Smals
ICT for society

E-mail Address Reliability

Maintenir la qualité des adresses e-mail

Vandy Berten
Isabelle Boydens

Décembre 2013

Deliverable 2013/TRIM4/01
@SmalsResearch
http://www.smalsresearch.be

www.smals.be

Management Summary (Français)

Management Summary (Nederlands)

Table des matières

1. Introduction

egovernment

benchmarks

egovernment

benchmarks

bounce mails

A

A

A

5

⚜ ⚜ A ⚜

2. Aspects syntaxiques et techniques

2.1. Décomposition d’une adresse

2.2. Vérification syntaxique

2.2.1. Syntaxe du nom de domaine

2.2.2. Syntaxe du nom d'utilisateur

undescore)

2.2.3. Longueur maximale d’une adresse e-mail

-
-
-
-
-
-

http://tools.ietf.org/html/rfc821

http://tools.ietf.org/html/rfc5321

http://tools.ietf.org/html/rfc3696

http://www.rfc-editor.org/errata_search.php?rfc=3696&eid=1003

http://www.rfc-editor.org/errata_search.php?rfc=3696&eid=1690

-

abcdefghijklmnopqrstuvwxyzabcdefghijklmnopqrstuvwxyzabcdefghijkl@abcdefghijklmnopqrstuvwxyz.abcdefghijklmnopqrstuvwxyz.abcdefghijklmnopqrstuvwxyz.abcdefghijklmnopqrstuvwxyz.abcdefghijklmnopqrstuvwx.be

-
-
-

2.3. Validation du « Top Level Domain »

14 15

சிங்கப்பூர்

中國

الـ جزائـ ر

2.4. Validation du nom de domaine

17

18

2.5. Validation d'une adresse

2.5.1. Serveur MX et protocole SMTP

```
C:\>nslookup -q=mx gmail.com
[...]
Non-authoritative answer:
gmail.com mail exchanger = 5 gmail-smtp-in.l.google.com.
gmail.com mail exchanger = 10 alt1.gmail-smtp-in.l.google.com.
gmail.com mail exchanger = 20 alt2.gmail-smtp-in.l.google.com.
[...]
```

```
C:\>telnet gmail-smtp-in.l.google.com. 25
Trying 173.194.78.26...
Connected to gmail-smtp-in.l.google.com.
[...]
EHLO bxl.mapetitesociete.be
250-mx.google.com at your service, [91.183.59.xxx]
[...]
MAIL FROM:<albert.leroy@bxl.mapetitesociete.be>
250 2.1.0 OK pn9si600796wjc.42 - gsmtp
RCPT TO:<leroy.mariecelestine@gmail.com>
550-5.1.1 The email account that you tried to reach does not exist.
[...]
RCPT TO:<mariecelestine.leroy@gmail.com>
250 2.1.5 OK pn9si600796wjc.42 - gsmtp
QUIT
221 2.0.0 closing connection pn9si600796wjc.42 - gsmtp
```

bounce mail

hards

softs

2.5.2. Difficultés

-

-

-

catch-all

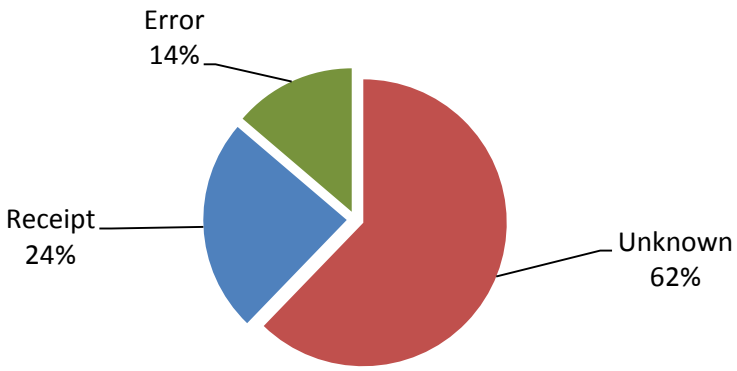
-

blacklister

2.5.3. Outils

2.6. Contrôle de consultation

: Exemple typique de situation suite à un envoi de campagne de communication : 24 % ont généré un accusé de réception, 14 % une erreur, et dans 62 % des cas, aucune information n'est disponible



2.6.1. Redirection de liens

http://mysite.com/track?m=albert@gmail.com&dst=www.smals.be

« albert@gmail.com;www.smals.be/a_page »

« YWxiZXJ0QGdtYWkuY29tO3d3dy5zbWFscy5iZS9hX3BhZ2U=»[19]

```
http://mysite.com/track?YWxi[…]tO3d3dy5zbWFscy5iZS9hX3BhZ2U=
```

```
<a href='http://mysite.com/track?YWxi[…]tO3d3dy
5zbWFscy5iZS9hX3BhZ2U='>http://www.smals.be/a_page</a>
```

```
<html><head><meta http-equiv="refresh" content="0;
URL="http://www.smals.be/a_page"></head></html>
```

```
www.smals.be/a_page
```

-
-
-
-
-

2.6.2. Image avec identifiant unique

2.6.3. Contrôle lors d'autres contacts

2.6.4. Réponse aux e-mails

2.7. Indicateurs de qualité et statistiques

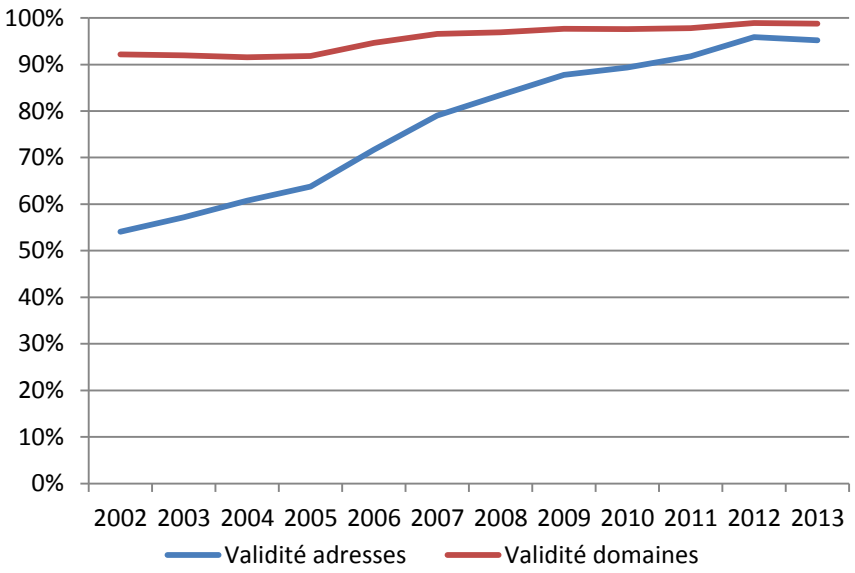
2.7.1. Erreurs syntaxiques

22

-

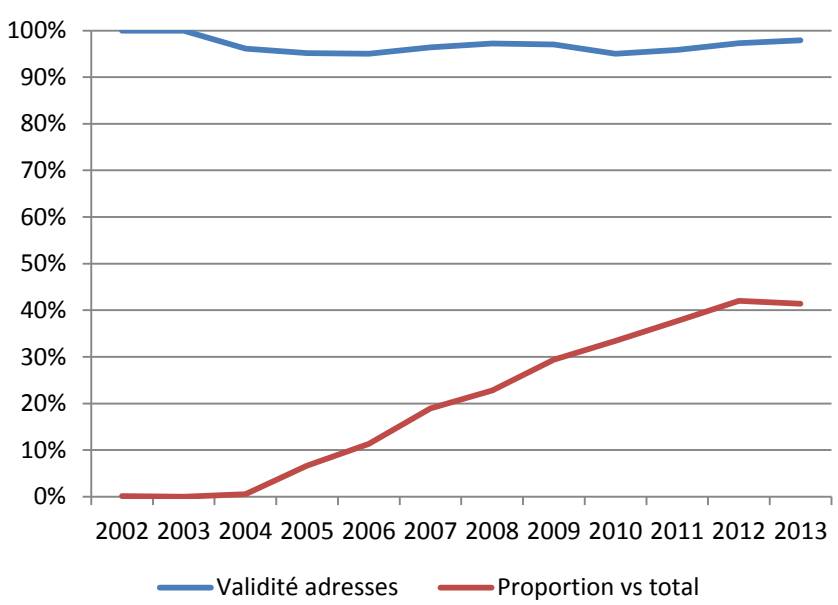
-

Validité des adresses e-mail et de leur nom de domaine associé d'une base de données de citoyens en fonction de l'année d'introduction



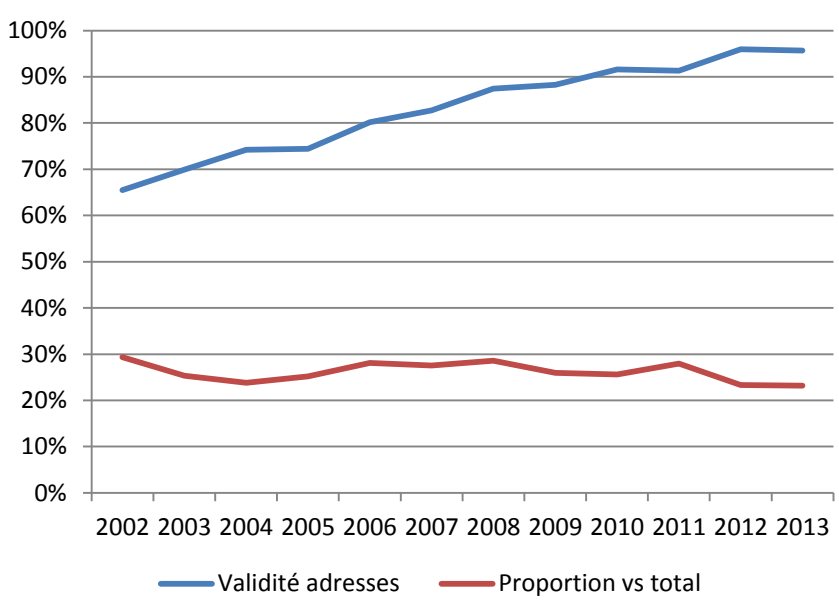
2.7.2. Dégressivité de la validité dans le temps

Validité des adresses Gmail et proportion dans la base de données



2.7.3. Dépendance au nom de domaine

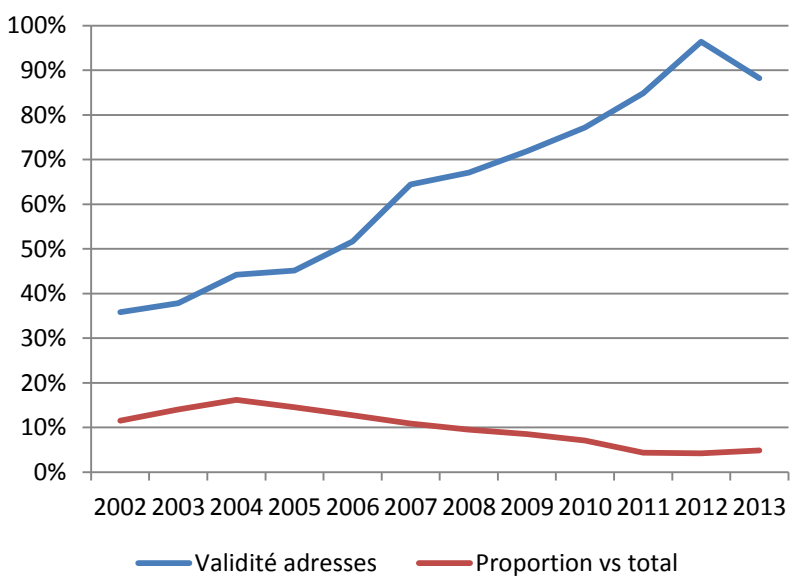
Validité des adresses Hotmail et proportion dans la base de données



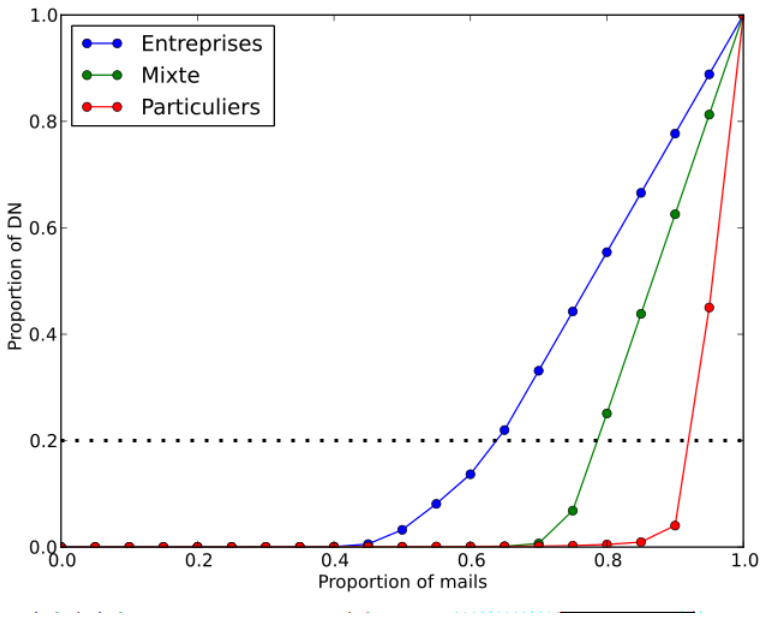
2.7.4. Proportion professionnels-particuliers

-
-
-

Validité des adresses Skynet et proportion dans la base de données



Distribution des adresses par rapport aux noms de domaine

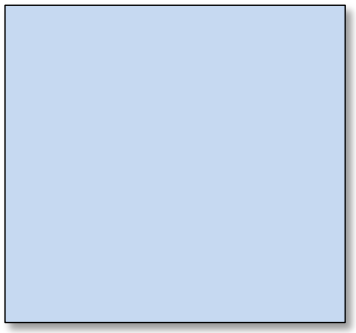


3. Stratégie de bonne gestion des adresses e-mail

3.1. Arbitrage stratégique

3.1.1. Rejeter les mauvais ou accepter les bons ?

3.1.2. Accepter l'exotisme ?



-
-
-
-

3.1.3. Quel contrôle sur les serveurs d'envoi d'e-mail ?

3.2. Aspects syntaxiques

3.2.1. Catégorisation

-
-
-

Adresses certainement fausses

```
^[^@,;:\s]+@[^@,;:\s]+$
```

```
^[^@,;:\s]+@([^@ ,;:\s+.-]+([-.][^@,;:\s+.-]+)*)$
```

```
[^@,;:\s]+@([\p{L}0-9]+([-][\p{L}0-9]+)*[.])*[\p{L}]{2,}$
```

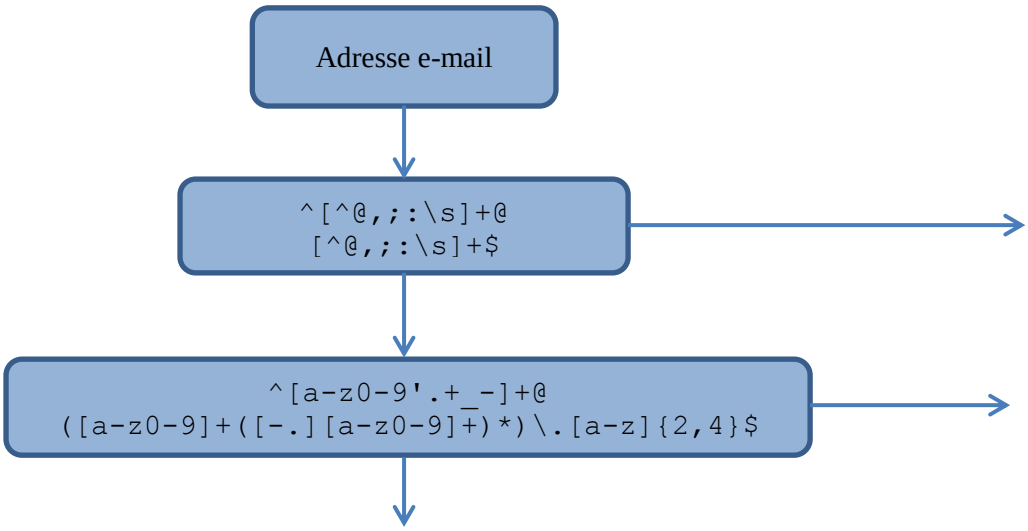
Adresses certainement correctes

```
^[a-z0-9'.+_-]+@([a-z0-9]+([-.][a-z0-9]+)*)\.[a-z]{2,4}$
```

undescore)

```
^[a-z0-9'+_-]+(\.[a-z0-9'+_-]+)*@
([a-z0-9]+([-.][a-z0-9]+)*)\.[a-z]{2,4}$
```

Adresses suspicieuses



3.2.2. Syntaxe spécifique

3.2.3. Suggestions de correction

-
-
  -
  -
-
  → →
-
-
-

3.3. Validation d’adresse

3.4. Suspicion d’erreurs

3.4.1. Par matching interne

underscore

-
-

3.4.2. Noms de domaine fréquents

			A	A

3.5. Matching et dédoublonnage

3.6. Batch ou on-line ?

3.7. Historique de la validité dans le temps et monitoring

3.8. Traitement on-line

3.8.1. Mise en place de tests en entrée

-
-
-

3.8.2. Suivi de la validité dans le temps

-
-
-
-
-
-
-

3.9. Traitement batch des fichiers existants

-
-
-
  -
  -
  -
  -
  -
-
-

3.10. Organisation

bounces

return on investment

-
-
-

-
-

4. Panorama d’outils existants sur le marché

4.1. Vérificateurs syntaxiques

.

.

.

.

-

-

```
^[a-zA-Z0-9.!#$%&'*+/=?^_`{|}~-]+@
[a-zA-Z0-9-]+(\.[a-zA-Z0-9-]+)*$
```

-

4.2. Testeurs d'existence

-
-
-
-
-

4.2.1. ServiceObjects

4.2.2. EmailVerify for .NET

4.3. Outils de suivi

4.4. Data Quality Tools

4.4.1. OpenRefine (Google refine)

4.4.2. IntoDQ (Trillium)

4.4.3. RedPoint

4.4.4. Human Inference

4.5. Outils CRM

bounces

5. Conclusion

return on investment

6. Bibliographie

[1] D. Clément, B. Laboisse, D. Duquennoy et A. Micheaux, «Non qualité de données & CRM : quel coût ?,» chez *QDC 2008*.

[2] I. Boydens, «Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium,» *Practical Studies in E-Government : Best Practices from Around the World,* pp. 113-130 (chapitre 7), 2011.

[3] Y. Bontemps, I. Boydens et D. Van Dromme, «Data Quality : tools,» Bruxelles, 2007.

[4] I. Boydens, Informatique, normes et temps, Bruxelles: Bruylant, 1999.

[5] I. Boydens, A. Hulstaert et D. Van Dromme, «Gestion intégrée des anomalies,» 2011. [En ligne]. Available: http://www.smalsresearch.be/publications/document?docid=62.

7. Annexes

7.1. Vérification syntaxique

7.1.1. Vérification syntaxique générale

```
^[^@,;:\s]+@([^@,;:\s+\.-]+([-\.][^@,;:\s+\.-]+)*)$
```

```
[^@,;:\s]+@
([\p{L}0-9]+([-][\p{L}0-9]+)*[.])*[\p{L}]{2,}$
```

```
^[a-z0-9\'\+_-]+(\.[a-z0-9\'\+_-]+)*@
([a-z0-9]+([-\.][a-z0-9]+)*)\.[a-z]{2,4}$
```

7.1.2. Vérification syntaxique spécifique

-
-

-

-

-

7.2. Typologie des évènements

7.3. Adresses e-mail officielles pour les citoyens

-
-
-
-

7.4. Éviter les risques liés au spam

spam

bots

honeypot

bots

honeypot

bots

honeypot

spamtraps

spamtraps »

8. Glossaire