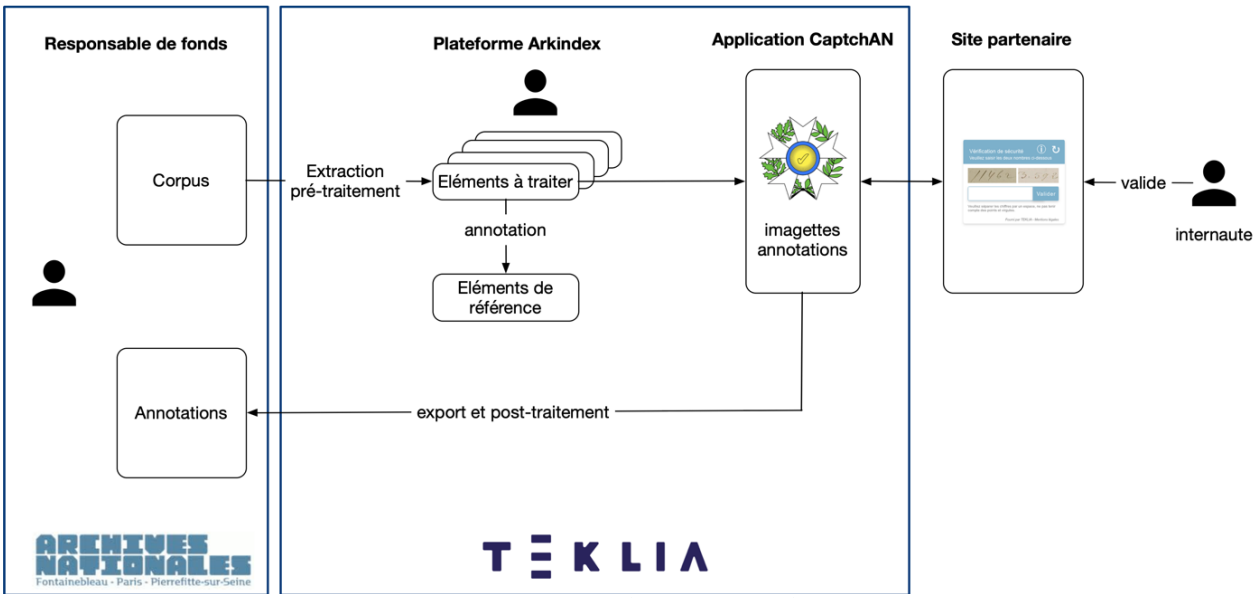

Figure 1 : Organisation d'un projet CaptchAN

Cas d'étude : le portail Decalog Pro/Essentiel

Decalog

Depuis trois décennies, la société Decalog met au service des institutions patrimoniales et de la connaissance des outils d'informatique documentaire ainsi que des services de conseil, de mise en œuvre des projets, de support et d'hébergement, dans un large périmètre fonctionnel. Decalog équipe de grands réseaux de musées comme les musées nationaux, des réseaux départementaux ou municipaux tout comme de petits musées publics ou privés ainsi que de nombreuses bibliothèques.

Intégration de CapthAN

Pour mieux sécuriser les formulaires accessibles depuis le portail, le système de captcha patrimonial a été intégré sur Decalog PORTAIL Pro/Essentiel.

Ce nouveau système de contrôle a été ajouté aux 3 emplacements suivants :

- Connexion abonné, en cas de 2 tentatives de connexion infructueuse
- Connexion gestionnaire, en cas de 2 tentatives de connexion infructueuse
- Écran de modification abonné

Ainsi, l'utilisateur est invité (dans les 3 situations indiquées précédemment) à reconnaître des images afin de s'assurer qu'il ne s'agit pas d'une tentative d'intrusion par des robots.

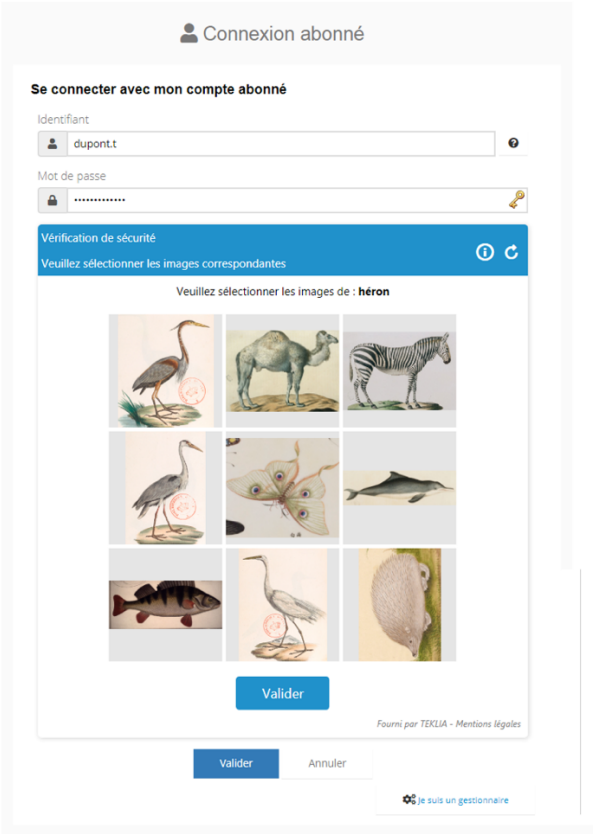

Figure 2 : Intégration de CaptchAN dans le portail Decalog

Les images proposées sur le captcha à l'utilisateur sont issues d'un corpus d'illustrés zoologiques de la BnF.

Lorsque l'utilisateur valide les images (par exemple, sélectionne les images de hérons), son action vient alimenter une base de données d'indexation des illustrations, permettant à la BNF d'améliorer la qualité de ses indexations. Une fois le formulaire validé, l'utilisateur est libre de poursuivre sa saisie.

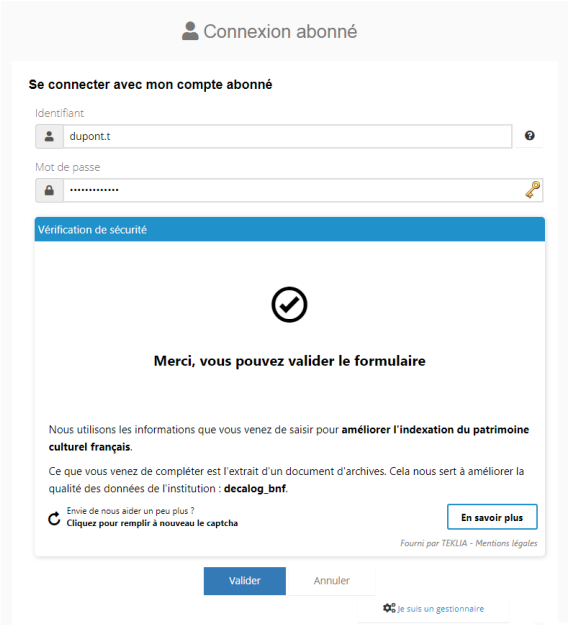

Figure 3 : Validation de la captcha sur le portail Decalog

Données d'usage

Le service CaptchAN a été lancé progressivement en production sur le portail de plus de 500 bibliothèques à partir du 26 janvier 2021. Nous avons choisi d'analyser l'usage du service sur une période d'un an après la période de mise en place progressive du service, entre le 1 avril 2022 et le 31 mars 2023.

Analyse des transactions par mois

La Figure 4 présente les statistiques des transactions sur le service de captcha sur une année entre le 01/04/2022 et le 31/03/2023. Sur cette période on décompte en moyenne 55 192 appels au service de captcha par mois. Les appels se décomposent ensuite en 4 catégories selon le déroulement de la transaction :

1. **Créé** : le captcha a été créé mais aucune réponse n'a été apportée. C'est le cas soit d'un utilisateur qui arrive sur une des pages comportant le captcha et qui navigue sur une autre page sans interaction, soit du moissonnage d'un robot.
2. **Rejeté** : une réponse a été apportée au captcha, par un utilisateur ou par un robot et la réponse n'est pas valide. L'accès à la page suivante est refusé.
3. **Arrêté** : l'utilisateur a rechargé la page du captcha soit car la tâche était trop complexe, soit pour changer de mode (par exemple passer en mode audio si l'utilisateur est déficient visuel).
4. **Validé** : une réponse a été apportée au captcha et a été validée. L'accès à la page suivante est autorisé.

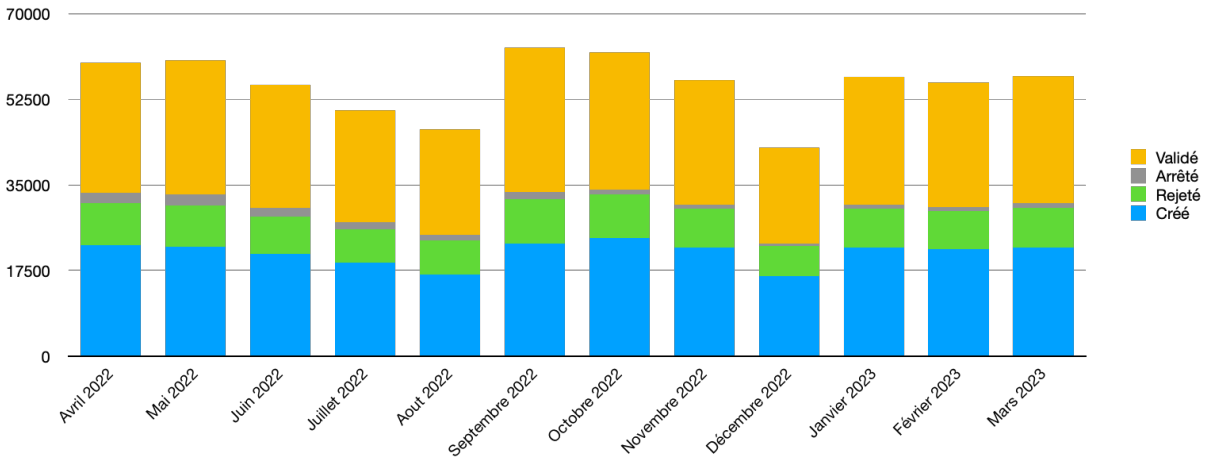

Figure 4 : Statistiques d'usage du service entre le 01/04/2022 et le 31/03/2022.

Le nombre de captcha validés représente en moyenne 76% du total des réponses. En moyenne, l'utilisateur donne la bonne réponse 1 fois sur 4, ce qui indique que le test n'est pas trivial. Une analyse des difficultés sera présentée plus bas. On note que le nombre de transactions arrêtées est très faible, 4% des transactions avec interaction en moyenne, ce qui indique que le captcha n'est pas a priori trop complexe.

En ce qui concerne l'affluence sur le service, on note que les mois de juillet, août et décembre sont les moins chargés, ce qui s'explique par les vacances scolaires. Les mois de septembre et octobre sont les plus chargés, ce qui peut s'expliquer par une connexion plus fréquente à leur espace personnel par les usagers pour réaliser des inscriptions ou des mises à jour de leur profil en début d'année scolaire.

La Figure 5 présente la répartition de l'usage du service sur les jours de la semaine. Comme attendu les dimanches et lundis sont les jours les moins actifs car ils correspondent à des jours de fermeture des bibliothèques et le mercredi est le jour le plus actif car les écoliers, collégiens et lycéens n'ont pas cours toute la journée et fréquentent plus les bibliothèques.

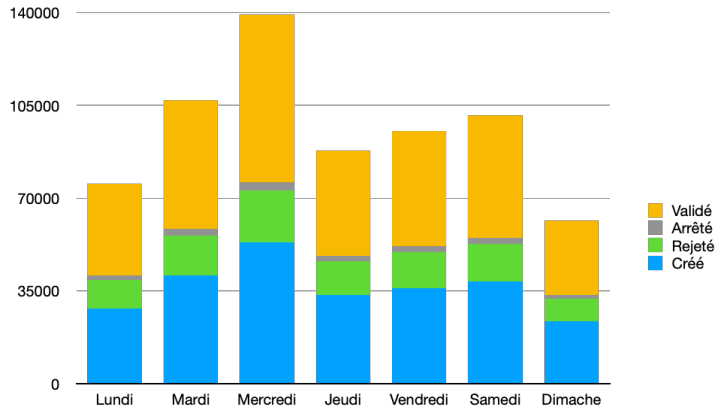

Figure 5 : Répartition de l'usage du service sur les jours de la semaine.

La Figure 6 présente la répartition des appels au service sur les heures de la journée. On note un pic d'activité à l'ouverture des services le matin et un autre pic l'après-midi qui correspond sans doute au pic de fréquentation des établissements. On remarque une plage d'utilisation des services élargie car seules les heures entre 0 et 3 heures du matin ont des activités très faibles.

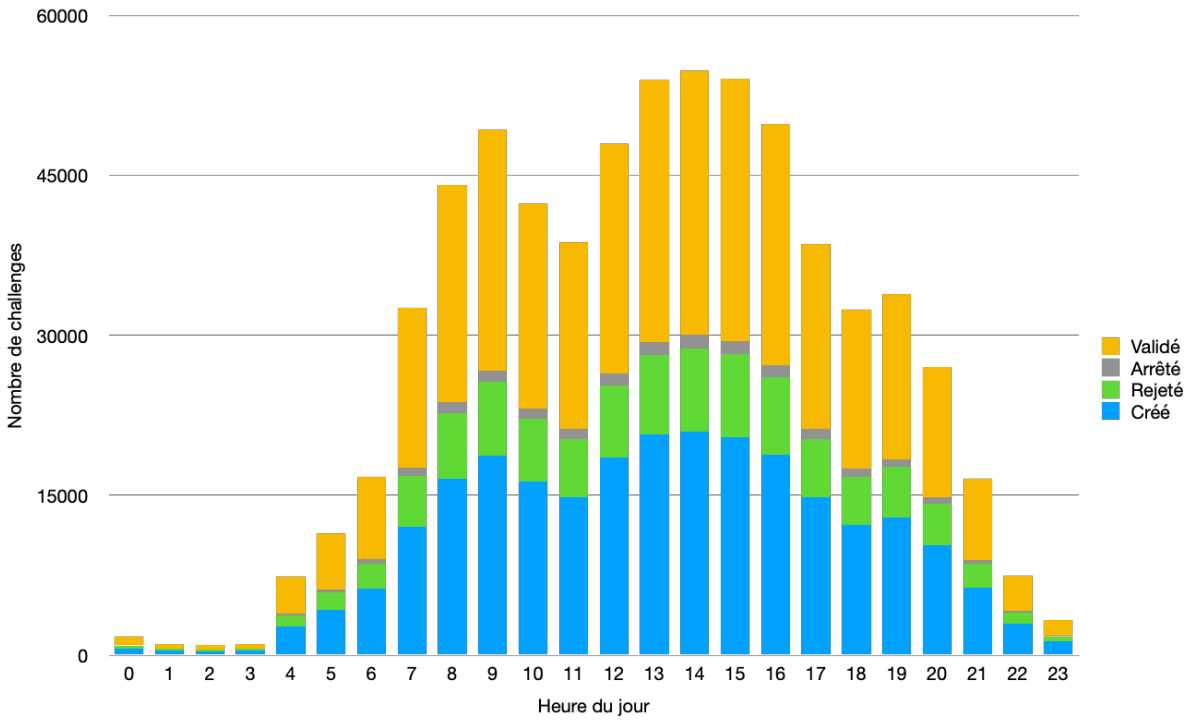

Figure 6 : Répartition de l'usage du service sur les heures de la journée

De l'analyse d'usage, il semble se dégager que le captcha est majoritairement utilisé par des humains, car l'activité du service est proportionnelle à l'activité humaine. Il n'est cependant pas exclu qu'une partie de l'activité soit due à des activités de moissonnage de robot : en effet, le taux de requêtes créées non finalisées est supérieur entre 22 heures et 3 heures du matin, période pendant laquelle l'activité humaine est la plus faible et pendant laquelle l'activité des robots est plus importante en proportion.

Analyse démographique

Une analyse plus précise, en particulier sur la démographie des utilisateurs (âges, CSP, position géographique) n'est pas possible avec les données collectées par le service CaptchAN. En effet, le service ne collecte aucune donnée personnelle sur l'utilisateur, n'utilise aucun cookie et ne stocke aucune information de navigation comme l'adresse IP ou la page de provenance. Si cette politique ne permet pas d'analyser finement l'usage du service, il est totalement respectueux de la législation et en particulier du « *Règlement Général sur la Protection des Données* » (RGPD) européen, alors que la CNIL avait mis en demeure le Ministère des Solidarités et de la Santé pour l'utilisation de la solution reCAPTCHA de Google au sein de l'application « StopCovid ».

Difficultés rencontrées par les utilisateurs

Les difficultés rencontrées par les utilisateurs du service CaptchaAN ne se situent pas au niveau de l'interaction ou de l'interface. Les tests réalisés avec des volontaires lors de la phase de mise au point du service avaient déjà permis d'améliorer l'ergonomie de l'interface.
La difficulté la plus souvent mentionnée par les utilisateurs était la complexité de la tâche d'identification de certains animaux. En effet, le corpus étant constitués d'ouvrages de zoologie, le vocabulaire présenté est celui de la zoologie. Même si nous avions opéré une sélection dans le vocabulaire de base pour supprimer les termes trop techniques, les animaux trop rares ou spécifiques, il reste que la connaissance des animaux est très variable dans un public aussi large que les usagers des bibliothèques. Les utilisateurs ont par exemple signalé qu'ils ne savaient pas ce qu'était un bivalve ou qu'il était difficile de faire la différence entre un fennec et un renard. Une solution pour s'adresser à un public plus large serait de mettre en place des catégories plus communes (chien, chat, cheval) ou plus larges (oiseau, poisson, reptile) mais la précision de l'indexation serait plus faible et l'intérêt pour l'institution patrimoniale se réduirait. Il est cependant possible de mélanger des tâches complexes et des tâches plus simples et de définir une stratégie qui présente d'abord une tâche complexe à l'usager et ensuite une tache plus simple s'il n'a pas réussi la première tâche ou s'il l'a passée.

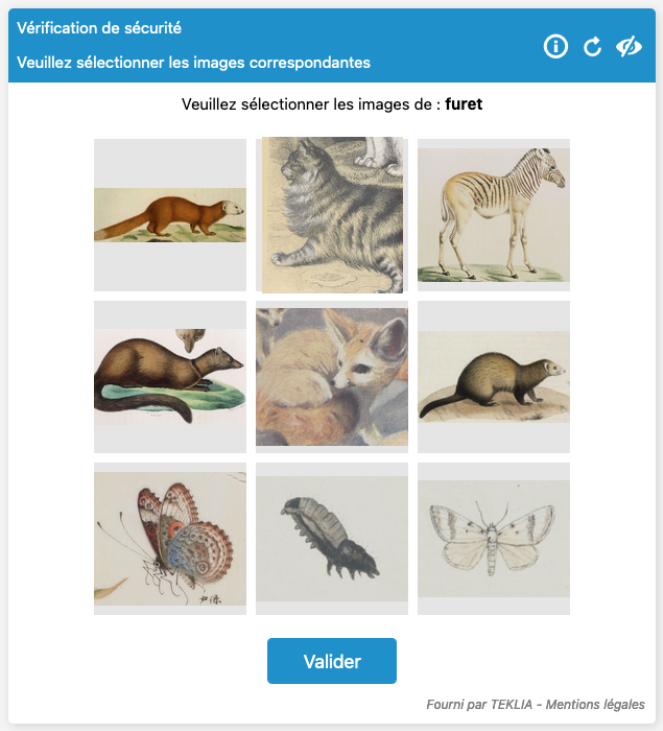

Figure 7 : Exemple de captcha "complexe" : identifier un furet.

Perspectives d'usage

Le projet CaptchAN a toujours suscité un intérêt immédiat chez les personnes concernées par la mise en place de captcha. Les institutions ou sociétés suivantes ont manifesté un intérêt pour le service et envisagent de le déployer ou ont testé un déploiement :

- La Bibliothèque Nationale de France pour son site https://arpenteur.bnf.fr/#/

- Les Archives Nationales pour leur site https://www.siv.archives-nationales.culture.gouv.fr/
- Le ministère de la culture pour les sites https://delia.culture.gouv.fr/ , https://histoiredesarts.culture.gouv.fr et https://archives-nationales-travail.culture.gouv.fr
- Sopra Steria pour remplacer la solution reCaptcha de google
- La direction numérique de la région Grand Est
- La société AG2R La Mondiale
- Wikipedia

Un des facteurs limitant le déploiement de la solution actuelle sur des sites institutionnels ou pour des entreprises est la demande d'une certification de sécurité de type ANSSI. Or ces certifications sont facturées plusieurs dizaines de milliers d'euros et ne sont valables que pour une version du logiciel. Le coût d'une telle certification est malheureusement incompatible avec un service quasi-gratuit tel que CaptchAN.