**#1**

안녕하세요. ‘지속 가능한 데이터를 바란다’ 발제를 맡은 이정윤입니다.

앞선 발표에서는 openapi와 보도자료 데이터를 다양한 방법으로 활용하며, 공공데이터의 중요성과 정확한 데이터 수집의 필요성을 다시 한번 확인했습니다.

**#2**

그러나 여전히 보건복지부에서 제공하고 있는 openapi의 데이터와 질병관리청의 일별 보도자료는 완전히 일치하지 않는다는 문제점이 있습니다. 따라서 이 발제에서는 보도자료와 공공데이터에 어떤 오류들과 차이가 있었는지 다뤄보려 합니다.

**#3**

우선, 처음 API 와 보도자료 데이터를 비교한 이유는, 감염현황 api에는 국내 첫 확진자 발생 시기인 1월 말의 데이터가 누락되어 있기 때문입니다. 그래서 이를 파악하고자 보도자료를 보았는데, 1월 2월에는 하루 다르게 분류 형태가 변한 날이 있을 정도로 일정한 체계를 갖추고 있지 않았습니다.

**/**

화면에 보이는 4개의 보도자료는 이러한 체계 변동을 대표적으로 보여주는 날짜입니다. 모두 ‘코로나 바이러스 감염증’ 사이트의 보도자료인데, 좌측 상단의 1/22 에는 조사대상 유증상자와 능동감시 대상자로 분류했지만, 우측의 2/7일에는 확진환자와 조사대상 유증상자로 분류하고, 불과 하루 뒤인 2/8에는 확진환자와 의사환자로 구분합니다. 일정한 체계 없이 수정이 필요한 순간마다 형태를 변동하고, 이에 관한 자세한 설명이 없어 처음 보도자료를 볼 때 이해에 어려움을 겪었습니다.

**/**

특히 가장 혼란스러웠던 부분은 ‘격리해제 수’ 였습니다. 지금 초록색으로 표시된 부분이 ‘격리해제수’ 인데, 1/22의 격리해제수는 11건, 2/7일의 격리해제수는 842건으로 두자리, 세자리 수인데, 2/8일 이후로는 2건, 3건 등 한 자리 수로 기록되기 때문입니다. 격리해제수는 누적값이기 때문에 수치가 감소할 순 없어서 이와같은 현상에 의아했는데, 자세히 살펴보고 나서야 그 이유를 알 수 있었습니다.

**#4**

2/8을 기점으로 이전에는 ‘격리해제수’ 가 ‘결과 음성수’를 의미했지만, 이후에는 ‘치료후 완치자’를 의미했던 것입니다. 컬럼 이름은 변하지 않았지만, 해당 데이터 값을 갑자기 바꾸니 이와같은 혼란이 생겼습니다.

**/**

또한, 이렇게 보도자료를 확인하는 과정에서 2월 초에 API에는 검사 진행수가 0건, 2건으로 집계되는데, 보도자료에는 87건, 61건등으로 보도되는 등 데이터가 일치하지 않는 날도 있다는 사실을 알게 되었습니다.

**/**

따라서 국내 첫 확진자가 발생한 2020년 1월부터 2021년 2월까지 전 기간의 보도자료 데이터 값을 다음과 같이 xlsx 파일로 만들어 같은 기간의 API와 일치/불일치 정도를 비교해보았습니다.

**#5**

그 결과 다음 4가지의 공통된 불일치 항목으로 정리할 수 있었습니다. 참고로 표시되어 있는 총 데이터 불일치 건수는 확인 시점에 따라 달라질 수 있다는 점을 미리 알려드립니다.

**/**

이 중 69건이라는 가장 많은 오류항목인 (4)번의 예시를 소개하겠습니다.

**#6**

화면의 두 보도자료는 2021년 1월 31일과 2월1일의 것 입니다. 매일 업데이트되는 보도자료에는 당일 데이터와 전날 데이터를 함께 표시하는데, 지자체 중복신고 및 오신고 등 ‘집계오류’의 원인으로 다음날 데이터가 수정되는 경우가 있습니다. 1/31 같은 경우에도 확진자가 31일에는 78205명이었지만, 1일 지자체 중복신고로 정정된 값은 2명이 적습니다. 또한 격리해제수도 31일에는 67876건이지만, 1일 지자체보고 오류로 정정된 값은 21건이 적습니다. 문제는, 보도자료에는 우측 자료처럼 빨간색으로 오류 수정을 표시하지만, API에는 수정값이 미반영 되는 경우가 있다는 것입니다.

**/**

아래의 표는 같은 시기의 API 데이터인데, 보시는 것처럼 확진자 수는 1일 수정된 값인 78203명으로 수정되어 있지만, 격리해제 수는 31일의 21건이 더 많은 수정 전 값이 그대로 표기되어 있습니다.

이 외에도, 같은 날의 API와 보도자료, 그 다음날 수정된 보도자료의 값까지 모두 다른 경우도 있고, 11/8 같은 경우에는 같은 날 같은 사이트에서 보도했음에도, ‘국내발생현황’ 보도자료와 ‘정례브리핑’ 데이터의 격리해제 수 값이 일치하지 않는다는 문제점도 있었습니다.

**/**

그렇다면, 국내 데이터들이 일치하지 않는다는 것을 확인한 이 시점에서, ‘해외에서 보도되고 있는 대한민국의 코로나19 감염 데이터는 국내 데이터와 일치한가?’에 대한 궁금증이 생깁니다.

**#7**

그래서 세계보건기구 WHO의 대한민국 데이터, 공공데이터 포털의 API, 질병관리청 보도자료 이 세 데이터도 한 번 비교해보았고, 그 결과 약 네 가지의 불일치 경우로 정리되었습니다. 특히 세 데이터가 모두 불일치하는 경우와, API와 보도자료는 일치하지만, WHO와는 불일치한 경우가 각각 25건, 127건이 있었습니다. WHO의 집계 기준은 정확히 알 수 없으나, 국내 데이터를 그대로 사용하고 있진 않다는 것을 알 수 있습니다.

**#8**

결국, 이렇게 데이터를 제공하는 기관에 따라 값들이 일치하지 않는 것은 데이터를 사용하는 소비자 입장에서는 ‘기준이 되는 데이터는 무엇인가?’ 하는 의문으로 남습니다. 최소한 국내 데이터 간에는 통합된 집계기준과 분류체계가 필요하지 않을까 생각되기도 합니다.

**#9**

이상으로 발표를 마치겠습니다. 감사합니다.