Document as a Data

Tidyverse Korea

이광춘

Table of Contents

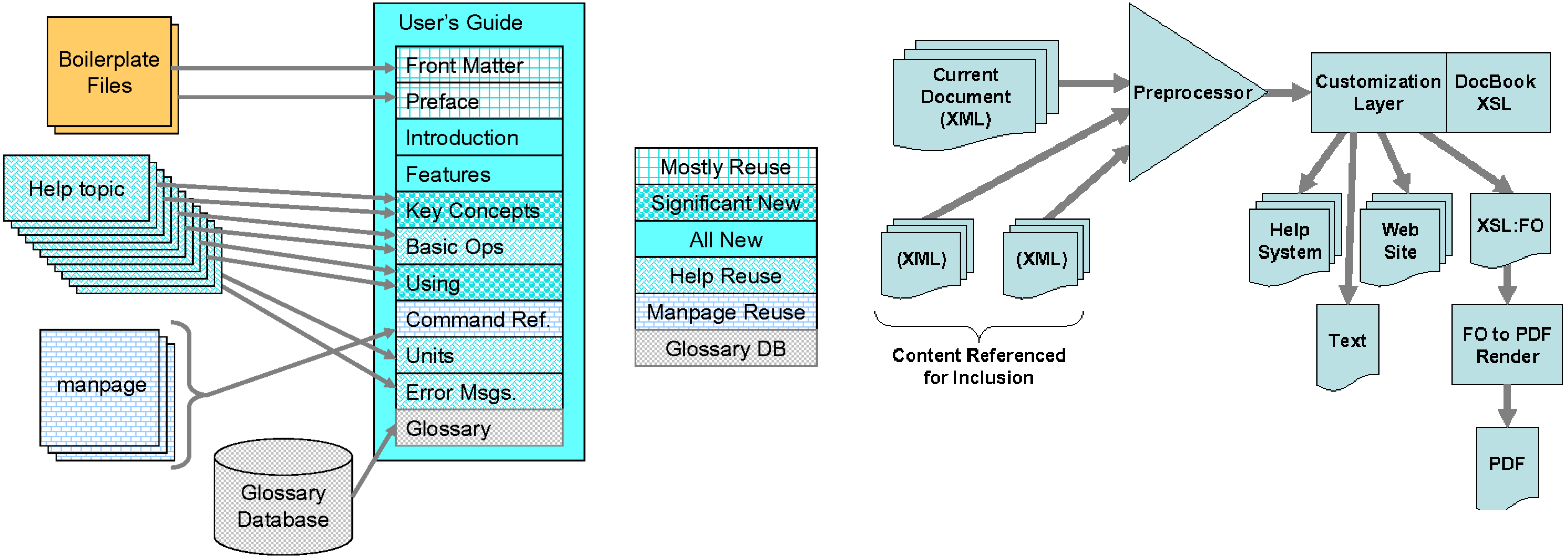
I. Document Anatomy

II. Compendium

III. OCR

Document Anatomy – Help Manual

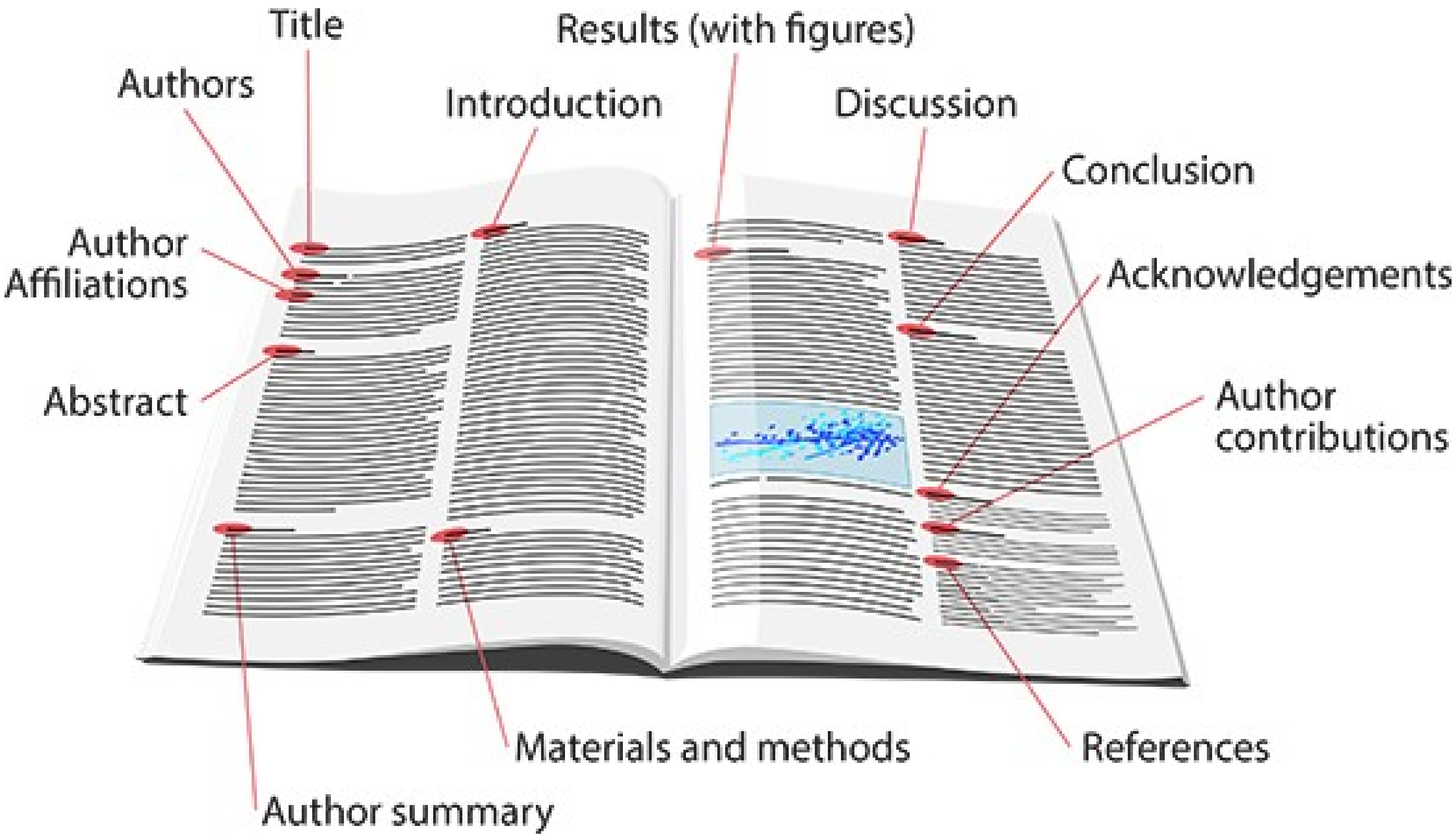
Anatomy of Hewlett-Packard User Manual Documents



[Kathy Haramundanis and Larry Rowland (Hewlett-Packard Company), "Experience paper: a content reuse documentation design experience", SIGDOC '07 Proceedings of the 25th annual ACM international conference on Design of communication Pages 229-233, El Paso, Texas, USA — October 22 - 24, 2007]()

Document Anatomy

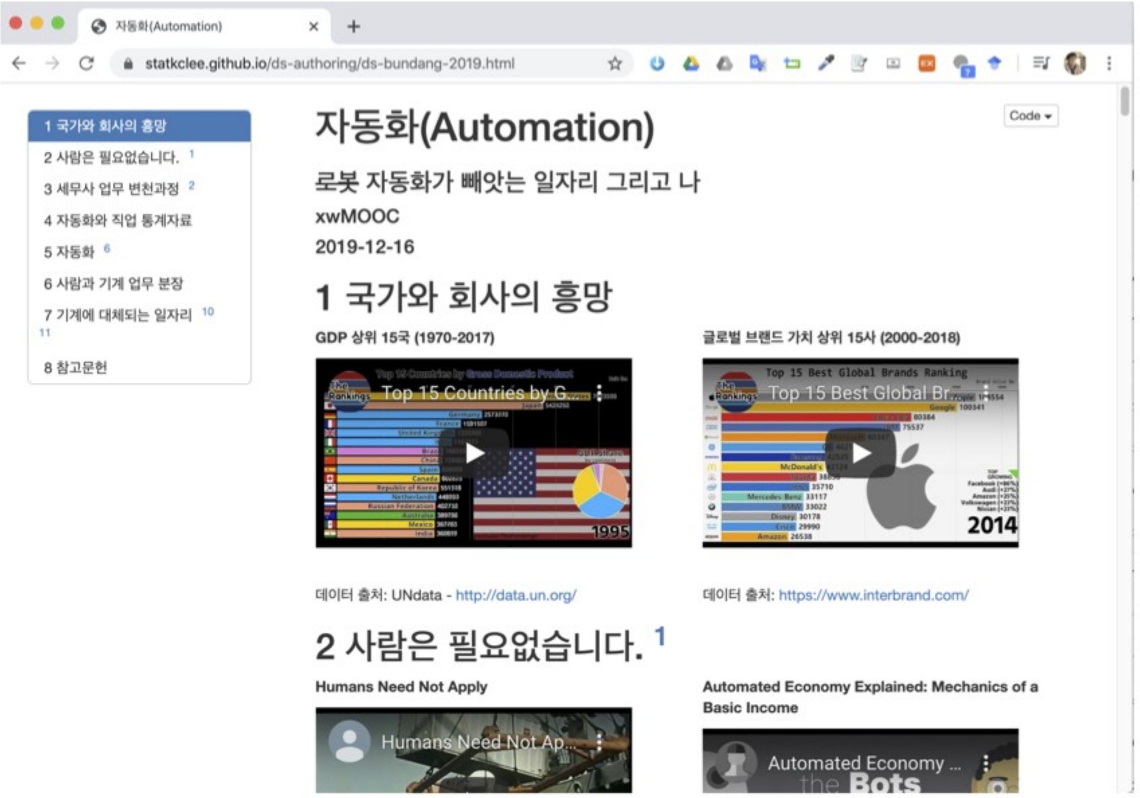
Anatomy of a research paper



출처: https://csevilla.uneportfolio.org/2017/03/26/what-is-a-scientific-paper/

Document Anatomy - Papers

Creating a research paper through Coding



자동화(Automation)

로봇 자동화가 빼앗는 일자리 그리고 나

xwMOOC

2019-12-16

1 국가와 회사의 흥망

GDP 상위 15국 (1970-2017)

데이터 출처: UNdata - http://data.un.org/

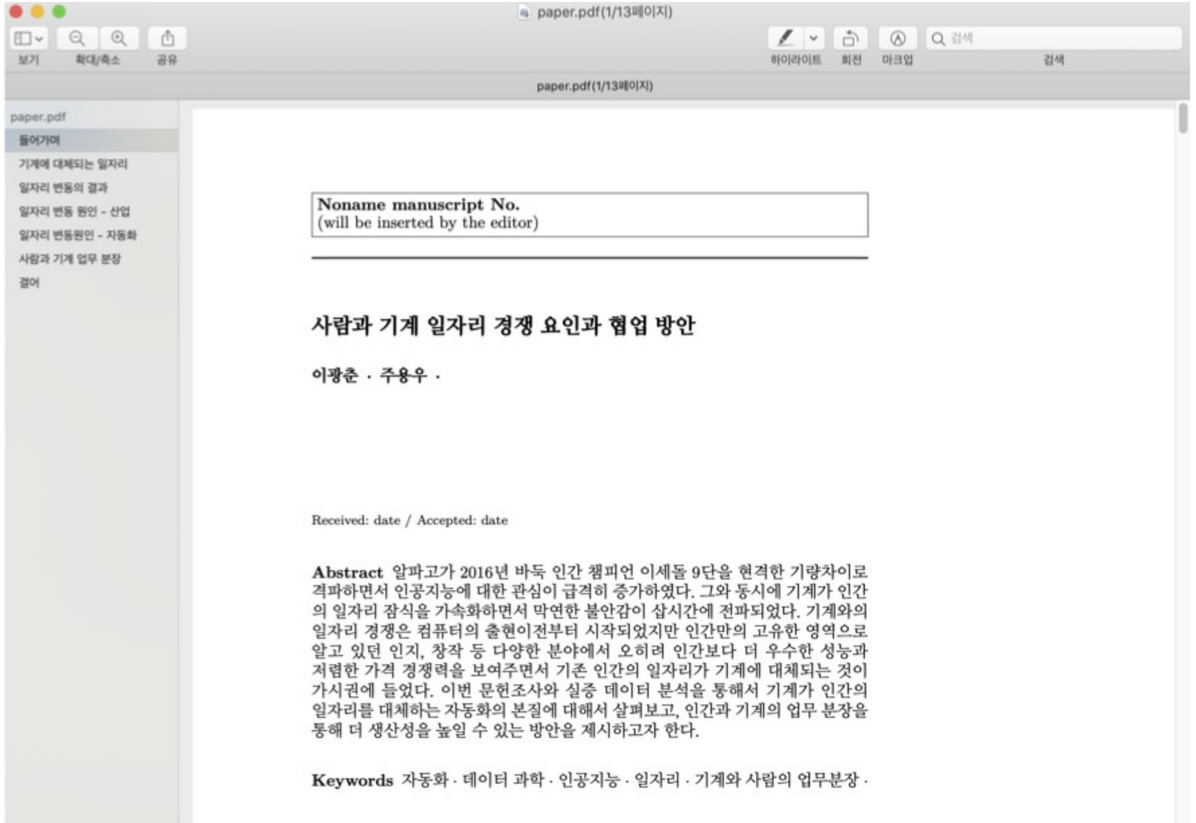
글로벌 브랜드 가치 상위 15사 (2000-2018)

데이터 출처: https://www.interbrand.com/

2 사람은 필요없습니다. 1

Humans Need Not Apply

Automated Economy Explained: Mechanics of a Basic Income



Noname manuscript No.
(will be inserted by the editor)

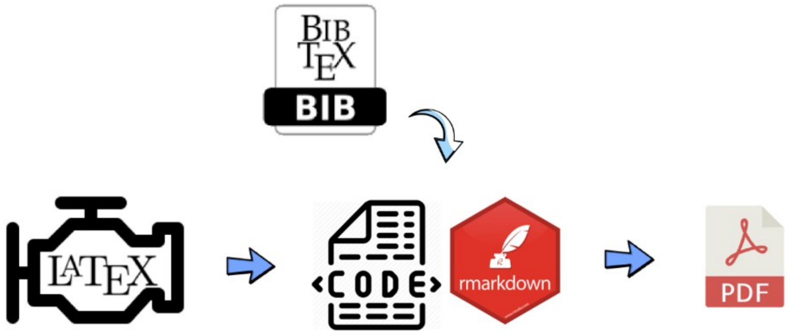
사람과 기계 일자리 경쟁 요인과 협업 방안

이광춘 · 주용우 ·

Received: date / Accepted: date

Abstract 알파고가 2016년 바둑 인간 챔피언 이세돌 9단을 현격한 기량차이로 격파하면서 인공지능에 대한 관심이 급격히 증가하였다. 그와 동시에 기계가 인간의 일자리 잠식을 가속화하면서 막연한 불안감이 삽시간에 전파되었다. 기계와의 일자리 경쟁은 컴퓨터의 출현이전부터 시작되었지만 인간만의 고유한 영역으로 알고 있던 인지, 창작 등 다양한 분야에서 오히려 인간보다 더 우수한 성능과 저렴한 가격 경쟁력을 보여주면서 기존 인간의 일자리가 기계에 대체되는 것이 가시권에 들었다. 이번 문헌조사와 실증 데이터 분석을 통해서 기계가 인간의 일자리를 대체하는 자동화의 본질에 대해서 살펴보고, 인간과 기계의 업무 분장을 통해 더 생산성을 높일 수 있는 방안을 제시하고자 한다.

Keywords 자동화 · 데이터 과학 · 인공지능 · 일자리 · 기계와 사람의 업무분장 ·



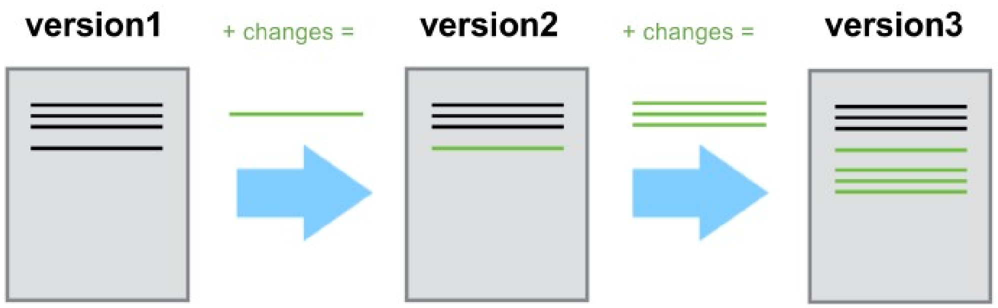
출처: https://statkclee.github.io/comp_document/automation-kasdba.html

Compendium

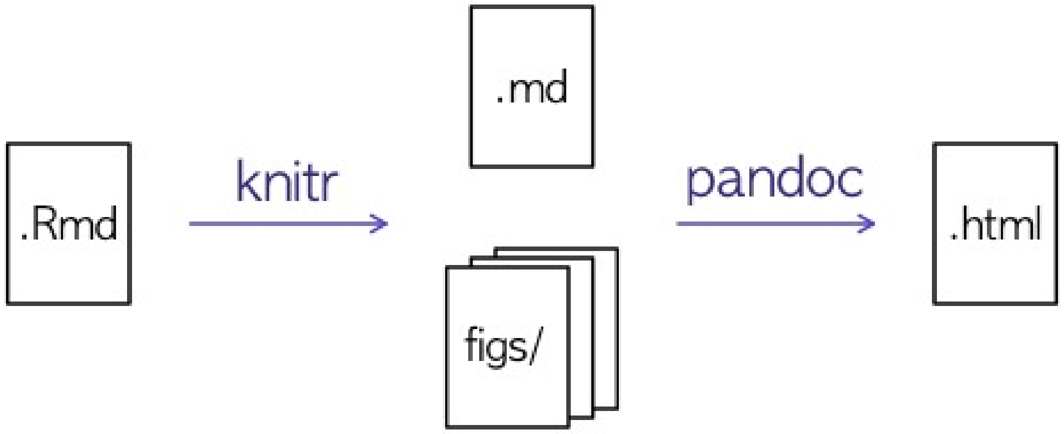
Document as a Code

```
.
├── .Rprofile
├── .gitignore
├── README.md
├── analysis
│   └── archive
│   └── markdown
├── data
│   ├── documentation
│   ├── handmade
│   ├── html_reports
│   ├── processed
│   ├── public
│   └── source
├── etl
├── publish
├── scratch
├── viz
└── {{cookiecutter}}.Rproj
```

Version Control

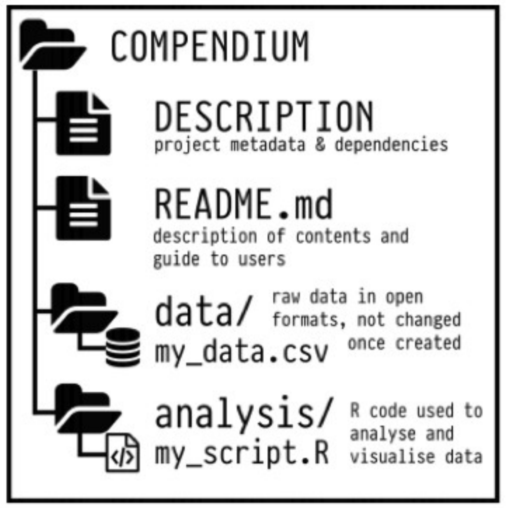


Literate Programming



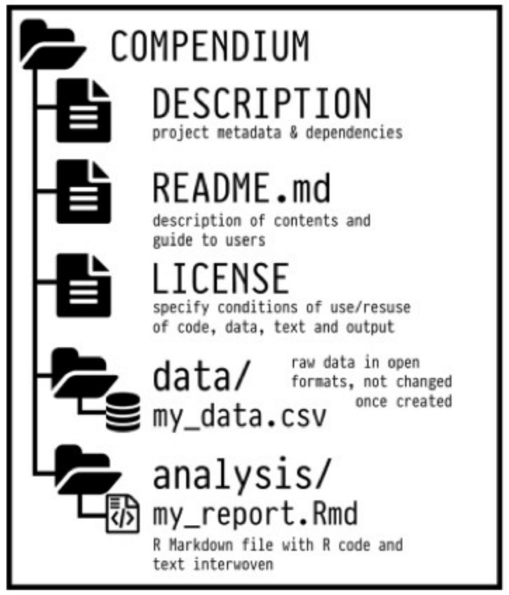
Compendium – Reproducible Research

데이터 분석



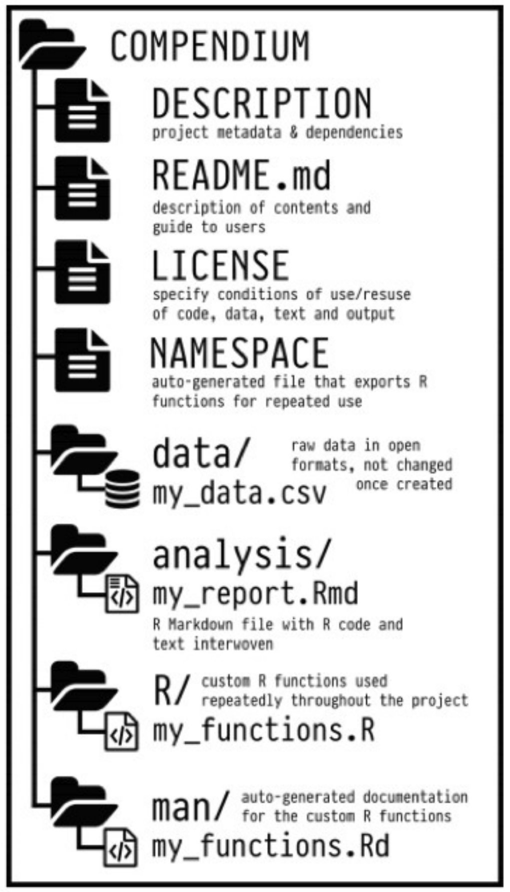
COMPENDIUM
DESCRIPTION project metadata & dependencies
README.md description of contents and guide to users
data/ my_data.csv raw data in open formats, not changed once created
analysis/ my_script.R R code used to analyse and visualise data

분석 보고서



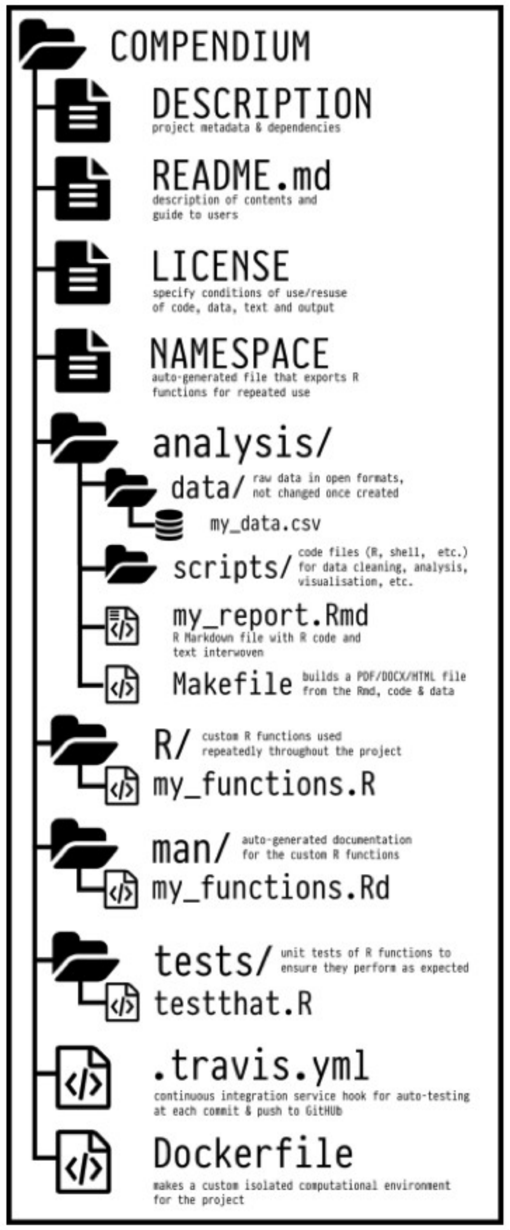
COMPENDIUM
DESCRIPTION project metadata & dependencies
README.md description of contents and guide to users
LICENSE specify conditions of use/resuse of code, data, text and output
data/ my_data.csv raw data in open formats, not changed once created
analysis/ my_report.Rmd R Markdown file with R code and text interwoven

분석 팩키지



COMPENDIUM
DESCRIPTION project metadata & dependencies
README.md description of contents and guide to users
LICENSE specify conditions of use/resuse of code, data, text and output
NAMESPACE auto-generated file that exports R functions for repeated use
data/ my_data.csv raw data in open formats, not changed once created
analysis/ my_report.Rmd R Markdown file with R code and text interwoven
R/ my_functions.R custom R functions used repeatedly throughout the project
man/ my_functions.Rd auto-generated documentation for the custom R functions

재현가능한 팩키지



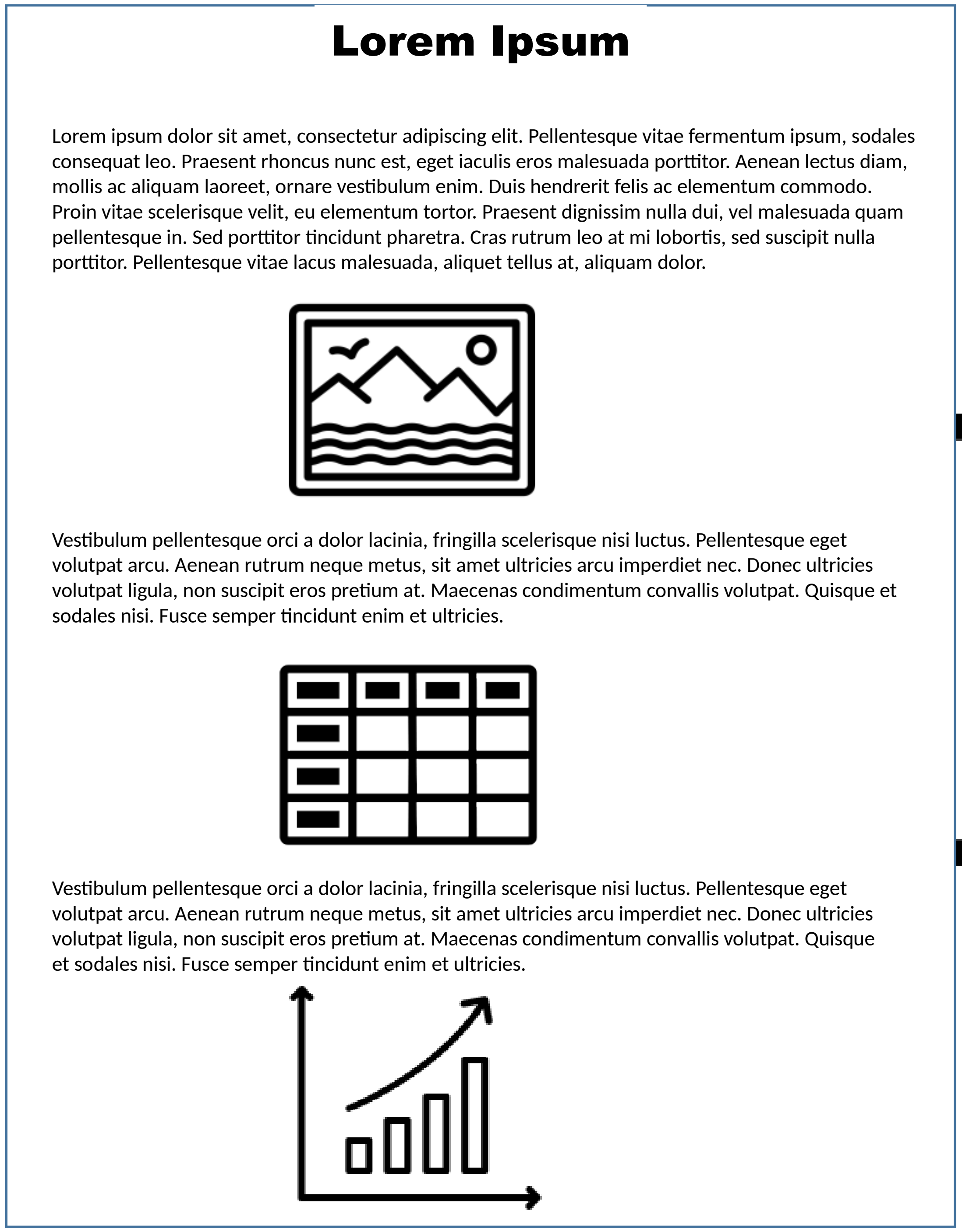
COMPENDIUM
DESCRIPTION project metadata & dependencies
README.md description of contents and guide to users
LICENSE specify conditions of use/resuse of code, data, text and output
NAMESPACE auto-generated file that exports R functions for repeated use
analysis/
data/ raw data in open formats, not changed once created
my_data.csv
scripts/ code files (R, shell, etc.) for data cleaning, analysis, visualisation, etc.
my_report.Rmd R Markdown file with R code and text interwoven
Makefile builds a PDF/DOCX/HTML file from the Rmd, code & data
R/ custom R functions used repeatedly throughout the project
my_functions.R
man/ auto-generated documentation for the custom R functions
my_functions.Rd
tests/ unit tests of R functions to ensure they perform as expected
testthat.R
.travis.yml continuous integration service hook for auto-testing at each commit & push to GitHUb
Dockerfile makes a custom isolated computational environment for the project

출처: https://statkclee.github.io/comp_document/cd_compendium.html

OCR

OCR (Optical Character Recognition)

이미지 형태의 문서를 텍스트 형태의 문서로 변환해주는 기술



Lorem Ipsum

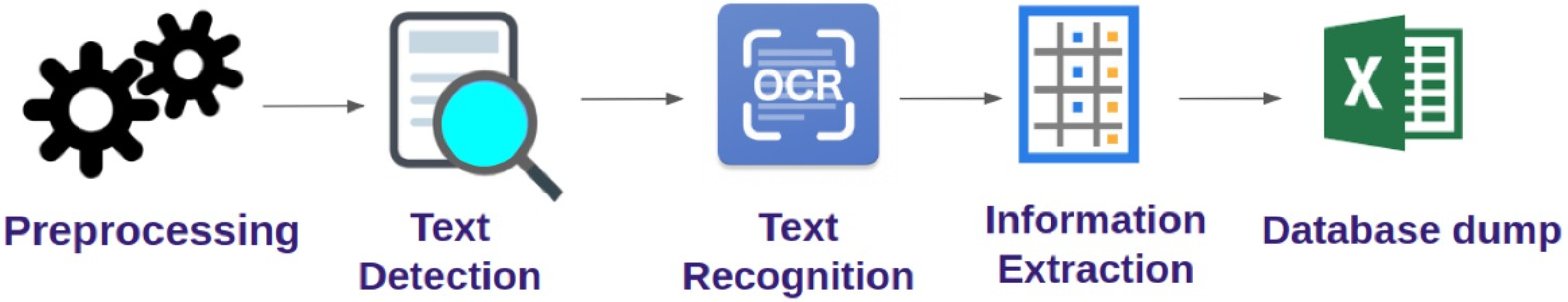
Lorem ipsum dolor sit amet, consectetur adipiscing elit. Pellentesque vitae fermentum ipsum, sodales consequat leo. Praesent rhoncus nunc est, eget iaculis eros malesuada porttitor. Aenean lectus diam, mollis ac aliquam laoreet, ornare vestibulum enim. Duis hendrerit felis ac elementum commodo. Proin vitae scelerisque velit, eu elementum tortor. Praesent dignissim nulla dui, vel malesuada quam pellentesque in. Sed porttitor tincidunt pharetra. Cras rutrum leo at mi lobortis, sed suscipit nulla porttitor. Pellentesque vitae lacus malesuada, aliquet tellus at, aliquam dolor.

Vestibulum pellentesque orci a dolor lacinia, fringilla scelerisque nisi luctus. Pellentesque eget volutpat arcu. Aenean rutrum neque metus, sit amet ultricies arcu imperdiet nec. Donec ultricies volutpat ligula, non suscipit eros pretium at. Maecenas condimentum convallis volutpat. Quisque et sodales nisi. Fusce semper tincidunt enim et ultricies.

Vestibulum pellentesque orci a dolor lacinia, fringilla scelerisque nisi luctus. Pellentesque eget volutpat arcu. Aenean rutrum neque metus, sit amet ultricies arcu imperdiet nec. Donec ultricies volutpat ligula, non suscipit eros pretium at. Maecenas condimentum convallis volutpat. Quisque et sodales nisi. Fusce semper tincidunt enim et ultricies.

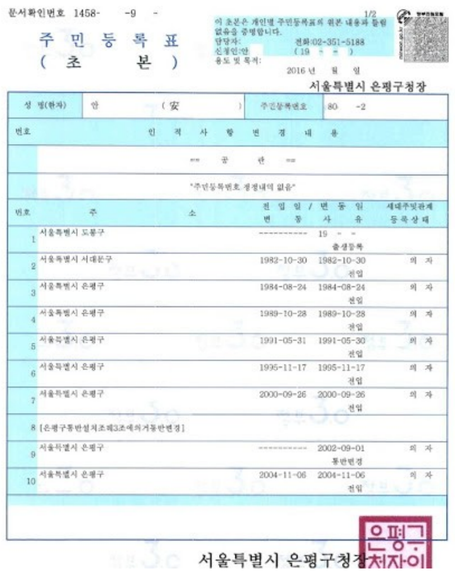
- 텍스트
- 이미지
- 표
- 그래프
- 문서구조
- 참고문헌
- …

OCR 프로세스



Preprocessing → Text Detection → Text Recognition → Information Extraction → Database dump

<원본 이미지>

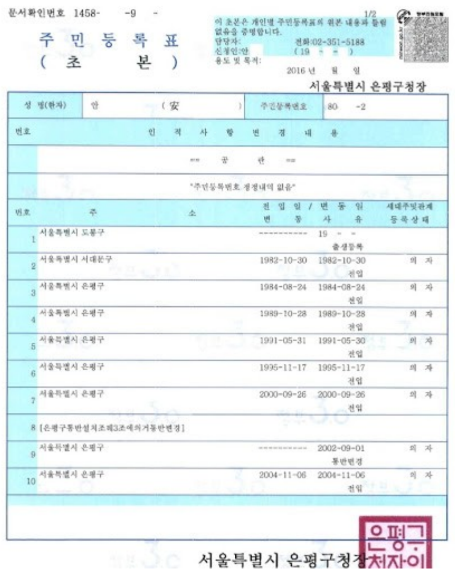


문서확인번호 1458- -9 -

주 민 등 록 표
(초 본)

서울특별시 은평구청장

<전처리 이미지>



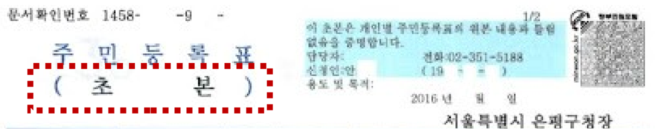
문서확인번호 1458- -9 -

주 민 등 록 표
(초 본)

서울특별시 은평구청장

<잘라낸 영역>

문서 종류



문서확인번호 1458- -9 -

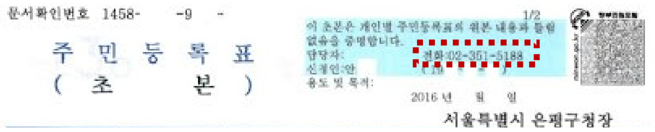
주 민 등 록 표
(초 본)

이 초본은 개인별 주민등록표의 원본 내용과 틀림없음을 증명합니다.
담당자: 전화:02-351-5188
신청인:안 (19 - -)
용도 및 목적:

2016 년 월 일

서울특별시 은평구청장

전화번호



문서확인번호 1458- -9 -

주 민 등 록 표
(초 본)

이 초본은 개인별 주민등록표의 원본 내용과 틀림없음을 증명합니다.
담당자: 전화:02-351-5188
신청인:안 (19 - -)
용도 및 목적:

2016 년 월 일

서울특별시 은평구청장

<글자 외곽선을 통해 추출한 이미지>

초 본

:02-351-5188

감사합니다.