

Real-time Labour Market Information on Skill Requirements

Setting up the infrastructure for EU system

Data ingestion

Ettore Colombo – Matteo Fontana – Andrea Scrivanti

Expert Workshop
Milan, 20-21 March 2018



Topics

1. Source selection strategy
2. Data ingestion techniques
3. Preliminary data and expected volumes
4. Scheduling, parallel processing and monitoring



Topics

1. Source selection strategy
2. Data ingestion techniques
3. Preliminary data and expected volumes
4. Scheduling, parallel processing and monitoring



Definitions

Data ingestion is the process of **obtaining** and **importing** data from web portals and **storing** in a database

Data ingestion mainly focuses on **volumes**, not on **quality**

The key point is to ensure a stable data flow preventing potential loss of data due to harvesting issues



Important notice

All the volumes reported hereafter are referred to **preliminary snapshot extractions**

The volumes are not conclusive nor complete, they have to be considered as an indication of potential data volumes



Definitions

Source: the job portal where the vacancy is retrieved from

Site: the website where the vacancy is published

i.e. the aggregator portal is the source of the vacancy, the company website is the site where the vacancy is published



Source selection strategy

4 Processing Steps



Augmentation



We analysed the results of the landscaping activity

- Completing the mapping of transnational sources
- Adding further transnational sources
- Adding the complete set of EURES sources

In order to define

- a priority list to define agreements
- a relevance order to realize data ingestion channels



Augmentation



22 distinct transnational sources

- 12 aggregators
- 2 job portals
- 7 recruitment agencies
- 1 advertisement portal

47 sources related to EURES



Agreements



We tried to establish a **direct agreement** with the **most relevant** sources in order to

- Obtain a stable data supply
- Share a data format
- Minimize the impact of the data ingestion activity



Agreements



If the agreement is **reached** the source is included in the data ingestion phase

If the agreement is **explicitly denied** the source is excluded from the data ingestion phase

If we get **no answer** an alternative plan to get data from the source is defined (crawling, scraping)



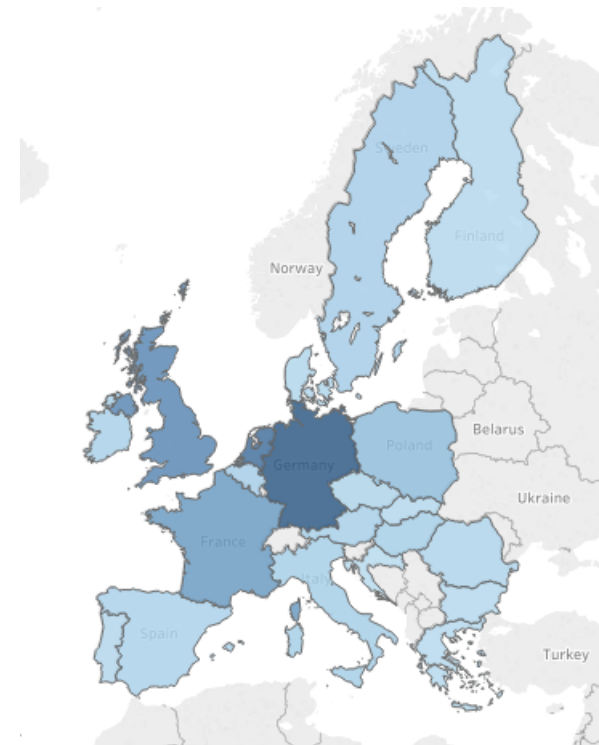
Who said YES



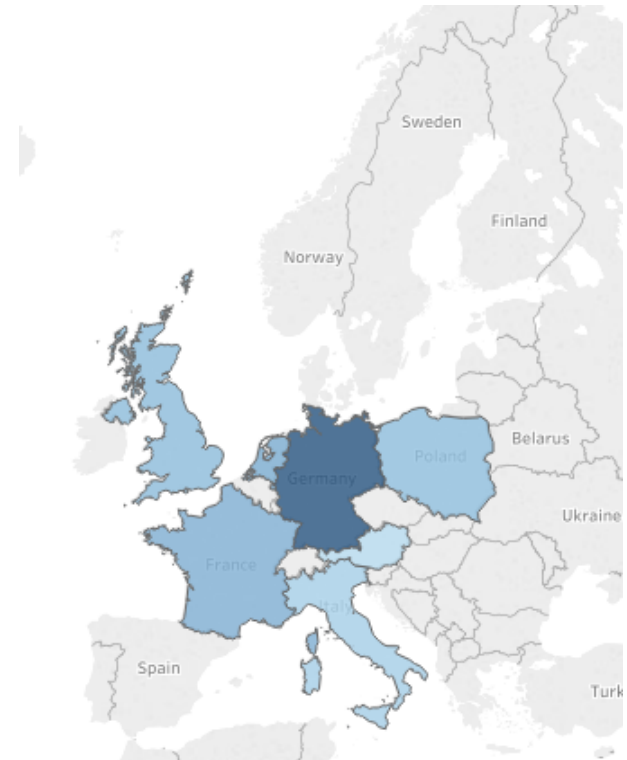
Jooble



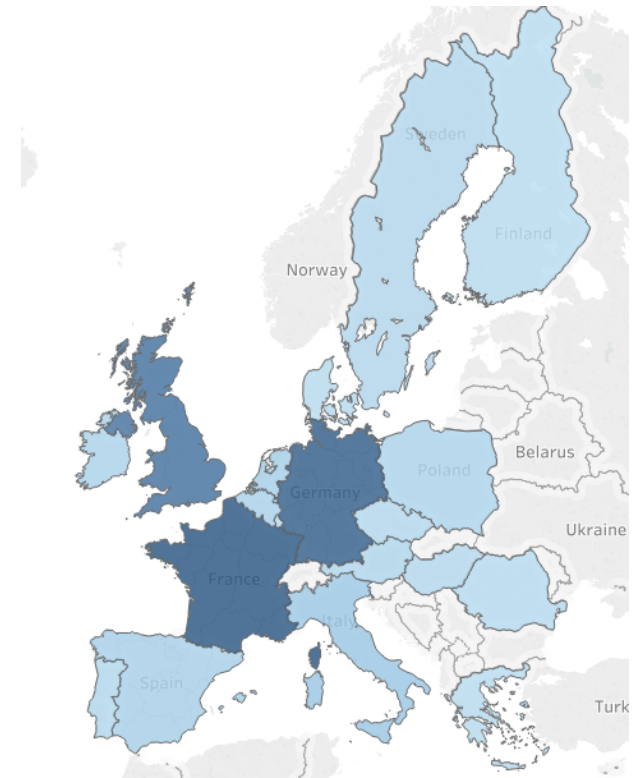
- 21 countries
- Over 3 millions OJV obtained from the first snapshot ingestion
- About 50% of the OJV published in the last 3 months



- 7 countries
- Over 3 millions OJV obtained from the first snapshot ingestion



- 19 countries
- Over 1 million OJV obtained from the first snapshot ingestion
- All the OJV published in the last month



Who said NO (or blocked us)



jobrapido

jobtome



job*ij*oba



CEDEFOP

European Centre
for the Development
of Vocational Training

Agreement in progress



indeed

jobbydoo

glassdoor™



CEDEFOP

European Centre
for the Development
of Vocational Training

Who we already get data from (trying to define an explicit agreement)



HAYS

15 countries



5 countries



Manpower®

11 countries

trenkwalder

11 countries



4 countries



CEDEFOP

European Centre
for the Development
of Vocational Training

Who we already get data from (trying to define an explicit agreement)



careerJET

only public API

InfoJobs

1 of 2 countries

randstad

17 countries

**CV
MARKET**

3 countries

Adecco

19 countries



CEDEFOP

European Centre
for the Development
of Vocational Training

Who we already get data from (local sources)



REED

XING


pôle emploi



 portal de empleo
empleate

topjobs

IRISH JOBS.IE
CREATE OPPORTUNITY

buscojobs
España

 Jobs Ireland
WHERE JOBSEEKERS GO



CEDEFOP

European Centre
for the Development
of Vocational Training

Coverage



At operating speed

- 28 countries covered
- Most of the sites from the landscaping list covered
- Many more sites coming from aggregators



Coverage



Issue: potential technical problems when gathering data from a source (unavailability, block, changes in data structure)

Risk: loss of data

Solution: redundancy

- Have the most important sites (by volume and/or coverage) ingested from two or more sources
- Avoid loss of data in case of troubles with a source
- Collect data from both primary and secondary sources



Topics

1. Source selection strategy
2. Data ingestion techniques
3. Preliminary data and expected volumes
4. Scheduling, parallel processing and monitoring



Overall Data Flow - Recap



Data
Ingestion

Pre-Processing

Information
Extraction

ETL

Presentation
Area



CEDEFOP

European Centre
for the Development
of Vocational Training

Data Ingestion



Data ingestion is the process of obtaining and importing **data** from **web portals** and storing in a **database**.



OJAs

Import OJVs



**STRUCTURED AND
UNSTRUCTURED
OJAS DATABASE**



CEDEFOP

European Centre
for the Development
of Vocational Training

Data Ingestion



Data ingestion is the process of obtaining and importing **data** from **web portals** and storing in a **database**.



Structured OJAs Database: a collection of Online Job Ads; each one represented by a list of values, each one with a specific meaning

Unstructured OJAs Database: a collection of web pages (i.e. HTML code); each one representing a Job Ad



Data Ingestion



Data ingestion is the process of obtaining and importing **data** from **web portals** and storing in a **database**.



Crawling



A **Web crawler** is a bot that systematically browses web portals for the purpose of **download all their pages**.

Crawling is the most common way to get information massively from the Internet: search engine spiders (e.g. GoogleBot)

JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom
Application deadline: Saturday, 30 September 2017
Reference number: 100

APPLY NOW

Home > Now Hiring: Software Developers > Junior Software Developer

Share: [in](#) [t](#) [g+](#) [e](#)

Description

As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful API's, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

Web page:

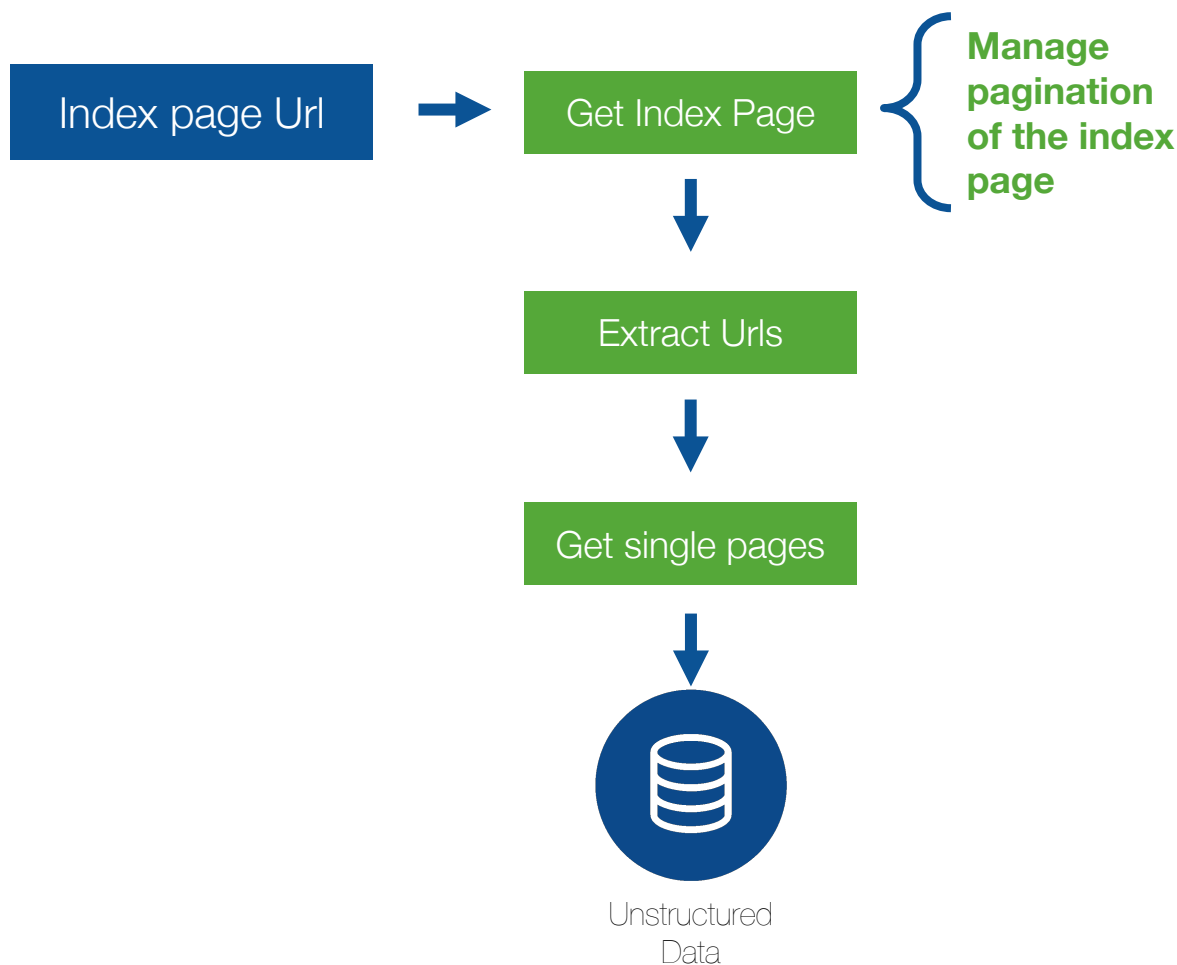
```
<!DOCTYPE html>
  <head>
    <meta name="title" content="Junior
Software Developer" />
  </head>
  <body>
    <header>
      <h2>Junior Software Developer</h2>
      <div><div>Location</div>United
Kingdom</div>
      ...
    </header>
    <div><div>Description</div>
    <span>As Junior Software Developer, you
will develop excellent software for use...
```



CEDEFOP

European Centre
for the Development
of Vocational Training

Redefining Crawling



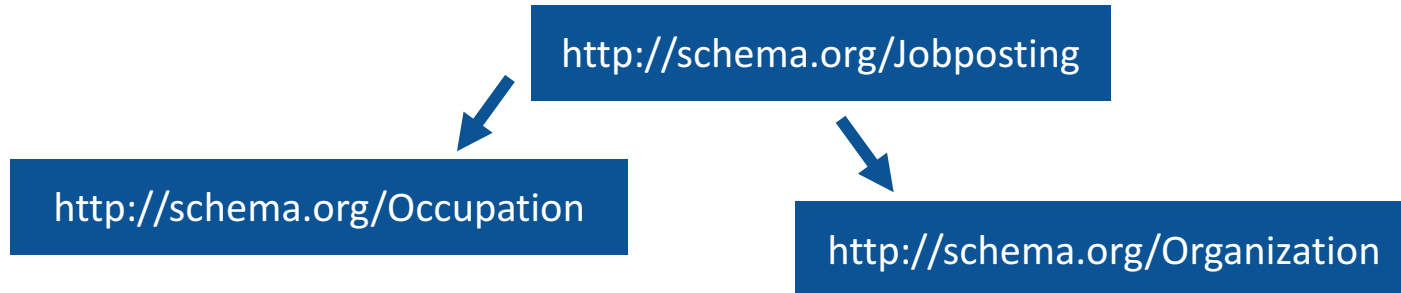
OJAs are **all** completely downloaded in **2 steps**

Issue: **managing pagination** could require little site-dependent configuration (no software development)

Structured data on the Web: Schema.org

In the last decades, some methods to introduce structured data in the web page has been proposed.

Nowadays, job-posting code snippets are used to improve automatic comprehension of web page by spiders (e.g. GoogleBot) and scrapers



A page without markup

```
<div>
  <h2>Software Engineer</h2>
  <p><strong>Location:</strong>
  Kirkland WA</p>
  <p><strong>Industry:</strong> Computer Software
  <br><strong>Occupational Category:</strong> 15-1132.00 Software Developers, Application
  <br><strong>Hours:</strong> Full-time, 40 hours per week
  <br><strong>Salary:</strong> USD 100000
  </p>
  <p>
    <strong>Description:</strong> ABC Company Inc.
    seeks a full-time mid-level software engineer to develop in-house tools.
  </p>
  <p><strong>Responsibilities:</strong></p>
  <ul>
    <li>Design and write specifications for tools for in-house customers</li>
    <li>Build tools according to specifications</li>
  </ul>
  <p><strong>Educational requirements:</strong></p>
  <ul>
    <li>Bachelor's Degree in Computer Science, Information Systems or related fields of study.</li>
  </ul>
  <p><strong>Experience requirements:</strong></p>
  <ul>
    <li>Mininum 3 years experience as a software engineer</li>
  </ul>
  <p><strong>Desired Skills:</strong></p>
  <ul>
    <li>Web application development using Java/J2EE</li>
    <li>Web application development using Python or familiarity with dynamic programming languages</li>
  </ul>
  <p><strong>Qualifications:</strong></p>
```

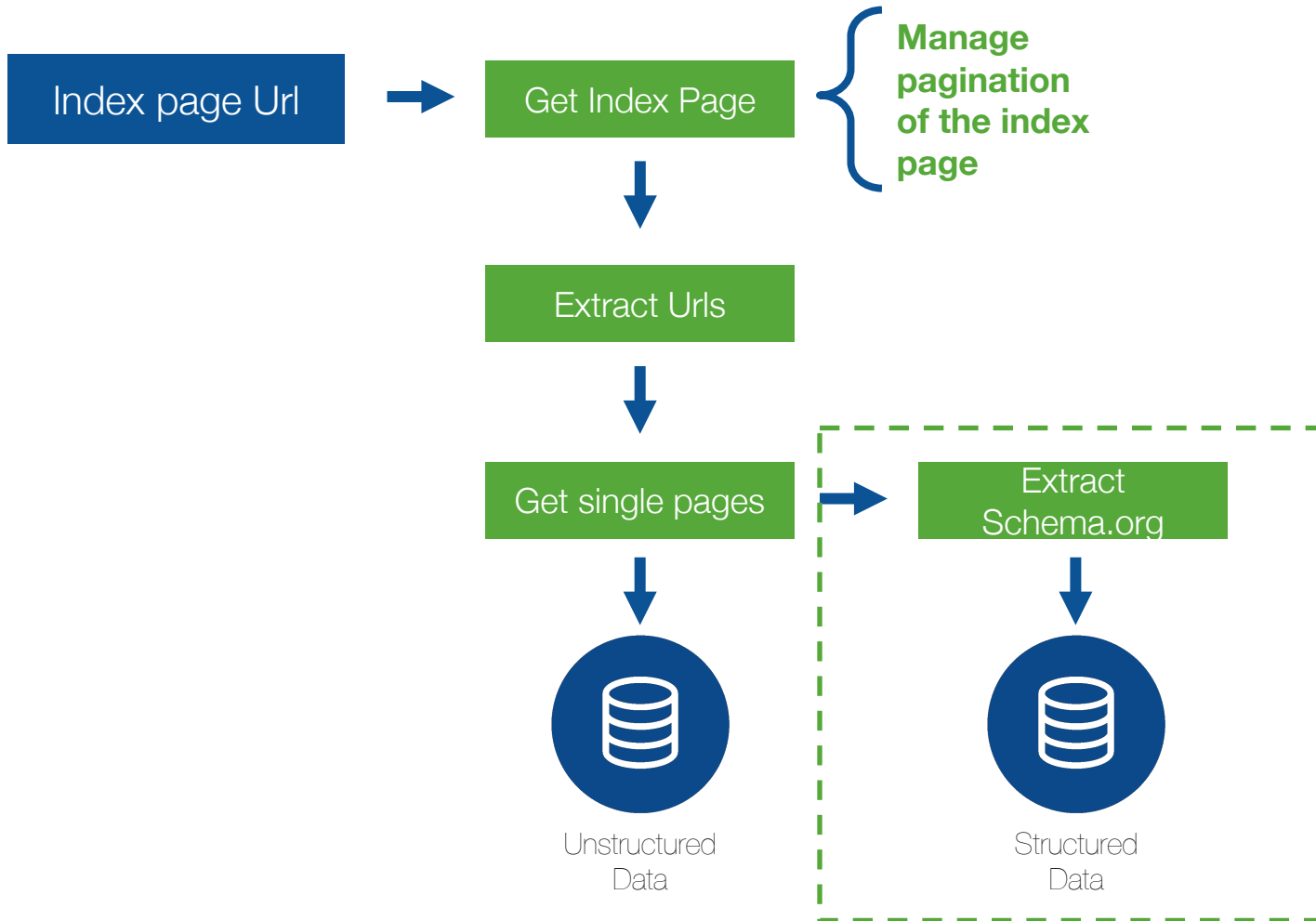


Using Microdata

```
<div itemscope itemtype="http://schema.org/JobPosting">
  <meta itemprop="specialCommitments" content="VeteranCommit" />
  <h2 itemprop="title">Software Engineer</h2>
  <span>
    <p><strong>Location:</strong> <span itemprop="jobLocation" itemscope itemtype="http://schema.org/Place"><span
      <span itemprop="addressLocality">Kirkland</span> <span itemprop="addressRegion">WA</span></span></span></p>
  </span>
  <p><strong>Industry:</strong> <span itemprop="industry">Computer Software</span>
  <br><strong>Occupational Category:</strong> <span itemprop="occupationalCategory">15-1132.00 Software Develop
  <br><strong>Hours:</strong> <span itemprop="employmentType">Full-time</span>, <span itemprop="workHours">40 h
  <br><strong>Salary:</strong> <span itemprop="salaryCurrency">USD</span> <span itemprop="baseSalary">100000</s
  </p>
  <p itemprop="description">
    <strong>Description:</strong> <span itemprop="hiringOrganization" itemscope itemtype="http://schema.org/Org
    seeks a full-time mid-level software engineer to develop in-house tools.</span>
  </p>
  <p><strong>Responsibilities:</strong></p>
  <ul itemprop="responsibilities">
    <li>Design and write specifications for tools for in-house customers</li>
    <li>Build tools according to specifications</li>
  </ul>
  <p><strong>Educational requirements:</strong></p>
  <ul itemprop="educationRequirements">
    <li>Bachelor's Degree in Computer Science, Information Systems or related fields of study.</li>
  </ul>
  <p><strong>Experience requirements:</strong></p>
  <ul itemprop="experienceRequirements">
    <li>Mininum 3 years experience as a software engineer</li>
  </ul>
  <p><strong>Desired Skills:</strong></p>
  <ul itemprop="skills">
    <li>Web application development using Java/J2EE</li>
    <li>Web application development using Python or familiarity with dynamic programming languages</li>
  </ul>
</div>
```



Adding Schema.org extraction capabilities to crawling



Considerations on Schema.org

The more Schema.org annotations are used, the more structured data are ingested, that improves:

Completeness
Complexity and Consistency



Crawling

Benefits of web crawling

1. Very low maintenance
2. High speed / High volume
3. High scalability

Problems of web crawling

1. Noise
2. **Unstructured data**

If Schema.org annotations are used...

Benefits of web crawling

1. Very low maintenance
2. High speed / High volume
3. High scalability
4. **Structured data**

Problems of web crawling

1. Duplication of OJAs



Scraping



Web scraping is data scraping used for extracting **structured** data from websites

JUNIOR SOFTWARE DEVELOPER

Location: United Kingdom
Application deadline: Saturday, 30 September 2017
Reference number: 100

APPLY NOW

Home > Now Hiring: Software Developers > Junior Software Developer

Share: [in](#) [t](#) [g+](#) [e](#)

Description

As Junior Software Developer, you will develop excellent software for use in field mapping, data collection, sensor networks, street navigation, and more. You will collaborate with other programmers and developers to autonomously design and implement high-quality web-based applications, restful APIs, and third party integration.

We're looking for a passionate, committed developer that is able to solve and articulate complex problems with application design, development and user experiences. The position is based in our offices in Harwell, United Kingdom.

Title:

Junior Software Developer

Area:

United Kingdom

Time:

Saturday, 30 September 2017

Description:

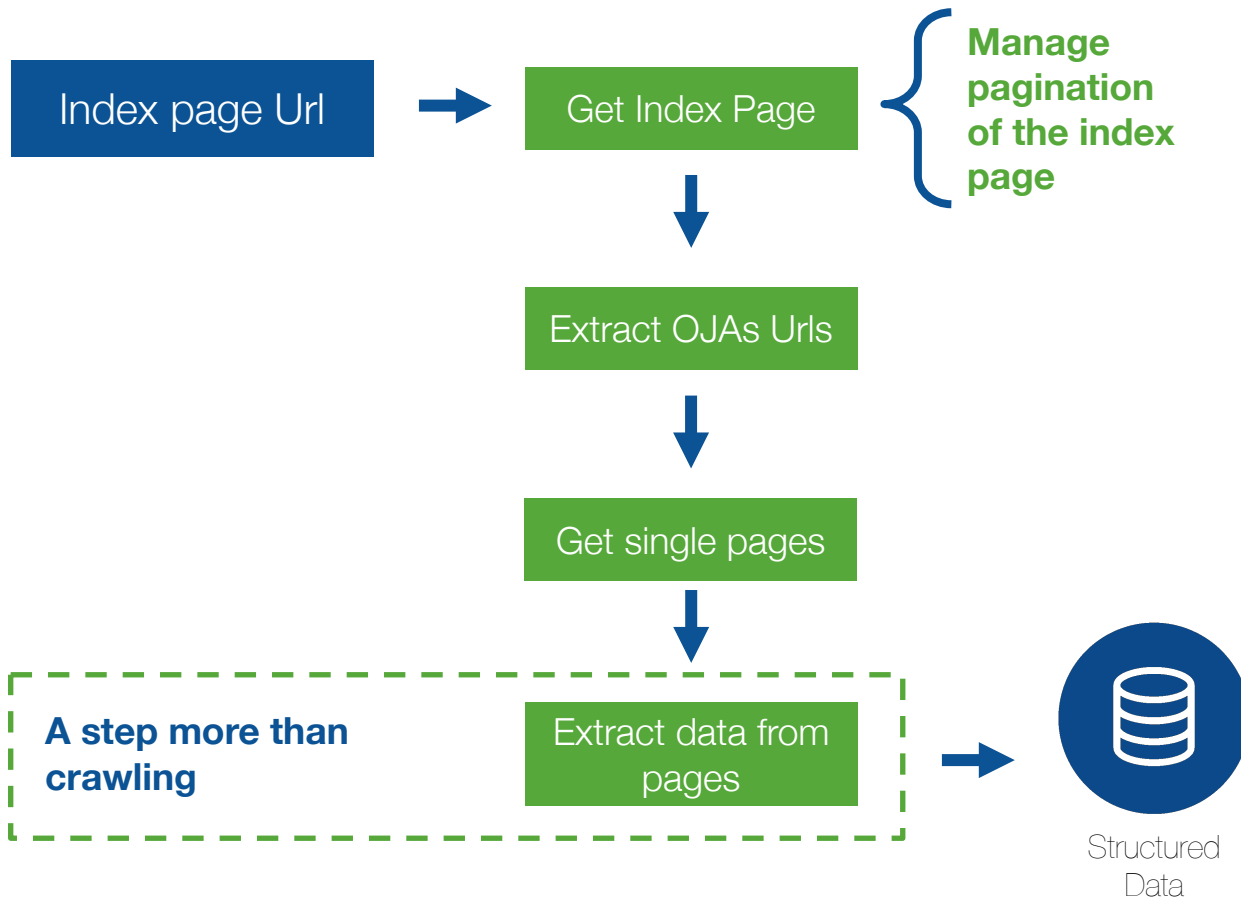
As Junior Software Developer, you will develop excellent software for use ...



CEDEFOP

European Centre
for the Development
of Vocational Training

Scraping process



All OJAs are downloaded

Only OJAs are downloaded: scrapers are set to download only job ads

Scrapers can get specific data from the pages and collect them in the database as structured data



Scraping

Benefits of web scraping

1. **Structured data**
2. High quality

Problems of web scraping

1. High maintenance
2. Low scalability

Scraping is effectively used on web-sites with a **well defined and uniform structure**

Secondary Job-portals links to **hundreds of other websites**: a scraper capable to get data from all of them cannot be developed...

... scraping can be performed only on Primary job-portals



Crawling and Scraping Fairness

Crawling and Scraping perform **lots of requests** to website servers

Performing «aggressive» crawling (or scraping) can be a problem if servers cannot manage large amount of requests. Websites could react banning and making access to their pages forbidden to crawler and scrapers .

To avoid overloading of the web-sites:

add delays according to Crawl-Delay directive in Robot.txt file (rarely indicated)

add delays to slow down access frequency on all the websites



Direct Access



Some job-portals provides mechanisms to enable **Direct Access** to (part of) their database to enable the integration of their OJAs in other websites. These mechanism could be:

APIs: interfaces through which interactions happen between a job-portal and our application that use its data.

RSS feeds: type of web feed which allows users to access updates to online content in a standardized, computer-readable format.

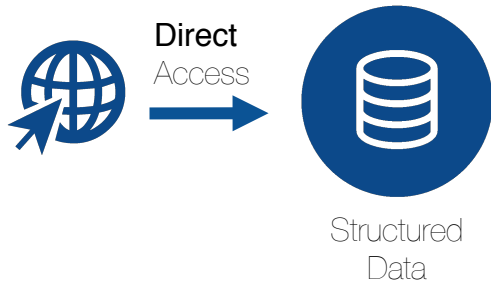
In other words, through APIs and RSS feeds we are **directly and online** connected with the **database** of the sources.

After signing a **formal agreement** with a job-portal, data can also be retrieved:

- using batch downloading of exports from their database, using for example **FTP** shared folders
 - using more specific and detailed **APIs** or **RSS feeds**



Direct Access: Partial Version



Title:

Junior Software Developer

Area:

United Kingdom

Time:

Saturday, 30 September 2017

Partial
Description:

As Junior Software Developer, you will develop excellent software for use ...

Job-post url:

<http://somesite.com/id=job-id>

Direct Access to the job-portal database

Partial view of all the attributes of the OJA:

- only main attributes are available
- description is provided in a short version

Job post url is always provided

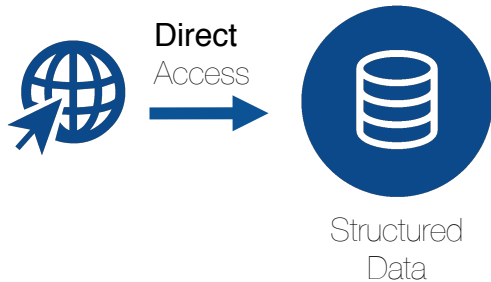
Crawling is required to get the complete pages with all the useful information



CEDEFOP

European Centre
for the Development
of Vocational Training

Direct Access: Full Version



Direct Access to the job-portal database

Complete view of all the attributes of the OJA

Title:

Junior Software Developer

Area:

United Kingdom

Time:

Saturday, 30 September 2017

Description:

As Junior Software Developer, you will develop excellent software for use ...

neuvoo



CEDEFOP

European Centre
for the Development
of Vocational Training

Direct Access



Benefits of **Partial** Direct Access

1. Low maintenance
2. High speed / High volume
3. High scalability
4. High quality
5. **Structured data (partially completed)**

Problems of **Partial** Direct Access

1. **Variety of structures**

Benefits of **Full** Direct Access

1. Low maintenance
2. High speed / High volume
3. High scalability
4. High quality
5. **Structured data**

Problems of **Full** Direct Access

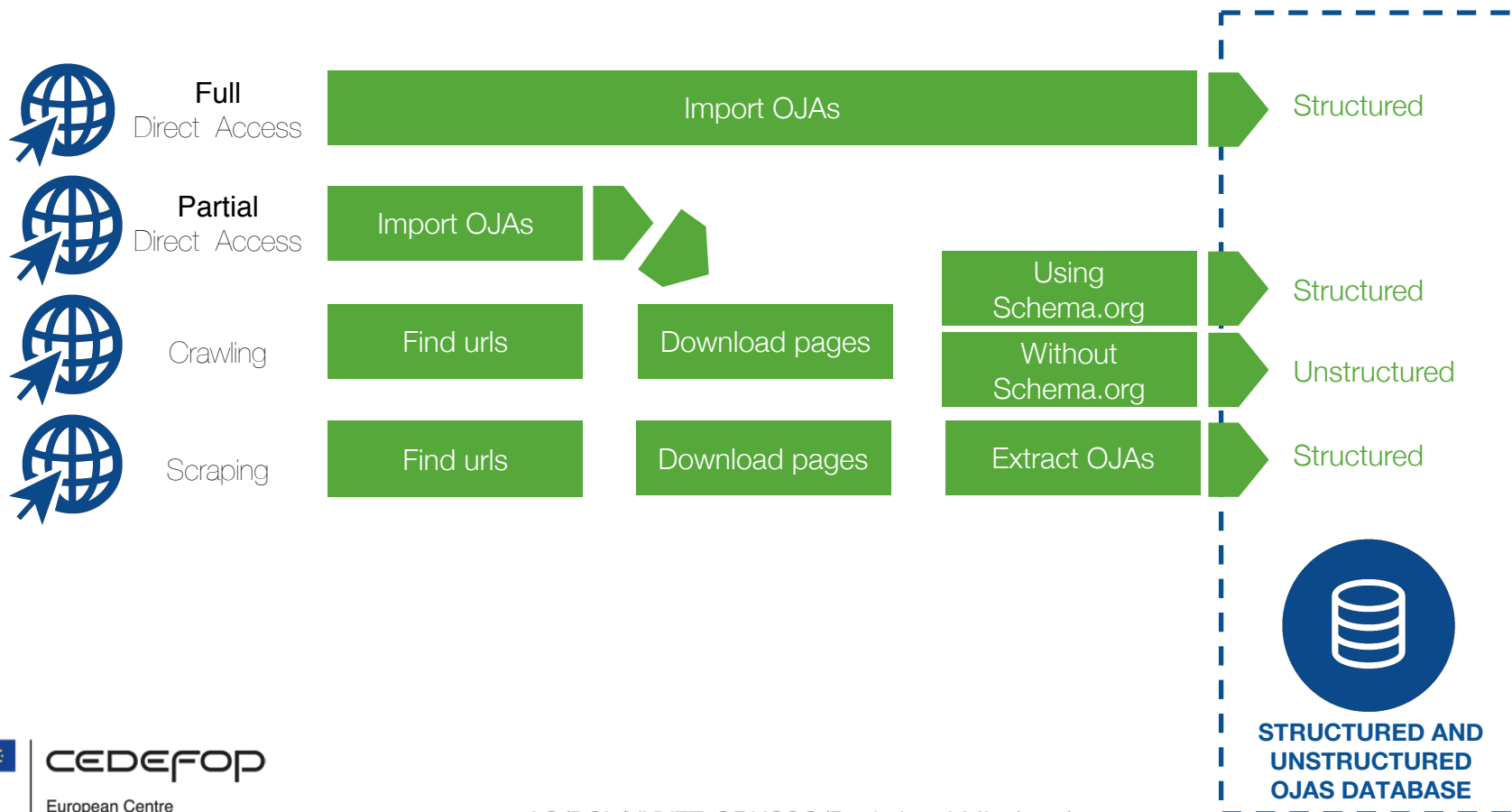
1. High level agreement
2. **Variety of structures**



Data Ingestion - Storage



Output of data ingestion step is a database with **structured and unstructured data** that will have to be cleaned and elaborated in the further steps to extract features and connect to classifications



Topics

1. Source selection strategy
2. Data ingestion techniques
3. Preliminary data and expected volumes
4. Scheduling, parallel processing and monitoring



Preliminary snapshot data

- Testing activities involved random (from single shot to continuous) ingestion from each source during the last month
- No time frame available, spot extractions

Early release

- 7 countries covered
- 70% of the landscaping list already covered (most relevant sources)
- 5.872.462 OJV collected

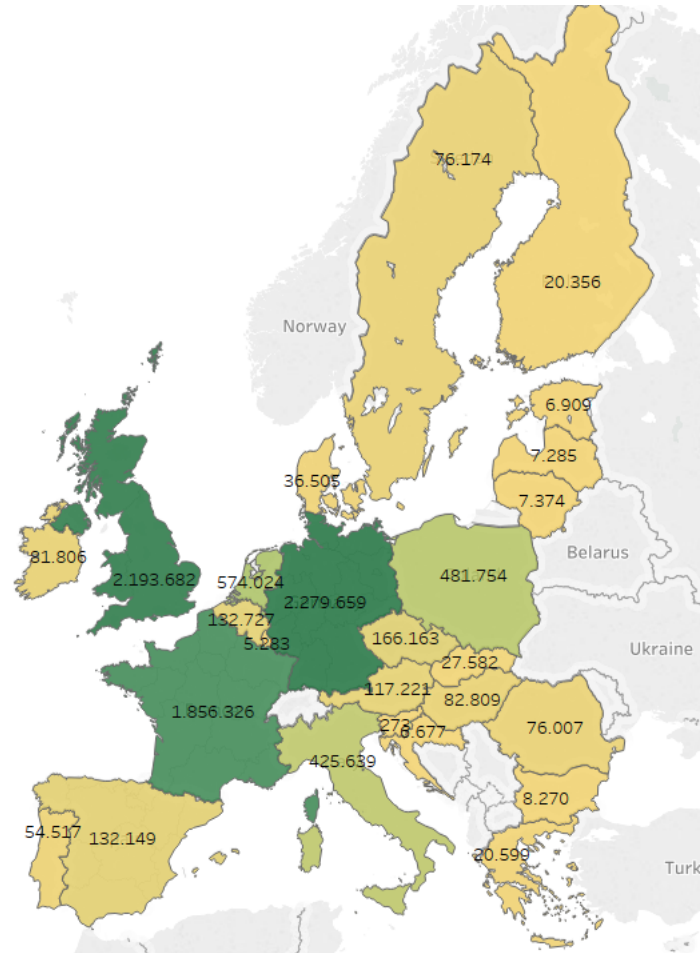
Full release

- 26 countries covered
- 65% of the landscaping list already covered (most relevant sources)
- 8.878.692 OJV collected



Preliminary snapshot data

Vacancies by country



CEDEFOP

European Centre
for the Development
of Vocational Training

Expected volumes



Ingestion: 200k job vacancies/day, 70 millions job vacancies/year



Analysis: 18+ millions job vacancies/year



400+ categorized occupations



1.500+ categorized skills



Topics

1. Source selection strategy
2. Data ingestion techniques
3. Preliminary data and expected volumes
4. Scheduling, parallel processing and monitoring



Managing Data Ingestion



Data ingestion is the process of obtaining and importing **data** from **web portals** and storing in a **database**.



OJAs

Import OJVs



**STRUCTURED AND
NO-STRUCTURED
OJAS DATABASE**



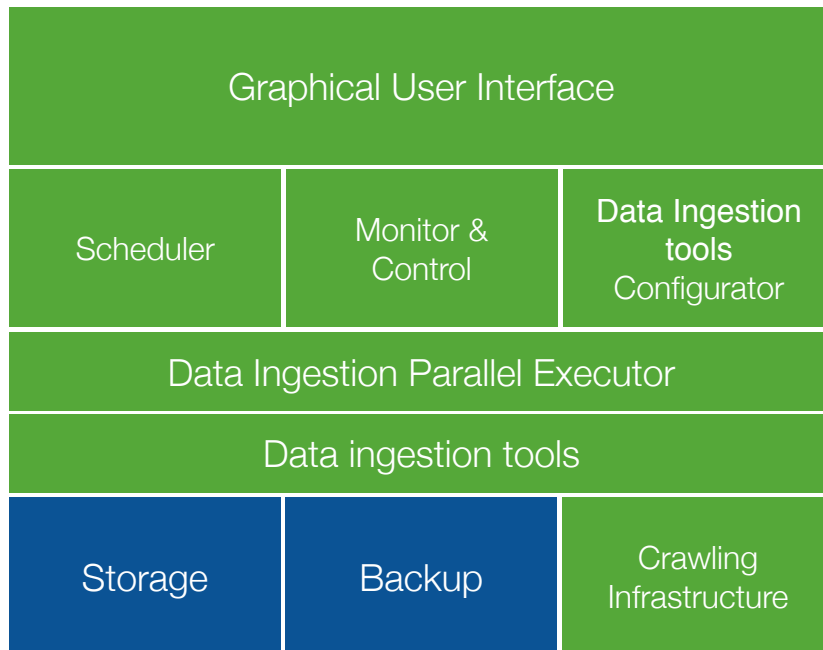
CEDEFOP

European Centre
for the Development
of Vocational Training

Managing Data Ingestion



Data ingestion is the process of obtaining and importing **data** from **web portals** and storing in a **database**.



Storage: a database capable to store **Big Data**

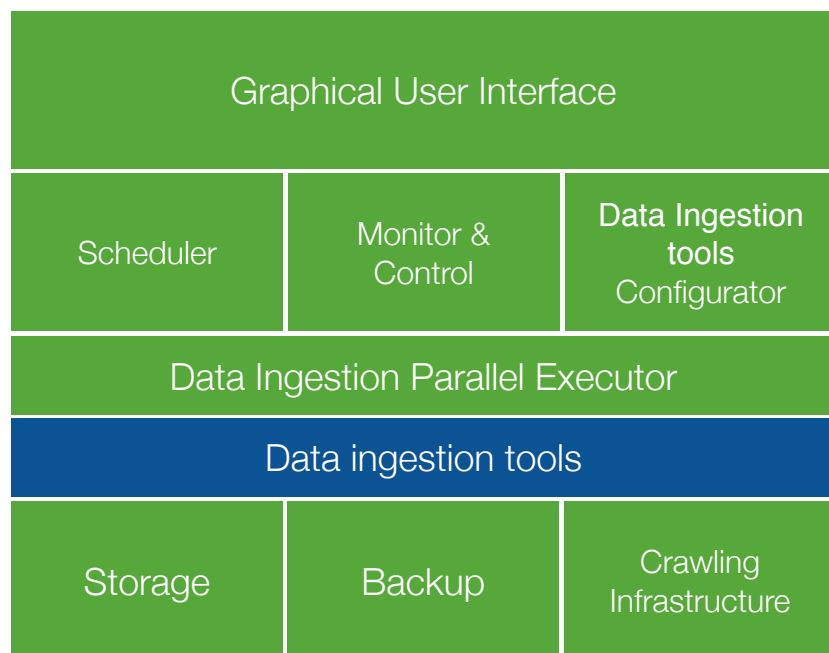
Backup: an infrastructure to store data for disaster recovery



Managing Data Ingestion



Data ingestion is the process of obtaining and importing **data** from **web portals** and storing in a **database**.



Custom software components
Direct Access connectors
Scrapers

Generic configurable component
Crawling



Data Ingestion Tools Development

What do we use Direct Access connectors for?

What do we use Scrapers for?

What do we use Crawler for?



Data Ingestion Tools Development

What do we use Direct Access connectors for?

What do we use Scrapers for?

What do we use Crawler for?

- **Direct Access and Scraping are better for sites with large amounts of OJAs**
 - **Full Direct Access is the best method but requires a formal agreement**

Maintenance costs and agreement costs are offset by the high quality of injected OJAs and by the high speed of ingestion

- **Partial Direct Access is better than scraping**

Even if some crawling efforts are required, data quality is higher and development is easier

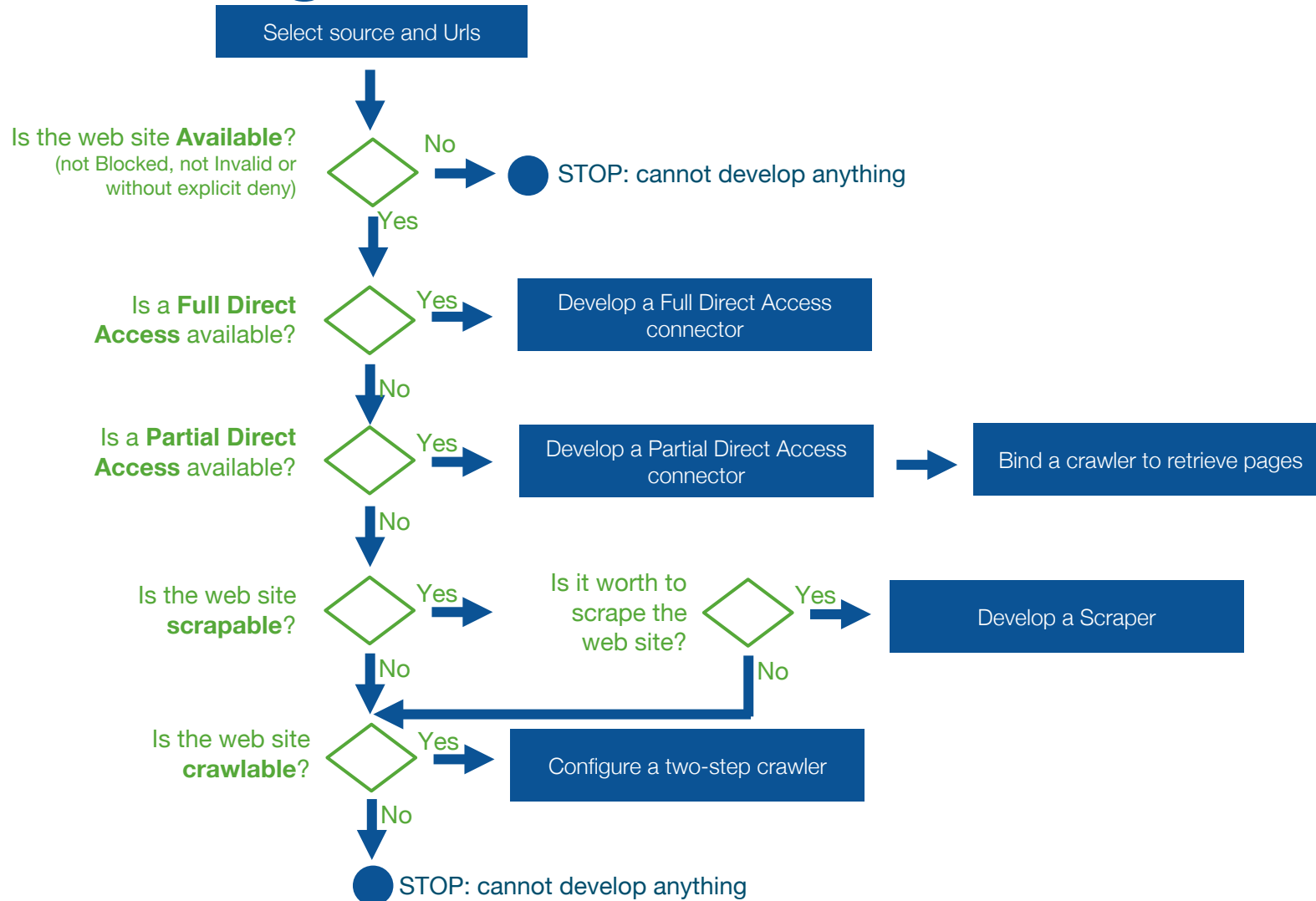
- **Crawling is easy to develop, scalable and robust**

Crawling can be used on many sites using any format (e.g. to complete Partial Direct Access)

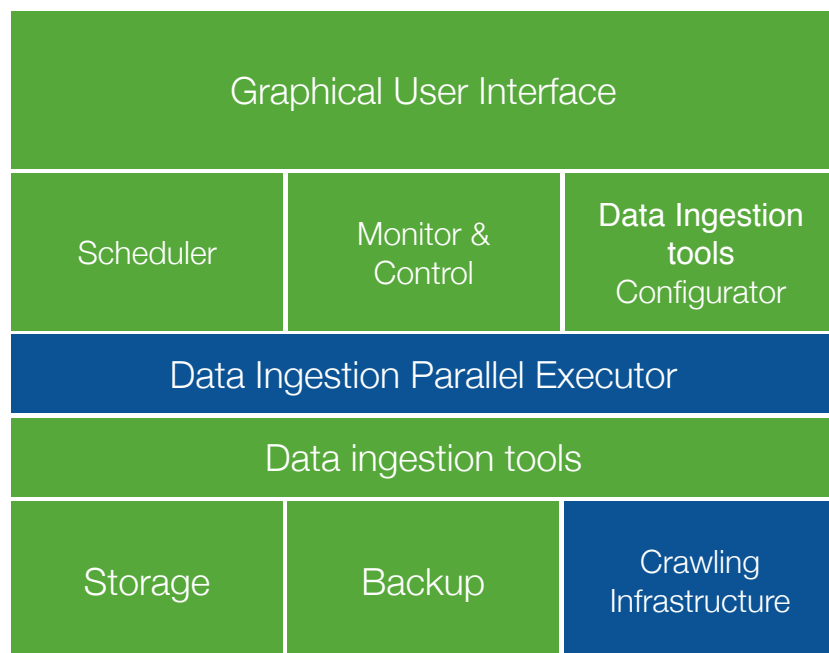
If the crawled source uses Schema.org annotations, data quality is really improved



Data Ingestion: a Decision Tree



How we perform page downloads



Parallel Executor

Volumes per day
(from preliminary data)
200.000 OJAs/day

Download speed
(from tests)
5 sec/OJA



1.000.000 sec/day !!!

Parallelization
with factor 11 (at least)

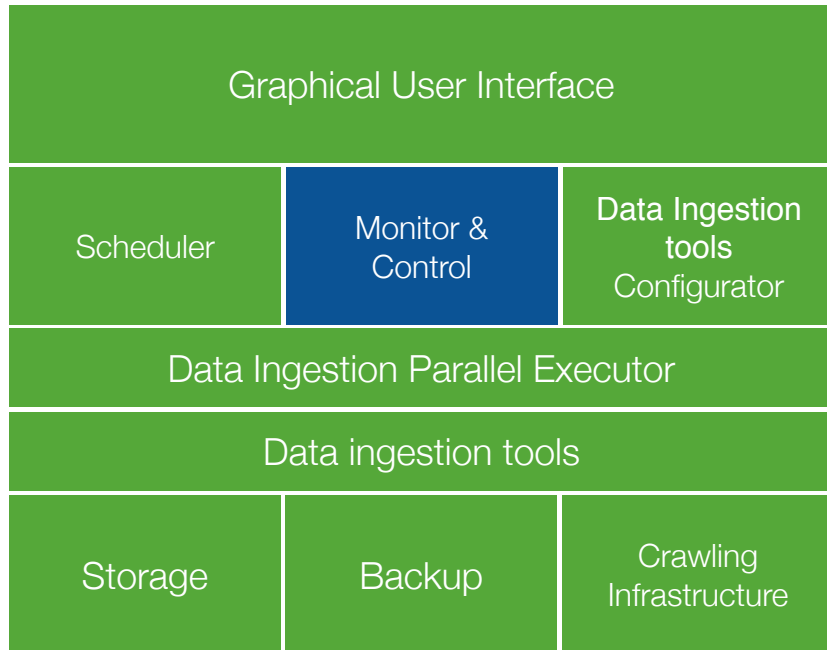
Data Ingestion Parallel Executor can run **20** parallel data ingestion processes
Crawling infrastructure uses up to **100** browsers



Monitor & Control

Data ingestion performance are affected by external elements (e.g. website speed; network bandwidth)

Data ingestion processes can fail for different causes (e.g. loss of connectivity; internal server errors from job portals; web-sites changes)



Monitor:

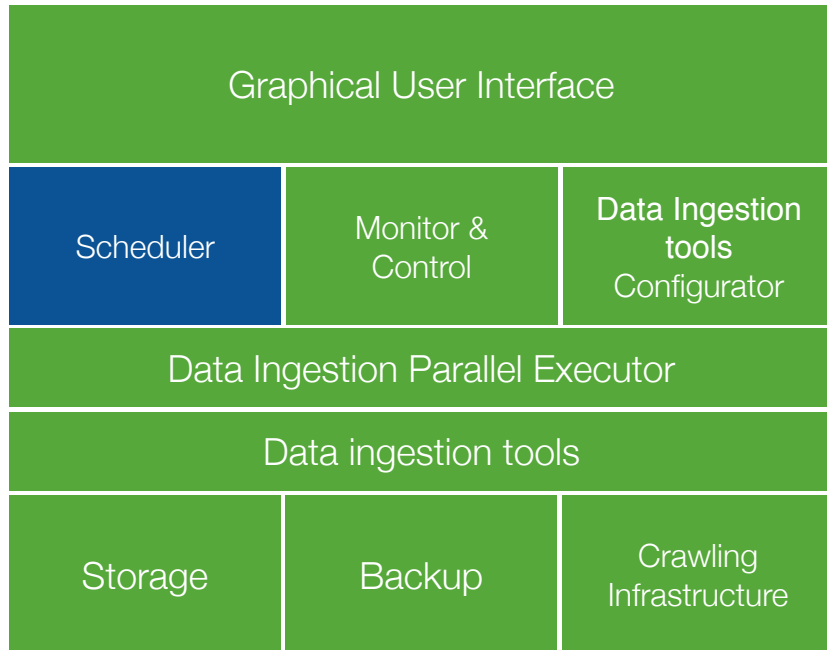
to know the state of the running and run processes (running, completed, failed, ...)

Control:

to launch and resume processes
to react to root causes



Scheduling Parallel Processes



Scheduler

Run specific data ingestion tool at a specific time

Distribute data ingestion to avoid queue overloading

Manage the automatic scheduling of activities after failures

Notify (via email) the status of the current data ingestion tasks

