LinkUp Raw Overview and User Guide

TABLE OF CONTENTS

About LinkUp Raw

ABOUT LINKUP RAW

LinkUp has built the world's most extensive, accurate, and up-to-date set of jobs direct from employer websites. With the largest and highest quality job listings sourced from 50,000 company websites globally, LinkUp has become a leading provider of job market data and analytics. From that robust and unique dataset of jobs, LinkUp has developed Raw, which delivers brilliant insights into the job market.

LinkUp Raw is a data product that contains the most complete set of LinkUp's job market data. It includes full job postings sourced directly from employer websites dating back to 2007, as well as statistics and analysis of the job postings. LinkUp Raw consists of four distinct components, each delivered daily.

Individual job records
Files contain every data field for each job except its full-text description.

Individual descriptions
Delivered daily, this contains the complete text-based job description for every job listing in our index. Unique job identifiers all for easy matching to other job record fields.

Core analytics
Raw contains core macro, sector, and micro analytics that are calculated and updated daily.

Reference Material
Our reference material includes a log of every scrape and modification to any scrape code we have done, information about company information, and identifiers that can be used to join to additional datasets (ie cusip, sedol, isin, ticker, and more).

LinkUp Raw snapshot

Historical job count	130 million
Archive start date	2007
Daily active job count	3.5 million
Global coverage	195 countries
Unique job and company attributes	20
Data access	FTP
File format	XML and CSV
Delivery frequency	Daily

Data Collection

Data Collection

LinkUp's web scraping technology and infrastructure collects data from thousands of employer websites daily. Each job posting url found is captured including its title, description, and location. Employer websites are checked at least once every 48 hours for any new postings, removed postings, or updated postings. This results in several million unique job postings being checked each day.

Job Attributes

This is a full list of attributes available for a job record in Raw.

Job Attribute Name	Description
City	This is the city the job is posted in. This is often directly from the page, but if not listed it is derived from other location information on the page.
Company ID	This is the unique identifier LinkUp uses for the company or employer.
Company name	This is the name LinkUp has listed for the employer.
Country	This is the country the job is posted in. This is often directly from the page, but if not listed it is derived from other location information on the page.
Created	This is the date and time the job's URL was first found by LinkUp.
Deleted	The date and time LinkUp found the job's URL to no longer be available.
Description	This is the full text description of the job posting from the employer's page. HTML formatting has been removed.
Hash	This is the unique identifier for a job. It is an md5 hash of the job's URL.
Last checked	This is the most recent date and time the job's URL was found to be available by LinkUp.
Last updated	The most recent date and time the job's title or description was changed.
SOC | O*NET occupation code	The Occupational Information Network (O*NET) is the occupational code associated to the job.

State	This is the region (state, province, territory, etc.) the job is posted in. This is often directly from the page, but if not listed it is derived from other location information on the page.
Title	This is the job's title, exactly as posted on the employer's site.
Unmapped Location	This Boolean field is an indicator for when we were not able to get a match for a location for that job.
URL	This field will provide each individual job's original URL.
Zip	This is the zip the job is posted in. This is often directly from the page, but if not listed it is derived from other location information on the page.

Data Collection (continued)

Company Attributes

This is a full list of company attributes provided in Raw. These can be associated with job postings via the company id.

Company Attribute Name	Description
Company ID	This is the unique identifier LinkUp uses for the company or employer.
Company name	This is the name LinkUp has listed for the employer.
Cusip	This stands for Committee on Uniform Securities Identification Procedures. A CUSIP number identifies most financial instruments.
Factset ID	This is the entity identifier for FactSet that can be used to map to Factset data structures.
Isin	An International Securities Identification Number (ISIN) is a code that uniquely identifies a specific securities issue.
LEI	This is the Legal Entity ID as provided by Factset.
NAICS	This is the primary NAICS code for the employer, if applicable.
Perm ID	This is the primary identifier used by PermId.org
SEDOL	This stands for Stock Exchange Daily Official List, a list of security identifiers used in the United Kingdom and Ireland for clearing purposes. The numbers are assigned by the London Stock Exchange, on request by the security issuer.

Ticker	This is the Stock Ticker associated with the employer, if applicable.
URL	This is the employer’s ‘homepage’. Often this is the corporate website.

Raw Contents

Raw Contents

LinkUp Raw contains a variety of job, company, and analysis files. The naming standard of each of these files indicates the content.

Job Files

File Name	Frequency	Description
Raw Full Job Records	Monthly	This file contains all of LinkUp's job records. Once a month, LinkUp generates a new full set of jobs to make it easy to rebase. This is an archive containing 16 total CSV files. Each file is a split of the full set based on job hash to make them more consumable.
Raw Full Job Descriptions	Monthly	This file contains all of LinkUp's job descriptions. Once a month, LinkUp generates a new full set of descriptions to make it easy to rebase. This is an archive containing 16 total XML files. Each file is a split of the full set based on job hash to make them more consumable.
Raw Daily Job Records And Descriptions	Daily	This file contains all of the changes to the job set for the day. At midnight UTC, LinkUp generates a file containing all new jobs and updates to existing job records that occurred that day. This is an archive containing both the job records as a CSV, and the descriptions as a separate XML file. This can be applied to the previous day's set to keep your data current.

Raw Contents (continued)

Analytics Files

File Name	Frequency	Description
Core Company Analytics	Daily	A full daily history of company-level job counts. Includes Company ID, Company Name, Stock Ticker, Stock Exchange Company, Day and Active Duration for all companies going back to 2007.
Core Ticker Analytics	Daily	A full daily history of stock-ticker counts of jobs. This uses FactSet data to track ticker over time. Uses only primary tickers.

Reference, Ticker, and Scrape Log Files

File Name	Frequency	Description
Raw Daily PIT Company Reference	Daily	A daily differential file showing company field changes each day.
RAW Full PIT Company Reference	Monthly	A file that shows all information LinkUp has or has derived about a company_id. This file shows company name, homepage url, legal identity, permID, and Naics code
FS Daily Company Reference	Daily	A file that shows all information LinkUp has derived from Factset ID from via the FactSet concordance process. This file shows ticker, stock exchange country, stock exchange name, ISIN, CUSIP, SEDOL and primary flag.
Raw Daily Company Scrape Log	Daily	A daily differential file showing any changes that have been made to scrapes.
Raw Monthly Company Scrape Log	Monthly	A log containing a field for every company ID (*scrape), a flag of whether the scrape ran each day, and a flag for any time the scrape has been changed.

Additional Raw Analytics files are available.

File Size and Delivery

File Size and Delivery

LinkUp delivers all Raw files through its data platform Job Data Engine (JDE). JDE has its own web interface to serve the files to end users. Alternatively, JDE uses a secure FTP server to store all data files. Read access to this FTP location is provided in addition to JDE accounts for all customers. Files are generated each morning and are available before 12:00 PM Central Time (18:00 UTC / GMT During US Standard Time (Nov 1 2020- Mar 14, 2021) and UTC / GMT 17:00 During US Daylight Saving Time (Mar 8th 2020 - Nov 1 2020)) Monthly full Raw job records and full descriptions may take longer.

The archive job description file is about 300 GB uncompressed and the job records file is 42 GB uncompressed. Zipped, those files are 100 GB and 17 GB, respectively.

FAQ

FAQ

Coverage

Do you have all public companies?

LinkUp scrapes 100% of publicly traded U.S. companies that have jobs on their website. About 15% of public companies currently don't post jobs on their website.

What percent of your companies are private?

About 75% of companies in our database are private.

Job Counts

What are drops?

LinkUp uses regular expressions to parse jobs data from web pages. Regular expressions allow us to search for specific text patterns on the page that we've manually verified to be a job posting. When an employer changes their careers page, this searching may need to be updated. If the text pattern we're searching for is no longer there, it results in 0 jobs. If the text pattern is now in a different place on the page, and not on every job post, we may report fewer than the correct amount but not 0. If the text pattern is now in more places than on every post we may get more than the correct amount.

Alternatively, we often require a job posting to have some location information to be considered valid. We also use regular expressions to parse that information. If the employer changes their postings to no longer have locations (or changes where/how that information is displayed) on every job, we may not report those jobs. Resulting in fewer jobs, but not necessarily 0.

What are spikes?

Spikes are typically when a company shows a dramatic increase in job count. Spikes can occur when a company changes their Applicant Tracking System, and our spiders find jobs in both places for a few days. This can be easily detected when a job count for a company appears to roughly double, and when it doesn't stay at that level for more than a few days. The company scrape log exists to provide additional context around spikes and drops in the data. If a return to normal job count correlates with a scrape change, it can be inferred that the change in job count was due to a career portal change rather than an organic change in labor demand.

FAQ (continued)

Job Counts (continued)

What is drift and what causes it?

Drift is the shifting of our historical job counts as previously undetectable job URLs come back online. This is reflected in the consistent downward trend in our dataset's overall job count for the most recent 10 days.

There are several reasons jobs can temporarily disappear from a career page. The site could go down temporarily for maintenance, they could be formatted incorrectly rendering them "unscrapable" for a time, our scrape system could have an issue with a certain CID or the company could have changed its ATS or career portal software.

On the day those jobs disappear or otherwise undetectable by our spiders, they are reported as deleted. However, because we collapse on job URL, when the exact same URL re-appears on the page at a later date, we consider that the same job, and the delete date is removed from the record, showing that the job was never taken down. This is good because we are not duplicating records from the same jobs, but it causes what we call drift.

What are 'Evergreen' jobs?

Evergreen jobs are jobs that never or rarely get removed from a company's career page. We recommend identifying job duration greater than 180 days as Evergreen jobs.

Quality

Why is LinkUp data higher-quality than other job data providers?

Completely unique in the industry, LinkUp is the only job search engine that indexes jobs exclusively from company websites. Updated daily, the 3.5 million jobs in our search engine are always current with no duplicates or job pollution. By indexing jobs everyday, directly from company websites, we have eliminated all of the 'noise' that pollutes other labor datasets, such as duplicate listings, expired jobs, and job board 'pollution' (scams, frauds, lead-gen, etc.). As such, our job market data is clear, accurate, and powerful.

FAQ (continued)

Scraping

What do you consider a unique company?
LinkUp assigns one company ID to a unique career page. If one company has separate career pages for their various divisions or countries, those would be separate scrapes, therefore separate company IDs.

What do you consider a unique job?
LinkUp collapses on the URL and therefore considers each unique URL one job. Most ATS' generate a new URL for a new job.

What if a job gets removed from a careers page and comes back on later? If the job has the same URL when it comes back, we remove the delete date.

Are breaks flagged and how long does it take to fix them?
Companies are frequently changing ATS' and updating job formats on their career pages, which causes breaks. As a result, we have created a 'broken' scrapes queue that notifies our scrape team when a company scrape breaks. Generally, public companies are flagged and fixed within 24 hours with or without escalation. Additionally, clients can contact us about a broken scrape, and we'll get to it right away, otherwise scrapes get fixed in the order they were flagged.

How are cities identified?
LinkUp gets the city from the job itself, which is not necessarily the company's headquarters location. Sometimes this is displayed clearly on the job, and sometimes it is 'hidden' but still capturable. If the jobs don't mention a location, we'll then check to see if the company has multiple addresses/offices/ locations. If they only have one, we make the assumption that all jobs are located there unless we have reason to believe otherwise.

What location fields are available in your dataset?
City, state, zip code, and country are included in the dataset. MSA can be provided in a feed.

FAQ (continued)

SOC | O*NET

What is SOC | O*NET?

LinkUp applies Standard Occupational Classification (SOC) codes and the Occupational Information Network (O*NET) system to all current and historical job openings in its dataset. The SOC | O*NET system features over 1,000 standardized occupational categories with detailed descriptions.

What version of O*NET is LinkUp using? LinkUp uses version 21.0 2016.

*You can learn more about SOC | O*Net here.*

Tickers

How are tickers mapped?

LinkUp partners with FactSet for mapping our companies to tickers. Our ticker reference file contains fields for ticker, exchange, and exchange country, along with start and end dates to reflect change over time.

How often are tickers updated?

Tickers are updated daily.

LinkUp

www.linkup.com | 866.359.9360 | info@linkup.com | 430 N 1st Ave, Suite 790, Minneapolis, MN 55401 U.S.