Exploring Software Solutions to Assist in Data Analysis

Pieter Vercruysse, Ministry of Finance, Belgium

My Background

- Pieter Vercruysse
- Ministry of Finance, Belgium
- pieter.vercruysse@minfin.fed.be
- Special Tax Directorate: specialized in Fiscal Fraud
- Head of Unit of an E-Audit Team
- Main tasks of my unit: analysis of accounting, mailboxes, office documents, pdfs, databases, bank statements, …
- Previous positions: corporate income tax, litigation service

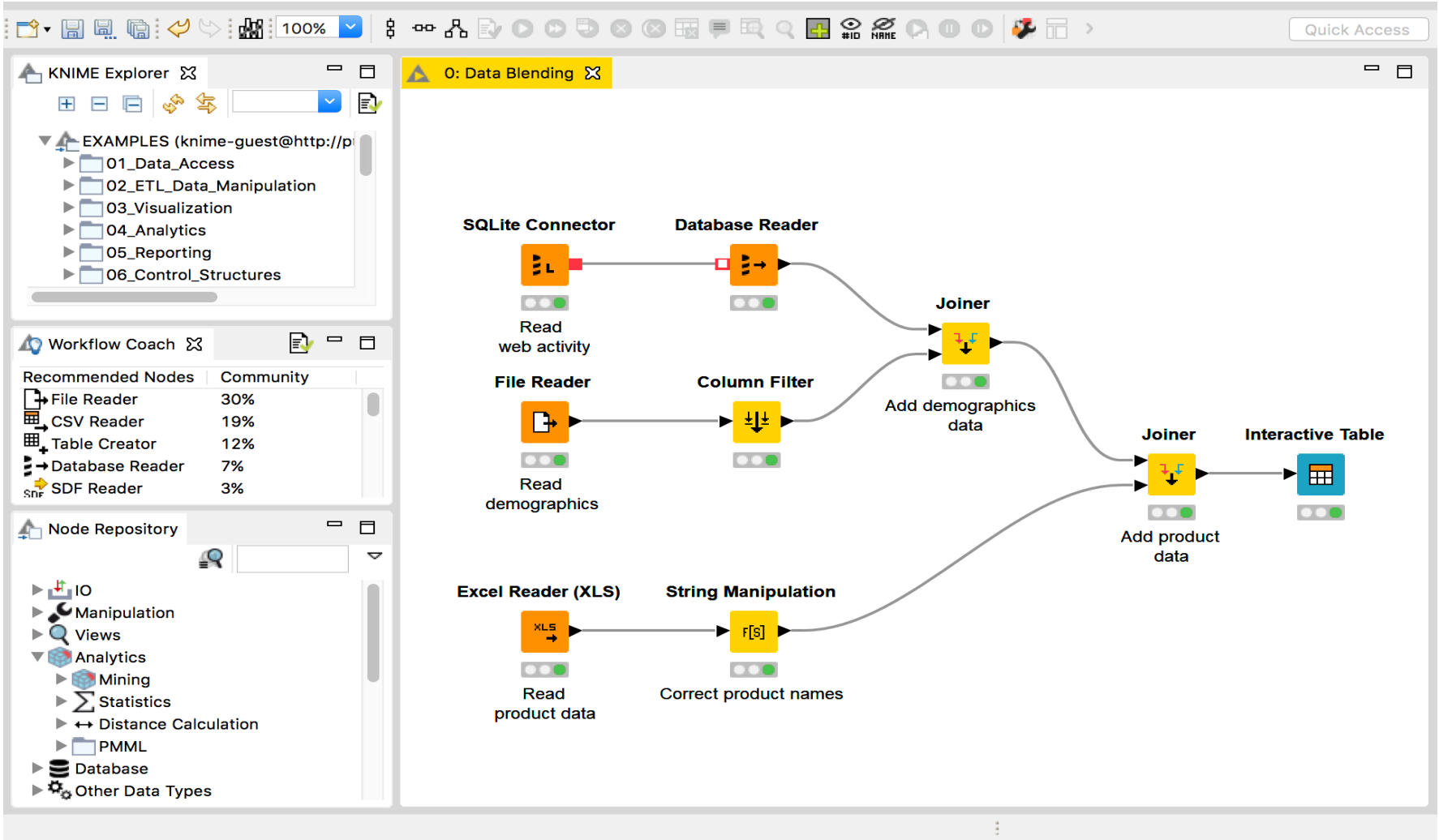
OECD / OCDE: Restricted Use - À usage restreint

Goal of this presentation

- Two important concepts:
  - *Workflows*
  - *Notebooks*
- Two important open source and free tools:
  - *KNIME*
  - *Pandas, a Python library*

OECD / OCDE: Restricted Use - À usage restreint

Example of a workflow in KNIME



OECD / OCDE: Restricted Use - À usage restreint

Example of a notebook in Pandas (Python)

jupyter Jupyter_Pandas_Cheat_Sheet Last Checkpoint: a few seconds ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

Jupyter Pandas Cheat Sheet

In [0]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
import pandas as pd

df = pd.read_csv('Data/IMDB_Movie_Data.csv' ,sep=",")
```

Exploring Data

In [3]:
```
df.head()        # first five rows
```

Out[3]:

	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
0	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0
1	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0
2	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0
3	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0
4	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02	40.0

In [0]:
```
df.tail()        # last five rows
```

Structure of this presentation

- ❑ General info
- ❑ “Nice to Know” Topics
- ❑ Demos
- ❑ Hyperlinks

OECD / OCDE: Restricted Use - À usage restreint

General information

Structured data versus unstructured data

- Structured data is organized and formatted in a specific way, typically in a table of a database or spreadsheet with rows and columns.
  This data is easy to search, sort, and analyze because it follows a schema.
- Unstructured data is not organized in a pre-defined manner. It is more difficult to process and analyze because it does not follow a fixed structure or schema.
  Examples of unstructured data include text documents, emails, social media posts, and images.
  Unstructured data is more complex and requires more sophisticated techniques to extract useful information from it.

OECD / OCDE: Restricted Use - À usage restreint

Traditional analysis software

- For analysis of **structured data** (tables, lists), for instance:
  - *Used by normal users: Excel, Access ;-)*
  - *Used by e-auditors: ACL, Arbutus, IDEA*
- For analysis of **unstructured data** (different types of files), for instance: ZyLAB ONE, an eDiscovery solution that helps to uncover and review critical information buried within the documents and files of your organization.
- For **forensic** analysis, for instance : Encase or FTK (Forensic Toolkit): analysis of a forensic image of a hard disk, and other storage media

OECD / OCDE: Restricted Use - À usage restreint

Open source and free alternatives

- For analysis of **structured data** (tables, lists): Knime, Python, R
  - KNIME Analytics Platform is an open source and free software with an intuitive, visual interface that lets you build analyses of any complexity level
  - Pandas, a Python library
  - R, a language and environment for statistical computing and graphics
- For analysis of **unstructured data** (different types of files): DocFetcher (a portable version exists!)
- For **forensic** analysis of images: Autopsy
- For **Windows**: Linux -> especially Linux Mint (see hyperlinks for installation guide)

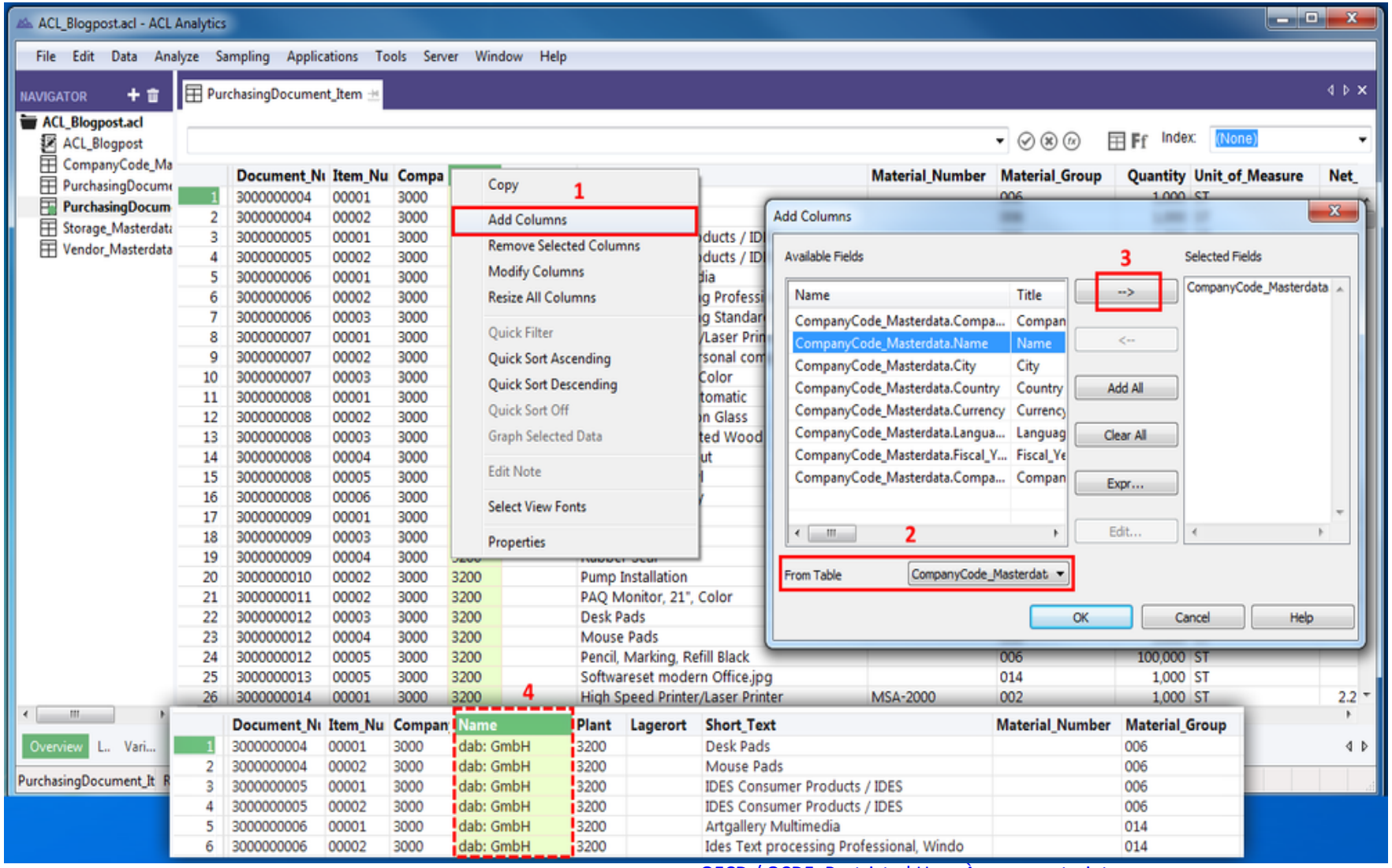
OECD / OCDE: Restricted Use - À usage restreint

ACL, Arbutus and IDEA, commercial software for data analysis of structured data

- ACL, Arbutus and IDEA: computer-aided auditing solutions designed for data extraction and analysis and for taking samples from populations
- These are typical tools that are used by e-auditors to determine the accuracy of account balances and transactions
- Advantages: industry standard, very fast, possible use of wizards
- Commercial software

OECD / OCDE: Restricted Use - À usage restreint

Printscreen of ACL, commercial e-audit software



OECD / OCDE: Restricted Use - À usage restreint

KNIME, open source and free alternative to ACL, Arbutus and IDEA (1)

Benefits of using a workflow in KNIME

- ❑ You get a good overview of all the tasks performed on the source data.
- ❑ If you have made a mistake somewhere, you just adjust the node or code and run the workflow again
- ❑ You get insight into the calculations by viewing the configuration screen and it's intuitive.
- ❑ Easy reuse: the workflow is applicable to other data with the same structure!
- ❑ Easy for automation: you can put a workflow in a loop and apply it to thousands of files if necessary!

OECD / OCDE: Restricted Use - À usage restreint

KNIME, open source and free alternative to ACL, Arbutus and IDEA (2)

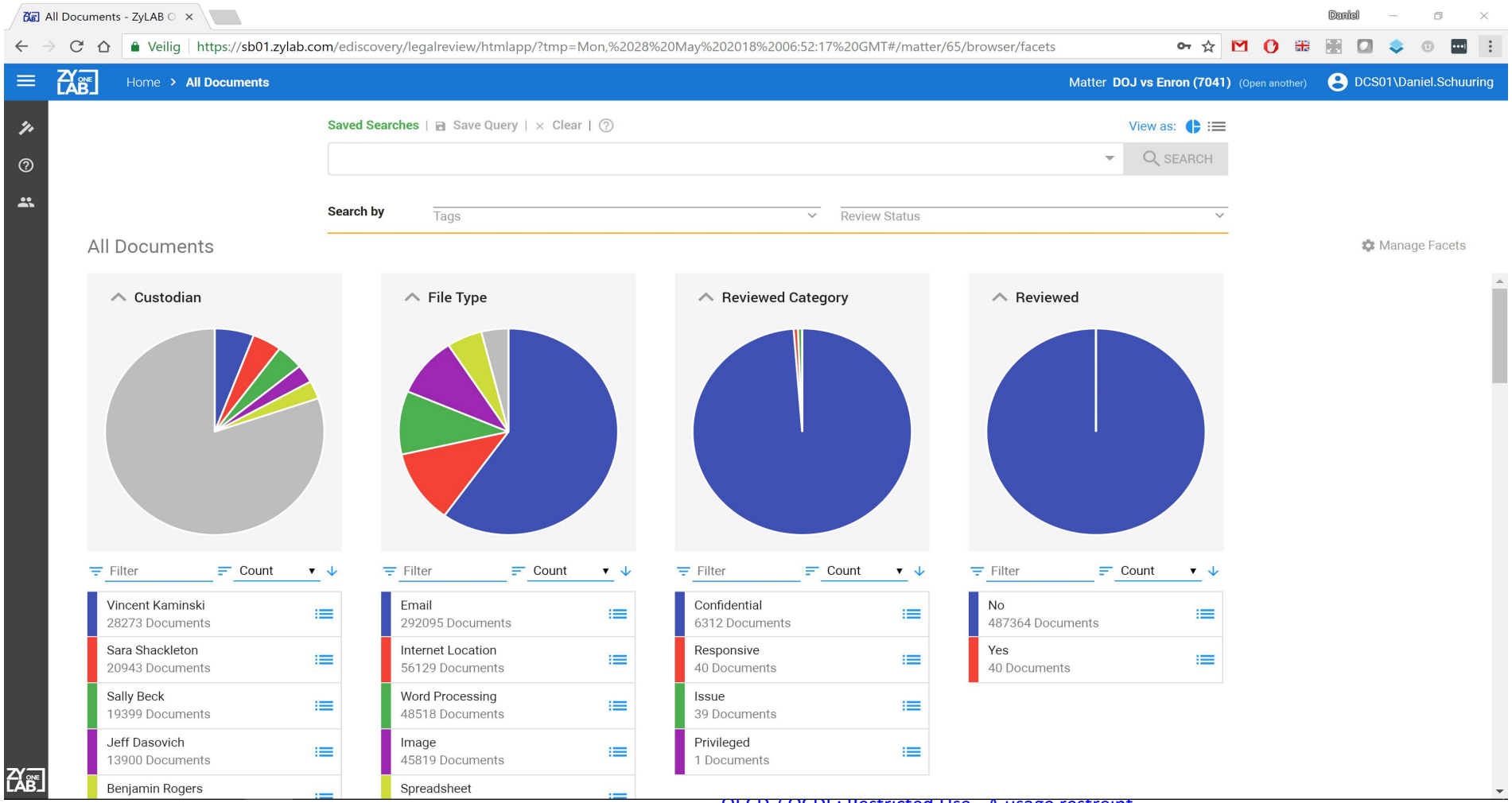
- ❑ You can document the workflow
- ❑ Enough examples of workflows on the KNIME Hub
- ❑ You can easily exchange workflows with other people: export, import of workflows
- ❑ You become more independent of the ICT department (you don't have to be a programmer!) and you can be creative: empowerment for tax officers ;-)
- ❑ Free webinars for example: https://www.knime.com/events/approaches-to-fraud-detection-autoencoder-isolation-forest

Disadvantages

- ❑ It is difficult to print out the workflow in Knime
- ❑ It is difficult to create reports in Knime. Python is better for this

OECD / OCDE: Restricted Use - À usage restreint

Printscreen of Zylab, commercial software for e-discovery



OECD / OCDE: Restricted Use - A usage restreint

Printscreen of DocFetcher, a free and open source application for e-discovery

DocFetcher

Overview | Download | Screenshots | More

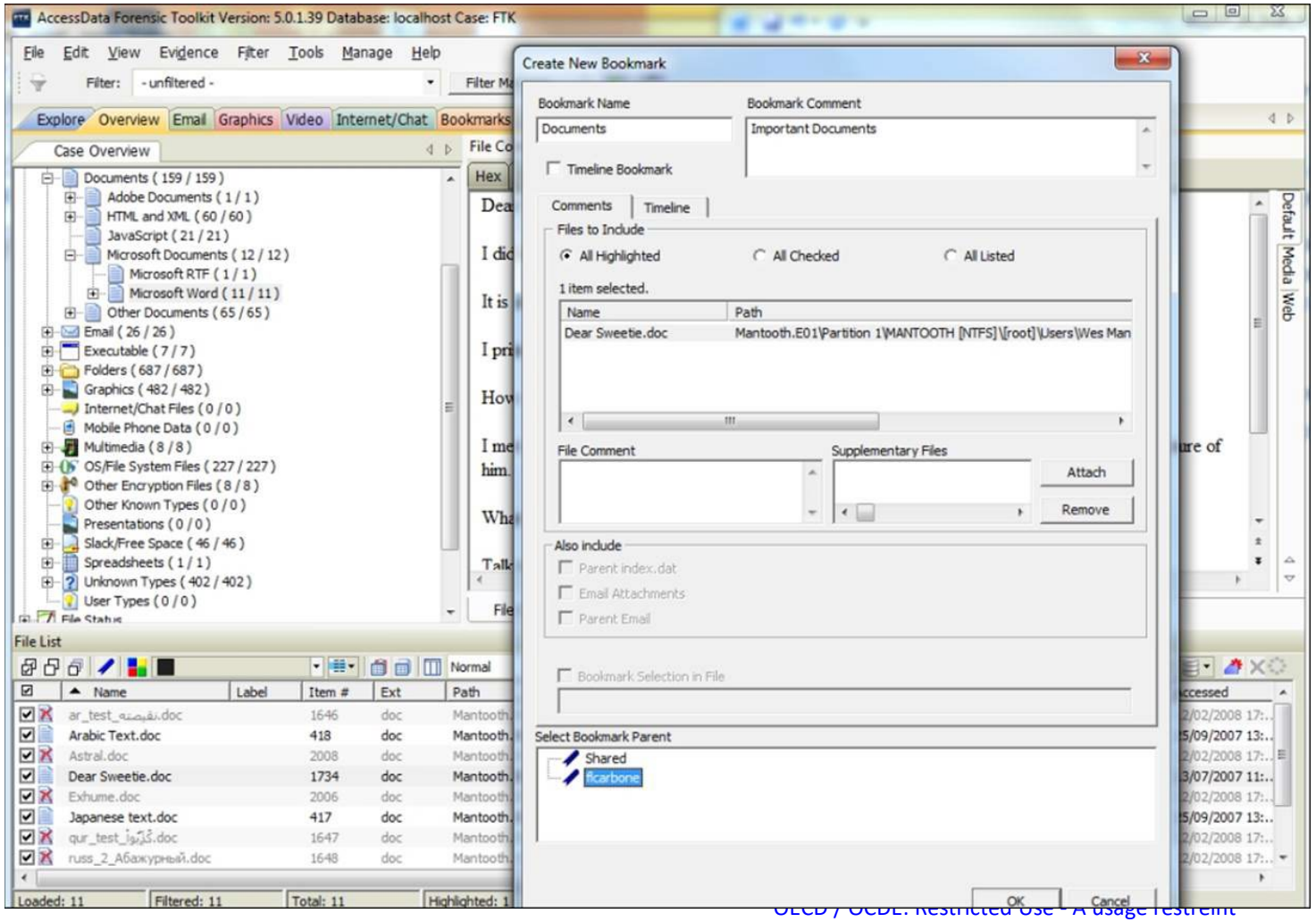
Note: You may be interested in DocFetcher Pro, the commercial big brother of DocFetcher with more features and fewer bugs.

Description

DocFetcher is an Open Source desktop search application: It allows you to search the contents of files on your computer. — You can think of it as Google for your local files. The application runs on Windows, Linux and OS X, and is made available under the Eclipse Public License.

OECD / OCDE: Restricted Use - À usage restreint

Printscreen of FTK (Forensic Tool Kit), a commercial software product



OECD / OCDE. Restricted Use - A usage restreint

Printscreen of Autopsy, a free and open source application for forensic analysis

1. Getting Started

Open Autopsy and create a new case.

OECD / OCDE: Restricted Use - À usage restreint

A better way of working with regard to data analysis

1) Workflows

There are several advantages to using workflows in data analysis:

- ❑ Improved efficiency: Workflows can help to AUTOMATE and streamline data analysis processes, **so ideal for repetitive tasks**
- ❑ Increased accuracy: Workflows can help to ensure that data analysis tasks are performed consistently and correctly
- ❑ Enhanced reproducibility: Workflows can be documented and shared with others

OECD / OCDE: Restricted Use - À usage restreint

2) Notebooks

Notebooks are a popular tool for data analysis, and they offer several advantages:

- **Interactive computing**: Notebooks allow data analysts to write and execute code, as well as include text, images, and other multimedia content, in a single document
- Reproducibility: Notebooks allow data analysts to document their analysis process step by step, making it easier to reproduce and share their results with others **(audit trail)**
- **You can make beautiful reports that can be exported to html or pdf!** Also reports without code cells!
- Collaboration: Notebooks can be shared with others and collaborated on in real-time
- Ease of use: Notebooks have a user-friendly interface and are easy to learn

OECD / OCDE: Restricted Use - À usage restreint

Some “Nice to Know” Topics

- ❑ ChatGPT
- ❑ Everything
- ❑ Regular expressions = “Advanced Wildcards”

OECD / OCDE: Restricted Use - À usage restreint

ChatGPT

ChatGPT is a chatbot developed by OpenAI. It is a machine learning model that is trained to generate human-like text. It can be used to ask questions and receive answers in real-time, making it a convenient and efficient way to find information.

It was launched on November 30, 2022.

Advantages of ChatGPT

- ❑ It can offer information in clear, simple sentences, rather than a list of Internet links
- ❑ It can explain concepts in a way that people can easily understand.
- ❑ You can easily copy and reuse the results

Disavantages of ChatGPT

- ❑ ChatGPT is only able to provide information and assistance based on the data that it has been trained on
- ❑ Not to be used for sensitive information

OECD / OCDE: Restricted Use - À usage restreint

Everything

Everything is a search tool that allows users to quickly locate files and folders on their computer. It's freeware and it has a portable version.

It has several advantages over other search tools

- Speed: this tool is designed to be fast, and it can search through millions of files in a matter of seconds
- Flexibility: Everything allows users to search for files and folders using a wide range of filters and options, including size, date modified, and file attributes
- It can index mapped network drives/NAS/Network shares and external drives

Disavantages of Everything

- The file content is not indexed

OECD / OCDE: Restricted Use - À usage restreint

Regular expressions (1)

A regular expression (regex for short) is a sequence of characters that defines a search pattern. Regular expressions are mainly used for pattern matching with strings.

- F.i. Find all files with OECD or OESO or OCDE in the file name: OECD|OESO|OCDE
- F.i. Check if there are Belgian IBAN numbers in the file names: BE\d{14}.txt

Example of a Belgian IBAN: BE68539007547034

Regular expressions are used in many programming languages and tools. They are powerful tools for extracting information from text, and they are often used in data analysis.

OECD / OCDE: Restricted Use - À usage restreint

Regular expressions (2)

Advantages of regular expressions

- Precision: Regular expressions allow users to specify exactly what they are looking for in the text: f.i. the text between two words
- Automation: Regular expressions can be used to automate text processing tasks, such as finding and replacing specific patterns of text

Disadvantages of regex

- Regular expressions can be difficult to read, understand and use
- It’s hard to remember the meaning of the symbols that are used in the patterns

OECD / OCDE: Restricted Use - À usage restreint

Examples

Introduction to KNIME + Workflows: Demos

1) Simple analysis on the Bob's Boats example spreadsheet
2) Analysis of financial statements with user-defined ratios

- ❑ Using xbrl files (extensible business reporting language, the open international standard for digital business reporting)
- ❑ Using the XPath query language

OECD / OCDE: Restricted Use - À usage restreint

Introduction to Python + Notebooks: Demos

1) Simple analysis on the Bob's Boats example spreadsheet
2) Time series and anomaly detection (machine learning)

Remark: the combination of KNIME and Python is also possible and very powerful!

OECD / OCDE: Restricted Use - À usage restreint

Introduction to Neo4J + Graph Databases: Demo

- ❑ Graph database => Focus on Data Relationships
- ❑ Application: Dynamic representation of the money flows
- ❑ Combination of KNIME and Neo4J is possible: no programming necessary!

OECD / OCDE: Restricted Use - À usage restreint

Hyperlinks (1)

- KNIME: https://www.knime.com/downloads
- KNIME Hub: https://hub.knime.com/
- KNIME Webinar on fraud: https://www.knime.com/events/approaches-to-fraud-detection-autoencoder-isolation-forest
- KNIME and pdfs: https://hub.knime.com/victor_palacios/spaces/Public/latest/Simple%20PDF%20Text%20Extraction~kjy6Q-3szxcH6716
- Pandas: https://medium.com/geekculture/the-basics-of-a-jupyter-notebook-pandas-cheat-sheet-for-data-engineers-c1099791a93c
- R Programming Language: https://www.projectpro.io/article/why-is-now-the-time-to-learn-r-programming/178

OECD / OCDE: Restricted Use - À usage restreint

Hyperlinks (2)

- Docfetcher: https://docfetcher.sourceforge.net/en/download.html
- Autopsy: https://www.autopsy.com/download/
- Linux Mint: https://itsfoss.com/install-linux-mint/
- Notebooks and Netflix: https://netflixtechblog.com/notebook-innovation-591ee3221233
- ChatGPT: https://openai.com/blog/chatgpt/
- Everything FAQs: https://voidtools.com/faq/#how_do_i_search_for_a_file_type
- Regular expressions: https://regex101.com/

OECD / OCDE: Restricted Use - À usage restreint

Hyperlinks (3)

- Neo4j: https://neo4j.com/
- Neo4j and Panama Papers: https://neo4j.com/case-studies/the-international-consortium-of-investigative-journalists-icij/
- Neo4j and fraud: https://neo4j.com/use-cases/fraud-detection/

OECD / OCDE: Restricted Use - À usage restreint

استكشاف حلول البرمجيات للمساعدة في تحليل البيانات

بيتر فركرويس، وزارة المالية، بلجيكا

معلوماتي الأساسية

- بيتر فيركرويس
- وزارة المالية، بلجيكا
- pieter.vercruysse@minfin.fed.be
- مديرية الضرائب الخاصة: مختصة بالاحتيال الضريبي
- رئيس وحدة فريق التدقيق الإلكتروني
- المهام الرئيسية لوحدتي: تحليل المحاسبة، وصناديق البريد، والوثائق المكتبية، وملفات PDF، وقواعد البيانات، والكشوفات المصرفية، ...
- المناصب السابقة: ضريبة الدخل على الشركات، قسم التقاضي

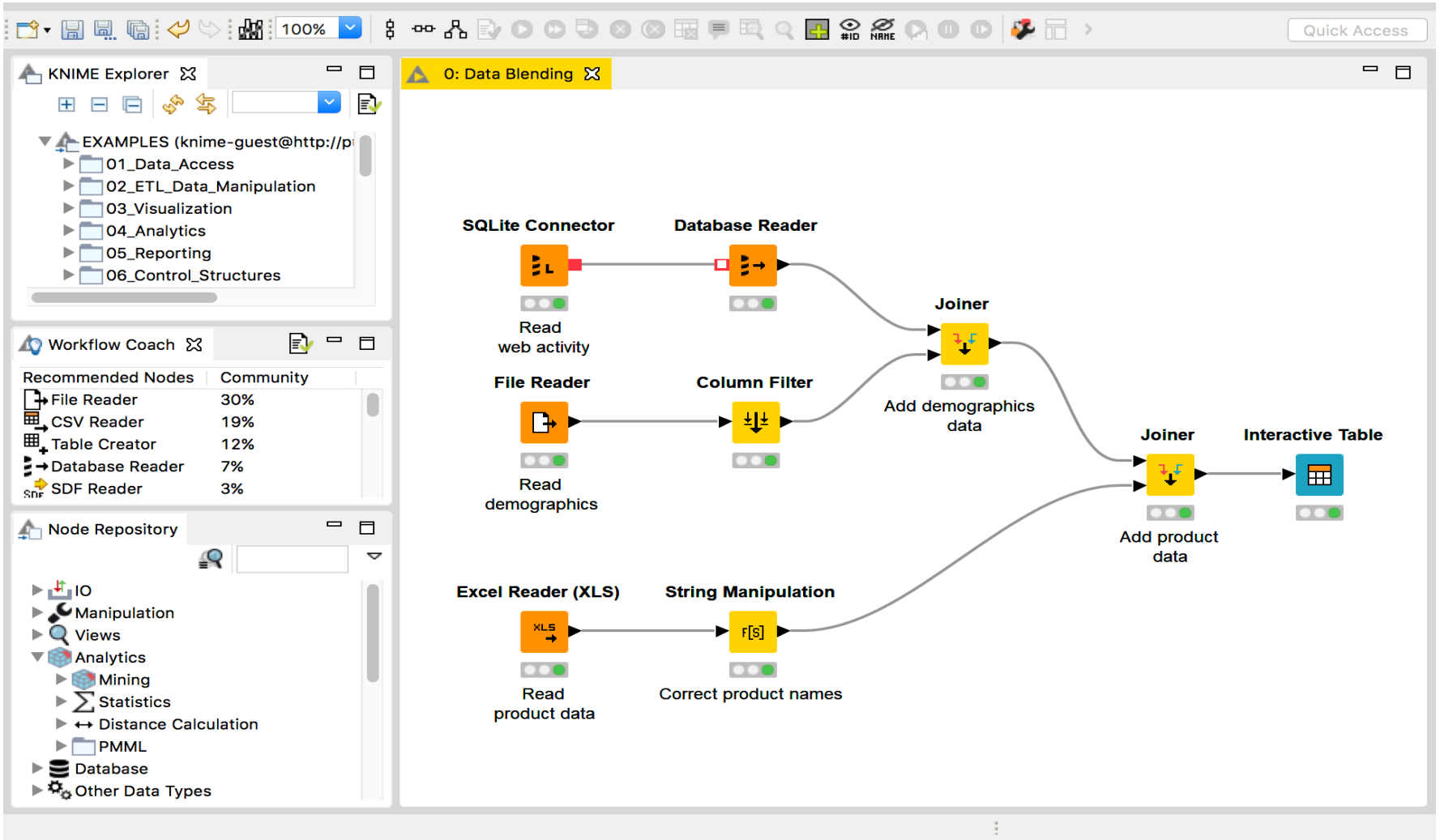
OECD / OCDE: Restricted Use - À usage restreint

الهدف من هذا العرض

- مفهومان مهمان:
  - *سير العمل*
  - *دفاتر الملاحظات*
- أداتان مهمتان مفتوحة المصدر ومجانية:
  - *KNIME*
  - *باندا Pandas، مكتبة برمجية مطوّرة بلغة بايثون Python*

OECD / OCDE: Restricted Use - À usage restreint

مثال على سير العمل في KNIME



OECD / OCDE: Restricted Use - À usage restreint

مثال على دفتر الملاحظات في باندا (بايثون)

jupyter Jupyter_Pandas_Cheat_Sheet Last Checkpoint: a few seconds ago (unsaved changes) Logout

File Edit View Insert Cell Kernel Widgets Help Not Trusted Python 3 (ipykernel)

Run Code

Jupyter Pandas Cheat Sheet

In [0]:
```
import pandas as pd
import numpy as np
```

In [2]:
```
import pandas as pd

df = pd.read_csv('Data/IMDB_Movie_Data.csv' ,sep=",")
```

Exploring Data

In [3]:
```
df.head()        # first five rows
```

Out[3]:

	Title	Genre	Description	Director	Actors	Year	Runtime (Minutes)	Rating	Votes	Revenue (Millions)	Metascore
0	Guardians of the Galaxy	Action,Adventure,Sci-Fi	A group of intergalactic criminals are forced ...	James Gunn	Chris Pratt, Vin Diesel, Bradley Cooper, Zoe S...	2014	121	8.1	757074	333.13	76.0
1	Prometheus	Adventure,Mystery,Sci-Fi	Following clues to the origin of mankind, a te...	Ridley Scott	Noomi Rapace, Logan Marshall-Green, Michael Fa...	2012	124	7.0	485820	126.46	65.0
2	Split	Horror,Thriller	Three girls are kidnapped by a man with a diag...	M. Night Shyamalan	James McAvoy, Anya Taylor-Joy, Haley Lu Richar...	2016	117	7.3	157606	138.12	62.0
3	Sing	Animation,Comedy,Family	In a city of humanoid animals, a hustling thea...	Christophe Lourdelet	Matthew McConaughey,Reese Witherspoon, Seth Ma...	2016	108	7.2	60545	270.32	59.0
4	Suicide Squad	Action,Adventure,Fantasy	A secret government agency recruits some of th...	David Ayer	Will Smith, Jared Leto, Margot Robbie, Viola D...	2016	123	6.2	393727	325.02	40.0

In [0]:
```
df.tail()        # last five rows
```

هيكلية هذا العرض

- ❑ معلومات عامة
- ❑ مواضيع مفيدة
- ❑ العروض التوضيحية
- ❑ الارتباطات التشعبية

OECD / OCDE: Restricted Use - À usage restreint

معلومات عامة

البيانات المهيكلة مقابل البيانات غير المهيكلة

- يتم تنظيم البيانات المهيكلة وتنسيقها بطريقة معيّنة، عادةً في جدول قاعدة بيانات أو جدول بيانات يحتوي على صفوف وأعمدة.
  يسهل البحث عن هذه البيانات وفرزها وتحليلها لأنها تتبع مخططًا.
- البيانات غير المهيكلة غير منظّمة بطريقة محددة مسبقًا. فمن الصعب معالجتها وتحليلها لأنها لا تتبع بنية أو مخططًا ثابتًا.
  تتضمن أمثلة البيانات غير المهيكلة المستندات النصية ورسائل البريد الإلكتروني ومنشورات الوسائط الاجتماعية والصور.
  البيانات غير المهيكلة أكثر تعقيدًا وتتطلب تقنيّات أكثر تعقيدًا لاستخراج معلومات مفيدة منها.

OECD / OCDE: Restricted Use - À usage restreint

برمجيات التحليل التقليدية

- لتحليل **البيانات المهيكلة** (جداول ، قوائم)، على سبيل المثال:
  - *مستعملة من قبل المستخدمين العاديين: Excel, Access ;-)*
  - *مستعملة من قبل المدققين الإلكترونيين: ACL, Arbutus, IDEA*
- لتحليل البيانات غير المهيكلة (أنواع مختلفة من الملفّات)، على سبيل المثال: ZyLAB ONE، حل eDiscovery يساعد في الكشف عن المعلومات المهمة الموجودة في المستندات والملفّات الخاصة بمؤسستك، ومراجعتها.
- للتحليل **الجنائي**، على سبيل المثال: Encase أو FTK (مجموعة أدوات التحقيق الجنائي): التحليل الجنائي للقرص الصلب، ووسائط تخزين المعلومات الأخرى

OECD / OCDE: Restricted Use - À usage restreint

بدائل مفتوحة المصدر ومجانية

- لتحليل **البيانات المهيكلة** (جداول ، قوائم) Knime, Python, R
  - KNIME Analytics Platform هو برنامج مجاني، مفتوح المصدر بواجهة مرئية سهلة الاستخدام، تتيح لك القيام بتحليلات من أي مستوى من التعقيد
  - باندا، مكتبة بلغة بايثون
  - R، لغة وبيئة للحوسبة الإحصائية والرسومات
- لتحليل البيانات **غير المهيكلة** (أنواع مختلفة من الملفات) DocFetcher (هناك نسخة محمولة!)
- للتحليل **الجنائي** للصور : Autopsy
- لـ **ويندوز**: Linux -> بشكل خاص Linux Mint (يرجى مراجعة الارتباطات التشعبية للحصول على دليل التثبيت)

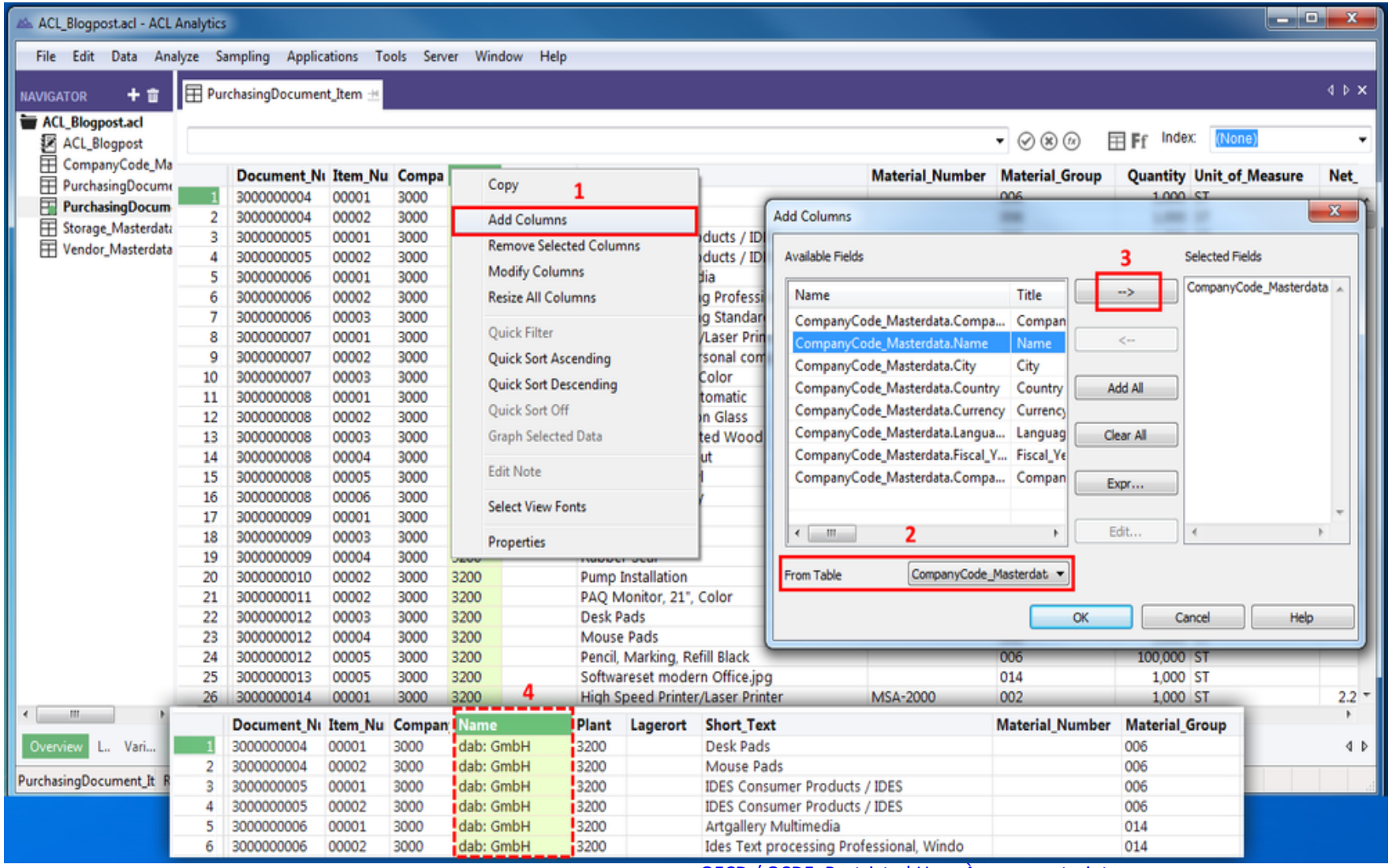
OECD / OCDE: Restricted Use - À usage restreint

ACL و Arbutus و IDEA، برمجيات تجارية لتحليل بيانات البيانات المهيكلة

- ACL و Arbutus و IDEA: حلول تدقيق بمساعدة الحاسوب، مصممة لاستخراج البيانات وتحليلها ولأخذ عيّنات عن السكان
- هذه أدوات نموذجية يستخدمها المدققون الإلكترونيون لتحديد صحة أرصدة الحسابات والعمليات
- المزايا: معيار قياسي، سريع جدًا، استخدام ممكن للمعالجات
- البرمجيات التجارية

OECD / OCDE: Restricted Use - À usage restreint

صورة لشاشةACL، برنامج تجاري للتدقيق إلكتروني



OECD / OCDE: Restricted Use - À usage restreint

KNIME، برنامج مجاني، مفتوح المصدر وبديل لـ ACL و Arbutus و 1IDEA [ ]

فوائد استخدام سير العمل في KNIME

- ❑ تحصل على نظرة عامة جيّدة عن جميع المهام التي يتم تنفيذها على بيانات المصدر.
- ❑ إذا كنت قد ارتكبت خطأ في مكان ما، فما عليك سوى ضبط العقدة أو الرمز وتشغيل سير العمل من جديد
- ❑ يمكنك الحصول على نظرة ثاقبة عن العمليات الحسابية من خلال عرض شاشة التكوين وهي سهلة الاستخدام.
- ❑ سهولة إعادة الاستخدام: سير العمل قابل للتطبيق على البيانات الأخرى بنفس الهيكل!
- ❑ سهولة الأتمتة: يمكنك وضع سير عمل في حلقة وتطبيقه على آلاف الملفات إذا لزم الأمر!

OECD / OCDE: Restricted Use - À usage restreint

KNIME، برنامج مجاني، مفتوح المصدر وبديل لـ ACL و Arbutus و 2IDEA [ ]

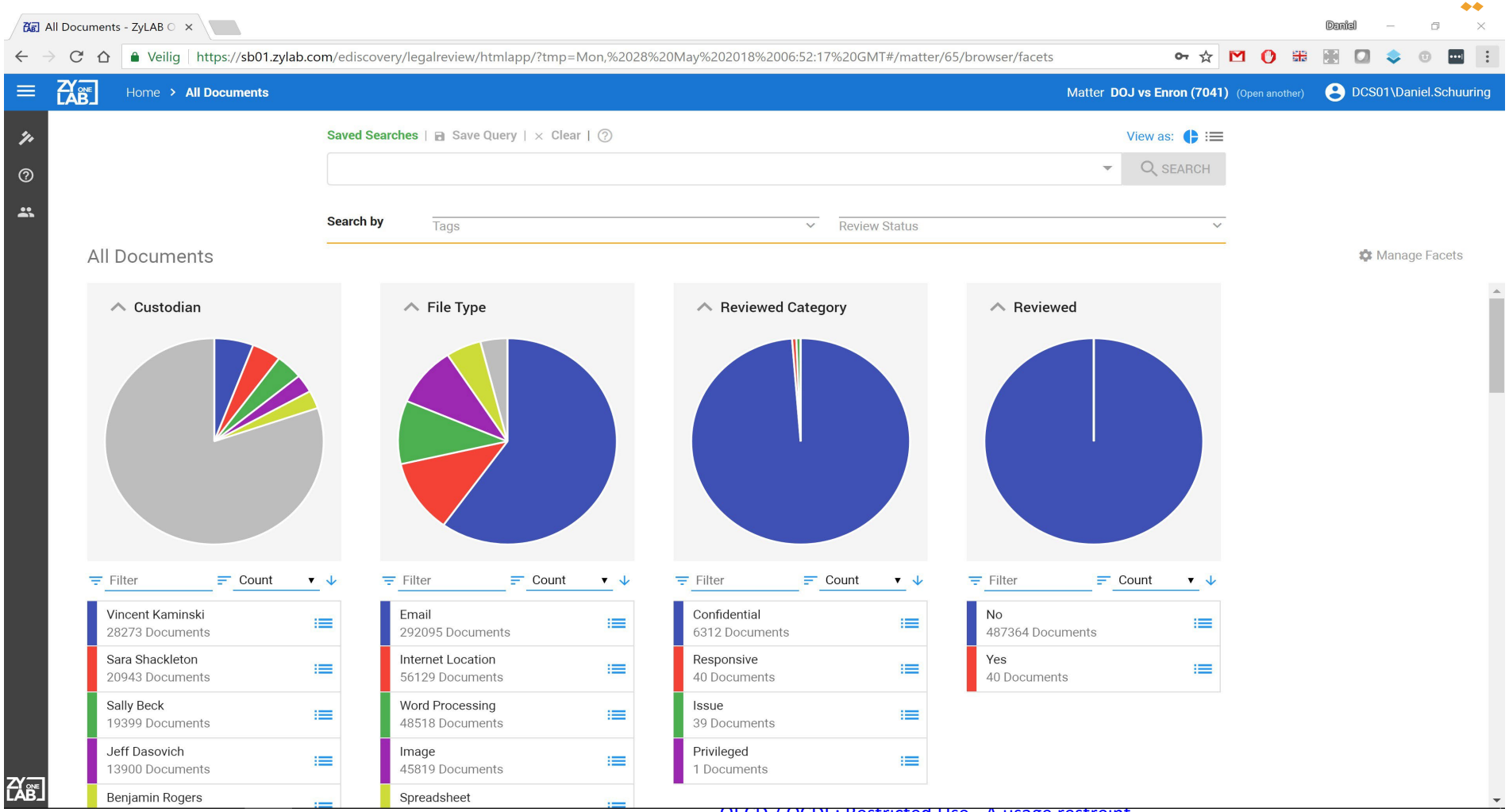
- يمكنك توثيق سير العمل
- أمثلة كافية لسير العمل على KNIME Hub
- يمكنك بسهولة تبادل مهام سير العمل مع أشخاص آخرين: تصدير واستيراد مهام سير العمل
- تصبح مستقل عن قسم تكنولوجيا المعلومات والاتصالات (ليس عليك أن تكون مبرمجًا!) ويمكنك أن تكون مبدعًا: تمكين موظفي الضرائب ؛-)
- ندوات مجانية على الإنترنت على سبيل المثال: https://www.knime.com/events/approaches-to-fraud-detection-autoencoder-isolation-forest

السلبيات

- من الصعب طباعة سير العمل في Knime
- من الصعب إنشاء تقارير في. Python برمامج Knime أفضل لذلك

OECD / OCDE: Restricted Use - À usage restreint

صورة لشاشة Zylab، برنامج تجاري للاكتشاف الإلكتروني



OECD / OCDE: Restricted Use - A usage restreint

صورة لشاشة Zylab، برنامج تجاري للاكتشاف الإلكتروني

صورة لشاشة DocFetcher، وهو تطبيق مجاني ومفتوح المصدر للاكتشاف الإلكتروني

DocFetcher

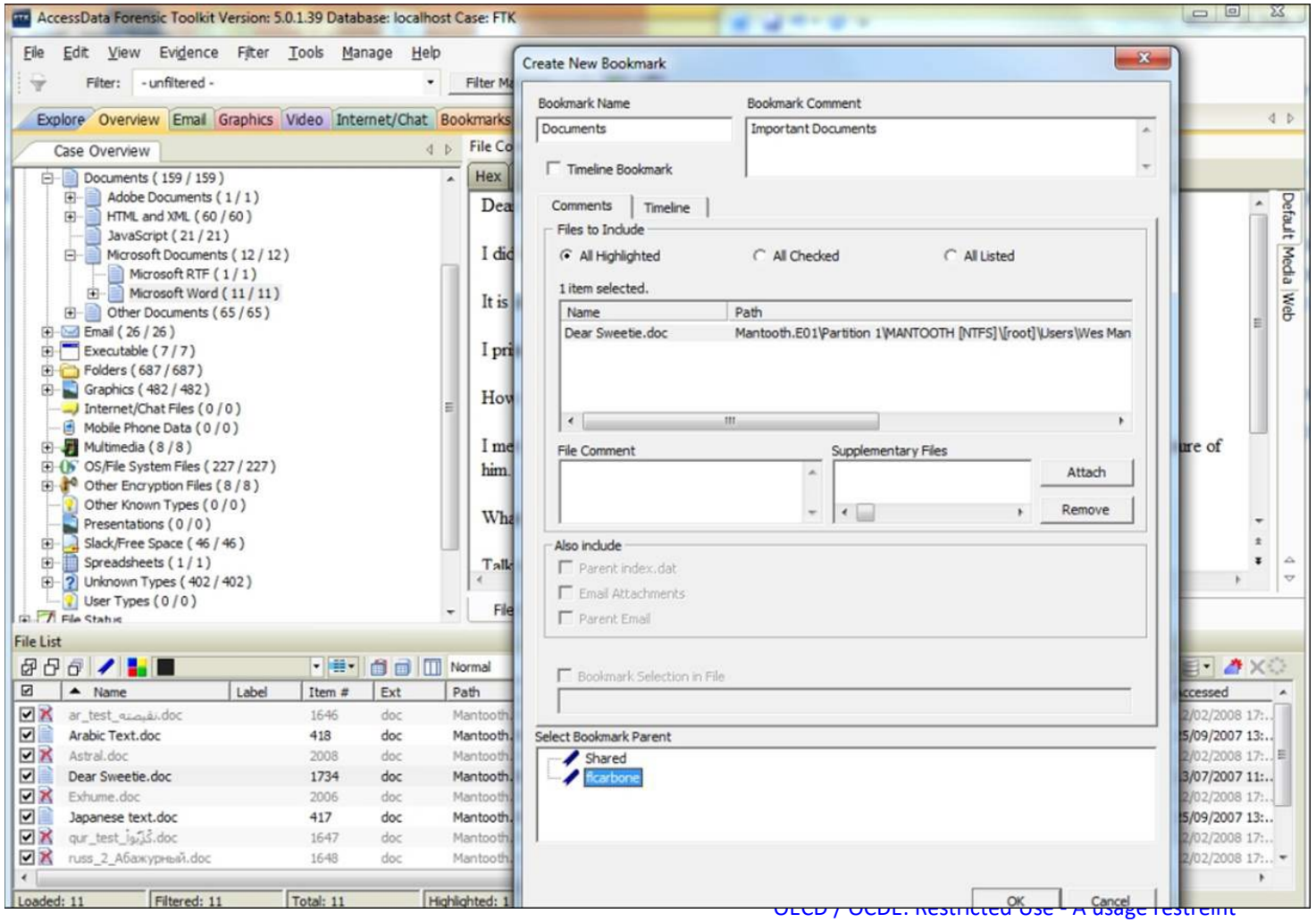
Overview | Download | Screenshots | More

Note: You may be interested in DocFetcher Pro, the commercial big brother of DocFetcher with more features and fewer bugs.

Description

DocFetcher is an Open Source desktop search application: It allows you to search the contents of files on your computer. — You can think of it as Google for your local files. The application runs on Windows, Linux and OS X, and is made available under the Eclipse Public License.

صورة لشاشة FTK (مجموعة أدوات جنائية)، منتج برمجي تجاري



OECD / OCDE. Restricted Use - A usage restreint

صورة لشاشة Autopsy، وهو تطبيق مجاني، مفتوح المصدر للتحليل الجنائي

1. Getting Started

Open Autopsy and create a new case.

OECD / OCDE: Restricted Use - À usage restreint

طريقة أفضل للعمل فيما يتعلق بتحليل البيانات

1) سير العمل

هناك العديد من المنافع لاستخدام سير العمل في تحليل البيانات:

- تحسين الكفاءة: يمكن أن تساعد مهام سير العمل في أتمتة عمليات تحليل البيانات وتبسيطها، **لذا فهي مثالية للمهام المتكررة**
- زيادة الدقة: يمكن أن تساعد مهام سير العمل في ضمان تنفيذ مهام تحليل البيانات بشكل متسق وصحيح
- تحسين إعادة الانتاج: يمكن توثيق مهام سير العمل ومشاركتها مع الآخرين

OECD / OCDE: Restricted Use - À usage restreint

2) دفاتر الملاحظات

تعتبر أجهزة الكمبيوتر المحمولة أداة شائعة لتحليل البيانات، وهي توفّر العديد من المزايا:

- **الحوسبة التفاعلية**: تسمح دفاتر الملاحظات لمحللي البيانات بكتابة التعليمات البرمجية وتنفيذها، بالإضافة إلى تضمين النصوص والصور ومحتويات الوسائط المتعددة الأخرى في مستند واحد
- إعادة الانتاج: تسمح دفاتر الملاحظات لمحللي البيانات بتوثيق عملية التحليل الخاصة بهم تدريجيًا، مما يسهّل إعادة إنتاج نتائجهم ومشاركتها مع الآخرين (مسار التدقيق)
- **يمكنك صياغة تقارير جميلة يمكن تصديرها إلى html أو pdf!** وصياغة تقارير أيضا من دون خلايا التعليمات البرمجية!
- التعاون: يمكن مشاركة دفاتر الملاحظات مع الآخرين والتعاون في العمل عليها في الوقت الحقيقي
- سهولة الاستعمال: تتميز دفاتر الملاحظات بواجهة سهلة الاستخدام وسهلة التعلّم

OECD / OCDE: Restricted Use - À usage restreint

مواضيع مفيدة

- ChatGPT ❑
- Everything ❑
- ”Regular expressions = Advanced Wildcards ❑

OECD / OCDE: Restricted Use - À usage restreint

ChatGPT

ChatGPT هو روبوت محادثة تم تطويره بواسطة OpenAI. إنه نموذج للتعلّم الآلي يتمّ تدريبه لإنشاء نص يشبه النص الذي يتم أنشائه بواسطة إنسان يمكن استخدامه لطرح الأسئلة وتلقي الإجابات في الوقت الحقيقي، مما يجعله طريقة ملائمة وفعّالة للعثور على المعلومات.

تمّ إطلاقه في 30 تشرين الثاني/نوفمبر 2022.

مزايا ChatGPT

- ❑ يمكن أن يقدّم معلومات في جمل واضحة وبسيطة، بدلاً من قائمة بارتباطات الإنترنت
- ❑ يمكنه شرح المفاهيم بطريقة يمكن للناس فهمها بسهولة.
- ❑ يمكنك بسهولة نسخ وإعادة استخدام النتائج

سلبيات ChatGPT

- ❑ يمكن لـChatGPT فقط توفير المعلومات والمساعدة على أساس البيانات التي تمّ تدريبه عليها
- ❑ لا يُستخدم للمعلومات الحساسة

OECD / OCDE: Restricted Use - À usage restreint

Everything

Everything عبارة عن أداة بحث تتيح للمستخدمين تحديد موقع الملفات والمجلدات بسرعة على أجهزة الحاسوب الخاصة بهم. إنها مجانية ولها نسخة محمولة.

لها مزايا عديدة مقارنة بأدوات البحث الأخرى

- السرعة: تم تصميم هذه الأداة لتكون سريعة، ويمكنها البحث في ملايين الملفات في غضون ثوانٍ
- المرونة: تتيح للمستخدمين البحث عن الملفات والمجلدات باستخدام مجموعة واسعة من المرشحات والخيارات، بما في ذلك الحجم وتاريخ التعديل وسمات الملف
- يمكنها فهرسة محركات أقراص الشبكة المعينة / شبكة التخزين المرفقة / مشاركة عبر الشبكة ومحركات الأقراص الخارجية

سلبيات Everything

- محتوى الملف غير مفهرس

OECD / OCDE: Restricted Use - À usage restreint

التعبيرات العادية (1)

التعبير العادي هو سلسلة من الأحرف التي تحدد نمط البحث. تُستخدم التعبيرات العادية بشكل أساسي لمطابقة الأنماط مع السلاسل.

- على سبيل المثال ابحث عن جميع الملفات التي تحتوي على OECD أو OESO أو OCDE في اسم الملف:
  OECD|OESO|OCDE
- على سبيل المثال تحقق مما إذا كانت هناك أرقام حساب مصرفي دولي بلجيكي في أسماء الملفات: txt.{14BE\d{

مثال على رقم حساب مصرفي دولي بلجيكي: 68539007547034BE

تستخدم التعبيرات العادية في العديد من لغات وأدوات البرمجة. إنها أدوات قوية لاستخراج المعلومات من النص، وغالبًا ما تستخدم في تحليل البيانات.

OECD / OCDE: Restricted Use - À usage restreint

التعبيرات العادية (2)

مزايا التعبيرات العادية

- الدقة: تسمح التعبيرات العادية للمستخدمين بتحديد ما يبحثون عنه بالضبط في النص: على سبيل المثال، النص بين كلمتين
- الأتمتة: يمكن استخدام التعبيرات العادية لأتمتة مهام معالجة النص، مثل البحث عن أنماط معيّنة من النص واستبدالها

سلبيات التعبيرات العادية

- قد يكون من الصعب قراءة التعبيرات العادية وفهمها واستخدامها
- من الصعب تذكّر معنى الرموز المستخدمة في الأنماط

OECD / OCDE: Restricted Use - À usage restreint

مثلًا

مدخل إلى KNIME + سير العمل: عروض توضيحية

1) تحليل بسيط لجدول بيانات نموذج Boats’Bob s

2) تحليل البيانات المالية بنسب محددة من قبل المستخدم

- ❑ استخدام ملفات xbrl (لغة تقارير الأعمال الموسّعة، المعيار االدولي المفتوح لتقارير الأعمال الرقمية)
- ❑ استخدام لغة استعلام XPath

OECD / OCDE: Restricted Use - À usage restreint

مدخل إلى Python + دفاتر الملاحظات: عروض توضيحية

1) تحليل بسيط لجدول بيانات نموذج Bob ’Boats
2) الكشف عن الفئة الزمنية والعيوب (التعلّم الآلي)

ملاحظة: الجمع بين KNIME و Python ممكن أيضًا وفيدًا جدًا!

OECD / OCDE: Restricted Use - À usage restreint

مدخل إلى J4Neo + قواعد بيانات الروسم البيانية: عروض توضيحية

- قاعدة بيانات الرسوم البيانية => التركيز على العلاقات بين البيانات
- التطبيق: التمثيل الديناميكي لتدفقات الأموال
- الجمع بين KNIME و J4Neo ممكن: لا حاجة للبرمجة!

OECD / OCDE: Restricted Use - À usage restreint

الارتباطات التشعبية (1)

- KNIME: https://www.knime.com/downloads
- /KNIME Hub: https://hub.knime.com
- KNIME ندوة عبر الإنترنت حول الاحتيال: https://www.knime.com/events/approaches-to-fraud-detection-autoencoder-isolation-forest
- KNIME وملفات pdfs: T%20PDF%20https://hub.knime.com/victor_palacios/spaces/Public/latest/Simple6716szxcH3Q-6Extraction~kjy%20ext
- Pandas: https://medium.com/geekculture/the-basics-of-a-jupyter-notebook-c93a1099791pandas-cheat-sheet-for-data-engineers-c
- R: https://www.projectpro.io/article/why-is-now-the-time-to-learn-r-لغة البرمجة 178programming/

OECD / OCDE: Restricted Use - À usage restreint

الارتباطات التشعبية (2)

- Docfetcher: https://docfetcher.sourceforge.net/en/download.html
- Autopsy: https://www.autopsy.com/download/
- Linux Mint: https://itsfoss.com/install-linux-mint/
- دفاتر الملاحظات و Netflix: https://netflixtechblog.com/notebook-innovation-3221233ee591
- ChatGPT: https://openai.com/blog/chatgpt/
- Everything: أسئلة وأجوبة عن https://voidtools.com/faq/#how_do_i_search_for_a_file_type
- التعبيرات العادية: https://regex101.com/

OECD / OCDE: Restricted Use - À usage restreint

الارتباطات التشعبية (3)

- Neo4j : https://neo4j.com/
- Neo4j و ICIJ Panama Papers: https://neo4j.com/case-studies/the-international-consortium-of-investigative-journalists-icij/
- Neo4j والاحتيال: https://neo4j.com/use-cases/fraud-detection/

OECD / OCDE: Restricted Use - À usage restreint