

# USPTO Patent Processor CLI Documentation

---

## Overview

The USPTO Patent Processor CLI is a command-line tool for downloading, extracting, and analysing patent data from the USPTO (United States Patent and Trademark Office) database. It supports both patent grants and applications, with capabilities for single-year and multi-year processing.

## Installation

### Prerequisites

- Python 3.7+
- Required Python packages:

```
pip install requests lxml beautifulsoup4 nltk tqdm aiofiles
```

### Setup

1. Clone the repository
2. Navigate to the project directory
3. Install dependencies

## Usage

### Basic Command Structure

```
python patent_cli.py [operation_mode] [options]
```

### Operation Modes

#### 1. Single Year Processing

Process patents from a specific year:

```
python patent_cli.py --year 2020 --kind grant
```

#### 2. Year Range Processing

Process patents from a range of years:

```
python patent_cli.py --year-range 2018 2020 --kind grant
```

3. Process Existing Files

Process previously downloaded patent files:

```
python patent_cli.py --input-dir ./my_patents --process-only
```

Optional Arguments

Argument	Description	Default
--kind	Patent type (grant or application)	grant
--output-dir	Output directory for downloads	./data
--workers	Number of worker processes	4
--download-only	Only download files	False
--unzip-only	Only unzip files	False
--process-only	Only analyse patents	False

Operation Flags

Download Only

Download patent files without processing:

```
python patent_cli.py --year 2020 --kind grant --download-only
```

Unzip Only

Unzip previously downloaded files:

```
python patent_cli.py --year 2020 --kind grant --unzip-only
```

Process Only

Analyse patents from existing files:

```
python patent_cli.py --year 2020 --kind grant --process-only
```

## Examples

1. Process grants from 2020 with custom output directory:

```
python patent_cli.py --year 2020 --kind grant --output-dir ./patent_data
```

2. Download applications from 2018-2020:

```
python patent_cli.py --year-range 2018 2020 --kind application --download-only
```

3. Process existing files with 6 workers:

```
python patent_cli.py --input-dir ./my_patents --process-only --workers 6
```

## Output Structure

### Directory Structure

```
output_dir/  
├─ patent_grant_YEAR_zip/    # Downloaded ZIP files  
└─ patent_grants_YEAR/      # Extracted and processed files
```

### Database Output

- Results are stored in SQLite database (**patents.db**)
- Two main tables:
  - **patent\_examples**: Individual patent examples
  - **patent\_statistics**: Aggregated patent statistics

## Error Handling

- The tool provides detailed error messages and progress updates
- Operations can be interrupted safely with Ctrl+C
- Failed operations can be resumed using operation flags

## Limitations

1. Years supported: 1976-2025
2. Memory requirements increase with concurrent workers
3. Processing time varies based on:
  - Data volume
  - Hardware capabilities

- Network speed

## Performance Tips

1. Adjust worker count based on available CPU cores
2. Use `--download-only` and `--process-only` for large datasets
3. Process years sequentially for memory-constrained systems

## Troubleshooting

### Common Issues

1. Download Failures

```
# Retry with single year
python patent_cli.py --year YEAR --kind grant --download-only
```

2. Processing Errors

```
# Reduce workers and retry
python patent_cli.py --year YEAR --kind grant --workers 2
```

3. Memory Issues

```
# Process in smaller chunks
python patent_cli.py --year YEAR --kind grant --process-only --workers 2
```

## Support

For issues and feature requests, please create an issue in the project repository.

## License

This project is licensed under the MIT License - see the LICENSE file for details.