



E-mail Address Reliability

Vandy BERTEN
Isabelle BOYDENS



Smals
ICT for society

fnrs
LA LIBERTÉ DE CHERCHER

ULB



Table des matières

Introduction

Contexte

Difficultés et incertitudes

Protocoles et mécanismes

Syntaxe

Contraintes syntaxiques générales et spécifiques

Méthodes traditionnelles

Propositions

Validation

Existence d'une adresse

Contrôle de consultation

Matching

Présomptions d'erreurs

Dédoublonnage

Bonnes pratiques & Conclusions



A hand is pointing at a glowing globe. The globe is overlaid with a network of white lines and dots, representing global connectivity. Several white envelope icons are scattered across the globe, symbolizing email or digital communication. The background is dark blue with a subtle pattern of white lines and dots.

Introduction

Table des matières

Introduction

Contexte & enjeux

Organisation

On-line vs Batch

Exemples

Difficultés et incertitudes

Protocoles et mécanismes

Syntaxe

Validation

Matching

Bonnes pratiques & Conclusions



Problématique

- De plus en plus d'administrations ou sociétés utilisent/veulent utiliser les **adresses e-mail** :
 - Contacts avec les citoyens/clients
 - Recommandé électronique (notifications)
 - V-ICT-OR veut les ajouter au Registre National
- Or :
 - Les DB d'e-mail sont souvent de **mauvaise qualité**
 - Les processus mis en place ne permettent en général pas de **maintenir un niveau correct**



Mauvaise qualité

- Campagnes de communication (e-gov) :
 - Jusqu'à 20% de « bounce »
 - Taux de lecture confirmé faible (\pm 20-25%)
- Certains organismes n'osent pas utiliser leur propres DB ...
- Pourquoi ?
 - Quasiment jamais de contrôle en entrée, ou minimaliste
 - Aucun suivi dans le temps
 - Incohérence de flux
 - Cumul d'incertitudes



Qualité : pourquoi l'améliorer

- Gains **directs** : ROI fonction des usages et du contexte:
 - Man-power réduit si bonne qualité
 - Communications officielles : diminutions des **envois papiers**
 - Gains potentiels à terme en **millions d'euros**
- Gains **indirects** :
 - Amélioration des services rendus et de l'efficacité
 - Crédibilité, législation



Qualité : comment l'améliorer

- Il existe de nombreuses techniques connues, pas/peu/mal appliquées :
 - Vérification syntaxique
 - Validation par envoi d'e-mail de confirmation
 - Suivi dans le temps par indicateurs de lecture
 - Rétroaction en cas d'erreur à l'envoi
- Il existe de nombreux outils ; en général ils tiennent mal compte de la complexité du problème



Qualité : comment l'améliorer

Apports originaux de notre étude :

- Amélioration notable de la **qualité** des tests
 - Syntaxe **spécifique** (↗ 15-20%)
 - Mise en évidence d'adresses **suspectes** (↗ 5-10%)
 - Amélioration des **techniques de validation** « batch »
- **Suggestions** de correction
 - Erreurs syntaxiques
 - Comparaison avec info annexes : nom, prénom
- Compilation de **best-practices**



On-line vs Batch

On-line	Batch
Sur un portail, à l'enregistrement	DB existante
Doit être rapide !	On a le temps, étude poussée possible
Interroger l'utilisateur → facile	Interroger l'utilisateur → plus difficile
Envoyer un e-mail + action → facile	Envoyer un e-mail → plus difficile
Ne dispense pas du batch par la suite!	S'applique sur des DB avec ou sans contrôle à l'entrée Si contrôle, se focalise sur l'évolution (DN, usage)

La plupart des méthodes s'appliquent aux deux

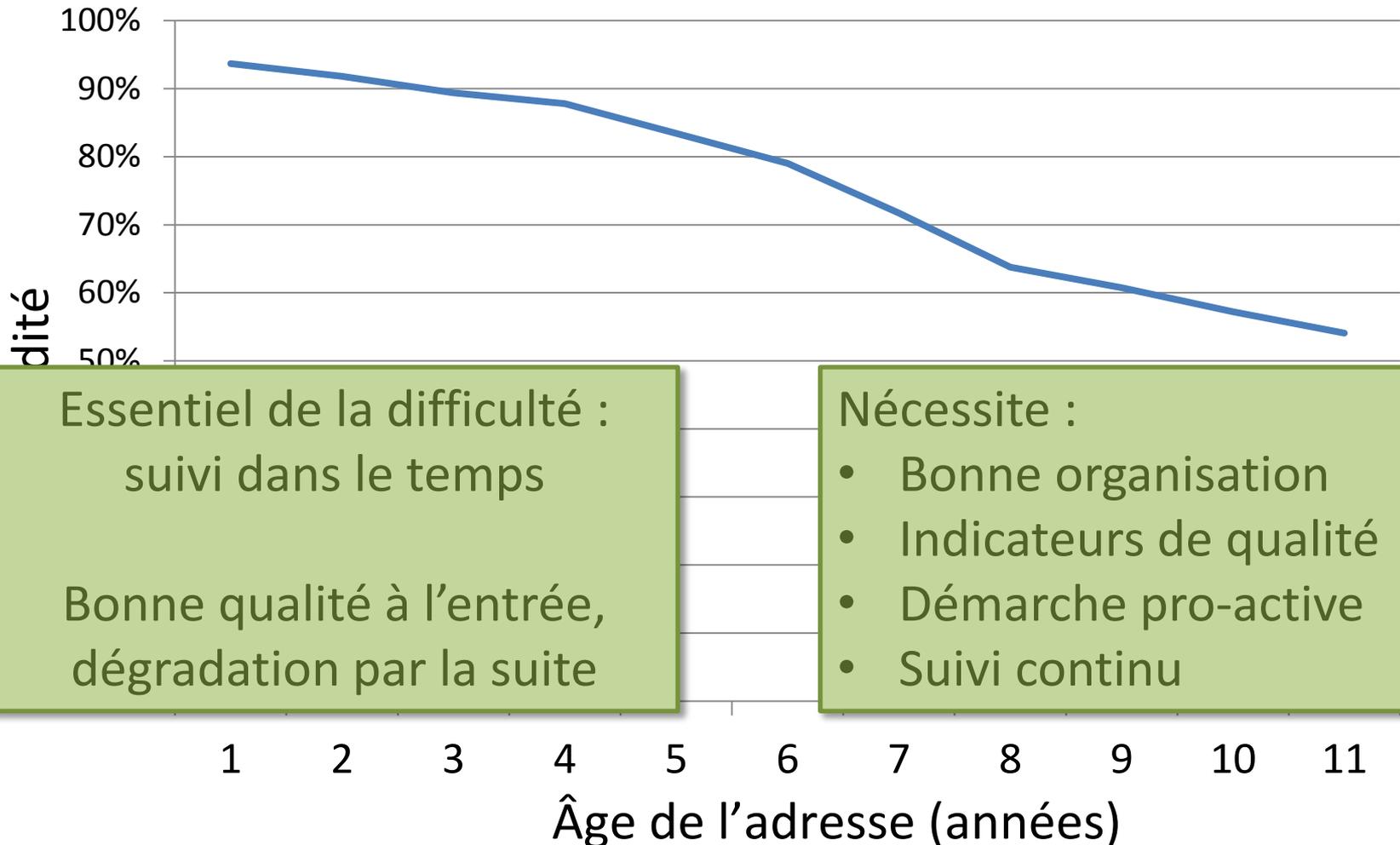


Étude basée sur ...

- La **littérature** (marketing, technique et e-gov à l'étranger)
- Des **tests et expériences** abondants sur des **grandes bases de données** (échantillons)
- Les **Data Quality Tools** et des **développements** propres
- 10 ans d'expérience en Data Quality (**DQ Cell**)
- Des contacts multiples avec le **terrain**, le développement, des services opérationnels et juridiques



Dégressivité de la validité



Vérification/validation : Étapes

Albert.Leroy@smals.be



Vérification/validation : Étapes

Vérification syntaxique

- Présence d'un (et un seul) « @ », absence d'espace, ...
- **Standards non respectés** : restrictions et extensions

Albert.Leroy@smals.be



Vérification/validation : Étapes

Albert.Leroy@smals.be

Top Level Domain (TLD)

- Aujourd'hui : plutôt statique, facile à valider (+/- 280)
- Prochainement : .brussels, .vlaanderen, .中国, .இந்தியா, .آزمایشی



Vérification/validation : Étapes

Nom de domaine (DN)

- Dans la pratique : a-z, 0-9, « - », « . »
- Dans le futur : accents, autres alphabets (IDN)
- Dynamique ; .be : **1300 changements/jour**, monde : **300k!**

Albert.Leroy@smals.be



Vérification/validation : Étapes

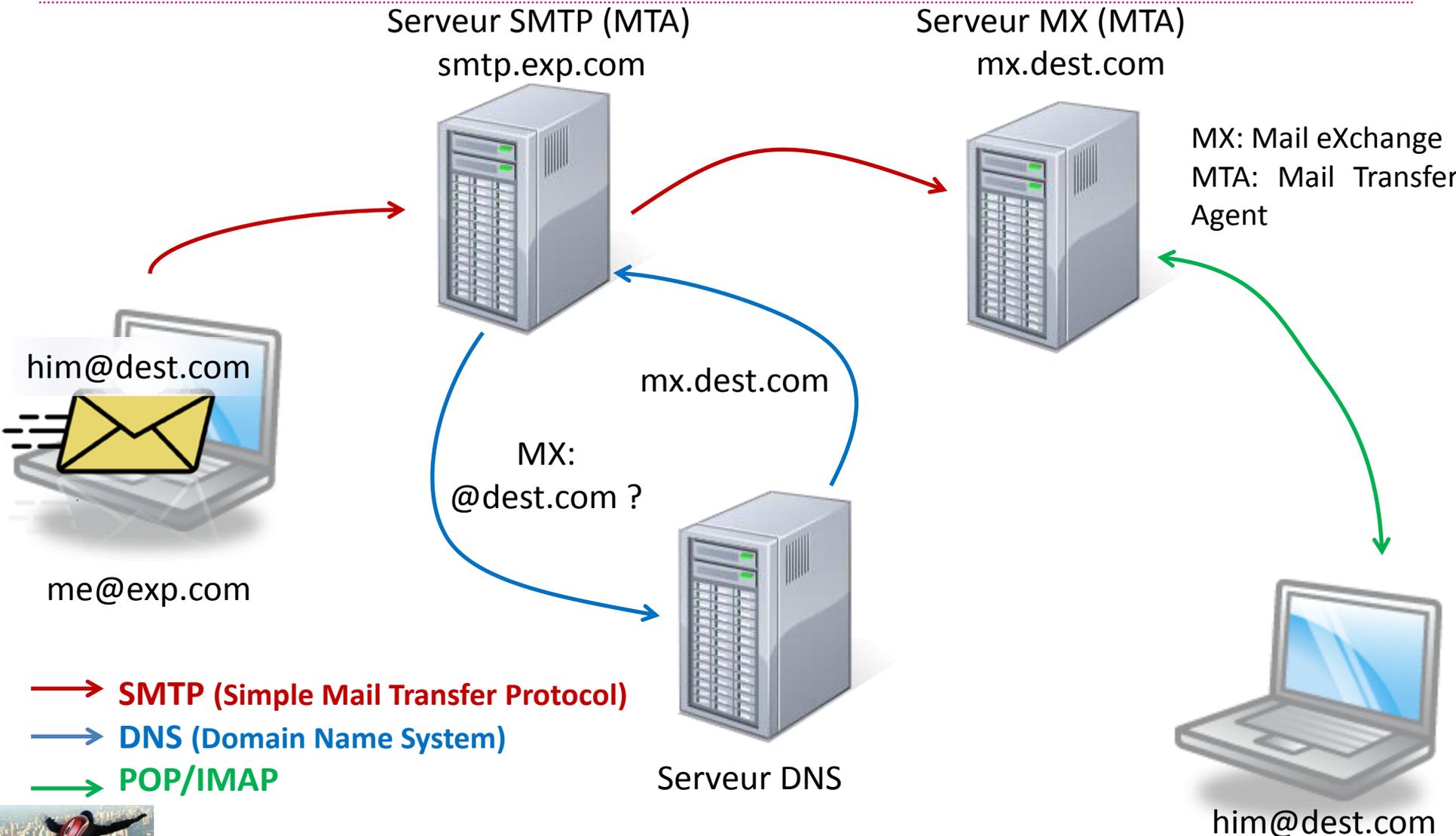
Albert.Leroy@smals.be

Username

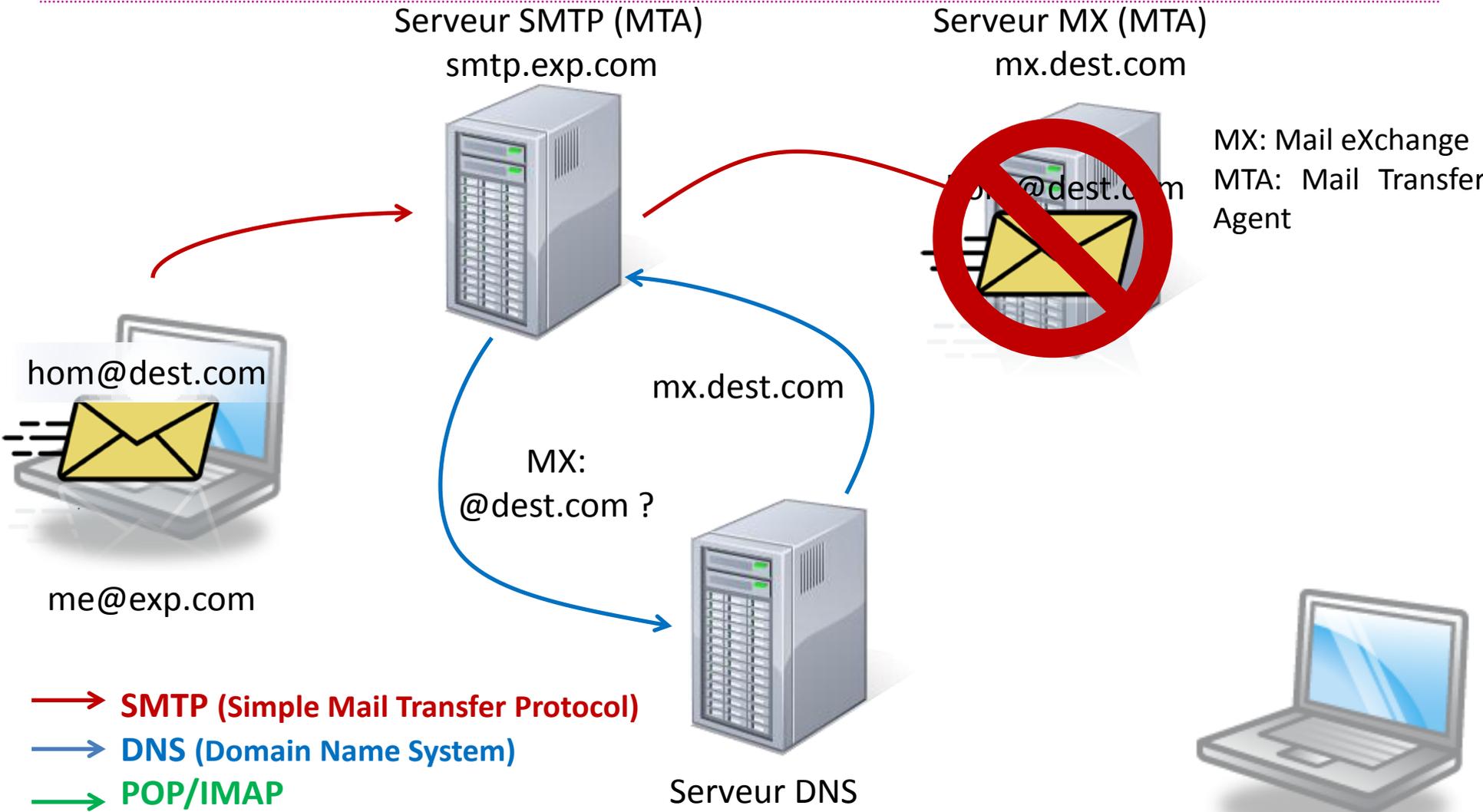
- Validation *a priori* (avant envoi) :
 - Utilisation du protocole SMTP
- Validation *a posteriori* (après envoi) :
 - Analyse de « bounce »
 - Contrôle de lecture (image, lien, ...)



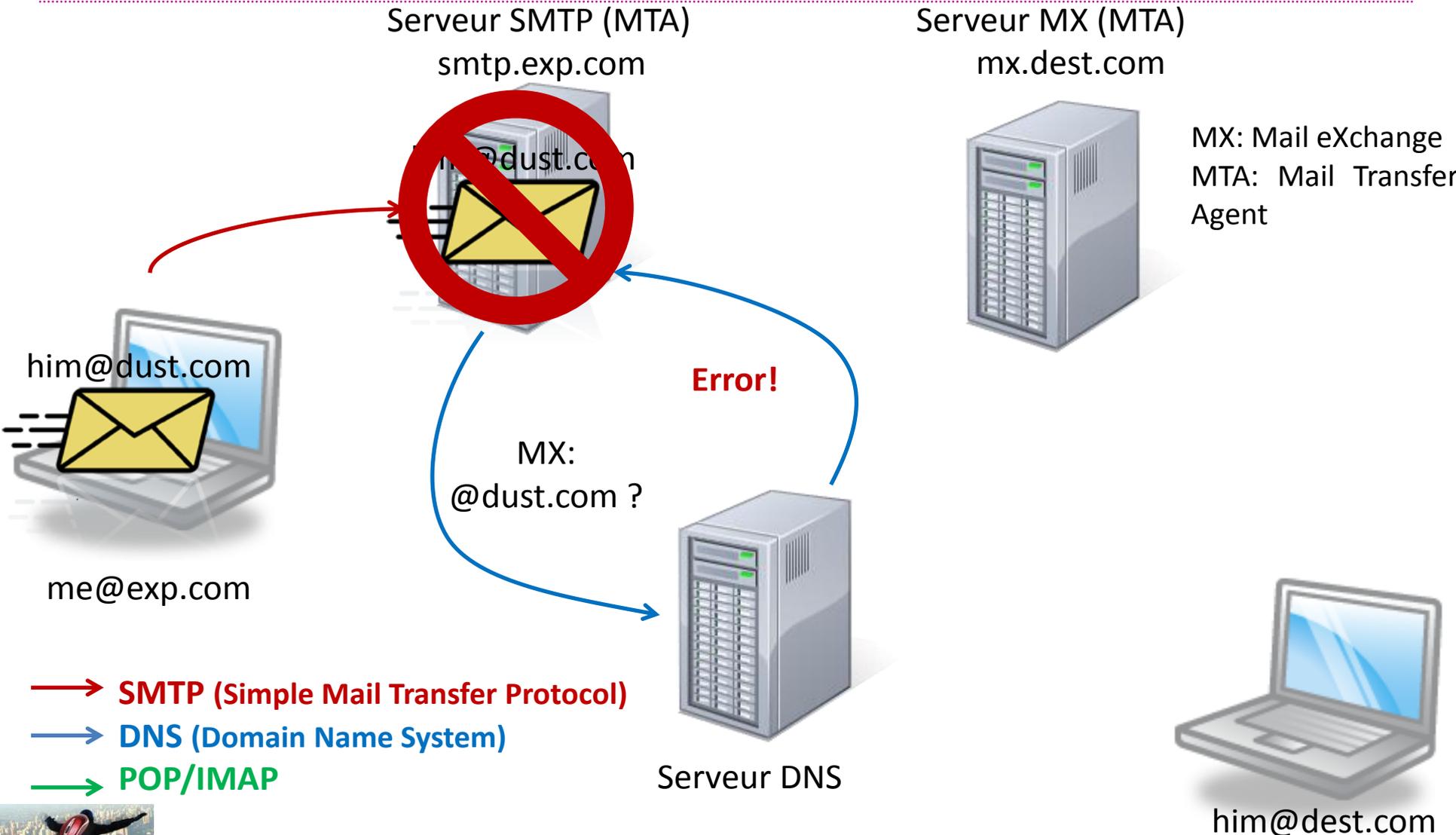
Mécanisme d'envoi d'un e-mail



Envoi d'un e-mail : bounce



Envoi d'un e-mail : bounce





Syntaxe

Table des matières

Introduction

Syntaxe

Contexte

Les standards et la pratique

Techniques de contrôle et limites

Syntaxes spécifiques

Propositions

Outils

Validation

Matching

Bonnes pratiques &
Conclusions



Contexte

- Syntaxe = vérification « orthographique » :
 - **Obligatoire** : un (et un seul) arobase (@)
 - **Interdit** : espace, virgule, point-virgule
 - Points non consécutifs, ...
- Ne dit pas si l'adresse/le domaine/le TLD **existe** !
- Analogie :
 - Code postal belge : 4 chiffres
 - 1234 respecte la syntaxe, mais n'est pas un CP !
 - Idem avec les numéros de téléphone
- Standards : RFC 5321 et 5322



Contexte

Erreurs évidentes		Erreurs ??
albert.leroy#smals.be	✓	albêrt.leroy@简体中文.com
albert.leroy@smals,be	✓	albert.leroy@be
albert leroy@smals.be	✗	albert-leroy@gmail.com
0471/257 800	✗	albert..leroy@smals.be
Rue Fonsy 20	✓	albert%leroy@blahblah.be
www.smals.be	✗	albert.leroy@a--b.be
	✓	-----@hotmail.com
	✗	albert@ma_boite.be



Syntaxe : que disent les standards ?

- Syntaxe « classique » :
 - [a-z], [0-9]
 - « . » et « - » (non consécutifs, pas en début ou en fin)
 - Dernière partie (TLD) : [a-z]{2,6}
- Nouveautés :
 - gTLD : .brussels, .vlaanderen, ...
 - IDN : ñandú.cl , 简体中文.com, ...
 - IDN ccTLD : .中国, .இந்தியா, . آزمايشی, .pф, ...

Albert.Leroy@smals.be

- Caractères normaux et accentués (case sensitive)
- .!#\$%&'()*+,-/=?^_`{|}~"
- Points non consécutifs, pas en début ou en fin



Dans la pratique ?

- Partie « nom de domaine » :
 - Respect imposé par la **structure « DNS »**
 - Un nom de domaine **non-conforme** n'apparaît **pas dans les tables**
- Partie « username » :
 - Uniquement évaluée par le serveur **mail de destination**
 - Username attribué par l'organisation qui le gère → une certaine **liberté** malgré les standards !
 - Dans la pratique : beaucoup plus **restrictif** que la norme
 - Rarement **case sensitive**
 - Certains acceptent des **extensions**

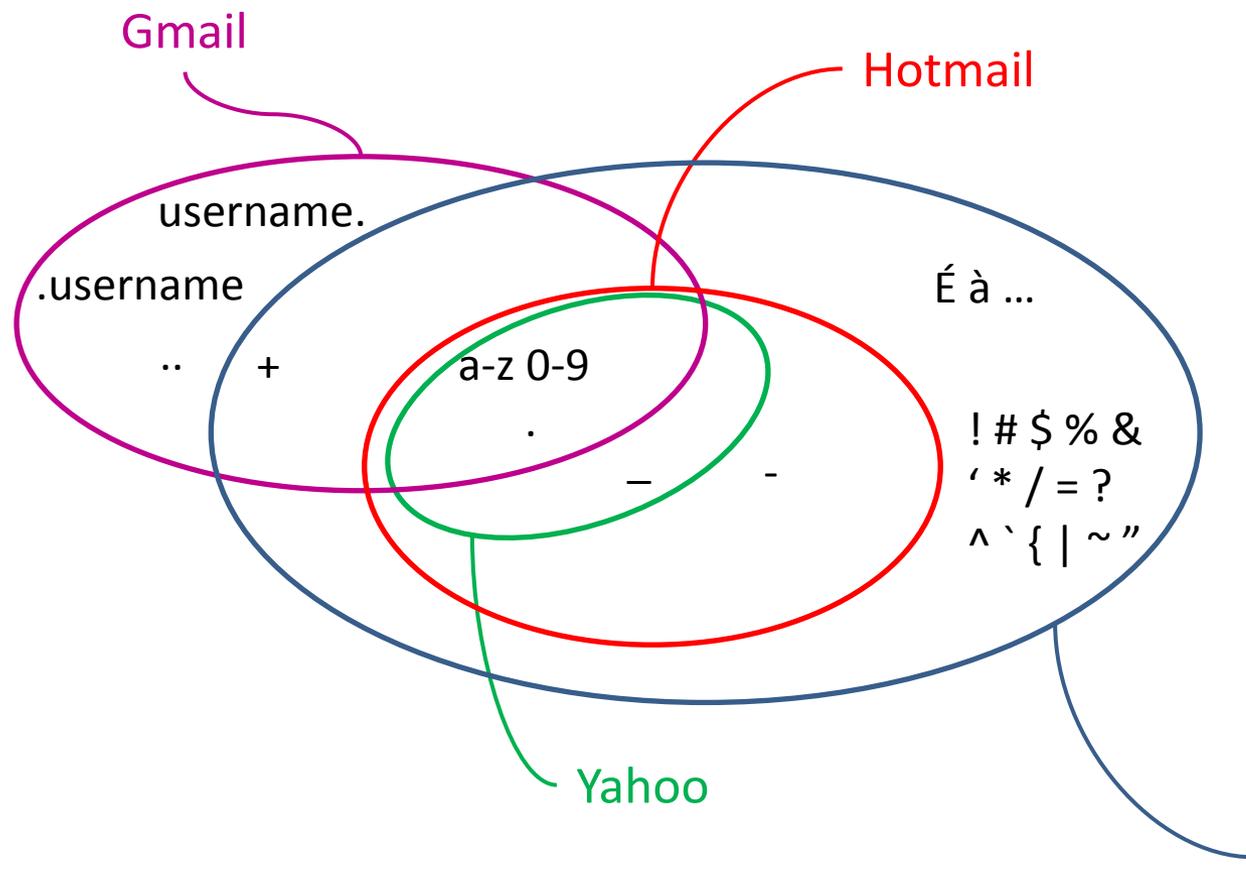


Syntaxe spécifique : intérêt ?

- Il est facile de trouver la **syntaxe spécifique** pour la plupart des grands fournisseurs (Gmail, Hotmail, Belgacom, Telenet, ...)
- Sur certaines DB étudiées : **85% des adresses !**
- Permet d'être **beaucoup plus restrictif** sur ce qu'on laisse passer, sans risquer les « **faux négatifs** »



Syntaxes spécifiques (username)



- Yahoo :
- Max 1 point
 - 4-32 caractères
 - 1^{er}: a-z, dernier : a-z, 0-9
 - ...

- Gmail :
- « . » et après « + » : ignorés
 - 6-30 caractères
 - ...

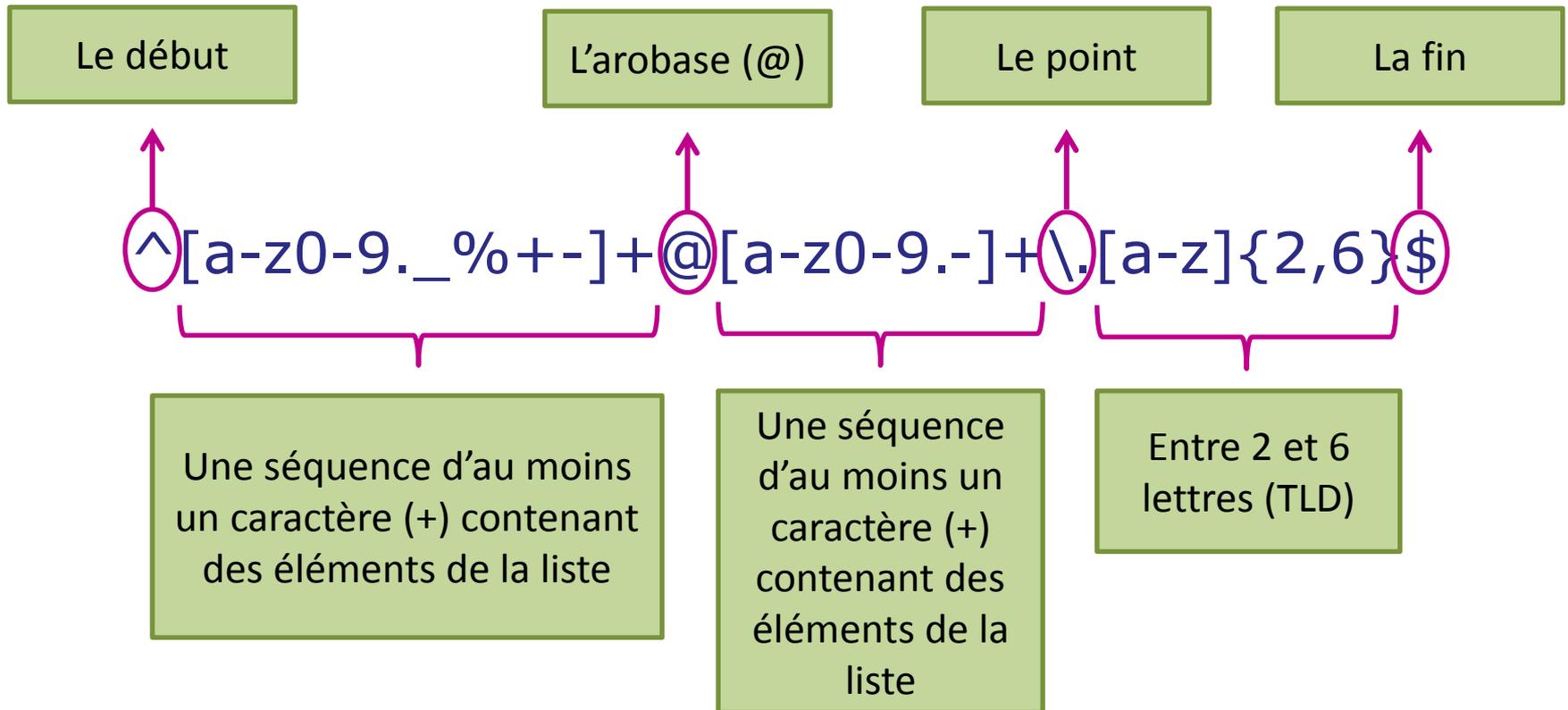


Comment vérifier ?

- Technique de vérification la plus répandue : les expressions régulières
- Sorte de mini-langage de programmation, utilisable de façon (quasi) standard par (quasi) tous les langages
- Très puissant pour vérifier qu'une chaîne de caractères rencontre bien certaines contraintes
- A malgré tout quelques limites



Expressions régulières



Expressions régulières

- On trouve beaucoup de variantes d'expressions régulières de vérification d'e-mail
- La **plupart** conviennent pour l'énorme **majorité** des adresses en cours
- Certaines **acceptent** beaucoup **trop**
- D'autres **refusent** des adresses **valides**
- Parfois : totalement **illisible**

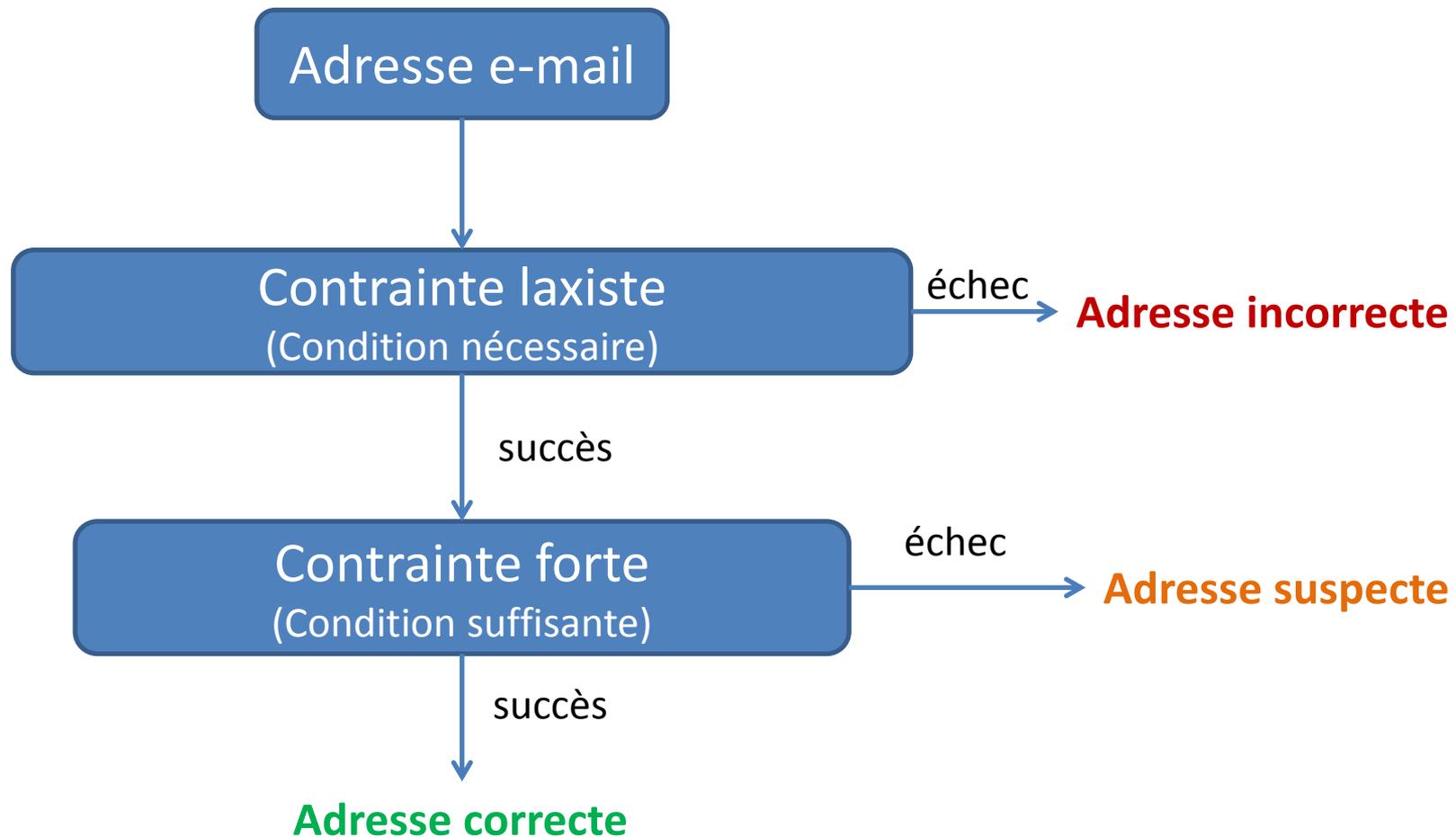


Expressions régulières : les limites

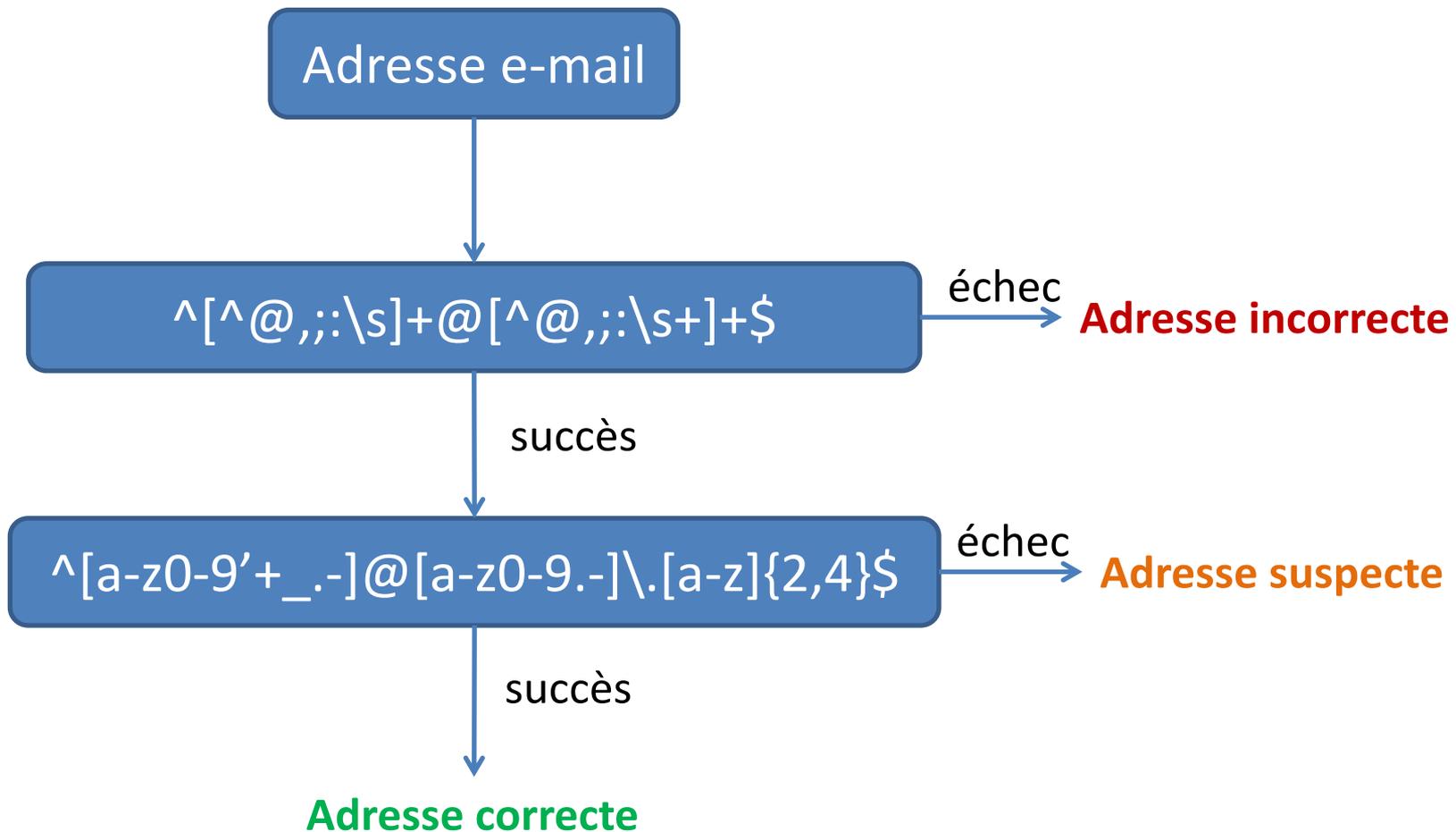
- Problème :
 - Soit trop contraignant
 - Soit trop laxiste
- Difficile de tout vérifier avec une expr. régulière
- Notre proposition : vérifier en **2 temps** :
 - Éliminer des adresses **certainement fausses** avec un test **laxiste**
 - Identifier des adresses **suspectes** avec un test très (trop) **contraignant**
 - On-line : demander confirmation
 - Batch : ajouter dans une liste « à contrôler »
 - Ce qui reste : **correct** (→ tests spécifiques)



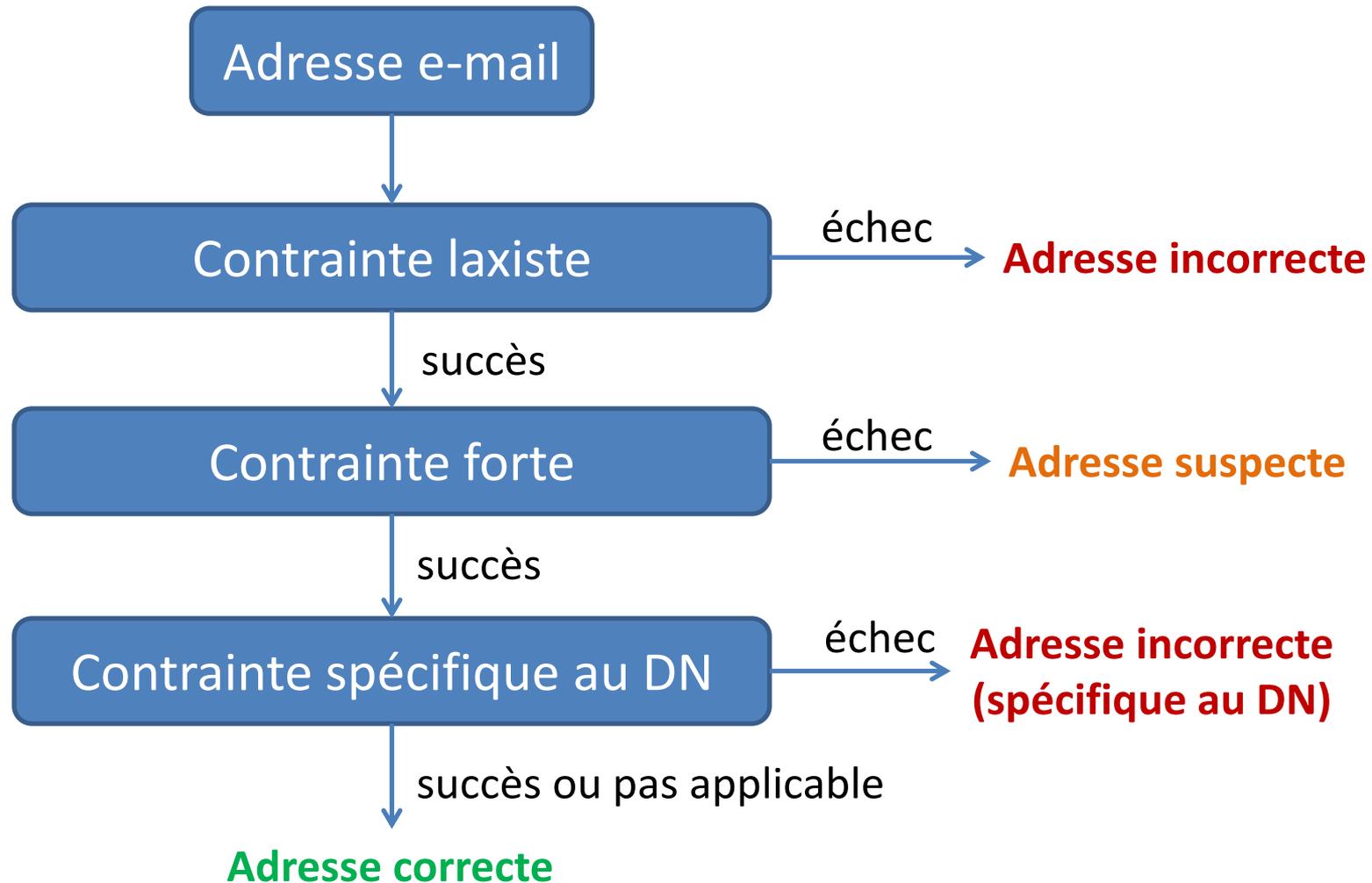
Proposition : Trois catégories



Proposition : Trois catégories



Proposition : Quatre catégories



PoC

Mail: Full check

Firstname:

Lastname:

Verification and Validation

Syntax check: Incorrect syntax

Matching

Database matching: No match
Name matching: No name provided
Domain name matching: No match

Submission

Cannot submit!

Suggestions

- alberleroy@smals.be

Address: alberleroy@smals.be
Firstname:
Lastname:

Verification and Validation

Syntax check: Unusual format, please double check!
 Suspicious characters : alberleroy@smals.be

TLD (.be): Passed
Domain (smals.be): Passed

Matching

Database matching:
Name matching:
Domain name matching:

Submission

We have some doubt,

Suggestions

- alberleroy@smals.be

Mail: Full check

Firstname:

Lastname:

Verification and Validation

Syntax check: Incorrect syntax
 Forbidden characters : alberleroy|@smals.be

Matching

Database matching: No match
Name matching:
Domain name matching:

Submission

Cannot submit!

Suggestions

- alberleroy@smals.be

Mail: Full check

Firstname:

Lastname:

Verification and Validation

Syntax check: Incorrect syntax for domain gmail.com

Matching

Database matching: No match
Name matching: No name provided
Domain name matching: Common domain name

Submission

Cannot submit!

Suggestions

- albert.leroy@gmail.com





Mail:

Full check



Firstname:

Submit

Lastname:

Reset

Syntaxe : l'essentiel

Plus
complexe
que ce que
l'on pense !

Tests
spécifiques au
domaine
⇒ ↗ précision

Tests binaires =
trop laxistes, ou
trop contraignants



Adresses
suspectes ⇒
limite les faux
positif/négatif





Validation

Table des matières

Introduction

Syntaxe

Validation

Validation du nom de domaine

Validation d'adresse

Contrôle de lecture

Limites et difficultés

Outils

Matching

Bonnes pratiques
& Conclusions

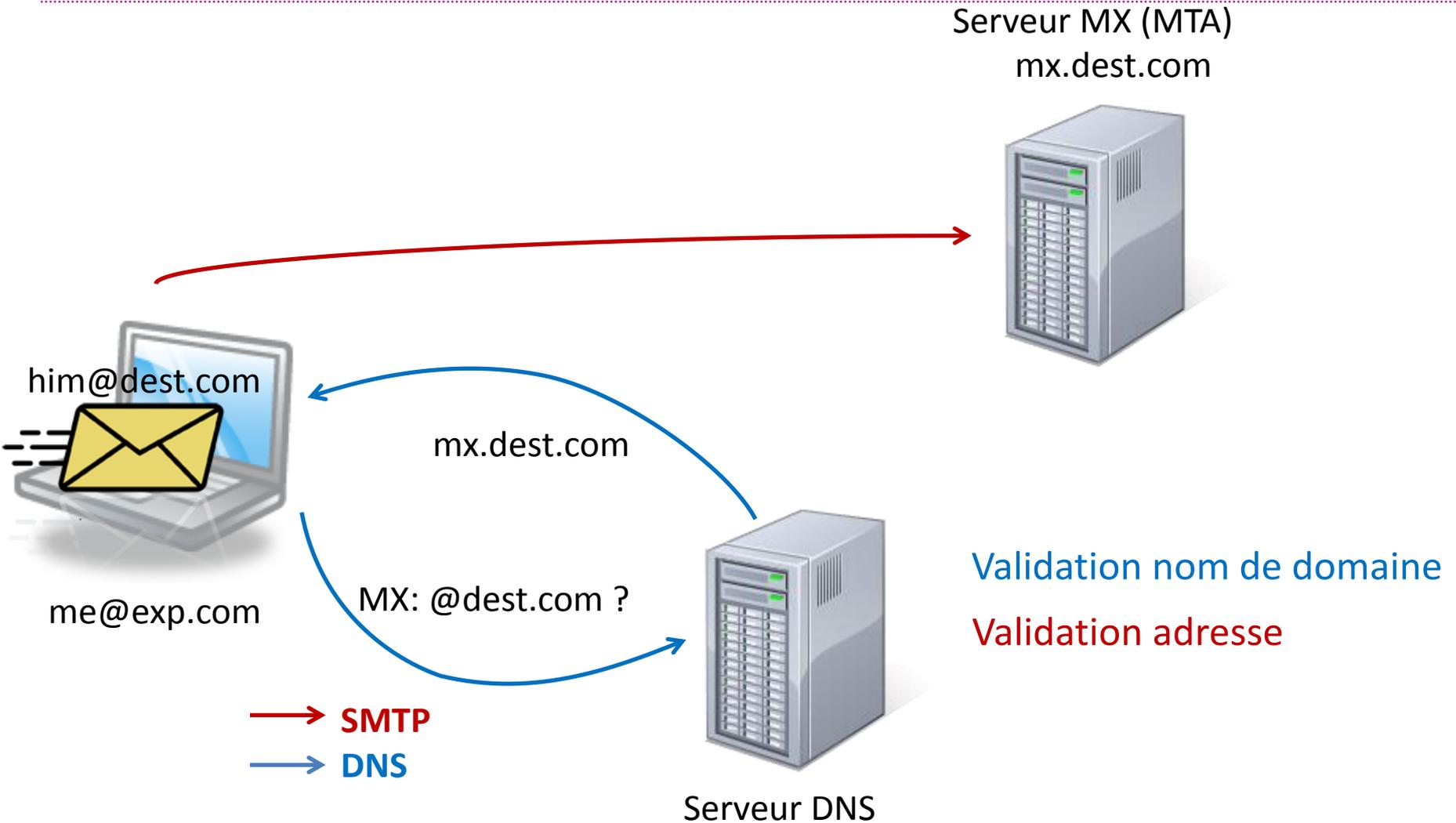


Contexte

- Une adresse syntaxiquement correcte **n'implique pas qu'elle existe !**
- Une adresse qui existe **n'est pas toujours consultée**
- **Existence** d'une adresse :
 - Le nom de domaine
 - L'adresse elle-même
- Contrôle de **lecture** :
 - Insertion d'image
 - Lien unique



Mécanisme de validation



Validation : conversation SMTP

```
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.xx.xx.xx][...]
MAIL FROM:<albert.leroy@bxl.mapetitesociete.be>
250 2.1.0 OK pn9si6796wjc.42 - gsmtip
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 pn9si6796wjc.42 - gsmtip
QUIT
221 2.0.0 closing connection pn9si6796wjc.42 - gsmtip
```



Limites et difficultés

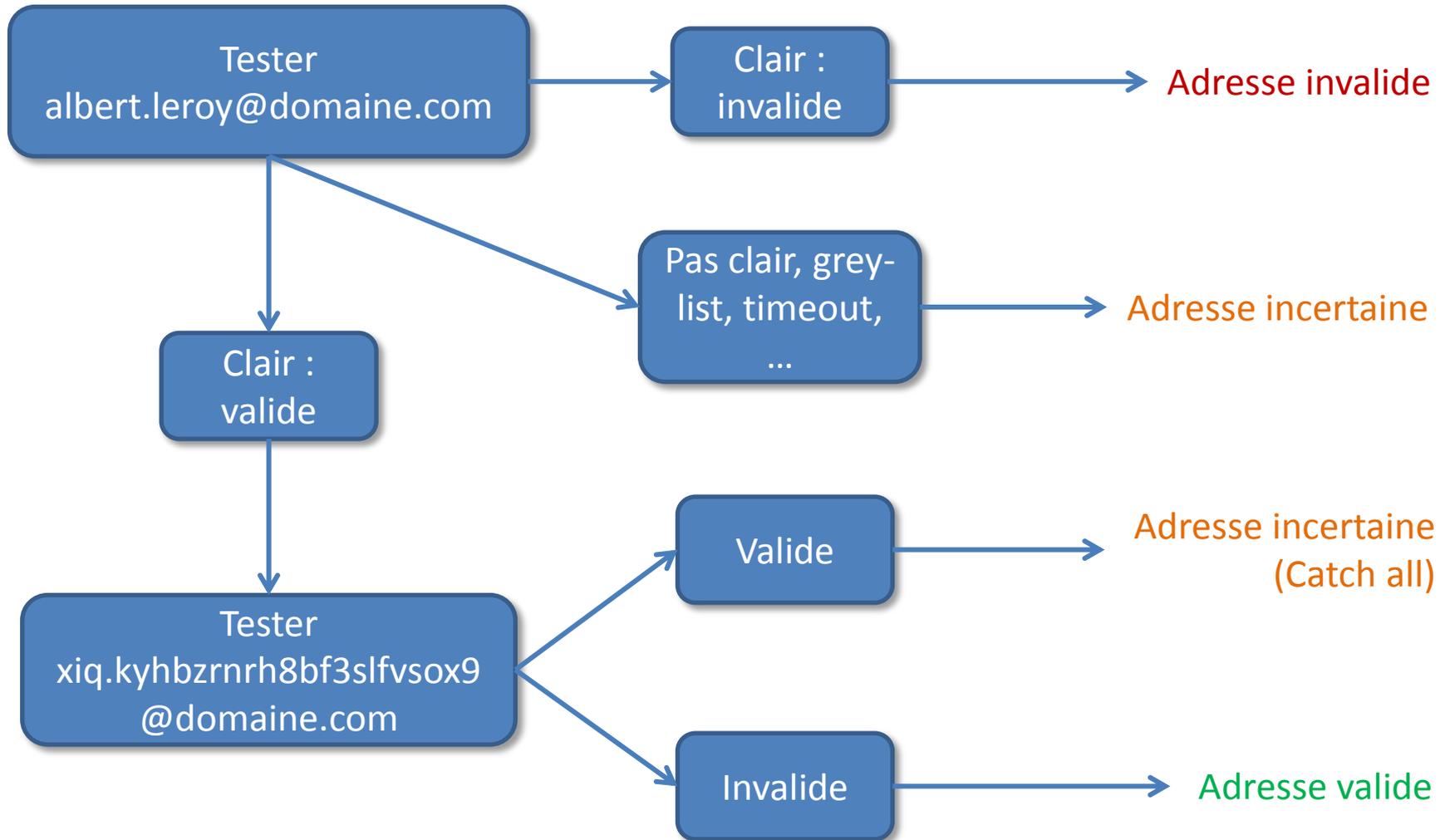
- Validation nom de domaine : très fiable
- Validation adresse : beaucoup d'incertitudes !

```
550 MAILBOX NOT FOUND 550 <john@foo.bar>... User unknown
550 5.7.2 This smells like Spam
550 <john@foo.de>. Benutzer hat zuviele Mails auf dem Server.
554 Delivery error Sorry your message to joe@foo.bar cannot be delivered. [#102]
```

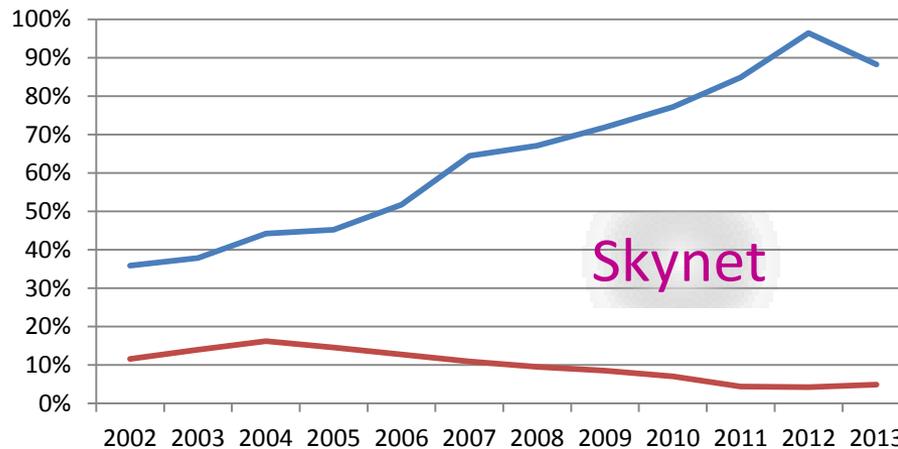
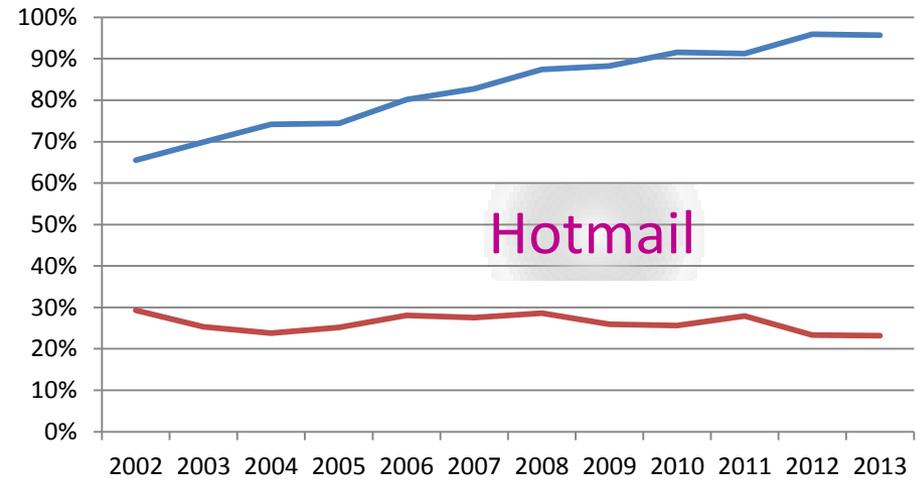
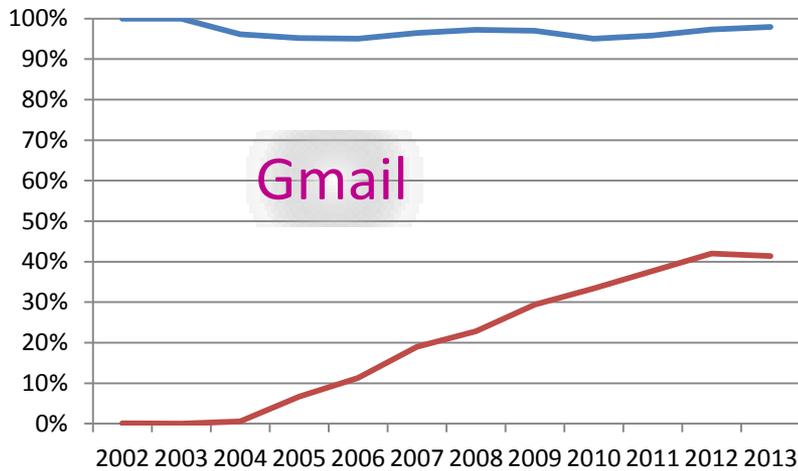
- Si source pas clairement identifiée :
 - Greylisting (explicite)
 - Catch-all (implicite) } > 20-25% depuis un PC « standard »
- En cas de requête massive : risque de blacklist !
- Recyclage des adresses



Validation d'adresse



Dégressivité par domaine (histoire des usages)



— Validité adresses — Proportion vs total



Contrôle de lecture

- Quelques techniques pour s'assurer qu'une adresse est toujours active, mais rien de très fiable
- Accusé de réception (Outlook & co) : pas standard, pas inter-plateforme
- Présence de lien « unique » à cliquer
 - Il faut une raison de cliquer !
- Présence d'une image « unique »
 - Il faut accepter d'afficher les images



Lien unique

```
<a href="http://mysite.com/redir?  
m=albert.leroy@smals.be&a=www.smals.be">  
www.smals.be</a>
```

To:albert.leroy@smals.be

www.mysite.com/redir

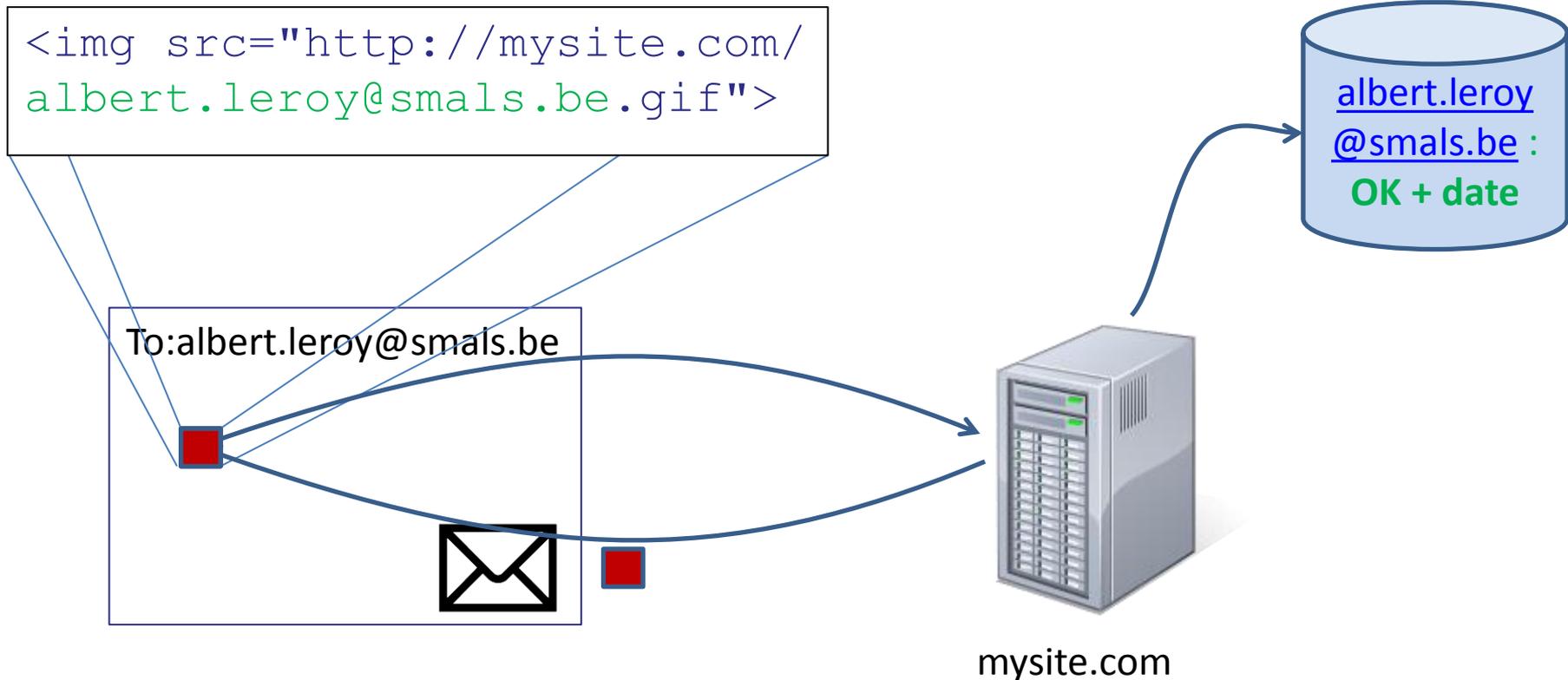
www.smals.be

albert.leroy@smals.be :
OK + date

- Variantes : crypter/masquer le contenu, passer via une base de données
- Il faut une bonne raison de cliquer !
- Indispensable à l'enregistrement (e-mail de confirmation)



Image unique



Remarque :

- Souvent un pixel blanc (invisible)
- Il faut accepter les images !

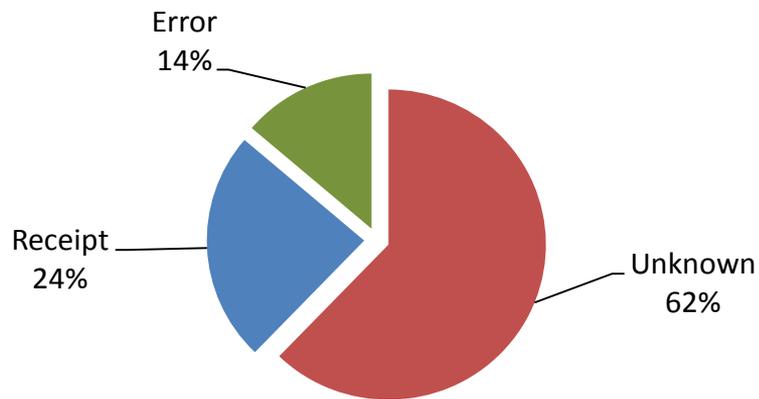


Image unique

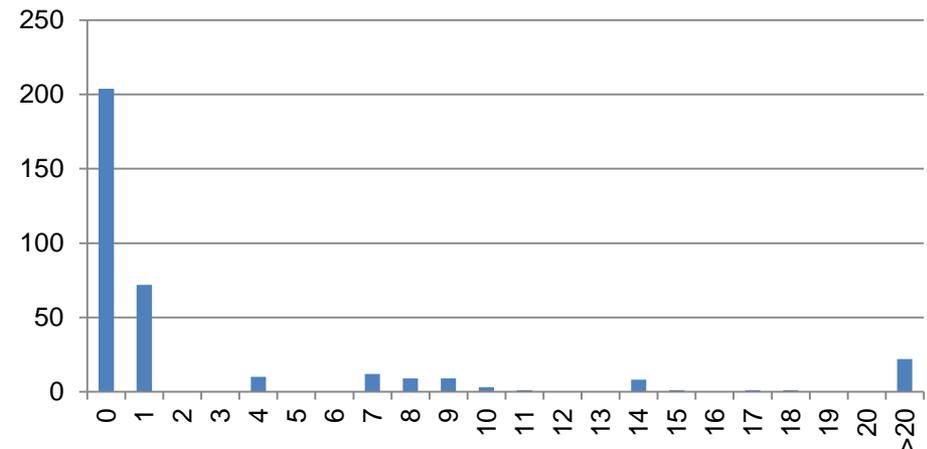
- Inconvénient : beaucoup de systèmes désactivent les images distantes par défaut

Security: To ensure privacy, images from remote sites were prevented from downloading. [Show Images](#)

 Images are not displayed. [Display images below](#)



Délai lecture



PoC

Address: albert.leroy@gmail.com
Firstname:
Lastname:

Verification and Validation

Syntax check: Passed
TLD (.com): Passed
Domain (gmail.com): Passed

Matching

Database matching: No match
Name matching: No name provided
Domain name matching: Common domain name

Submission

Address: albert.leroy@smals.be
Firstname:
Lastname:

Verification and Validation

Syntax check: Passed
TLD (.be): Passed
Domain (smals.be): Passed
Address check: This mail server is known to ignore all requests (catch-all server) ... Cannot tell whether this address exists or not!

Matching

Mail: Full check

Firstname:

Lastname:

Submission

Verification and Validation

Mail: Full check

Firstname:

Lastname:

Verification and Validation

Syntax check: Passed
TLD (.be): Passed
Domain (sams.be): Wrong DN

Matching

Database matching: No match
Name matching: No name provided
Domain name matching: Found match: sams.be ⇔ smals.be

Submission

Cannot submit!

Suggestions

- albert.leroy@smals.be





POC Email Addresses Reli x



vandyssh.smalsrech.be/mail



Mail:

Full check

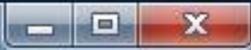
Firstname:

Submit

Lastname:

Reset





POC Email Addresses Reli x vandyssh.smalsrech.be/lo x Inbox - vandy.berten@gm x



vandyssh.smalsrech.be/mail



Mail: Full check

Firstname:

Lastname:

Validation : l'essentiel

Existence du domaine :
très fiable

Existence de l'adresse :

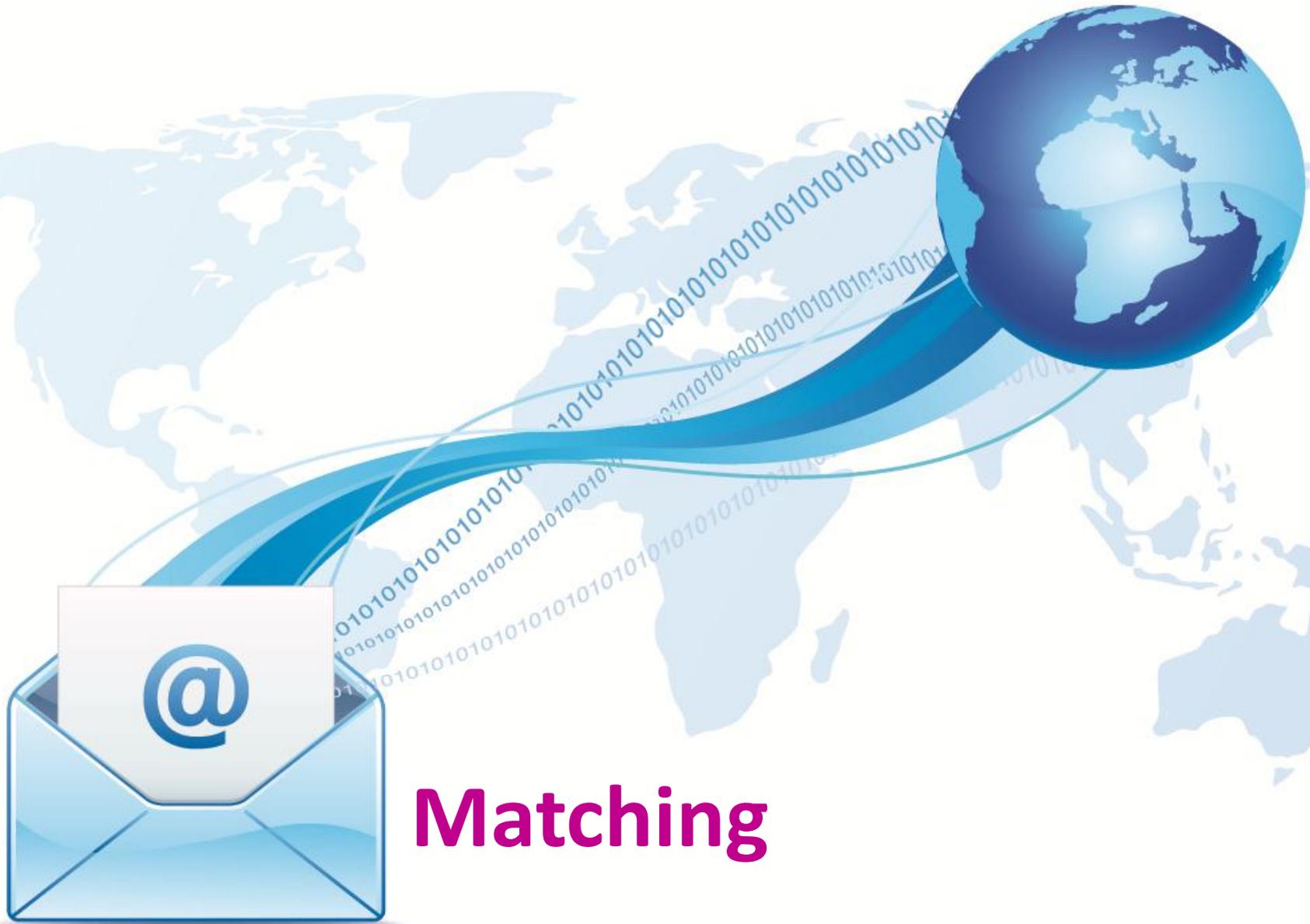
- Fausse/Correcte : \pm fiable
- Beaucoup d'inconnu (catch-all, greylist, ...)

Contrôle de lecture :

- Lu : très fiable
- Pas d'accusé = pas d'information

Seule validation fiable : envoi d'un e-mail avec action obligatoire





Matching

Table des matières

Introduction

Syntaxe

Validation

Matching

Contexte

Matching interne (informations annexes)

Matching sur nom de domaine

Dédoublonnage

Suggestions

Difficultés

Outils

Bonnes pratiques &
Conclusions



Contexte

- Matching : classique en Data Quality
- But : comparer des infos identiques ou similaires
- Différentes utilités :
 - Matching interne : dans un « record », utiliser la redondance pour croiser des info → **suspicion d'erreurs** (nom/prénom/e-mail)
 - Dédoublonnage : « records » similaires → **suspicion de doublons**
 - Domaine connu : similitude avec des domaines fréquents → **suspicion d'erreurs** (nom de domaine)
- Ne permet pas de détecter des erreurs, mais de les **soupçonner**



Matching interne : info annexes

Nom	Prénom	E-mail
Leroy	Albert	albert.leroy@smals.be

Nom	Prénom	E-mail
Leroy	Albert	ablert.leroy@smals.be

Nom	Prénom	E-mail
Leroy	Ablert	albert.leroy@smals.be

Nom de société	E-mail
Les meilleurs asbl	lesmeilluers@gmail.com

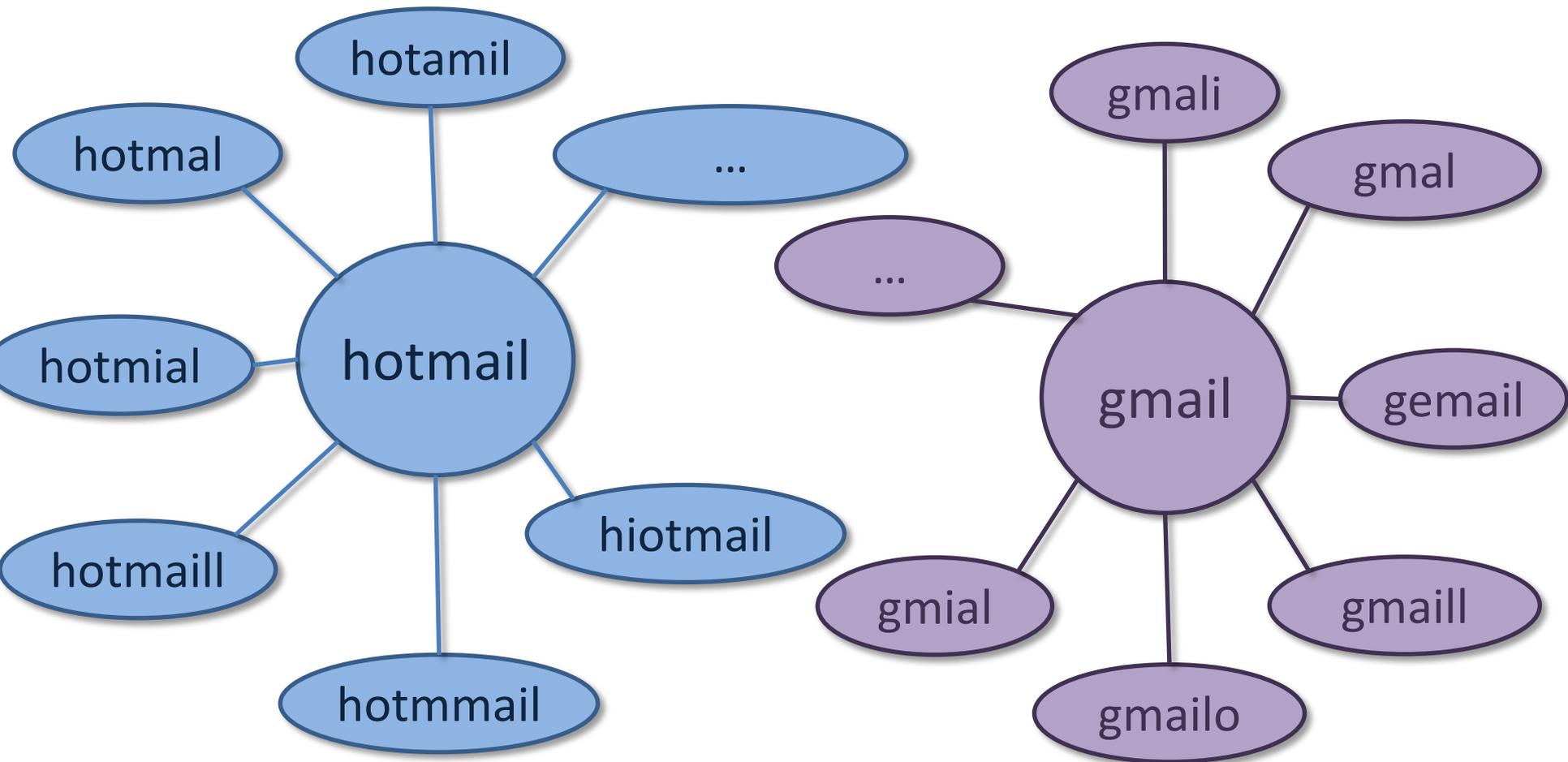


Exemple réel (anonymisé)

Nom	Prénom	E-mail
El l arkani	Ahmine	ahmine.el l arkani@hotmail.com
De t immerman	Alain	alain.de t immerman@gmail.com
De s aux	Adèle	adelede s aux@swing.be
Vanden l berghe	Arnaud	arnaud.vanden l berghe@skynet.be
De s medt	Geert	geertdes s medt@hotmail.com
Piotrowski	Alexander	alexander l .piotrowski@telenet.be
Lanoy	Caroline	carol l anoy@hotmail.fr
Michel	Charles-Édouard	charlesedou.michel@yahoo.fr
Hammoud	Haffsa	haf l ssa.hammoud@gmail.com
Mabrouk	Tarik	mabrouk.tare l @gmail.com
Wouters	Idalie	wouterso l dalie@yahoo.fr



Matching interne : domaines connus



Dédoublonnage

- Les **techniques classiques** de dédoublonnage peuvent être utilisées pour détecter des **doublons** dans une DB
- Il faut se servir de **différentes informations** pour établir un double
- En général, on établit un « **classement** » : double très probable (beaucoup d'information similaire) → double moins probable
- Parfois **beaucoup plus que deux !**



Outils

- Le matching et dédoublonnage sont des tâches **complexes**, nécessitant beaucoup de « **tuning** »
 - Seuls des **outils spécialisés de Data Quality** y arrivent correctement
 - Gratuit : OpenRefine
 - Propriétaire :
 - IntoDQ (Trillium)
 - RedPoint
 - HumanInference
- Pas de modules spécifiques aux e-mails mais largement paramétrables



PoC

Address: albert.leroy@smals.be
Firstname: Albert
Lastname: Laroy

Verification and Validation

Syntax check: Passed
TLD (.be): Passed
Domain (smals.be): Passed

Matching

Database matching: No match
Name matching: Mail contains firstname
Suspecting typo: leroy ⇒ laroy: l[e/a]roy
Domain name matching: Common domain name

Submission

We have some doubt, please double check!

Suggestions

- albert.laroy@smals.be (Albert Laroy)
- albert.leroy@smals.be (Albert Leroy)

Address: albert.leroy@smalz.be
Firstname:
Lastname:

Verification and Validation

Syntax check: Passed
TLD (.be): Passed
Domain (smalz.be): Passed

Matching

Database matching: No match
Name matching: No name provided
Domain name matching: Found match: smalz.be ⇔ smals.be

Submission

We have some doubt, please double check!

Suggestions

- albert.leroy@smals.be





Mail:

Full check

Firstname:

Submit

Lastname:

Reset



Dédoublonnage

Matching : l'essentiel

But :
suspecter/identifier
des cas douteux

Purement empirique !
⇒ Traitement humain
souvent nécessaire
(facilité par les suspicions)

Matching interne :
erreur dans le
nom/prénom/e-mail

Domaine connu :
erreur dans le nom
de domaine

Dédoublonnage :
données similaires

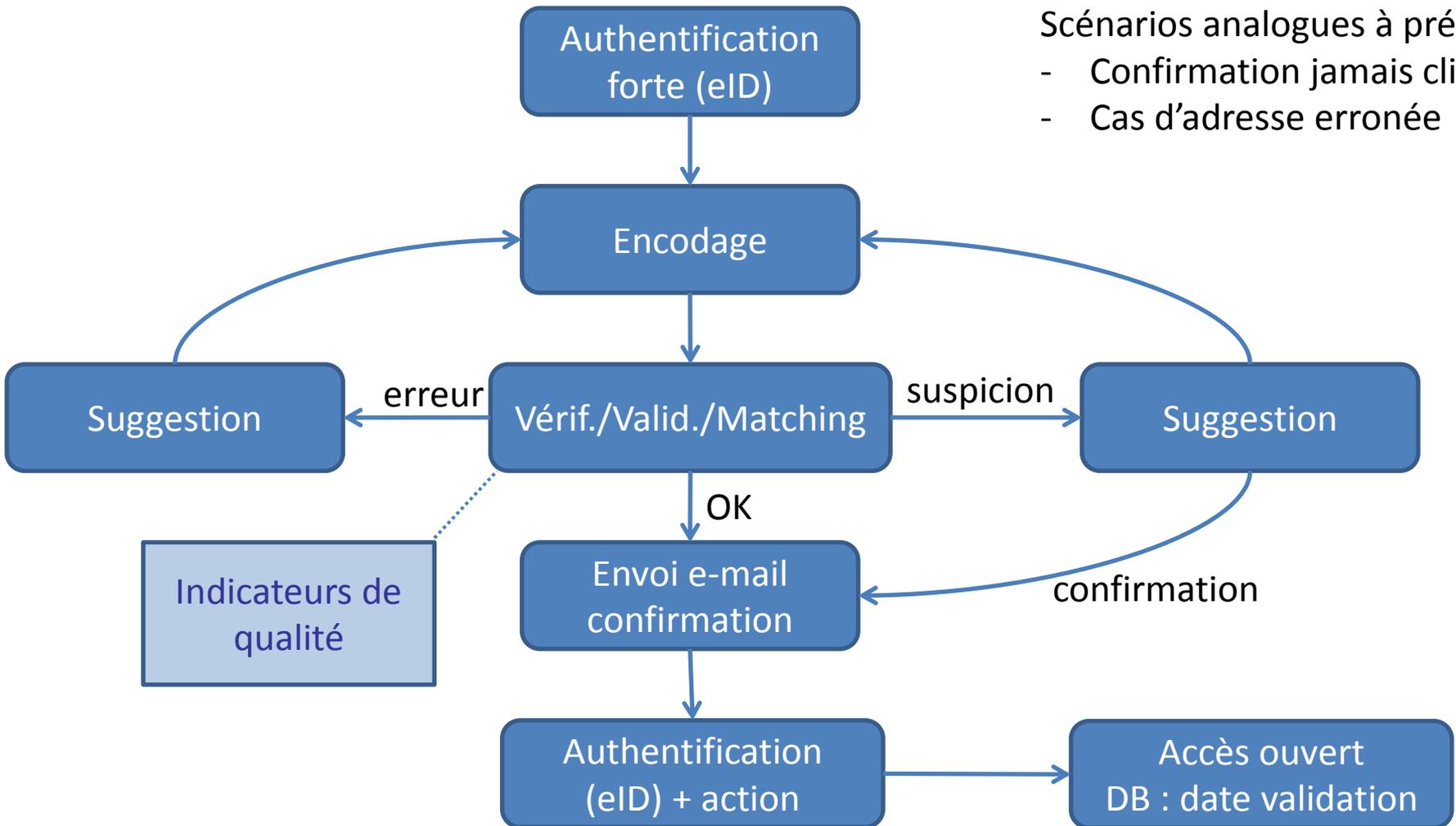




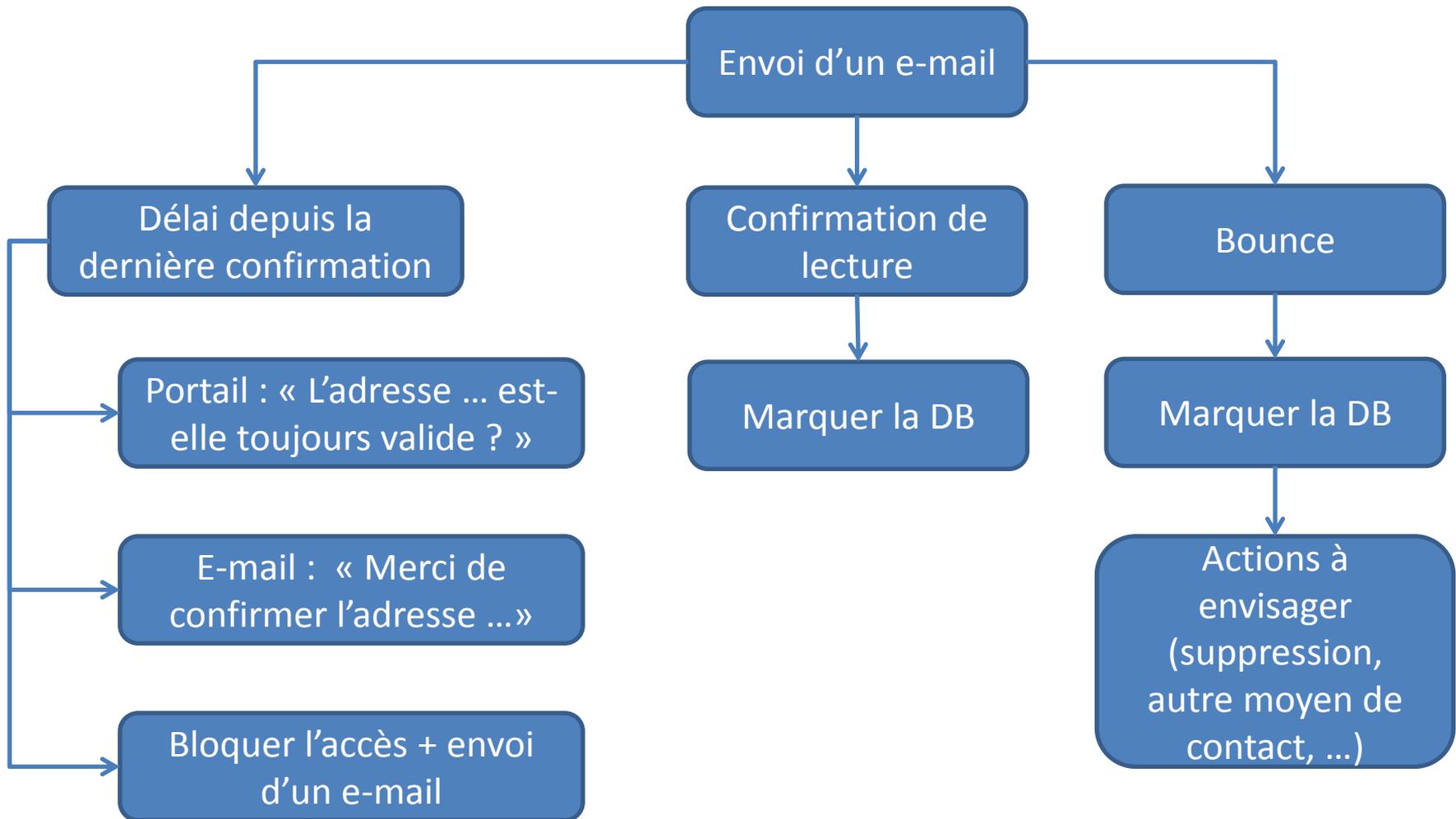
Bonnes pratiques & Conclusions

Bonnes pratiques : encodage/update

- Scénarios analogues à prévoir :
- Confirmation jamais cliquée
 - Cas d'adresse erronée



Bonnes pratiques : envoi d'un e-mail



Adresse e-mail pour chaque citoyen ?

- Proposition souvent entendue : attribuer une adresse e-mail à chaque citoyen
- Il est alors responsable de la consulter ou de la transférer
- Répandu ? À notre connaissance, seulement en :
 - Turquie
 - Iran
 - Malaisie
- Est-ce si simple que ça ?



Adresse e-mail pour chaque citoyen ?

- Hypothèse 1 : albert.leroy@citizen.be
- Problèmes :
 - Quid des homonymes ?
 - Peut-on réattribuer (décès, perte de nationalité ?)
 - Plus possible de garder son adresse privée
 - Que faire en cas d'harcèlement, de spam ? Comment changer de « vie numérique »
 - Comment fait Di Rupo pour recevoir ses documents ?



Adresse e-mail pour chaque citoyen ?

- Hypothèse 2 : 80.09.10-125.57@citizen.be
- Problèmes :
 - Pas très convivial ...
 - On peut spammer toute la Belgique, ou tous les hommes de + de 50 ans
- Hypothèse 3 : 3548680686@citizen.be
- Problèmes :
 - Encore moins convivial, difficile à diffuser (Banque, boulot, privé...)
 - Longueur > 8, sinon spam facile
 - Presque nécessaire de faire un transfert (« .forward »), donc autant renseigner son adresse dans une « e-box »



Conclusions

Peu de déterminisme,
beaucoup
d'incertitude

Souvent,
gains
importants

Difficulté :
dégressivité
dans le temps

Syntaxe
spécifique :
améliore la
qualité

Complexité
de la gestion

Catégorisation
binaire : laxiste
ou restrictive
→ Suspensions

Importance d'un suivi continu
et d'une bonne organisation

Pro-action !



Vandy Berten
vandy.bersten@smals.be



Isabelle Boydens
isabelle.boydens@smals.be



More on Smals Research
Website Smals : www.smals.be
Website Research : www.smalsresearch.be
Twitter : [@SmalsResearch](https://twitter.com/SmalsResearch)



Annexes



Références

- I. Boydens, «*Strategic Issues Relating to Data Quality for E-government: Learning from an Approach Adopted in Belgium*» In « Practical Studies in E-Government: Best Practices from Around the World », New York, Springer, pp. 113-130 (chapitre 7), 2011.
- Y. Bontemps, I. Boydens et D. Van Dromme, «*Data Quality: tools*», Bruxelles, Smals, 2007.
- I. Boydens, «*Informatique, normes et temps*», Bruxelles, Bruylant, 1999.
- I. Boydens, A. Hulstaert et D. Van Dromme, «*Gestion intégrée des anomalies*», Bruxelles, Smals, 2011.



Glossaire

- DNS : Domain Name System
- gTLD : generic Top Level Domain
- IDN : Internationalized Domain Name
- MTA : Mail Transfer Agent
- MX : Mail eXchange
- SMTP : Simple Mail Transfer Protocol
- TLD : Top Level Domain

