

## HireDiversity Spidering Specifications

### *Introduction*

Spiders are often a convenient way for a programmer to get data off of a website.

Convenient as they are, they are also slow, hard to develop, and fragile.

For optimum reliability, a file-based or XML-based transfer system is recommended.

A spider will allow HireDiversity to generate a batch file from your career website. The batch file will be generated at our end, so there will be no need to deliver it. This will facilitate a great deal of the batch posting process, moving a lot of the work from your end to our end.

The only information available to us will be the information you make publicly available on your career site. All applications will be referred back to that job page on your career site (if possible.) No tracking of applications as "sourced from HireDiversity" will be done unless your career site supports such a thing.

### *Technical Specifications*

#### *User Agent*

Our spider will identify itself with the following "User Agent" string in all HTTP requests:

```
Hispanic-Business-Inc-Spider/1.1
```

Note that the /1.1 is a version number and is subject to change. (Hispanic Business Inc. is the parent company of HireDiversity.)

If your web server keeps a log of web activity that includes the user agent, you can filter your log using this user agent to see all activity from our spider.

#### *IP address*

All spider activity will come from the Hispanic Business Inc. corporate IP address range. At the time of writing (8/18/2004) this is the Class C network

```
63.149.249.0/24
```

```
63.149.249.0 - 63.149.249.255
```

```
63.149.249.0 sub-net mask 255.255.255.0
```

#### */robots.txt*

Our spider respects the base "robots.txt" rules detailed at <http://www.robotstxt.org/>

If you have a /robots.txt file on your career site that tells robots not to spider your jobs, then our spider will not spider your jobs. This is a feature.

Note: some popular career site vendors (such as Monster as of 8/18/2004) have /robots.txt files that exclude all job-related spider activity. HireDiversity respects these vendors' wishes and will not spider their sites while they have these rules in effect. We will NOT spider any site in violation of their posted robots.txt rules – such an action would expose us to litigation.

If you would like to allow our spider access while keeping other spiders away, use the User-Agent directive as documented on <http://www.robotstxt.org/>

For example (note that it is incorrect to include the version number:)

```
# example robots.txt

# tell Hispanic Business Inc. they can spider anything
User-Agent: Hispanic-Business-Inc-Spider
Disallow:

#tell Googlebot they can spider anything but the /jobs/ directory
User-Agent: Googlebot
Disallow: /jobs/

# tell everyone else not to spider at all
User-Agent: *
Disallow: /
```

### *Sleep*

Our spider will sleep for a default period of six seconds between HTTP requests. This is to avoid putting undue load on your server. If you like, we can lengthen this sleep period. However, if there are lots of jobs, a lengthy sleep period may make it difficult to keep our HireDiversity.com job base up-to-date in a timely fashion.

### *Schedule*

Spidering "runs" are usually scheduled nightly. If your site has a lot of jobs (over 1000, say,) this may be scaled back to three times a week in order to conserve load on our server. The night-time scheduling should coincide with a period of low overall activity on your career site, which should ameliorate any load problems due to activity from our spider.

### *Associated Content*

Most user agents will download not only the HTML of the page, but also all the associated content – external style sheets, referenced images, perhaps a hit to /favicon.ico, external javascript, etc. Our spider is usually interested in just the HTML of the page. It will not download images, javascript, external stylesheets, etc., unless such is necessary to successfully navigate the list of jobs. (This hasn't happened yet.)

### *Features*

Our spider supports session-state cookies. If you use a session-state cookie to remember user information, our user agent will successfully echo it back to you. If you perform a memory-intensive action at the start of every session, you will be very happy that we do this.

Our spider does not currently support persistent cookies. If you really really want to drop a persistent cookie and have our spider pick it up, that can probably be done, but it would require extra development time.

Our spider supports secure http (https.) If your career site only accepts https, we can still spider you.

Our spider does not support Human Interactive Proofs such as visual recognition of text from a distorted image. This is a feature – such things are used to keep spiders out. Our spider will not go somewhere it isn't wanted.

Our spider does not support javascript directly. However, we're pretty good at reverse-engineering the javascript and constructing the appropriate URL from its component pieces directly.

Our spider does not support forms directly. However, it does support POST requests. We're also pretty good at extracting form information and constructing the appropriate POST or GET request from the extracted info.

Our spider does not support Flash, Shockwave, or Java interaction. If your career site requires this, some serious development time would be required to extract your jobs. If you have a text-only or low-bandwidth version of the site, that might be a way to go.

## **Approximations**

### *Job Function/Category*

HireDiversity maintains a list of "Job Functions" or "Categories" such as Accounting, Computers – Software, Real Estate, etc.

Your site probably has a different list.

As part of the spider, we will develop a map from your list to our list.

If your list changes, that will break the spider.

If you have no such list, we will rely on job keywording to assign your jobs to our "Job Functions". This is not always 100% accurate. For best performance we recommend that you develop a list so that we can have something to map from.

### *Location*

HireDiversity tracks jobs by ZIP code or Canadian Postal Code. We have a list of city/state/ZIP codes. If you display the city and state for each job, but not the ZIP code, we can infer the ZIP code from the listed city and state. This is not always 100% accurate. If the city name is mistyped, we may not be able to accurately place the job in the correct location. If a city is newer than our list, we may not know about it.

## **Changes**

Spiders rely on pattern matching to extract the relevant information. As such they are very sensitive to any content changes or design changes on the site. Any change to your career site may lead to any of a variety of errors. These can range from things like certain jobs being assigned to the wrong locations, to things like no jobs being pulled down at all.

Occasional minor changes to your career site, if noticed, will be changed at no charge. Frequent or major changes to your career site, as well as changing career site vendors, may incur a production charge.

## **Changelog**

10/14/2005... changed user-agent string to be compliant with RFC 2616 section 3.8

Was: Hispanic Business Inc. Spider/1.0

Now: Hispanic-Business-Inc-Spider/1.1