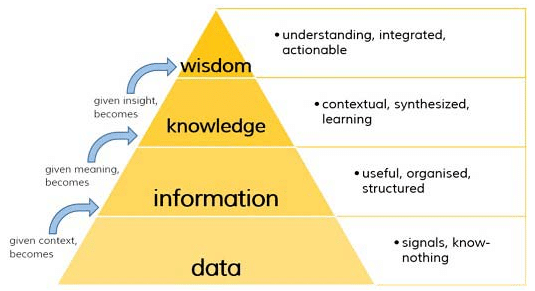
2장 데이터 수집

1. **화공 산업 데이터 소개**

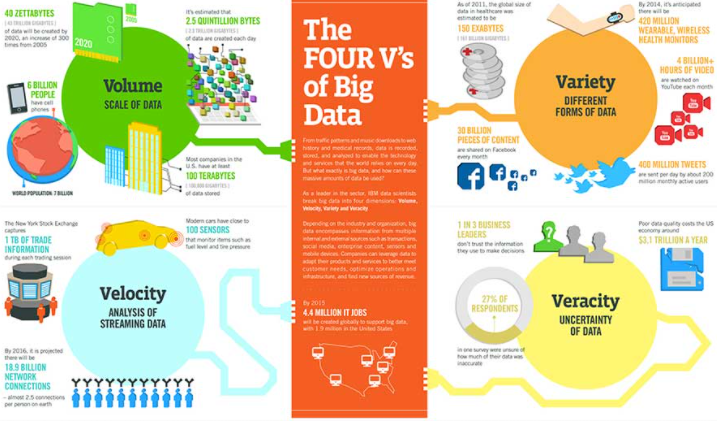
화공 산업에서는 실제 공정 · 실험 · 시뮬레이션 등과 같은 다양한 출처로부터 데이터가 발생한다. 이 데이터들은 **형태**와 **수치적 의미**에 따라 분류가 다르며 적절한 수집 · 저장 · 변환 · 이용방법이 다르다. ‘화공 산업 데이터 소개’는 데이터의 원시적인 형태와 화공 산업에서 이용되는 데이터를 소개한다.

1.1. DIKW 피라미드



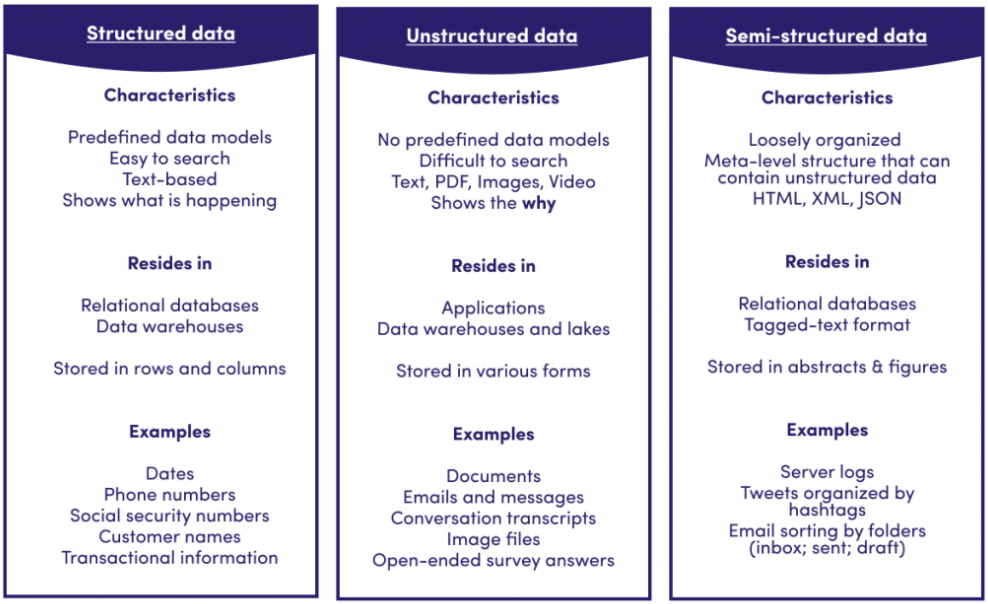
* Data, Information, Knowledge, Wisdom으로 이루어진 계층도는 과학 및 공학 분야에서 널리 인용되는 형태이다.
* 데이터는 관찰, 측정을 통해서 수집된 사실이나 값, 수치, 문자 등 가공되지 않은 날 것을 의미한다.
* 정보는 필요에 의해 정제 및 가공된 데이터를 의미한다.
* 지식은 일반화하고 체계화된 즉시 적용 가능한 데이터 형태이다.
* 지혜는 지식의 근본원리에 대한 이해를 바탕으로 도출되는 아이디어이다.

1.2. 빅데이터의 4V



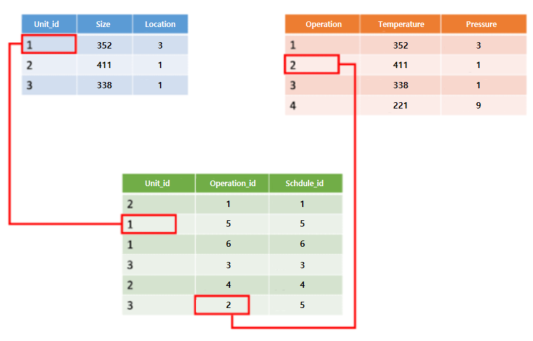
* 빅데이터는 단순히 많은 데이터를 의미하는 것이 아닌 편집 (Compilation), 저장 (Storage), 사용 (Exploitation)이 가능한 데이터 들을 의미한다.
* 빅데이터는 보통 4가지 큰 요소가 있는데, 이것들은 크기 (Volume), 속도 (Velocity), 다양성 (Variety), 정확성 (Veracity)이다.
* 빅데이터에서 크기는 정통적인 의미에서 데이터 크기를 의미한다. SNS나 특정 장치로부터 매우 많은 양의 데이터의 축적이 가능해야한다.
* 많은 양의 데이터는 데이터 베이스나 분석을 위한 계산 엔진으로 이동이 원활해야한다. 데이터의 수집 · 저장 · 변환 · 이용을 위해서 데이터의 전송/계산 속도는 빅데이터에서 역시 중요한 요소다.
* 다양성은 데이터의 범용적인 사용을 위해 필요한 요소다. 다양한 관점으로부터 객관적이고 명확한 정보 (Information)를 추출하여 가치판단을 도울 수 있어야한다.
* 빅데이터는 종종 데이터의 유실이나 변형이 생길 수 있다. 데이터의 정확성은 데이터 의존이 가능한 기준을 제시하므로 데이터의 사용과 직결되는 요소이다.

1.3. 데이터의 형태



데이터의 형태는 **정형 데이터** (Structured data), **비정형 데이터** (Unstructured data), 그리고 **준정형 데이터** (Semi-structured data)로 분류된다. 이는 데이터의 형식과 구성에 따른 분류로, 데이터의 저장 및 이용에 중요하다. 보통 화공 산업에서 발생하는 데이터는 센서와 같은 검출 장비에 의해 해석된 데이터인 정형 데이터를 의미한다. 하지만 많은 잠재적 **데이터 근원** (data source)은 사람의 **의도**나 **해석**이 아직 개입하지 않아 형식과 구성이 없는 비정형 데이터를 생성한다. 앞으로도 화학공학자들은 화공 산업에서의 정형 데이터 확보에 심혈을 기울일 것이지만, 데이터의 비정형 수준에서의 해석 또한 현재 수준에서 중요하다. 아래는 실제 화공산업에서 발생하는 데이터의 예시를 데이터 형태에 따른 분류로 정리했다.

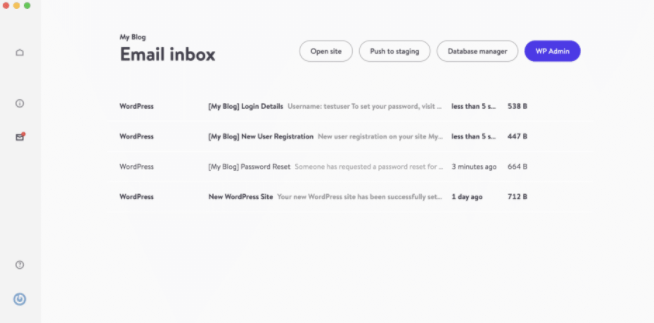
* 정형 데이터 (Structured data)



* 정형 데이터는 보통 열/행으로 구성된 테이블 형태의 데이터를 의미한다.
* 데이터 베이스에 저장될 때, 데이터 관계 (Relation)와 서열 (Hierarchy)를 규명한다.
* 데이터 베이스에 저장된 데이터는 수치형 또는 범주형 값을 포함한다.
* 비정형 데이터 (Unstructured data)

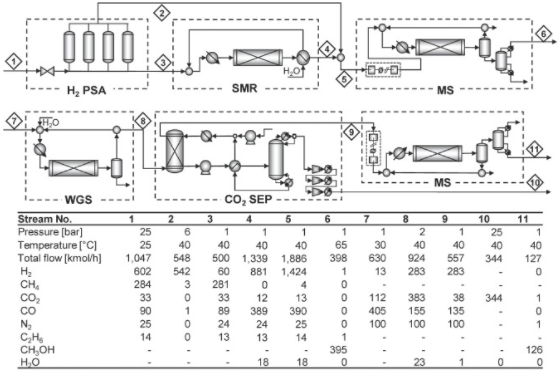


* 비정형 데이터는 사진 · 음성 · 비디오 등을 포함하며, 정형 데이터가 아닌 거의 모든 것을 의미한다.
* 데이터 사이 관계나 형식이 없어 저장 · 이용을 위해서 임시적인 형태를 정의해야 한다.
* 예를 들어, 사진은 명함 사이즈, 비디오는 1080p 수준 화질 등을 의미한다.
* 준정형 데이터 (Semi-structured data)

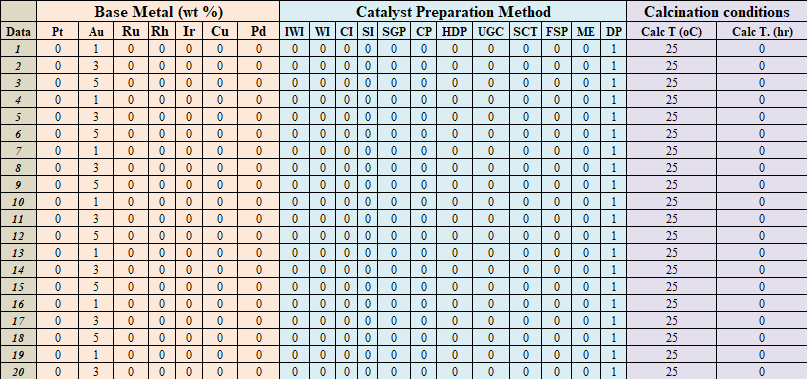


* 준정형 데이터는 정형 데이터와 비정형 데이터의 중간에 해당하며, 데이터에 형식이 존재하지만 정보 추출을 위한 단계가 필요한 데이터이다
* 일반적으로 이메일은 작성 방식 (보낸이, 받는이, 인사말 등) 이 존재한다. 따라서, 데이터 수집을 위한 특정 알고리즘만으로 데이터 전처리 과정 없이 불 필요한 데이터 제거를 통해 중요한 텍스트 만을 추출할 수 있다.

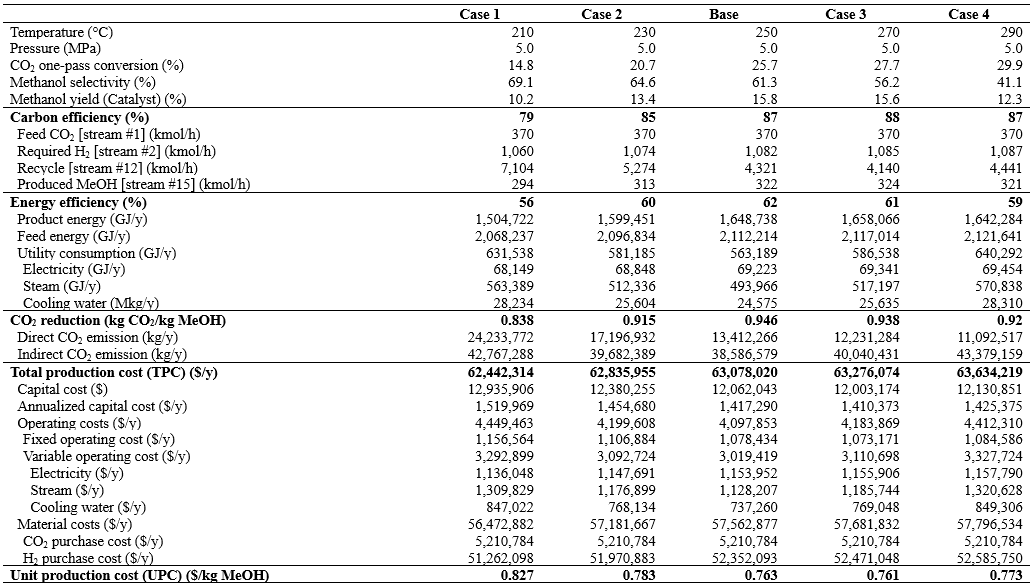
정형 데이터의 예시)



* 공정 흐름도 (Process flow diagram, PFD)의 흐름 정보 (Stream information)은 각 세부 항목에 대한 수치적 정보가 있는 정형 데이터 이다.

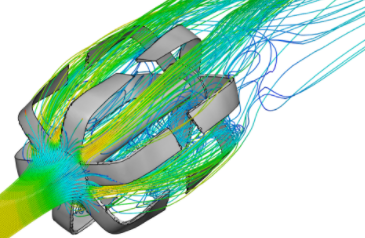


* 촉매 설계를 위한 실험의 독립변수 · 종속변수가 설정 · 측정 · 기록된 데이터는 정형 데이터이다.

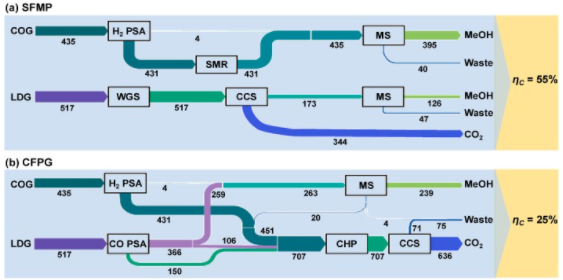


* 공정 평가는 공정들의 평가 요소들 (환경성, 기술성, 경제성)을 수치적으로 기록된 정형 데이터이다.

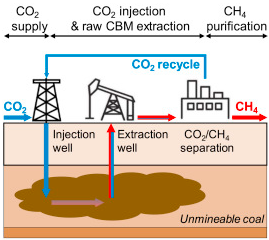
비정형 데이터의 예시)



* 전산 유체역학 이미지는 물체 내 · 외부에서 유체의 거동에 따른 열과 유체의 흐름을 보여주는 대표적인 비정형 데이터이다.

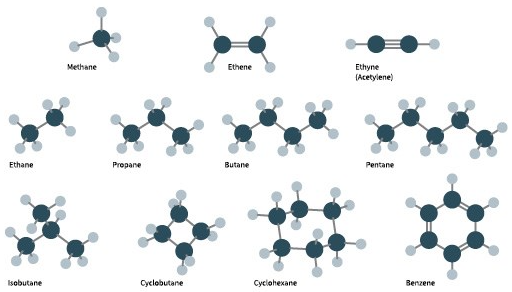


* 공정의 Block-diagram은 공정의 구성요소와 생산품을 표현하는 이미지이며 비정형 데이터이다.



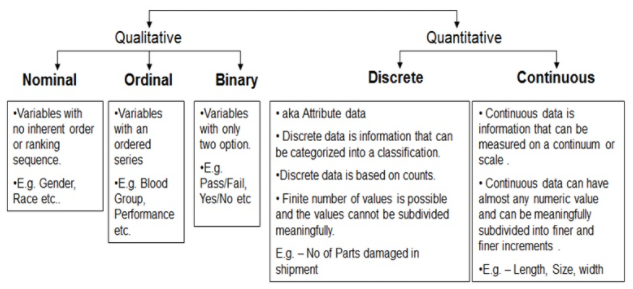
* 시추 공정의 개요도는 지반으로부터 메탄가스 추출을 위한 정보를 함축하고 있는 이미지이며 비정형 데이터이다.

준정형 데이터의 예시)



* 화합물의 형태와 기술은 특정 방식을 따르지만 정보 획득 (이중결합 여부, 고리형 여부 등)은 정보 인식 절차가 필요한 비정형 데이터이다.

1.4. 데이터의 수치적 의미

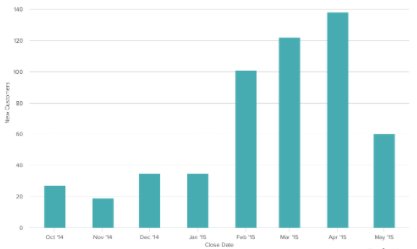


데이터는 수치적 의미에 따라 **질적 (Qualitative) 데이터**와 **양적 (Quantitative) 데이터**로 나뉜다. 이는 화공 산업에서 발생하는 데이터도 마찬가지이며, 데이터 범주와 특징에 따라 적절한 데이터 수집 및 저장 방법이 나뉜다.

* 질적 데이터 (Qualitative data)

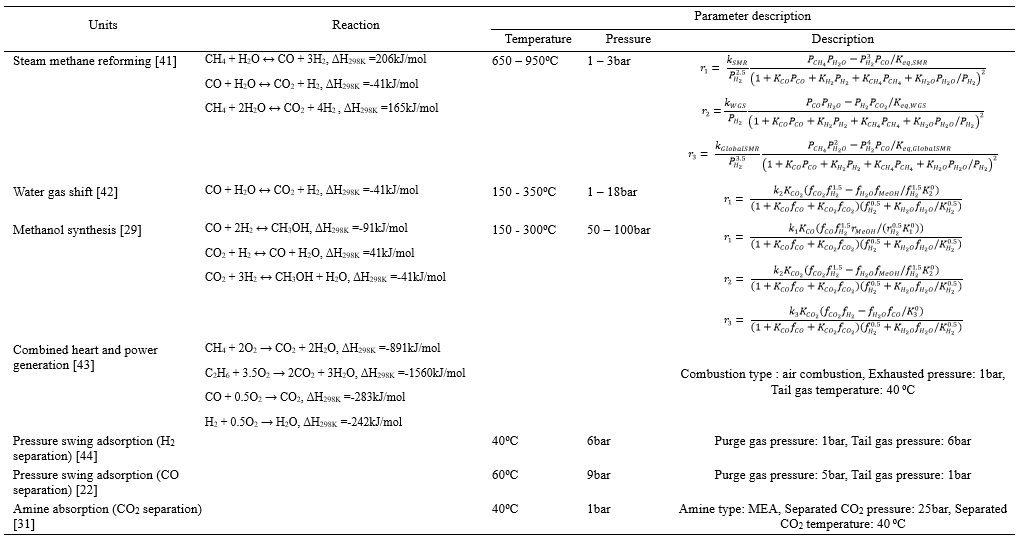


* 질적 데이터는 수집된 데이터 항목들에 수치적인 의미가 없는 데이터로, 성별 · 인종 · 혈액형 · 시험의 당락 (Pass/Fail) 등을 의미한다.
* 또한, 데이터의 서열/순차적 특징이나 범주의 수에 따라 ‘Nominal’, ‘Ordinal’, ‘Binary’의 형태로 분류된다.
* 양적 데이터 (Quantitative data)



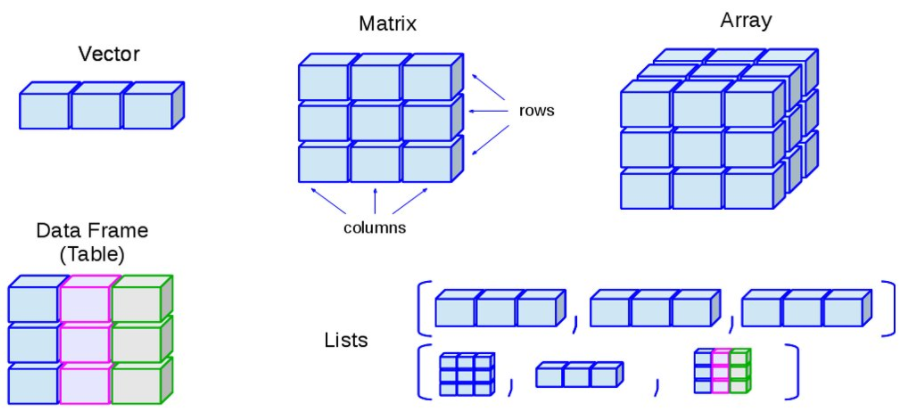
* 양적 데이터는 수치적인 의미를 갖는 데이터로, 시험 등수 · 물건의 길이 같은 것을 의미한다.
* 이는 다시 데이터의 순차적 특성에 따라 ‘Discrete’과 ‘Continuous’로 나뉜다.

질적 데이터와 양적 데이터의 예시)



* 위의 테이블은 공정에 사용된 반응을 위한 장치 (Unit), 반응 (Reaction), 운전 조건 (Temperature, Pressure), 그리고 반응식 (Description)을 포함한다.
* 장치는 Stream methane reforming, water gas shift등 반응을 위한 것으로, 이름에 해당하는 데이터이다. 이는 항목들 사이 연관이 없고 순차적인 특징이나 제한된 선택지가 아니므로 Normial 형태의 질적 데이터이다.
* 장치 이름 옆에는 [41], [42]와 같은 참고문헌이 있고 항목들 사이 순서나 연관이 중요하지 않다. 따라서 이들은 Discrete 형식의 양적 데이터이다.
* 운전 조건은 데이터 사이 수치적 연관성이 있는 온도와 압력으로 구성됐다. 따라서, 이들은 Continuous 형식의 양적 데이터이다.

1.5. 데이터의 구조



* 벡터는 벡터 데이터 구조는 숫자, 문자 등 원소들의 집합으로 이루어진 배열이다. 동일한 자료형의 값이 여러 개 연속되어 있다. 가장 단순한 데이터 구조로, 원소가 숫자면 숫자 벡터, 문자면 문자 벡터가 된다. 원소 중 문자가 하나라도 있을 경우엔 문자 벡터가 된다.
* 벡터는 원소들의 집합인 1차원이지만, 행렬은 행렬은 행과 열로 이루어진 2차원의 데이터 형태이다. 명령어 matrix를 통해 선언된다.
* 데이터 프레임 또한 2차원 형태의 데이터 구조로, 행렬은 모든 데이터 유형이 동일해야 하나 데이터 프레임은 서로 다른 유형의 데이터가 저장될 수 있다.
* 배열은 2차원 데이터 구조를 쌓아 올린 형태로, 행렬 형태의 데이터를 층위별로 저장한다.
* 데이터 프레임은 행렬과는 다르게 각 열이 서로 다른 데이터 타입을 가질 수 있으나 각 열의 데이터 타입은 동일해야 한다.
* 리스트는 터와 비슷하나 벡터와 달리 여러 자료형의 데이터를 섞어서 저장할 수 있다.(벡터나 data frame 등도 원소로 저장 가능)

**2. 데이터 수집 방법론**

데이터의 수집은 사용 목적과 작업 환경에 따라 상이하다. 이를테면, 1) 실제 규모 공정의 운전 전략 수립, 2) 신물질 개발을 위한 실험, 그리고 3) 핵분열 혹은 핵융합 실험은 시행착오에 의존할 수 없는 각기 다른 특징들 (비용, 효율, 위험도)이 존재한다. 따라서, 사용 목적과 작업 환경에 적합한 데이터 수집 방법론이 요구된다. ‘데이터 수집 방법론’은 다양한 데이터 수집 환경을 소개하고 데이터 수집을 위한 방법을 소개한다

2.1. 센서 네트워크



센서 (Sensor)는 공정이나 실험환경에서의 물질이나 상태의 환경적 변화를 신호로 치환하는 소자나 장치를 의미한다. 센서가 수집하여 치환하는 데이터는 보통 온도, 압력, 유량, 속도 등 다양한 물리량을 의미한다. 공정에서는 보통 수많은 센서로 하여금 공정에서의 반응물이나 생성물의 상태 확인을 통해 앞으로의 운전 전략을 결정한다. 결과적으로 공정을 효과적이게 제어해 효율이나 수율 등 공정 개선에 전반적인 활용이 가능하다.

* 다양한 센서

**(예시)**

2.2. 문헌 수집

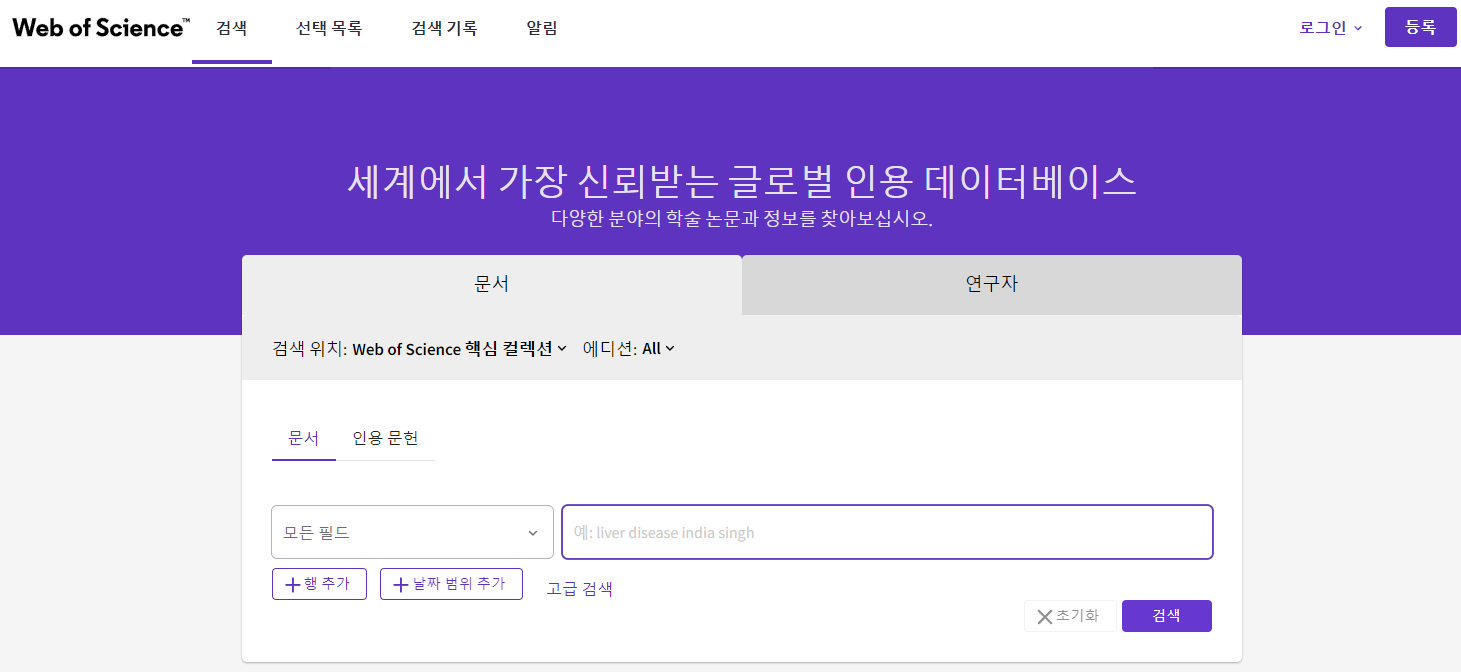


문헌 조사 및 수집 (Literature review)은 논문이나 서적으로부터 데이터 · 정보 · 지식 등을 얻는 과정이다. 대게 실험 설계를 위한 선행 연구 조사를 목적을 위해 문헌 조사 및 수집이 행해지지만, 화공 인공지능 개발을 위한 데이터 수집을 위해 진행될 수 있다. 특히, 신물질 개발을 위해 선행적으로 진행됐던 실험 데이터를 수집하여 물질 합성 전략이나 중요한 물성의 예측에 사용될 수 있다. 이를테면 고성능 촉매 발견을 위한 많은 시도를 논문과 같은 문헌으로부터 찾아 데이터 구조를 설정하고 축적할 수 있다. 문헌 수집을 위한 과정을 데이터 사이언스에서는 웹 크롤링 (Web-crawling)과 웹 스크래핑 (Web-scraping)이라고 하며 기초적인 프로그래밍 언어 지식으로 이를 시도할 수 있다.

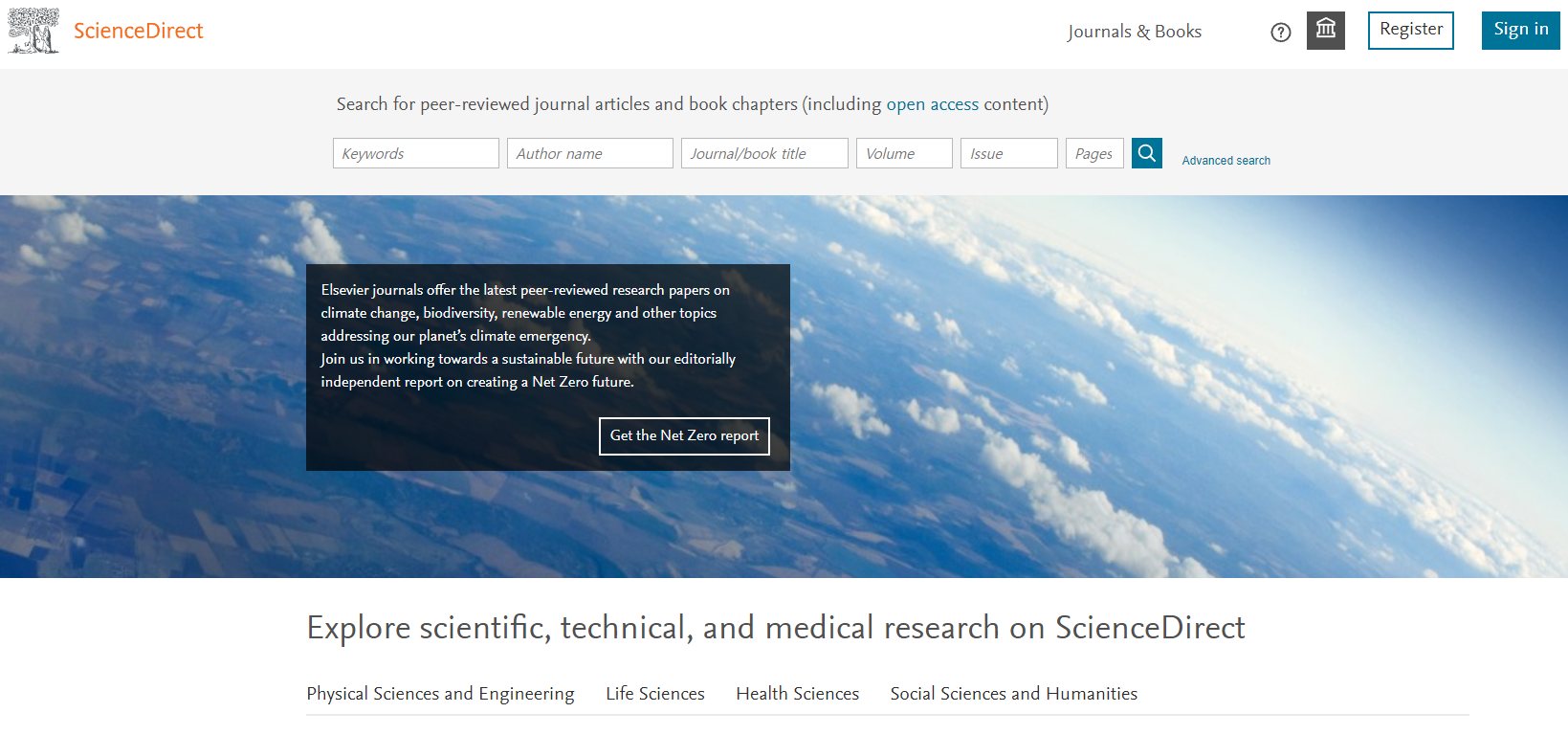
* 웹 크롤링 (Web-crawling)

웹 크롤링은 데이터 수집을 위해 적합한 데이터 근원을 탐색하는 과정을 의미한다. 신물질 개발을 위한 화공 데이터 수집을 위해 보통 Web of Science, SCOPUS, Science Direct와 같은 학술지 검색 엔진들을 사용한다. 해당 검색 엔진을 사용하면 학술지에 수록된 논문들과 실험 정보 · 결과를 확보할 수 있다.

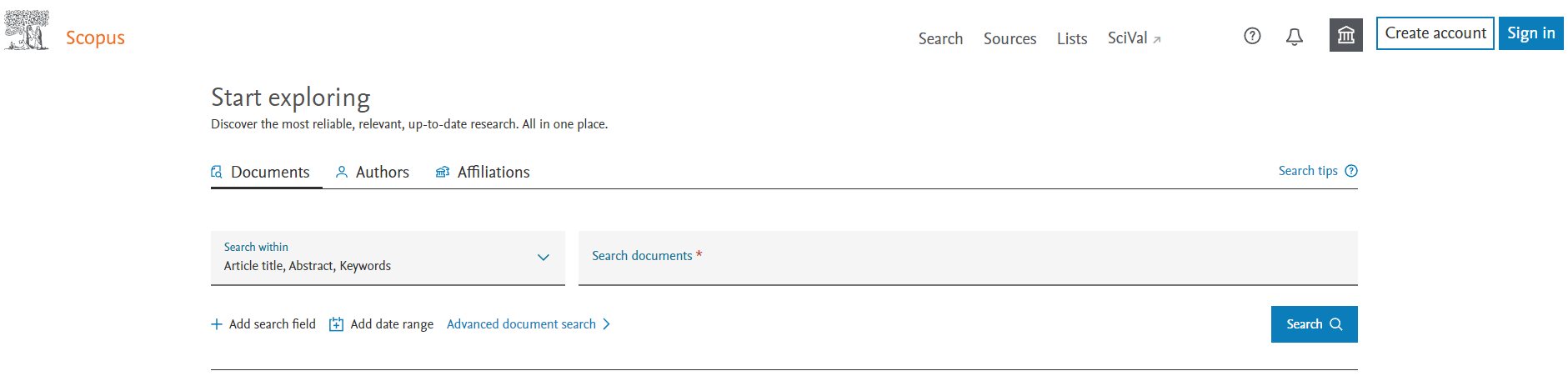
* Web of Science (www.webofscience.com)



* Science Direct (www.sciencedirect.com)



* SCOPUS (www.scopus.com)



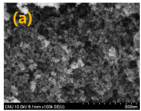
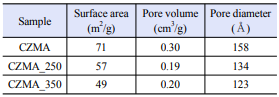
* 웹 스크래핑 (Web-scraping)

웹 스크래핑은 데이터 수집을 위한 자동화된 조직적 탐색 방법을 의미한다. 사람이 직접 데이터를 찾거나 컴퓨터 프로그램을 개발하여 인터넷으로부터 데이터를 체계적으로 수집하는데 사용된다. 물질 합성과 같은 분야의 데이터는 논문을 통해 테이블이나 그림으로부터 추출해낼 수 있다.

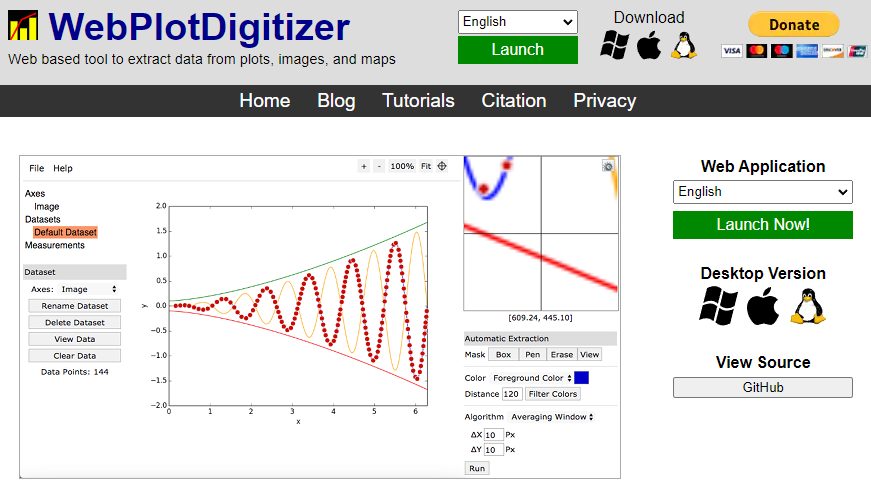
예를 들자면, 촉매 데이터 수집을 위해 문헌에서 촉매 합성 및 결과에 대한 정보를 찾을 수 있다. 해당 문헌 (DOI:10.7316/KHNES.2019.30.2.95)은 WGS 촉매 개발에 관한 내용으로 촉매 합성을 위한 정보를 포함하고 있다.

“Cu/ZnO/MgO/Al2O3 촉매는 공침법을 이용하여 합성하였다. Cu(NO3)2·3H2O (99%, Junsei), Zn(NO3)2·6H2O (95.0%, Junsei), Mg(NO3)2·6H2O (99%, Junsei), Al(NO3)3·9H2O (98.0%, Junsei)를 전구체로 사용하여Cu/Zn/Mg/Al의 비율이 45/45/5/5 mol%가 되도록 증류수에 혼합하여 전구체 수용액을 제조하였다. 제조된 전구체 수용액을 침전제인 Na2CO3 (99.0%, Samchun Chemical) 수용액에 한 방울씩 떨어뜨리면서 침전을 유도하였다. 침전이 진행되는 동안 1.0 M NaOH (≥97.0%, KANTO) 수용액을 첨가하여 pH를9.0으로 유지하였다. 침전이 끝난 후에 침전물은 60℃에서 지속적으로 교반하면서 36시간 동안 숙성되었다. 숙성이 끝난 용액은 상온에서 필터를 이용하여 침전물을 걸러내었다. 걸러진 침전물을 증류수에 3차례 세척하고 회수하여 110℃에서 24시간 동안 건조되었다. 건조된 침전물은 공기 분위기에서 4시간 동안 400℃에서 소성되었다. “

해당 문헌에서 촉매의 중요 성질에 대한 데이터와 촉매 형태에 대한 정형 데이터와 비정형 데이터를 얻을 수 있다.



종종 데이터는 그래프의 형태로 제공되며, 데이터 추출을 위해 그래프 전용 툴 (tool)들이 사용될 수 있다. 이를테면 ‘WebPlotDigitizer’는 그래프 데이터 추출에 사용되는 무료 툴 이다.



문헌으로부터 중요 그래프를 이미지로 저장한 후, 그래프 데이터 추출 툴을 사용해 데이터를 확보할 수 있다.

