정형 데이터(Structured Data)

정형 데이터는 고정된 필드에 저장된 데이터를 말하며 관계형 데이터베이스와 스프레드시트 등을 예로 들 수 있다. 정형 데이터의 경우는 데이터베이스를 설계한 기술자에 의해 수집되는 정보의 형태가 정해지게 된다.

반정형 데이터(Semi-Structured Data)

반정형 데이터는 고정된 필드에 저장된 데이터는 아니지만 XML, HTML 텍스트 등 메타데이터 및 스키마를 포함하는 데이터이다.

여기서 반정형 데이터에서 중요한 위치를 차지하고 있는 HTML의 변화에 대해서 말하고자 한다. 인터넷의 확산으로 HTML 자료들이 방대해지고 있는 상황에서 정보 탐색을 위한 요구사항들이 점차 늘어나고 있다. 웹 문서를 보다 쉽게 탐색하고 정확하게 해석하여 의미 있는 정보를 추출하기 위해서이다.

비정형 데이터(Unstructured Data)

비정형 데이터는 고정된 필드에 저장되어 있지 않은 데이터를 의미하며 페이스북과 트위터, 유튜브 영상, 이미지 파일, 음원파일, 워드 문서, PDF 문서 등을 예로 들 수 있다. 비정형 데이터의 경우는 페이스북, 트위터, 네이버, 다음 등에서 생성되는 실시간 정보들을 통해서 더 많은 정보들을 수집하고 분석할 수 있다. 예를 들면 특정 지역의 날씨 정보, 유동 인구의 수, 이들의 판매 정보 등을 수집할 수 있다. 형태가 정해지지 않는 정보속에서 분석 방향에 따라 다양한 정보를 수집할 수 있는 것이다.