ALIS勉強会
～岡崎市立図書館事件、通称Librahack事件について

2010/10/27

筑波大学メディアユニオン3F

共同研究会議室

- 当日はここで、杉谷智宏氏（@Vipper_The_NEET）が作成されたスライド、「よくわかる岡崎市立中央図書館事件の流れ」に基づいて事件の概要説明が行われました
- 杉谷氏によるスライド：
  - http://www26.atwiki.jp/librahack/pages/26.html

設問1

librahack氏のような
アクセスの仕方は特殊なのか？

そう思う

そうは
思わない

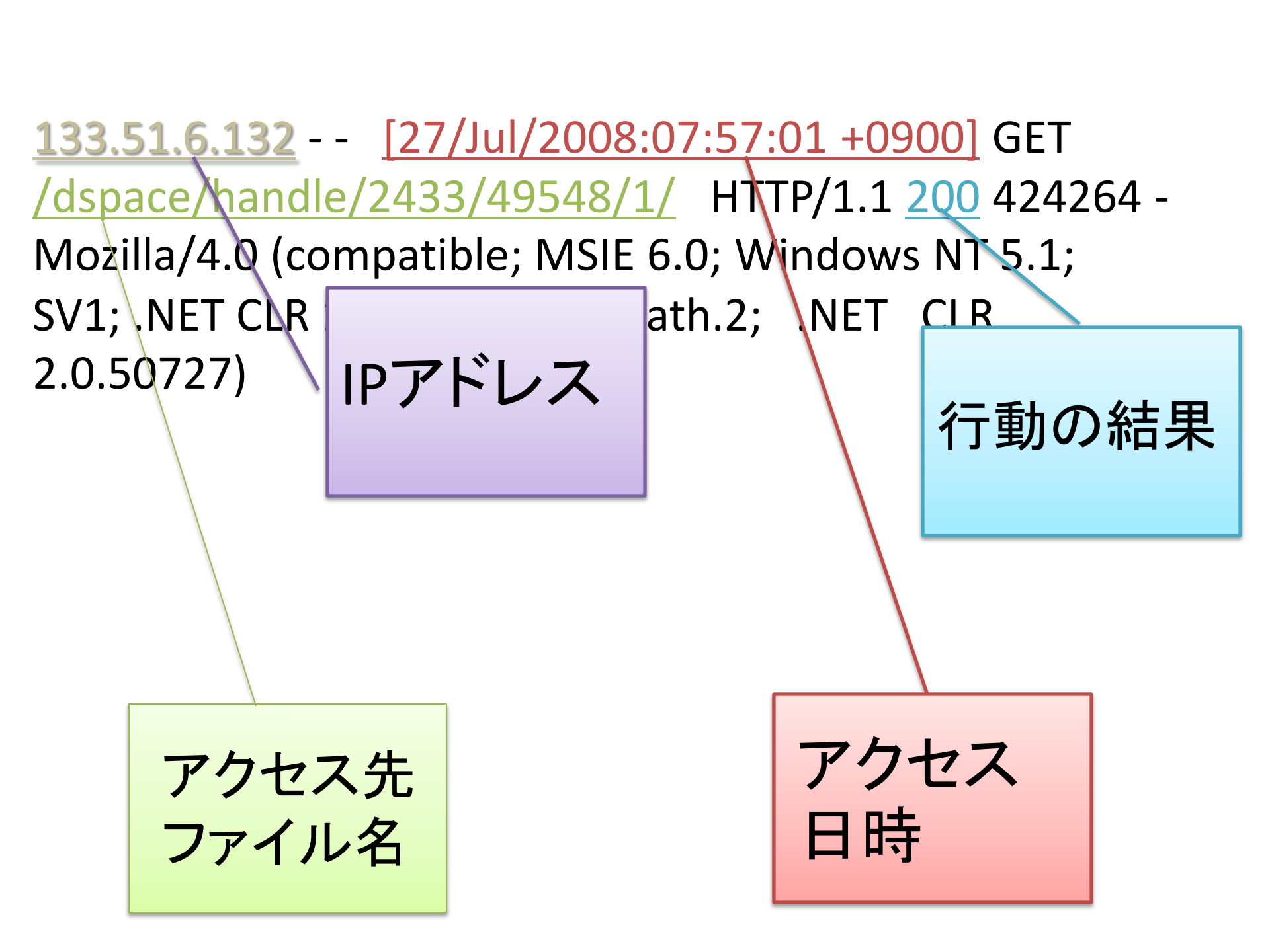
アクセスログとは？

133.51.6.132 - - [27/Jul/2008:07:57:01 +0900] " GET /dspace/handle/2433/49548/1/ HTTP/1.1 " 200 424264 " - " “Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322; InfoPath.2; .NET CLR 2.0.50727)”

66.249.70.136 - - [01/Jul/2008:04:02:10 +0900] "GET /dspace/simple-search?query=purification&start=90 HTTP/1.1" 200 32835 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)”

.
.
.

133.51.6.132 - - [27/Jul/2008:07:57:01 +0900] GET /dspace/handle/2433/49548/1/ HTTP/1.1 200 424264 - Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR ath.2; .NET CLR 2.0.50727)



アクセスログとは？

133.51.6.132 - - [27/Jul/2008:07:57:01 +0900] " GET /dspace/handle/2433/49548/1/ HTTP/1.1 " 200 424264 " - " “Mozilla/4.0 (compatible; MSIE 6.0; Windows NT 5.1; SV1; .NET CLR 1.1.4322; InfoPath.2; .NET CLR 2.0.50727)”

66.249.70.136 - - [01/Jul/2008:04:02:10 +0900] "GET /dspace/simple-search?query=purification&start=90 HTTP/1.1" 200 32835 "-" "Mozilla/5.0 (compatible; Googlebot/2.1; +http://www.google.com/bot.html)”

.
.
.

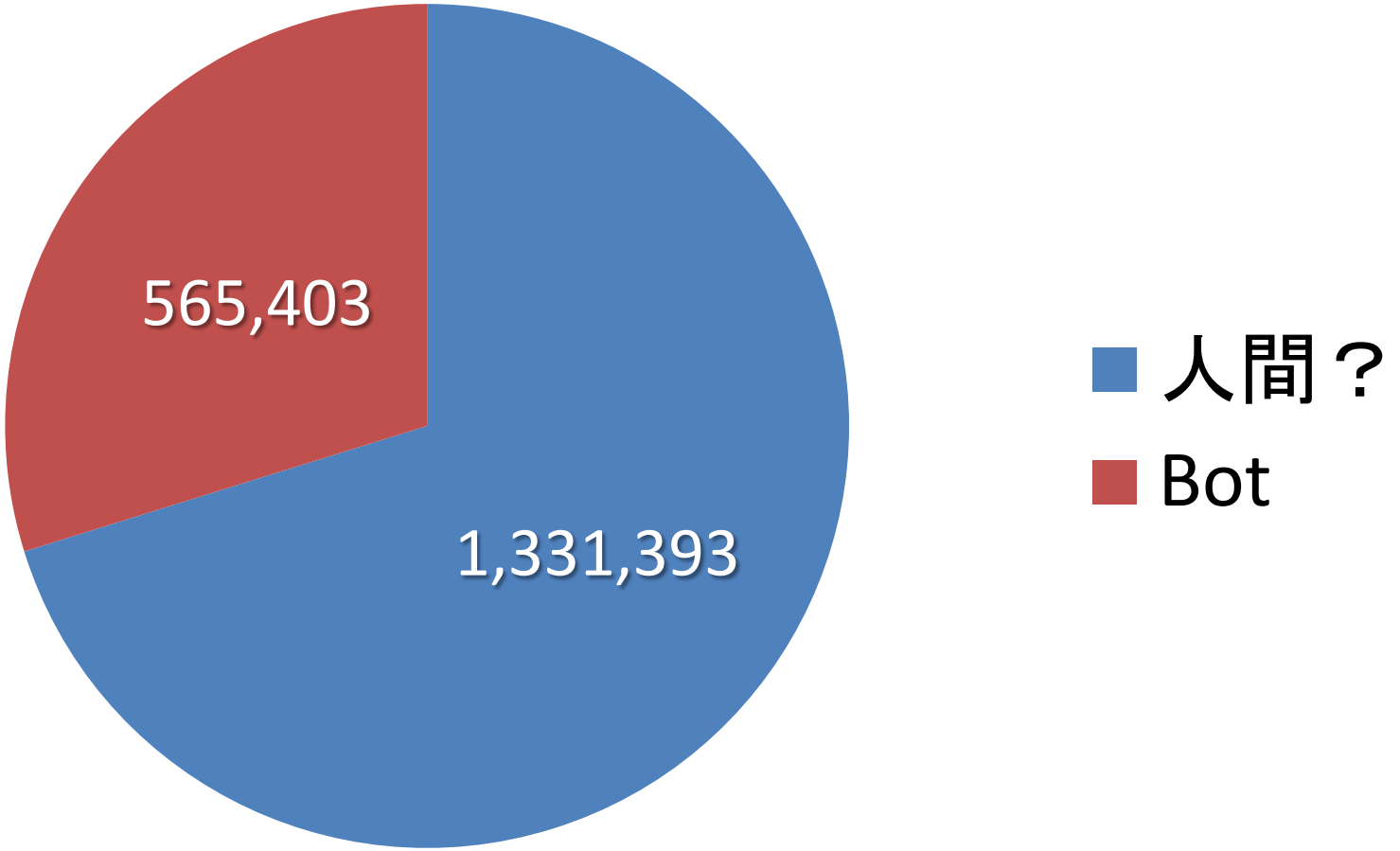


図. 筑波大学附属図書館Webサイト：2009.11.30-12.15のアクセス状況

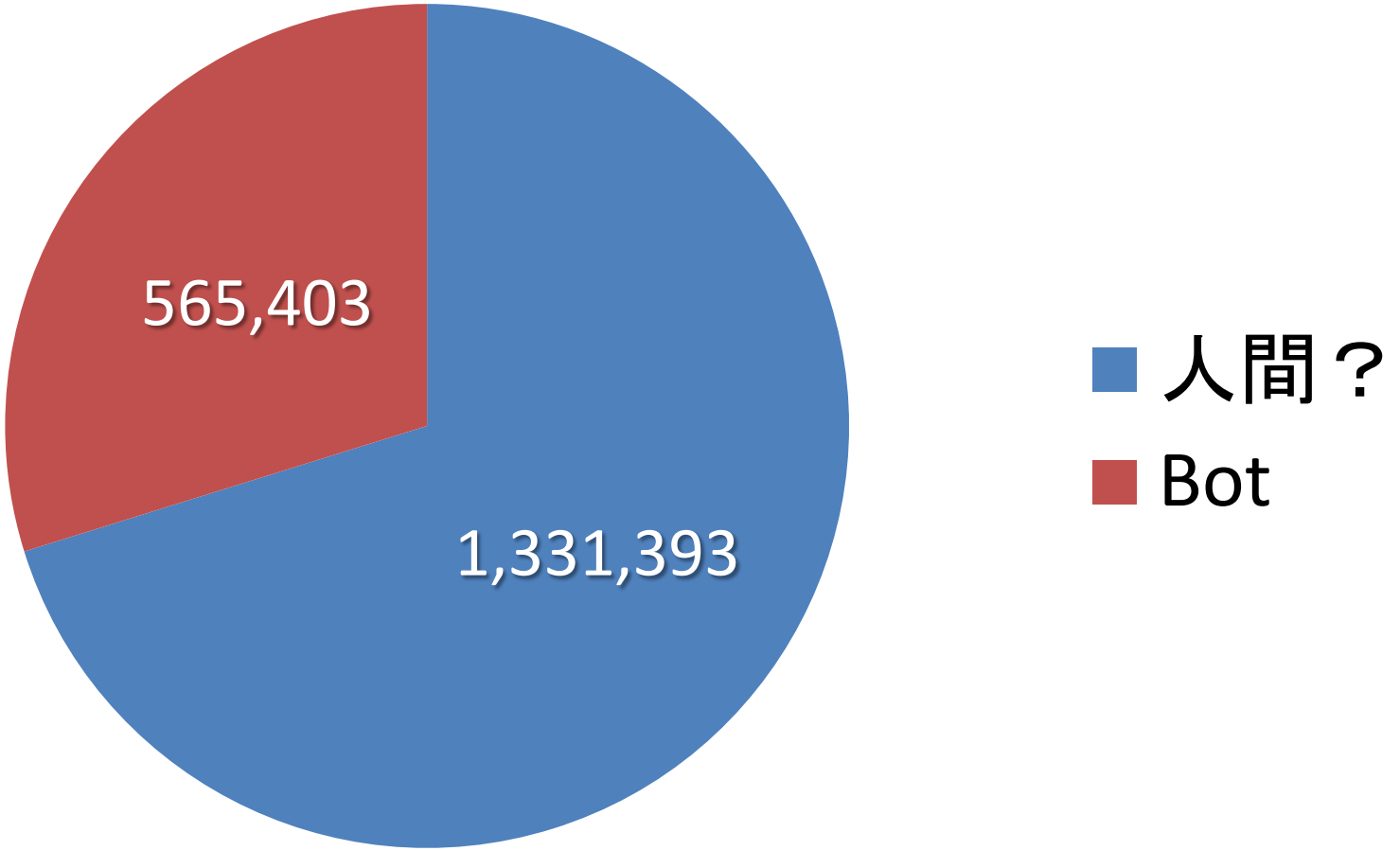


図. 筑波大学附属図書館Webサイト：2009.11.30-12.15のアクセス状況

- 人間：ボットは約２：１

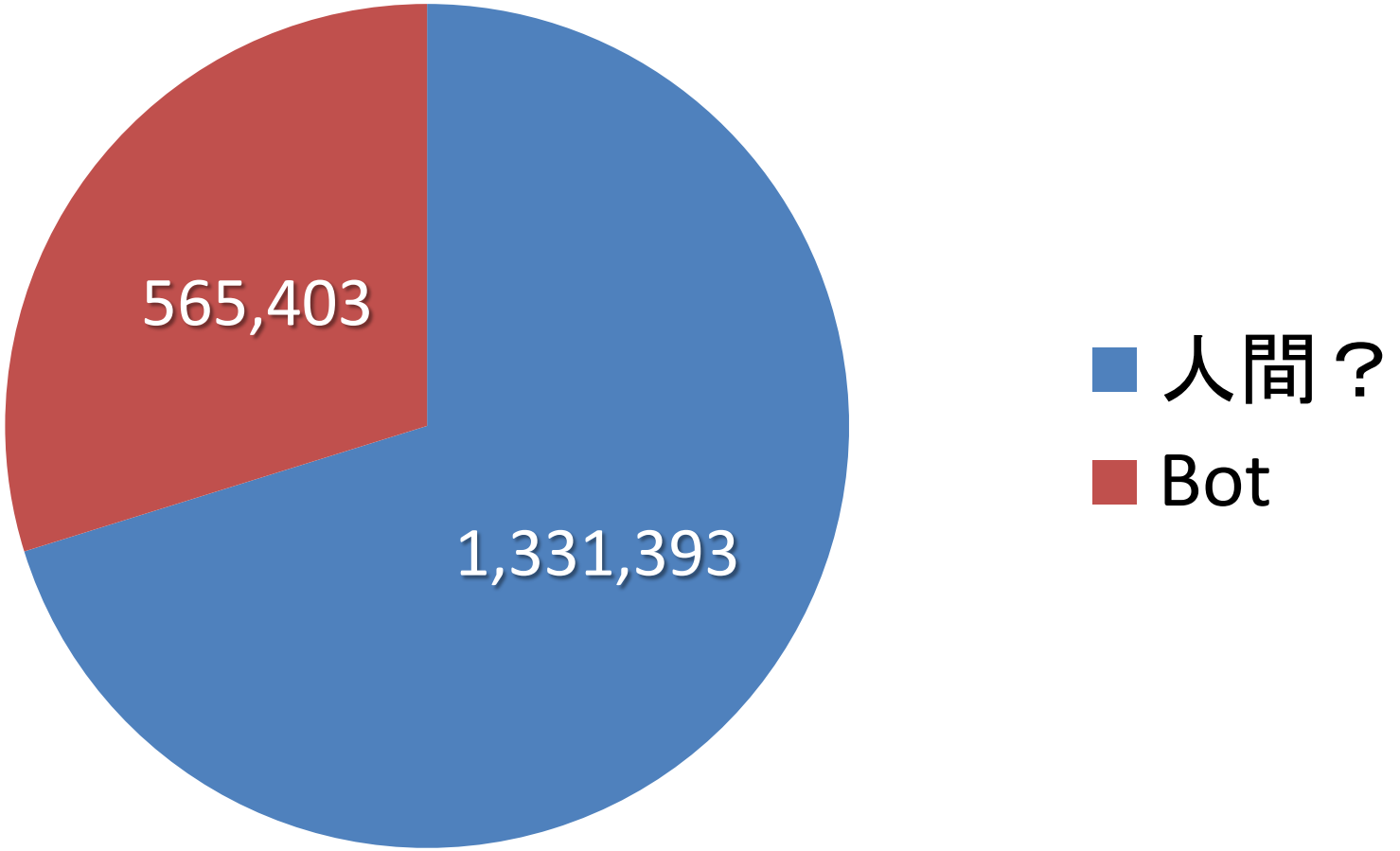


図. 筑波大学附属図書館Webサイト：2009.11.30-12.15のアクセス状況

- 人間：ボットは約２：１
- ボットはありふれている

アクセス数の多いボット

- 参考：Librahack氏
  - –14日で33,465回・・・2,390回/日

アクセス数の多いボット

- 参考：Librahack氏
  - –14日で33,465回・・・2,390回/日
- IPアドレスベース：Yahoo! Slurp
  - –41,563回（2,598回/日）

アクセス数の多いボット

- 参考：Librahack氏
  - –14日で33,465回・・・2,390回/日
- IPアドレスベース：Yahoo! Slurp
  - –41,563回（2,598回/日）
- User Agentベース：Googlebot
  - –250,500回（15,656回/日）

アクセス数の多いボット

- 参考：Librahack氏
  - –14日で33,465回・・・2,390回/日
- IPアドレスベース：Yahoo! Slurp
  - –41,563回（2,598回/日）
- User Agentベース：Googlebot
  - –250,500回（15,656回/日）

アクセス数の多いボット

- 参考：Librahack氏
  - –14日で33,465回・・・2,390回/日
- IPアドレスベース：Yahoo! Slurp
  - –41,563回（2,598回/日）
- User Agentベース：Googlebot
  - –250,500回（15,656回/日）

66.249.67.47 - - [30/Nov/2009:23:55:06 +0900] "GET /scripts/location-finder?loc=100&cal=DA01336-1994&kk
66.249.68.137 - - [30/Nov/2009:23:55:09 +0900] "GET /scripts/pubsearch/namazu.cgi?query=%1B%24B%24D
66.249.67.85 - - [30/Nov/2009:23:55:10 +0900] "GET /scripts/location-finder?loc=100&cal=DA00927-1991&kk
66.249.67.85 - - [30/Nov/2009:23:55:16 +0900] "GET /dspace/handle/2241/1/items-by-author?author=%E3%
66.249.67.85 - - [30/Nov/2009:23:55:18 +0900] "GET /scripts/pubsearch/namazu.cgi?query=tulips&whence=2
66.249.67.49 - - [30/Nov/2009:23:55:21 +0900] "GET /scripts/location-finder?loc=100&cal=295.3-C84&kk=00
66.249.67.49 - - [30/Nov/2009:23:55:25 +0900] "GET /dspace/handle/2241/1/items-by-author?author=%E3%
66.249.67.13 - - [30/Nov/2009:23:55:25 +0900] "GET /scripts/location-finder?loc=100&cal=V-C25720&kk=&cc
66.249.67.49 - - [30/Nov/2009:23:55:27 +0900] "GET /dspace/handle/2241/17952?mode=full HTTP/1.1" 200 1
66.249.68.136 - - [30/Nov/2009:23:55:28 +0900] "GET /scripts/pubsearch/namazu.cgi?query=%1B%24B%24J%
66.249.68.101 - - [30/Nov/2009:23:55:33 +0900] "GET /scripts/location-finder?loc=100&cal=H-P24580&kk=&c
66.249.67.216 - - [30/Nov/2009:23:55:35 +0900] "GET /dspace/handle/2241/1/items-by-author?author=%E3%
66.249.67.50 - - [30/Nov/2009:23:55:37 +0900] "GET /scripts/location-finder?loc=100&cal=DA01065-1992&kk
66.249.68.48 - - [30/Nov/2009:23:55:37 +0900] "GET /scripts/pubsearch/namazu.cgi?query=%1B%24B%24K%
66.249.67.216 - - [30/Nov/2009:23:55:42 +0900] "GET /portal/calendar/ HTTP/1.1" 302 361 "-" "Mozilla/5.0 (c
66.249.67.216 - - [30/Nov/2009:23:55:44 +0900] "GET /dspace/handle/2241/1/items-by-author?author=%E3%
66.249.68.47 - - [30/Nov/2009:23:55:47 +0900] "GET /scripts/pubsearch/namazu.cgi?query=%1B%24B%24D%
66.249.67.97 - - [30/Nov/2009:23:55:07 +0900] "GET /dspace/handle/2241/1/browse-author?top=%E3%82%E
66.249.67.97 - - [30/Nov/2009:23:55:54 +0900] "GET /dspace/handle/2241/1/items-by-author?author=%E3%
66.249.67.101 - - [30/Nov/2009:23:55:56 +0900] "GET /scripts/location-finder?loc=100&cal=B-J19300&kk=&c
66.249.67.13 - - [30/Nov/2009:23:55:57 +0900] "GET /scripts/pubsearch/namazu.cgi?query=doaj&whence=0
66.249.68.48 - - [30/Nov/2009:23:56:00 +0900] "GET /scripts/location-finder-e?loc=113&cal=A-%A5%C82310
66.249.67.98 - - [30/Nov/2009:23:56:04 +0900] "GET /scripts/location-finder?loc=100&cal=V-J50515&kk=&cc=

66.249.67.47 - - [30/Nov/2009:23:55:06 +0900] "GET /s
66.249.68.137 - - [30/Nov/2009:23:55:09 +0900] "GET
66.249.67.85 - - [30/Nov/2009:23:55:10 +0900] "GET /s
66.249.67.85 - - [30/Nov/2009:23:55:16 +0900] "GET /
66.249.67.85 - - [30/Nov/2009:23:55:18 +0900] "GET /s
66.249.67.49 - - [30/Nov/2009:23:55:21 +0900] "GET /s
66.249.67.49 - - [30/Nov/2009:23:55:**25** +0900] "GET /
66.249.67.13 - - [30/Nov/2009:23:55:**25** +0900] "GET /s
66.249.67.49 - - [30/Nov/2009:23:55:27 +0900] "GET /
66.249.68.136 - - [30/Nov/2009:23:55:28 +0900] "GET
66.249.68.101 - - [30/Nov/2009:23:55:33 +0900] "GET
66.249.67.216 - - [30/Nov/2009:23:55:35 +0900] "GET

設問2

自分のアクセスが原因でサーバがダウンしたことを，プログラマなら気が付くべきではなかったのか？

そう思う

そうは思わない

HTTP

ってなんなの？

By 常川真央

HTTP

= HyperText

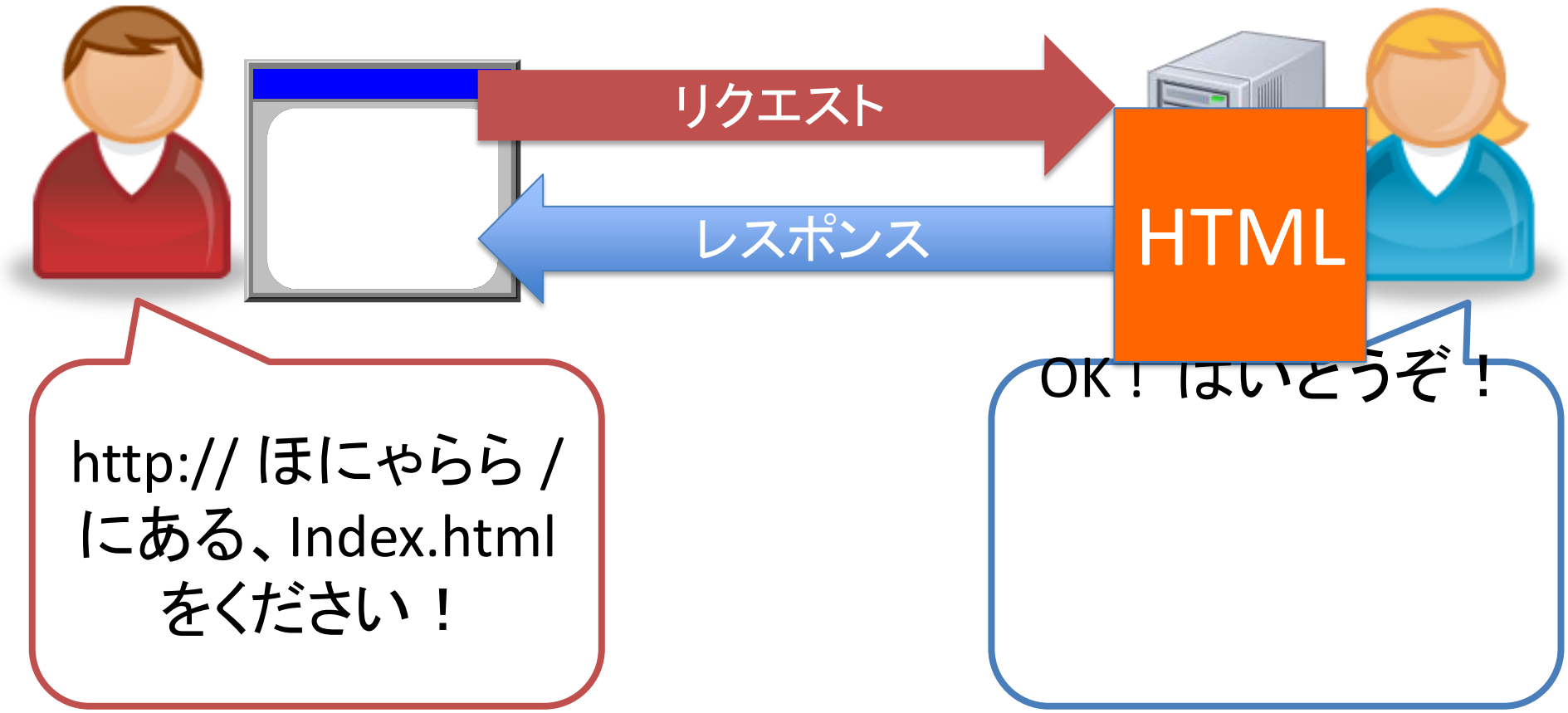
Transfer Protocol

＝ **HTMLなどをやりとりする会話のルール**

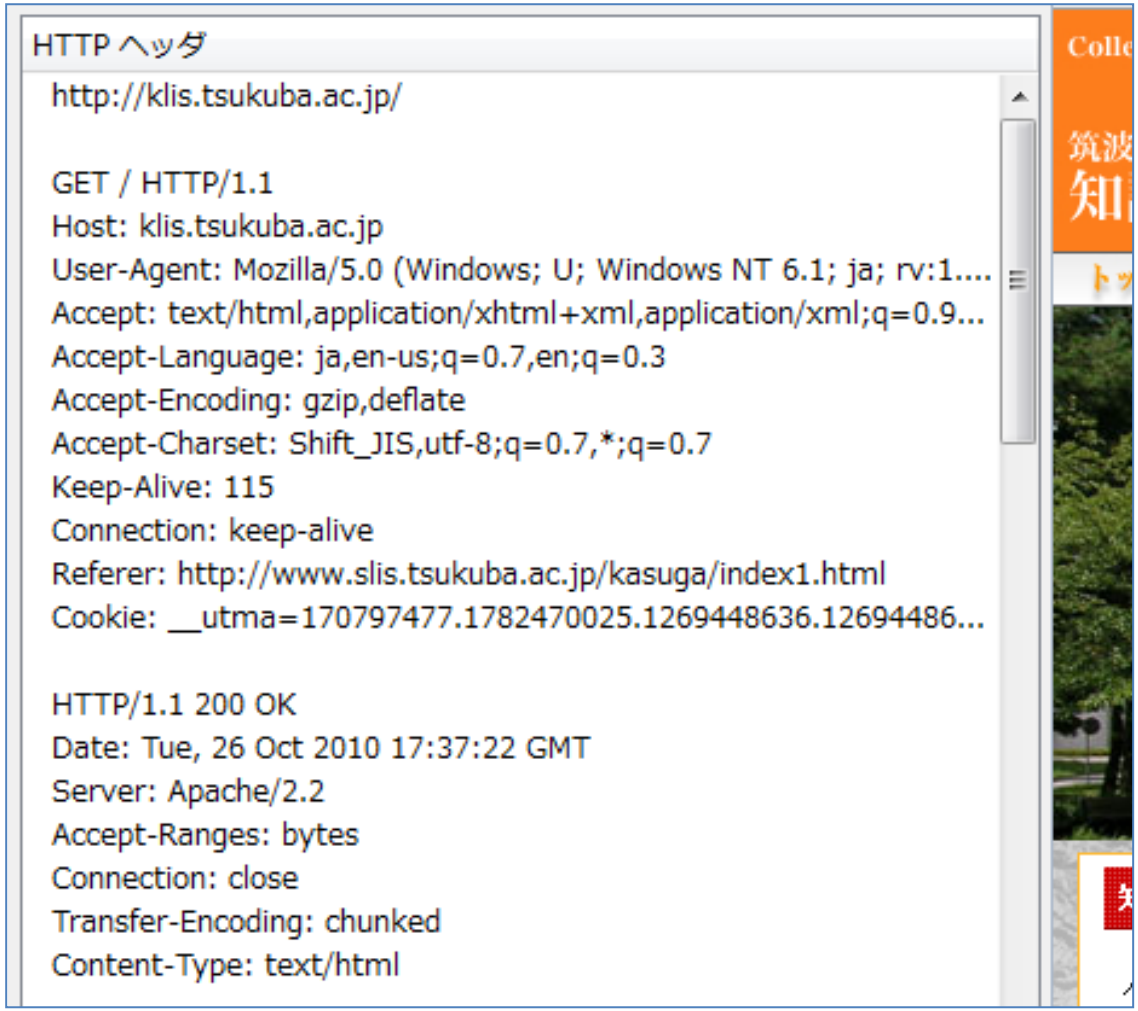
（そのほか、画像データなどHTMLに含まれるデータも対象となる）

HTTPは
Webで文書をやりとりするときの“世界共通語”

- Webは世界中の人が利用している
- 例えば日本からインドのサーバーの文書を入手することもあるため、言語の壁が生じる


リクエスト
レスポンス
HTML
http:// ほにゃらら /
にある、Index.html
をください！
OK！ はいどうぞ！

具体的には・・・・・・？ [デモ]



(FirefoxのLiveHTTP Headerを使ったデモ予定)

GET /index.html



メソッド



パス

パス ・・・・・・ 何を

メソッド ・・・・・・ どうしたいのか？

メソッドの種類 – 主に3種類

GET ― ××をください！

POST ― ××を作成/編集したい！

HEAD ― ××のヘッダをください！

（そのほか、PUT・DELETE・OPTIONなど全8種類）

ブラウザでの閲覧やクロールはGET

200 OK



ステータスコード

ステータスコード

…… お願いに対する返事

ステータスコードの種類

（略）

まとめ

- HTTPは「HTML・XMLをやりとりする会話のルール」

- リクエスト ― ユーザーからのお願い
  - 具体的には「メソッド」と「URL」でお願いごとを表現
  - GET ―「〜を送って！」

- レスポンス ― サーバーからのお返事
  - 具体的にはステータスコードとボディで返事を表現

補足:HTTPの性質

- 実はHTTP以外にも同様のプロトコルは存在しているが、HTTPほどには使われていない
- その理由は複雑すぎたため
- HTTPはシンプルで簡単に使えることから一気に普及した

補足:HTTPの注意点

- HTTPはシンプル
- 裏を返すと些細なメッセージも重要な意味を持つ
- よって、いい加減なステータスコードで返事をすると、とんでもない誤解が発生することもある

- HTTPを使うときは、普段の会話と同じように、相手にどんな気持ちを伝えたいかを配慮してメッセージを書くことが大事

「HTTPってなんなの?」

おわり

ステータスコード劇場

代表的なステータスコード

200　OK

404　Not Found

403　Forbidden

500　Internal Server Error

503　Service Unavailable

について，簡単に覚えて貰うために

寸劇をやります！

ステータスコード劇場

わかりやすくするために，今回は

「もしもリクエストが

告白だったら」

という仮定でやります．

参考：

As Sloth As Possible. “**告白されたときの正しいステータスコードの返しかた**”. (online), available from〈 http://blog.livedoor.jp/faulist/archives/1296865.html 〉, (accessed 2010-10-26).

200
OK

良かったね。

404
Not Found

私，有里じゃなくて

有元だけど？

403 Forbidden

二度と近寄らないでって言ったよね？

500
Internal Server Error

予期せぬ事態にテンパって
しまい思考回路はショート寸前.

503
Service Unavailable

ごめん，ちょっと今…

返事できないの．

まとめ

このように，ステータスコードには
色々な種類がありますが，明確な「拒否」を
示すのは

403 Forbidden

だけです.

自分が原因だと
わからなかったのか？

今回の事件のシステムではステータスコードは
500を返していたため,
「自分のせいだとは思わなかった」.

設問3

とはいえ、特定サーバに負荷をかけるのであれば，事前に許可くらい取ってしかるべきだったのでは？

そう思う

そうは
思わない

ユニークボット数

- IPアドレス単位：1,759
- User Agent単位：452
- 一々断りに来たら・・・

許可くらい取るべきだったのでは？

普通のシステムでは落ちるような

アクセスではなかった

→システムの問題点をざっくり説明します！

岡崎図書館MELILの問題点

当日は，高木浩光氏のブログエントリー

三菱図書館システムMELIL旧型の欠陥、アニメ化 - 岡崎図書館事件(7)
http://takagi-hiromitsu.jp/diary/20100829.html#p01

を参考に以下のことを説明しました。

- Webアクセスのしくみ（アニメ1・4）
- データベースと連携したWebアプリケーション（アニメ2・5）
- 岡崎市立図書館システムの挙動と問題点（アニメ3・6・8）

※（）内は該当するアニメーション

高木氏のブログエントリーと併せてご覧ください。

Webアクセスのしくみ（アニメ1）

- ブラウザからWebサーバに接続する
  - HTTPで通信する。
  - Webサーバ側に「セッションオブジェクト」を生成，それぞれに「セッションID」という受付番号がCookie*で設定される。

  *Cookie：Webサーバとブラウザの間でやりとりする必要な情報をブラウザ側に保存するしくみ

- アクセスごとにHTTP接続は切断される
  - 次の接続時は，Cookieのセッション IDを目印に先ほどと同じセッションオブジェクトへ接続する。

データベースと連携した Webアプリケーション（アニメ2）

- ブラウザからWebアプリケーションに接続
  - Webアプリケーション：Webサーバ上のアプリ
  - ここまでは「Webアクセスのしくみ」と同じ
- WebアプリケーションからDBサーバに接続
  - HTTP接続を維持したままDBサーバに接続。
  - DBサーバに接続後SQL*を実行。
  - SQL実行後，結果をブラウザに返すとともに DB接続とHTTP接続をそれぞれ切断する。

  *SQL：ソフトウェアからデータベースを操作するための命令文
- とりあえず「都度接続方式※」と呼ぶことにする

※高木氏のブログエントリーから。

岡崎市立図書館の
システムの挙動（アニメ3）

- ブラウザからWebアプリケーションに接続
  - ここまでは同じ…
- Webアプリケーションから DBサーバに接続
  - セッションオブジェクト生成と同時にDBに接続。
  - そのセッションオブジェクト専用のDB接続になる。
  - セッションオブジェクトが消えるまで残り続ける。
  - DB接続の上限数までこれを繰り返す。
- とりあえず「メリル方式※」と呼ぶことにする

※高木氏のブログエントリーから。

岡崎市立図書館システムの問題点

- 何が問題？
  - HTTP接続がなくなってもDB接続は生きたまま
  - DB接続の上限数を超えると…
    それ以降の端末では接続エラー
    - 返されたステータスコードは 500 “Internal Server Error”
  - DB処理の重さに関係なく発生する

- セッションのタイムアウトは10分間だった
  - タイムアウトが来ると…
    セッションオブジェクトが消えDB接続も切れる
  - DB接続に空きができるので接続できるようになる

Cookieが無効になっている ブラウザでアクセスする（アニメ4）

- ほとんどのクローラはCookie機能を持たない
  - セッションオブジェクトは作られるが，Cookieが無効なのでセッションIDは設定されない。
  - サーバにアクセスするたびにセッションオブジェクトが生成される。
- アクセスごとにHTTP接続は切断される
  - セッションオブジェクトは無尽蔵に生成できる。
  - Webサーバへの影響はほとんどない。

Cookieが無効なクローラで Webアプリにアクセスする

- 「都度接続方式」の場合（アニメ5）
  - SQL実行後，結果をブラウザに返すとともに DB接続とHTTP接続をそれぞれ切断する。
  - 「シリアルアクセス」である限りDB接続の上限数を上回ってしまうことはない。
    - DB接続とSQL実行はひとつ
- お行儀のよいクローラなら問題にならない
  - Librahack氏のプログラムでも問題にならない

Cookieが無効なクローラで Webアプリにアクセスする

- 「メリル方式」の場合（アニメ6）
  - SQL実行後，結果をブラウザに返したあともDB接続は残り続ける
  - セッションオブジェクトの数だけDB接続も生成
  - あっという間にDB接続の上限数を使い切る
  - 上限数以降はすべてエラー接続（アニメ8）
- （開発当時の）2005年でも古すぎる設計
  - あの当時すでにクローラは存在していました
  - 今まで問題にならなかったことが不思議

結論

- MDISが開発したシステムの欠陥
  - システム設計の致命的なミス
  - あまりにも時代遅れ過ぎる設計
    - 2005年…いや2000年…当時でもちょっと…
    - 2000年にはコネクションプーリング方式というものが
      すでに知られていた。（アニメ7）

- 他の図書館ではこの不具合を修正していた
  - 岡崎図書館には連絡もせず放置したまま
  - MDISはこの欠陥を隠蔽しごまかし続けた
  - これでもLibrahack氏に問題があったと思う？