INRIA
INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE
Projet M.I.S.T.R.A.L.

Stage de D.E.A.

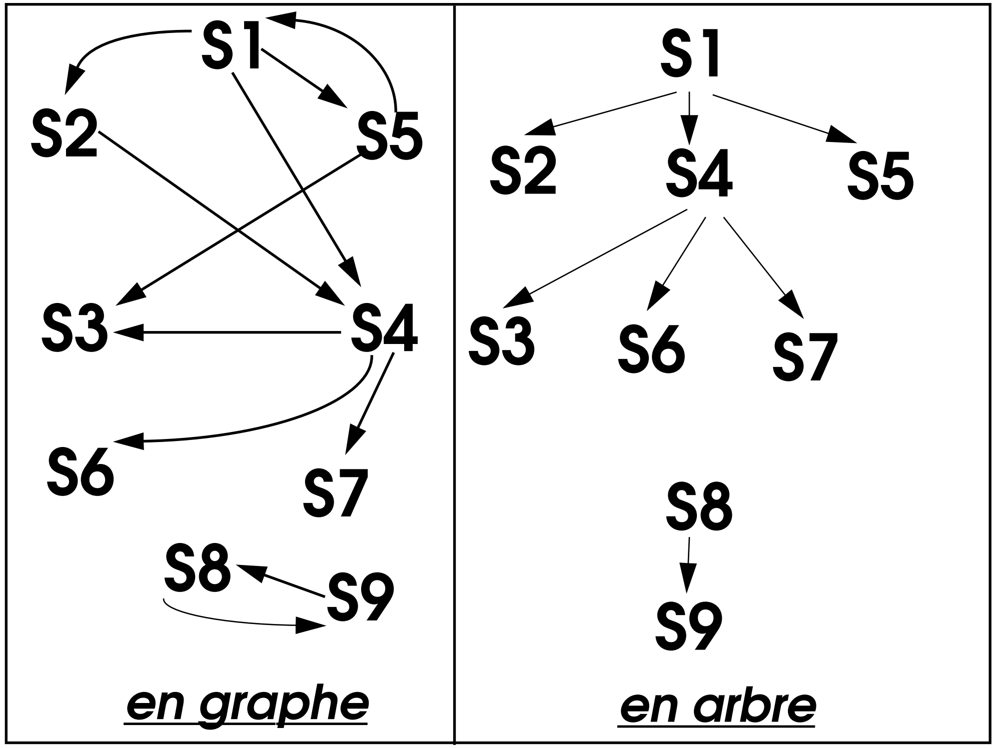
Moteur de recherche d'informations sur le Web

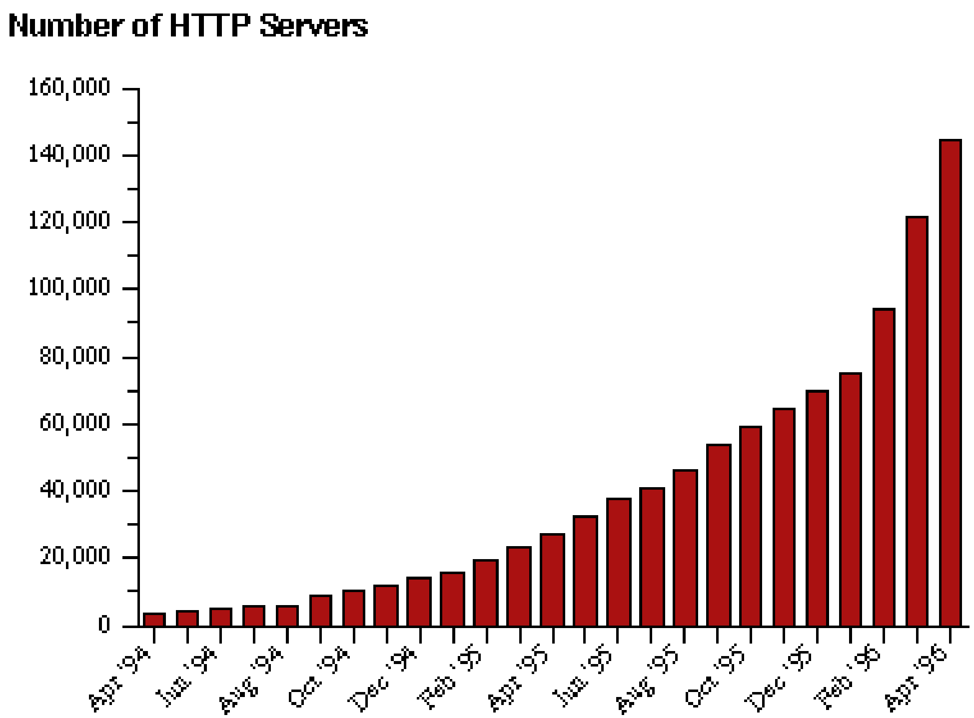
Vincent Morgera

Juin 1997

Introduction : l'accroissement du WORLD WIDE WEB

- Une formidable réussite
- Un outil commerçial
- L'*anarchie* du Web
- Une représentation simplifiée





INRIA

1. Vers l'utilisation des moteurs de recherche

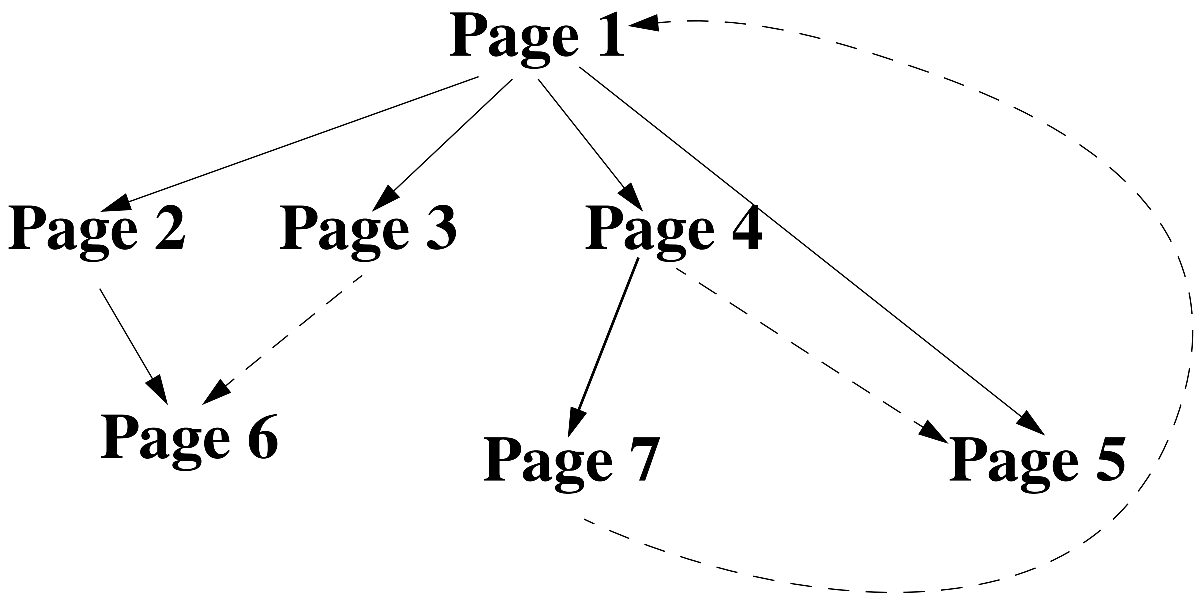
- Les recherches de WebCrawler
- Le travail d'un robot
  - *HyperText Transfert Protocol*
  - *L'analyse d'une page*

```
<html>
  <head><title> exemple de page HTML</title></head>
    <body>
      <h1> paragraphe 1</h1>
      ...
      <a href="page1.html">lien sur une page</a>
      <a href="page2.html">lien sur une autre page</a>
      ...
    </body>
</html>
```

  - *Les techniques d'indexation*
  - Les limites : le problème de politesse des robots

2. Le travail demandé

- Le développement d'un robot
  - Les recherches de W3 Consortium
  - Un exemple de parcours

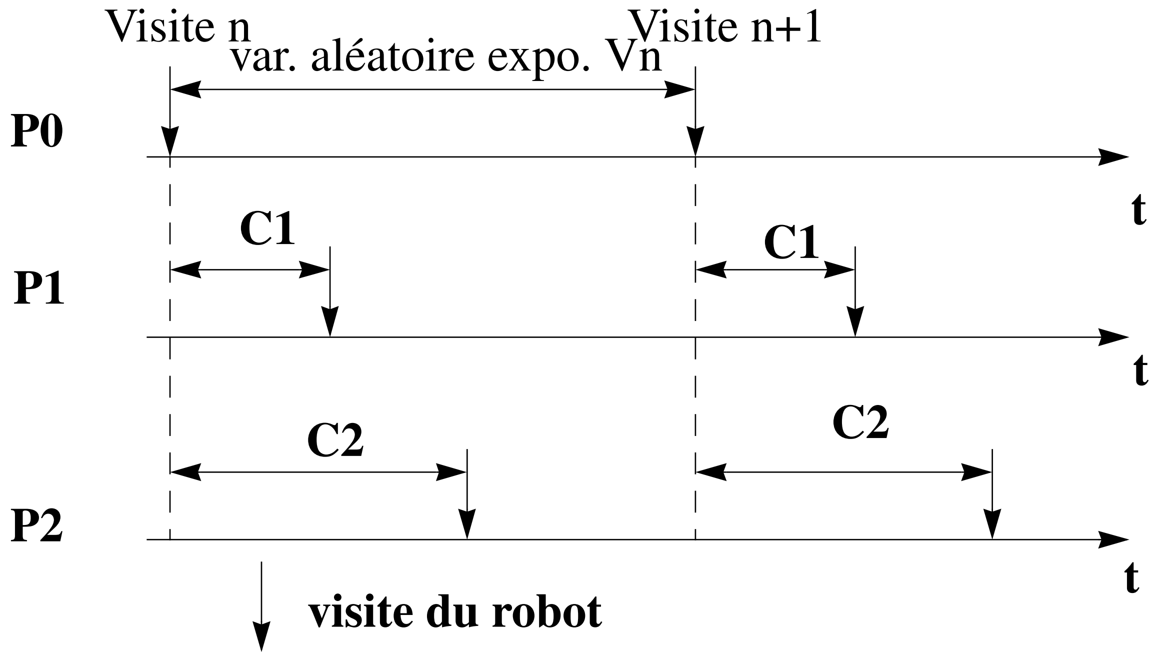


2. Le travail demandé (suite)

- Que doit-on visiter ?
  - Se limiter au même site
  - Les formats HTML
  - Les images
  - Les programmes CGI

- les spécificités du robot
  - Les données rapatriées
  - Le stockage dans la Base de Données
  - La notion de processus Poisson

Le processus Poisson

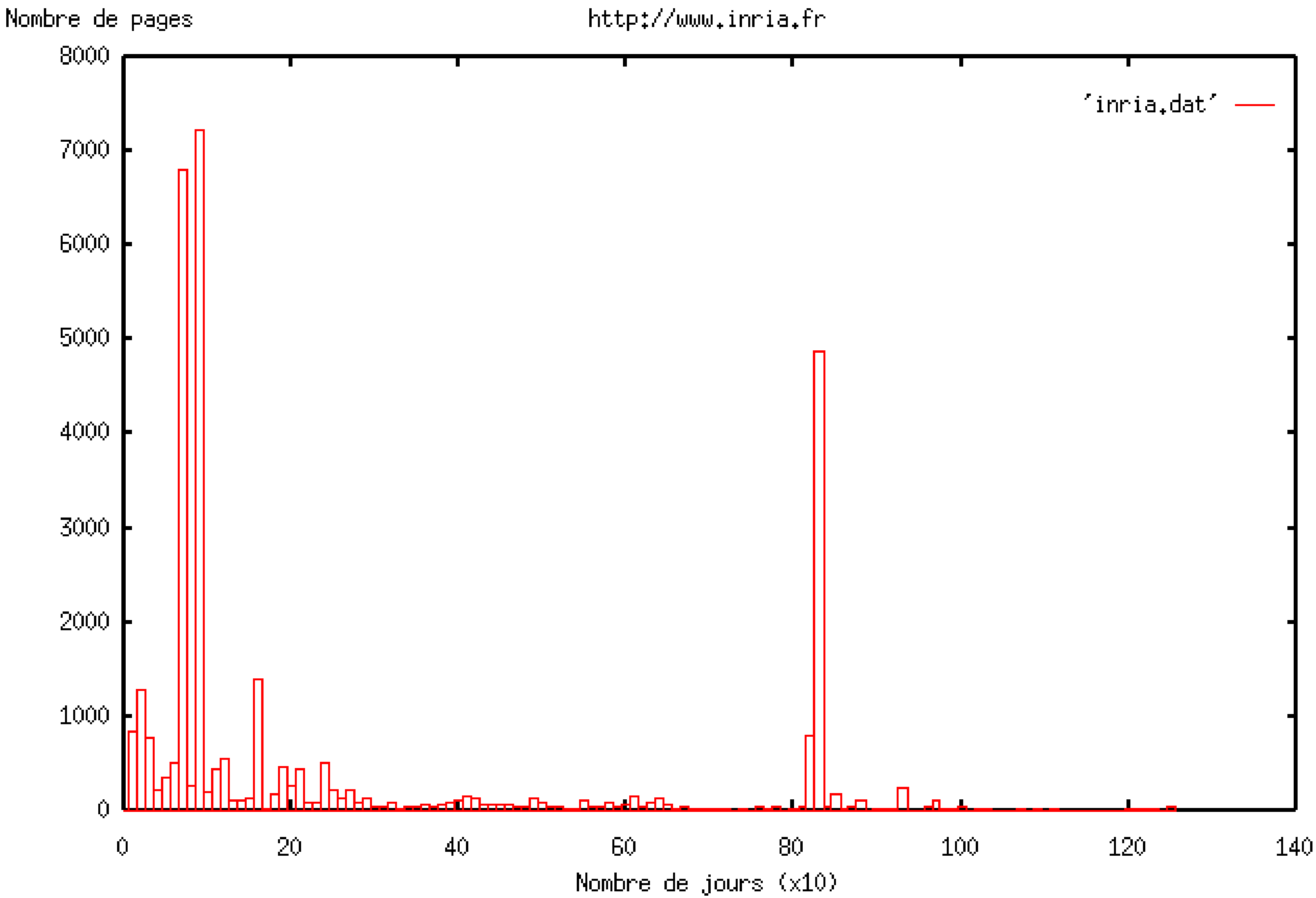


- Pourquoi ?
- Comment faire ?
  - La hiérarchie d'un site
  - Synchronisation des visites
  - Exécution du robot suivant des durées aléatoires exponentielles

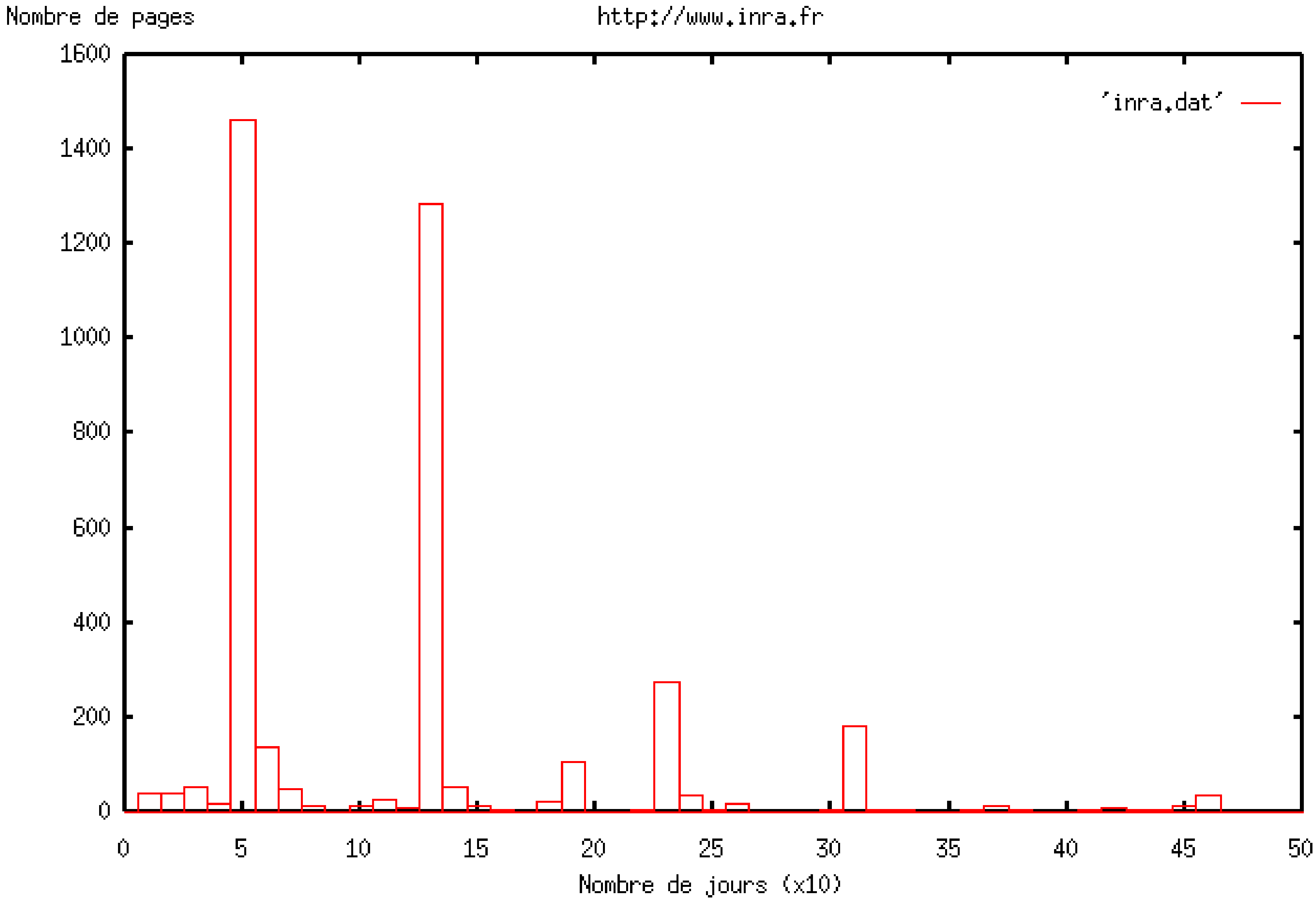
3. Les résultats obtenus

- L'I.N.R.I.A.
  - Taille : 16640 pages
  - Espérance : 271 jours
- L'I.N.R.A.
  - Taille : 2718 pages
  - Espérance : 113 jours
- France Télévision
  - Taille : 808 (France2) et 1027 (France3)
  - Espérance : 117 et 120 jours
- Nice Matin
  - Taille : 167 pages
  - Espérance : 100 jours

Histogramme de fréquence du serveur de l'I.N.R.I.A.

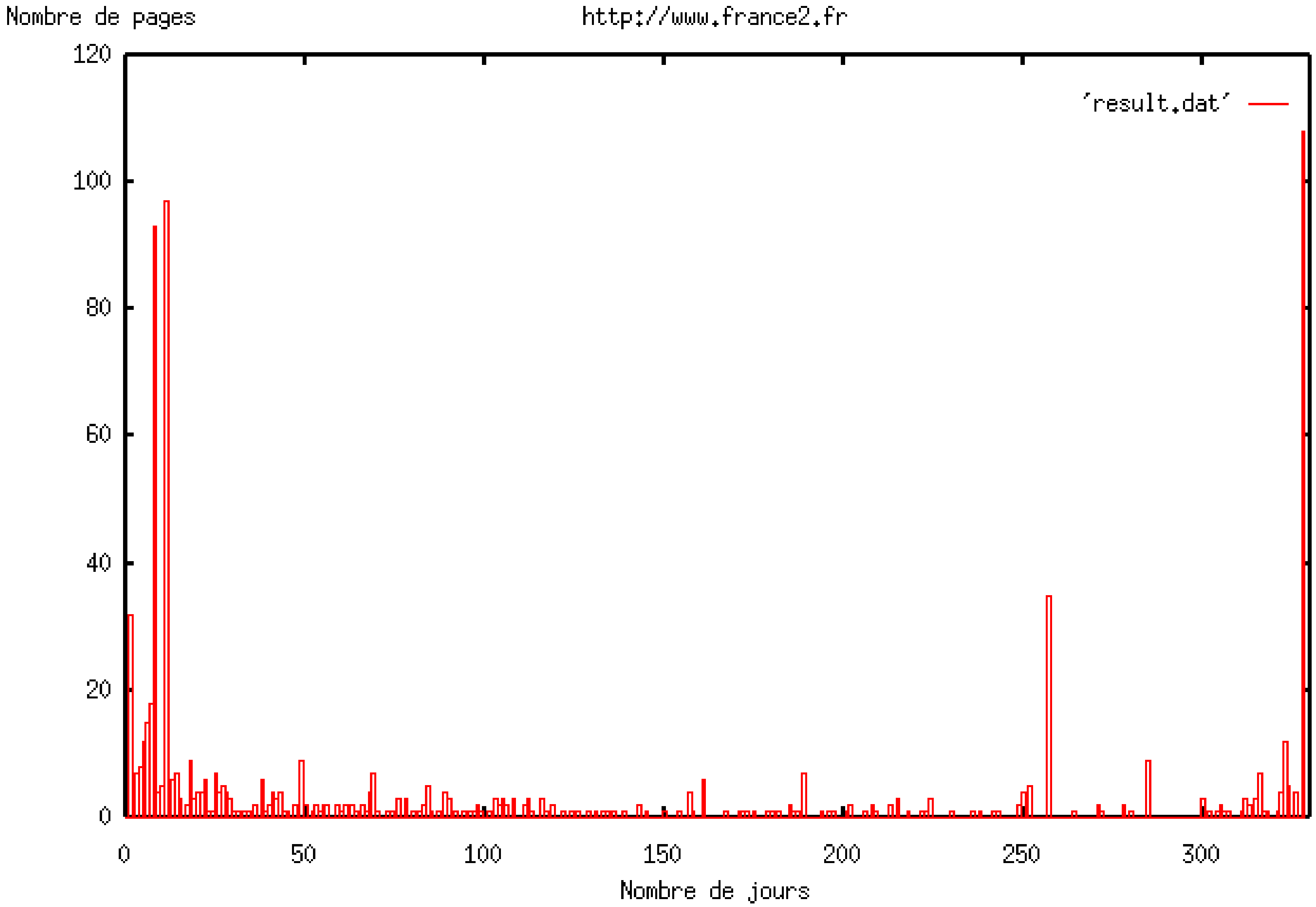


Histogramme de fréquences du serveur de l'I.N.R.A.



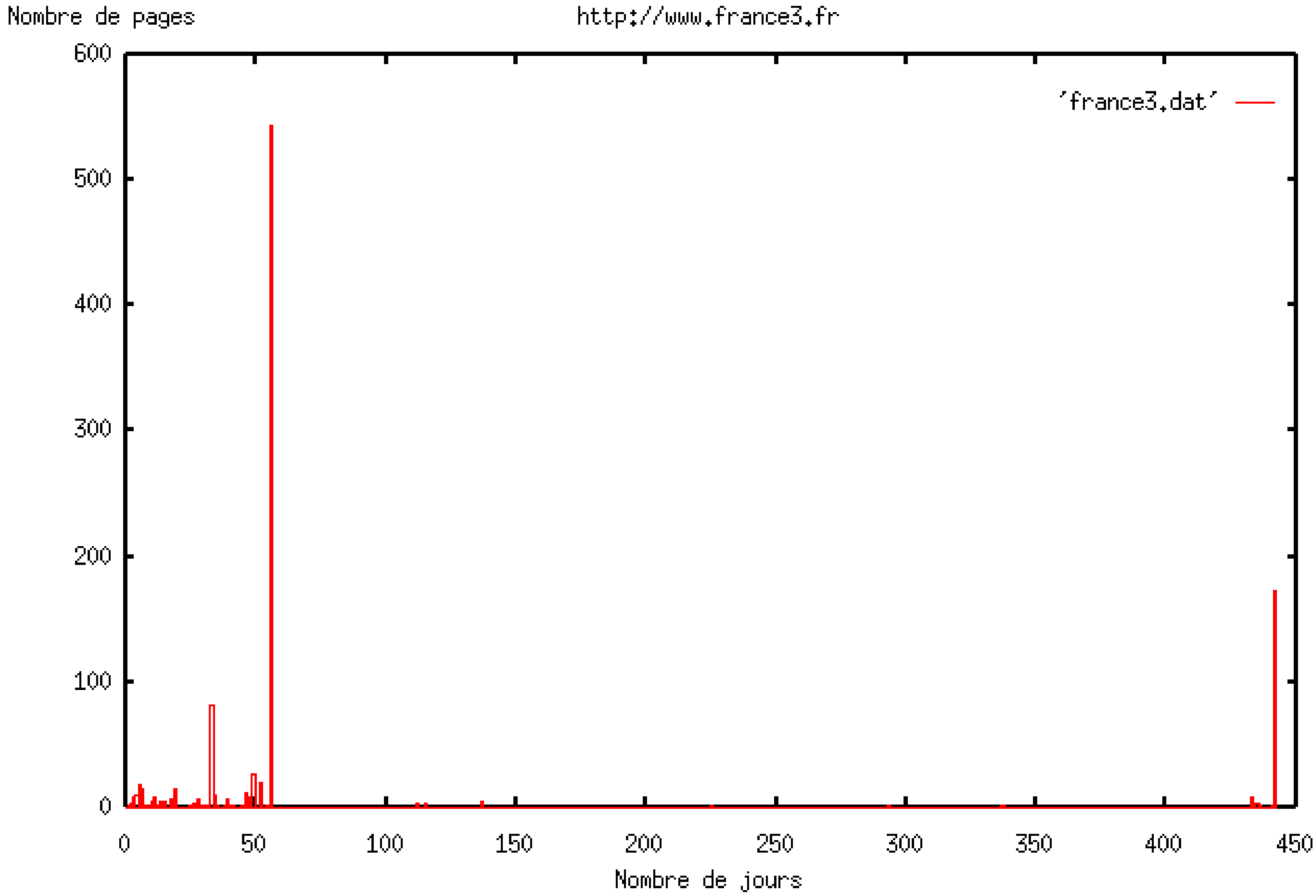
INRIA

Histogramme de fréquences du serveur de France 2

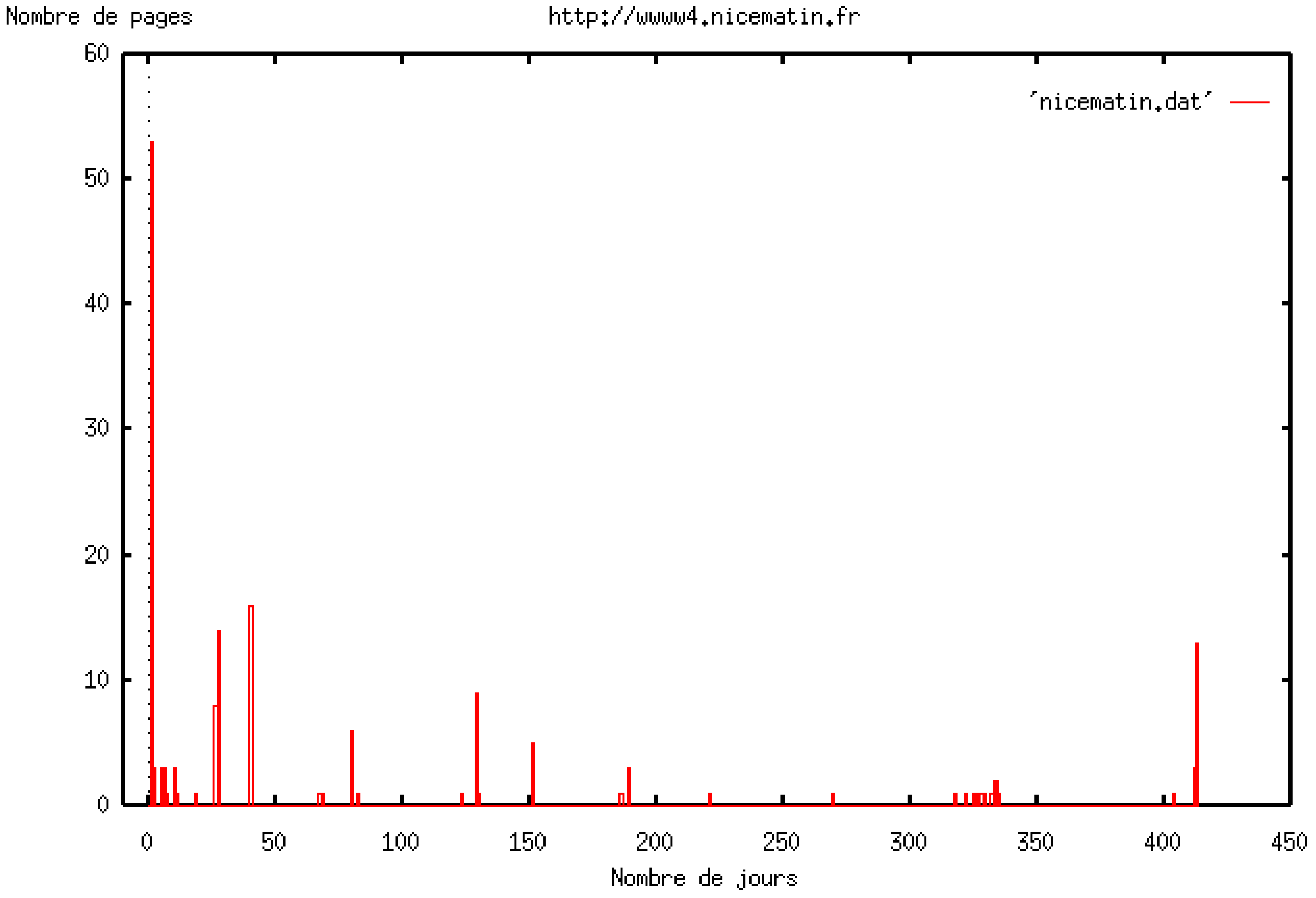


INRIA

Histogramme de fréquences du serveur de France 3



Histogramme de fréquences du serveur de Nice Matin



INRIA

4. Les conclusions

- Les fréquences obtenues ne sont pas représentatives des serveurs
- Solution : décomposer les sites
  - de façon manuelle (étude des histogrammes)
  - de façon statistique (étude sur une plus longue période)