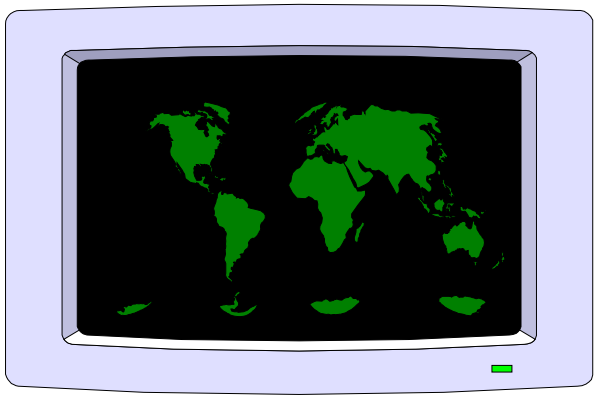
Internationalisation et multilinguisme du WWW



François Yergeau

Alis Technologies inc.

SÉMINAIRE ARISTOTE « WWW: PROSPECTIVE ET PERSPECTIVES »
24 Octobre 1996, École Polytechnique, Palaiseau, France

alis

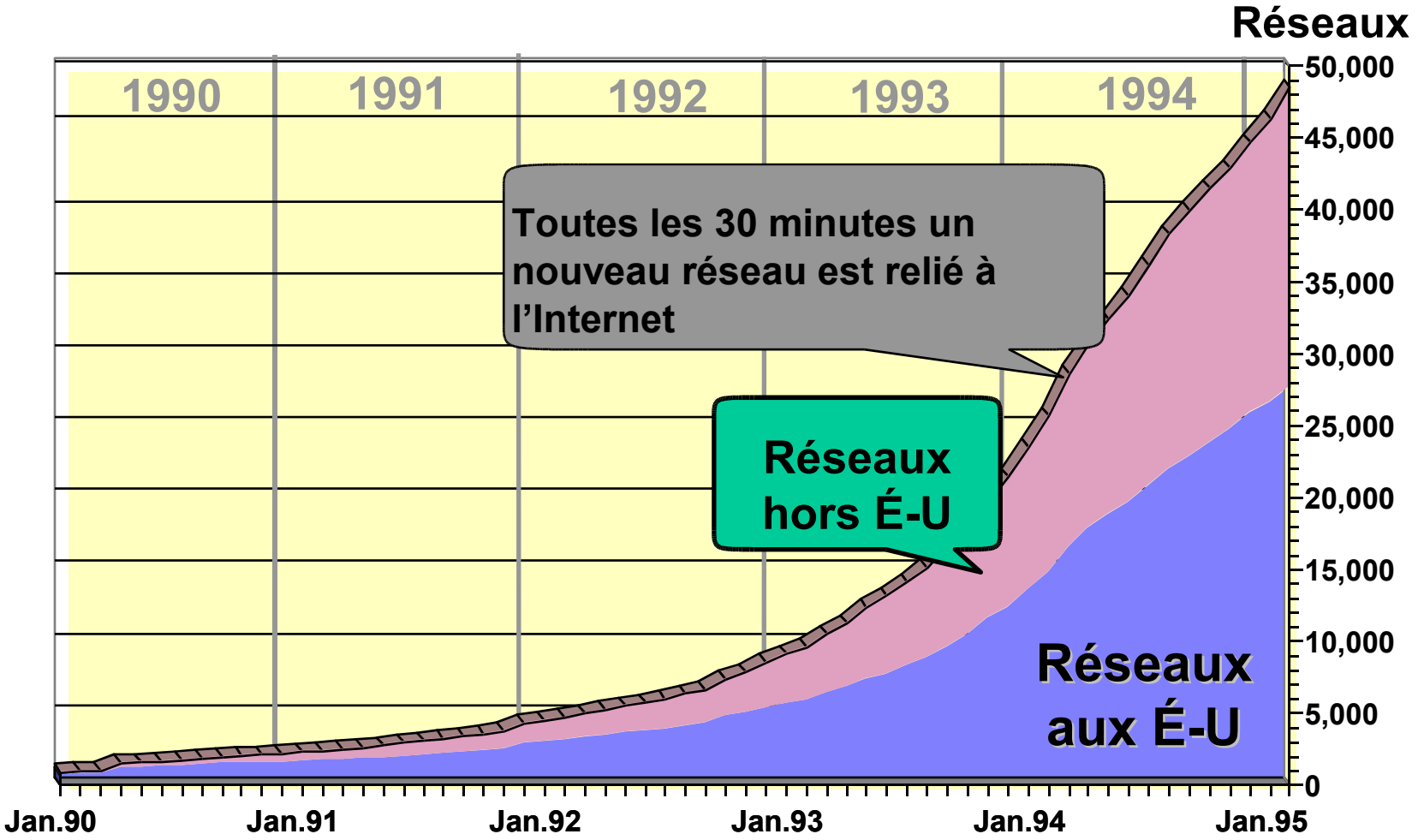
Introduction

- Le WWW est de plus en plus planétaire
- nombre croissant de non-anglophones
- peu de gens sont des bilingues fonctionnels
- apparition d'un ressentiment croissant quant à l'hégémonie d'une seule langue ou culture

➨ la diversité linguistique du Web devra bientôt réfléter celle de la planète.

alis

Internationalisation croissante



Source: ftp://nic.merit.edu/statistics/nsfnet

alis

Une analogie ...

Les conférences internationales unilingues :

- 50% des scientifiques n'y assistent pas
- 50% des assistants sont réduits à la passivité par leur mauvaise maîtrise de l'anglais
- 25% seulement peuvent pleinement participer

➨ **Perte pour la communauté scientifique et les exclus**

Source : M. & J. Guillemat, étude commanditée par l'association internationale des villes francophones de congrès, septembre 1991.

alis

Le décor

- Nombreuses langues, écrites en différents alphabets
- HTML, HTTP fondés sur Latin-1
- URL fondés sur ASCII !
- Résultat : rafistolages, trucs, incohérence, incompatibilité
- Seul l'anglais est garanti de passer partout

alis

HTTP

- HTTP a au moins la vertu de garantir la transmission à 8 bits (intégrité des données)
- Au début (CERN), on a réglé le problème des JdC incompatibles en imposant Latin-1
- Cette norme ne suffit plus aujourd'hui, mais HTTP la retient toujours comme défaut
- Conséquence : les serveurs ne savent pas identifier le JdC (paramètre *charset*)
- Au moins, les clients d'aujourd'hui tolèrent ce paramètre

HTML

- Conçu à l'origine pour Latin-1 seulement
- Existence d'entités (*é et C^ie^*) qui font croire que seul l'ASCII est admis
- Exigences de l'i18n :
  - *Tous* les caractères du monde
  - Compatibilité arrière
  - SGML : un seul JdC de document dans la déclaration SGML
  - HTTP : transmission efficace
  - Utilisateurs : jeux de caractères locaux
- Manque aussi du balisage pour application vraiment globale

alis

i18n d'HTML - RFC XXXX

- Adoption de l'ISO 10646 (Unicode) comme seul jeu de caractères de document pour HTML
- Modèle de référence de traitement
- Compatibilité : n'interdit pas les autres JdC
- Balisage de langue
- Balisage supplémentaire pour rendu BIDI et contextuel
- Quelques « cadeaux »

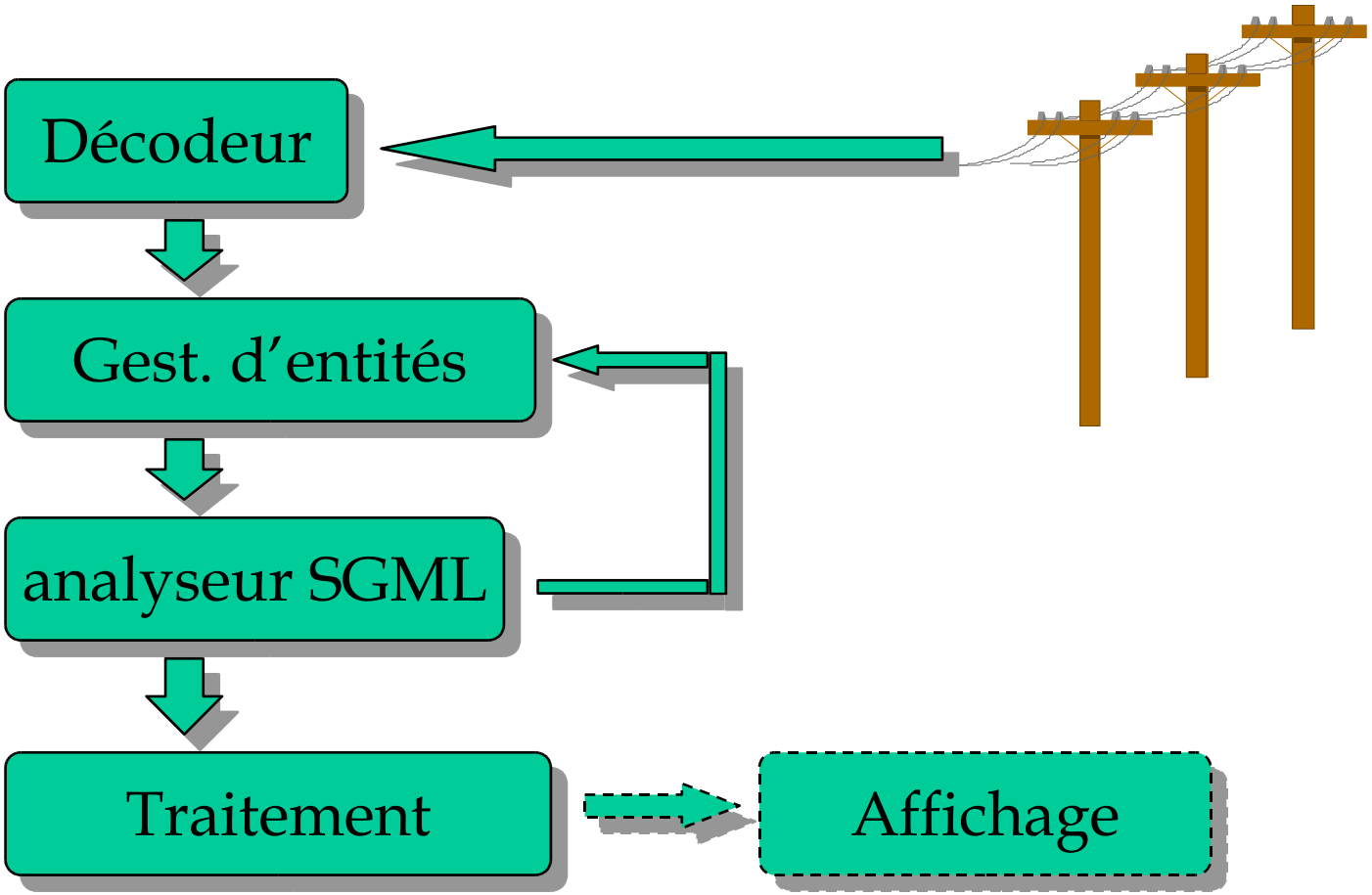
alis

Codage d'un document

- Séquence de caractères abstraits : l'information à coder
- Codage par un ou plusieurs jeux de caractères (association d'un répertoire avec des nombres entiers) ➠ séquence de nombres entiers
- Codage multi-octets : assemblage des JdC avec transitions explicite (ESC, 8e bit, etc.) ➠ séquence d'octets prête à transmettre, *charset* MIME
- Parfois, surcodage de transport (*compress*, *QP*, chiffrement, etc.) ➠ séquence d'octets brute

alis

Modèle de référence



alis

Modèle de référence (suite)

- **Rend univoques les références numériques :**
  - **Référence à Unicode seulement**
  - ***é* reste toujours un E ACCENT AIGU, jamais un IOU cyrillique (*ю*) ou un IOTA grec (*ι*)**
- **Permet le transcodage sans analyse lexicale**
- **Permet d'utiliser n'importe quel caractère, sans égard au *charset* :**
  - **caractères cyrillique, chinois, etc. dans document Latin-1 ;**
  - **lettres latines accentuées dans document japonais**

alis

Références numériques de caractères

- **Les RNC réfèrent au JUC**
- `&#149;` **est illégal. Pas d'entités pour :**

132 8222 „	140 338 Œ	149 8226 •
133 8230 …	145 8216 ‘	151 8212 —
134 8224 †	146 8217 ’	153 8482 ™
135 8225 ‡	147 8220 “	155 8250 ›
139 8249 ‹	148 8221 ”	156 339 œ

alis

Balises de langue

- **La plupart des éléments HTML admettent le nouvel attribut** LANG

  <SPAN LANG="it">**Grazie**</SPAN>

- <SPAN> **est un nouveau conteneur en-ligne générique, qui n'a de sens que celui de ses attributs** (LANG, DIR, ID **et** CLASS)

- **L'ancien** <LANG> **(HTML3) disparaît**

alis

Balises de langue (suite)

- LANG **aide à:**
  - **lever les ambiguïtés sur les glyphes (« défaire » l'unification Han)**
  - **contrôler la césure, les guillemets, l'espacement, les ligatures**
  - **permettre la synthèse de parole, le Braille, etc.**
  - **contrôler la classification, la recherche et le tri**

alis

Écritures cursives

- **Liaison cursive : introduction de** `&zwj` **(liant sans chasse) et** `&zwnj` **(anti-liant sans chasse) pour comportement « anormal »**
- **HEH ARABE = ه, ressemble trop à ٥ (5), on prend plutôt la forme initiale هـ =** `ه&zwj;` **pour abbrévier Hijri (le calendrier islamique)**

alis

Écritures bidirectionnelles

- **Les caractères ont tous une direction implicite, mais...**
- **...certains cas ont besoin de balisage pour suppléer l'algorithme implicite :**
  - **caractères neutres en contexte ambigü**
  - **enchâssements multiples**
  - **copier/coller**
  - **numéros de pièce**
- `&lrm;` `&rlm;` **attribut** `DIR`**, élément** `<BDO>` **suppléent les caractères de formattage Unicode**

alis

Indices et exposants

- **Les exposants sont nécessaires en français :**

M^{**Ile**} **= M^{Ile}**

- **En prime, les indices :**

x_{**i+1**} **= x$_{i+1}$**

alis

Justification de paragraphes

- **La justification des paragraphes est très importante en certaines langues**
- **D'où l'attribut** `ALIGN` **sur certains éléments de type bloc :** `P`, `H1...H6`, `LI`, **etc...**
- `ALIGN = LEFT|RIGHT|CENTER|JUSTIFY`
- **La valeur implicite dépend de** `DIR` **:**
  - `DIR=LTR` ➡ `LEFT`
  - `DIR=RTL` ➡ `RIGHT`

Citations

Les courtes citations en-ligne sont rendues différemment en différentes langues, à divers niveaux d'imbrication.

En contexte français :

<Q LANG=en>He said: <Q>ja!</Q></Q>

« He said: “ja!” »

alis

Citations (suite)

»Dansk ’da’ Danois«

„Deutsch ‘de’ Allemand”

“English ‘en’ Anglais”

« Français « fr » Français »

‘Nederlands “nl” Néerlandais’

«Norsk ’no’ Norvégien»

«Russe „ru” Russe»

alis

Formulaires

- **Le client doit connaître le(s) codage(s) que le serveur est prêt à accepter : attribut** `ACCEPT_CHARSET` **sur** `<INPUT>` **and** `<TEXTAREA>`
- **Exemple:**

```
<INPUT TYPE=text ACCEPT_CHARSET=
   "iso-8859-1,cp1252">
```

Formulaires (suite)

- **Le serveur doit savoir quel codage il reçoit :**
  - GET inutilisable, pas de *charset* dans les URL (entité HTTP possible, mais pas supporté)
  - POST mieux, si x-www-url-encoded avait un paramètre *charset*, mais :
    - **nuit aux mandataires (réparable)**
    - **pas de signets POST (réparable)**
  - RFC 1867 (multipart/form-data) encore mieux : chaque partie a étiquettes de *type* et *charset*

alis

Formulaires (suite)

Attention : la valeur implicite d'un champ de formulaire (attribut `VALUE`) peut revenir comme une séquence d'octets différente :

- Mêmes caractères, différents octets
- Séquence de caractères composite différente mais équivalente, par exemple E ACCENT AIGU versus E + DIACRITIQUE ACCENT AIGU

alis

Replis stratégiques

- **Attribut** `CHARSET` **sur les liens :**

  `<A HREF=... CHARSET="UTF-8">...</A>`

  → Aussi dans les signets

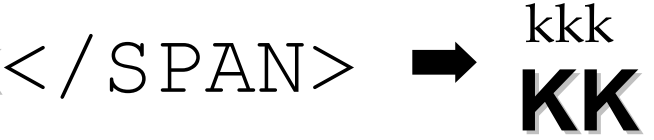
- **Document auto-étiqueté :**

  <META HTTP-EQUIV="Content-Type"
  CONTENTS="text/html; charset=UTF-8">

  ✗ Ne marche pas toujours

alis

Pour la soif : Ruby

- **Les caractères Ruby sont de petites aides à la prononciation placées au-dessus des idéogrammes**
- `<SPAN RUBY="kkk">KK</SPAN>` ➡ 

- **Cette syntaxe a des problèmes avec l'association kana-kanji et le bris de ligne**
- **Voir draft-duerst-ruby-00.txt**

alis

Pour la soif : césure

- **Certaines langues ont besoin de plus que** `&shy;` **pour la césure ; il faut un algorithme ou même un dictionnaire**
  - ↙ **Zucker ➡ Zuk - ker**
- **Propositions tardives, toutes problématiques, trop tard pour notre RFC**
- `&shy;` **même pas correctement supporté**

alis

Négociation HTTP

- **HTTP 1.1 permet (enfin !) la négociation de type MIME, de langue et de surcodage**
- **Un URL = un document, plusieurs versions**
- **En-tête Accept-Language :**

```
Accept-Language: fr, it;q=0.9
```

- **La plupart des fureteurs modernes le transmette**
- **Peu de serveurs s'en occupent**

alis

Serveur W³ internationalisé

- **Alis offre gratuitement un serveur NCSA, httpd 1.4 (bientôt 1.5) internationalisé**

`http://www.alis.com/P_NET/ncsa.html`

  - permet l'envoi de documents dans la langue préférée de l'utilisateur
  - transmet correctement le paramètre *charset*

- **Bientôt, Apache fera de même (déjà négociation)**

alis

Serveur W³ internationalisé

GET /dir1/dir2/index.html HTTP/1.0

Accept-Language:

fr ; q=1.0
en ; q=0.9
ru ; q=0.8

- index.fr.html
- index.en.html
- index.ru.html

Pas trouvé ? Repli sur :

- index.html

alis

Serveur W³ internationalisé

- **Le serveur connaît *charset* par le biais d'un en-tête ad hoc greffé aux fichiers texte :**

```
<!--Des_champs_pseudo-MIME_suivent
Content-Type: text/html;charset=xxx
Content-Language: fr
-->
```

- **Cet en-tête n'est pas transmis**
- **Valeur implicite ajustable pour les fichiers sans en-tête**

alis

Internationalisation des URL

`http://www.truc.dom/cgi/form?var1=val1&var2=val2`

- `http:` = protocole, ASCII
- `//www.truc.dom` = nom de domaine, DNS, ASCII
- `/cgi/form` = chemin d'accès
- `?var1=val1&var2=val2` = requête

8 bits ? JdC indéfini, doit utiliser %XX

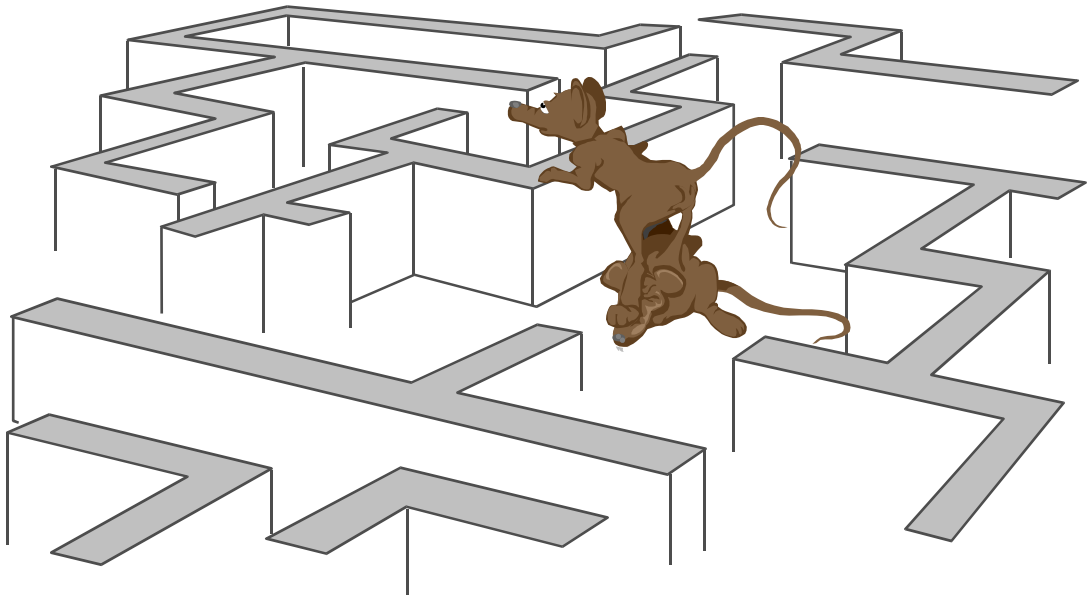
alis

Internationalisation des URL

- 2 types de solution :
  - Indiquer le codage
  - Codage universel
- Compatibilité arrière impossible dans le premier cas
- Deuxième cas, UTF-8 assure la compatibilité en conservant l'ASCII
- Consensus sur UTF-8, avec ou sans %XX, lors de la dernière conférence Unicode

alis

Trouver l'information



Recherche dans le Web

Pas de *charset*, pas de recherche fiable sauf en anglais

Manque de balisage linguistique :

Un instituteur allemand cherche « New-York » pour ses élèves : hécatombe d'information inutilisable !

alis

Raisons d'internationaliser

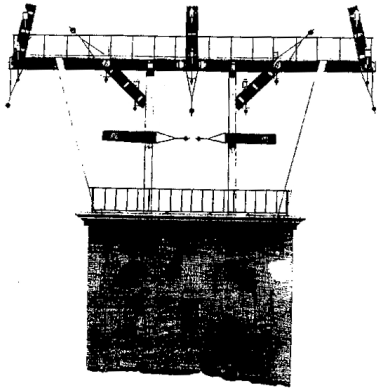
- Un logiciel multilingue permet à une langue de tenir sa place ailleurs que là où elle est dominante
- Pour éviter la prolifération des versions
- Pour rentabiliser l'investissement
- Internationalisation signifie :
  - localisation rapide et économique
  - universalité, vendable partout
  - aucune niche laissée à la concurrence

alis

Retrouver l'esprit de la première «autoroute» numérique !

Claude Chappe

- Né a Brûlon (200 km de Paris) en 1763
- Créateur du ***premier*** réseau télégraphique optique international



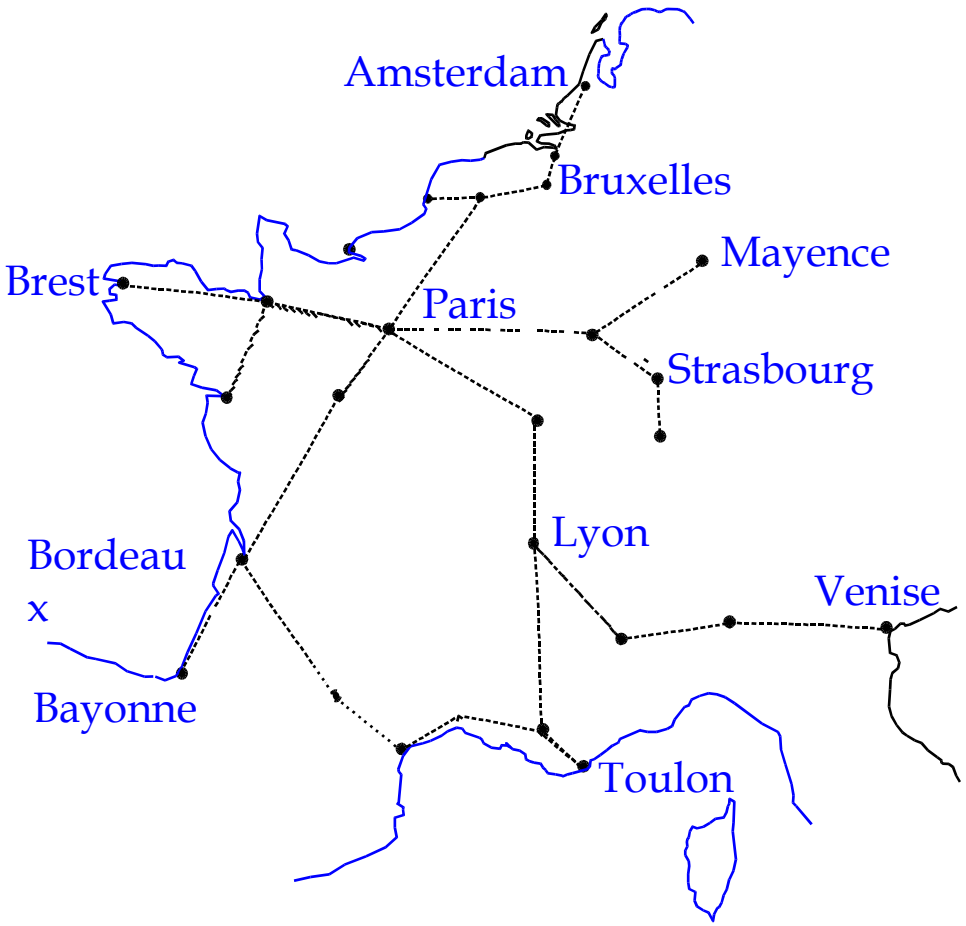
Message transmis le 2 mars 1791 sur 16 km:

« L'Assemblée Nationale récompensera les expériences utiles au public ».

Durée de transmission: 6 minutes 20 secondes

alis

Une leçon d'histoire...



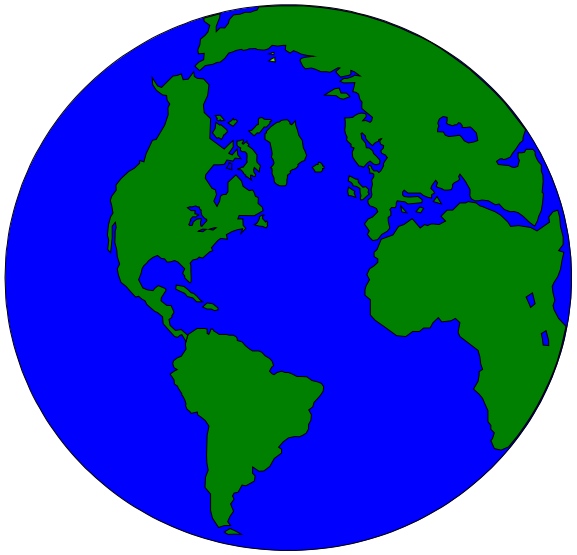
En 1852...

- 556 stations
- 4800 km
- 27 grandes villes reliées

Obstacles vaincus :

- scepticisme de ses concitoyens
- absence de consensus
- appui politique aléatoire
- financement initial insuffisant

alis



Dank u

Merci

Grazie

Gracias

Danke

Alis Technologies inc.

100, boul. Alexis-Nihon
Bureau 600
Montréal QC H4T 1A7

+1 (514) 747-2547
+1 (514) 747-2561 (fax)

info@alis.com

alis