

성웅 이순신장군을 검색하던 이야기...

미국의 세계최대의 데이터뱅크인 다이알로그(Dialog)에는 거의 모든 분야를 망라한 데이터베이스를 확보하여 무려 600여종의 데이터베이스가 수록되어 있다. 약20년 전에도 200여종의 데이터베이스가 있었던 것을 지금도 기억한다. 이 시스템은 요즘 흔히들 사용하는 (메뉴선택방식)의 데이터베이스가 아니라 모두 [키워드검색과 논리검색]을 겸한 고급스러운 검색시스템이다.

1978 이니까 19년전의 일이었다. 텔렉스로서 온라인검색을 성공한 이래 전화로서 검색해 보려고 애를 쓰던 시절이었다. 검색요령은 [키워드(주요단어)]로서 찾고 너무 많으면 논리검색(AND, OR, NOT)으로 검색범위를 제한하는 기법을 겨우 터득하였을 때의 이야기였다.

신문기사에 빈번히 남직한 유명한 사람의 이름은 즉각 키워드로 활용할 수 있어서 인명으로서 찾는걸 즐겨 했었다. 정치가도 찾아보고 학자도 찾아보고 진짜 학위 소지자인지 아닌지도 찾아보곤 했었다.

이 모습을 보던 한 친구가 말했다.

“요술방망이 같구먼...나오너라 똑딱하면 금방 나오고...참 희한하네요. 그나저나 당신이 아는 사람만 찾으니까 재미가 없잖아? 내가 존경하는 역사적인 인물을 한번 찾아 보십시오. 제대로 찾아내는지 아닌지 테스트를 좀 해야겠어...”

“그 사람이 누군데..”

“이순신 장군!”

“이름 한번 넣어보지요. 그런데 그어른의 영문 철자가 무엇이지요?”

“그야 나도 모르지...”

이순신의 성씨가 LEE일까 RHEE일까 LIE일까 REE일까 YI일까? 우선 생각나는 것만 해도 5가지가 넘는다. 다음 이순신의 순자의 영문자 철자

는 어떻게 표기하는지 알 수가 없다. SOON일까 SUN일까? 적어도 2가지는 되는 것 같다. 그럼 마지막 글자 신은 SHIN일까 SIN일까?

그 다음을 생각했다. 어느 데이터베이스를 고를까? 화학이나 물리분야의 데이터베이스에서 이순신을 찾아봐야 나올턱이 없다. 목록을 뒤져보니 역사데이터베이스가 있었다. 그러나 이나마 미국역사에 관한 데이터베이스였다. 일단 미국역사 데이터베이스를 불러놓고 나서 이순신이라는 키워드로 찾아 들어가 볼까 했는데 이름에서 막혔다.

성을 먼저 쓰고 이름을 쓰는지 이름을 먼저 쓰고 성을 쓰는지도 확실하지 않았다. 두번째 이름의 두자 사이를 띄울 것인지 붙일 것인지 하이픈으로 이을 것인지도 확실하지가 않다. 성을 먼저 쓰더라도 이름앞에 콤마를 붙일 것인지 그대로 넘어갈 것인지도 확실하지가 않았다. 결국 수많은 표기가능한 모든 방법을 총동원해서 키워드로 삼으려고 했는데 결국 실패하여 버렸다. 모두 120가지의 다른 표기방법이 있었기 때문이었다.

친구에게 변명을 했다.

“미국역사 데이터베이스에서 이름(이순신)으로 검색하는 것은 불가능합니다. 우리나라 고유명사의 로마자 표기법이 통일되지 않아서 정보검색을 할 수가 없네요. 그렇지만 이렇게 한번 찾아보겠습니다.”

니다.”

“아무리 요술방망이라도 어려우리라 생각은 했어요. 그런데 어떻게 찾는다고 했나요?”

“먼저 미국역사 데이터베이스에서 키워드를 코리아로 찾겠습니다. 이말은 한국역사에 관한 자료가 몇 건이 들어있느냐라는 질문과 같지요. Korea란 키워드를 넣으니까 1,200건이 들어있다고 응답이 나왔지요? 이 1,200건 안에 분명히 이순신 장군의 자료가 포함되어 있을 것인데 1,200건을 모두 다 읽어 볼 수는 없지요. 범위를 좁혀서 보지요. 그 다음 키워드로서 장군이란 말의 영어단어가 무엇입니까?”

“그야 제너럴(General) 아닙니까?”

“그럴수도 있지만 해군제독이란 말은 어드미럴(Admiral)이란 말이 더 적합한 것 같습니다. 들다 써 보지요.”

20건이 검색되었다. 즉 한국의 해군제독이라는 질문에 20건이 나왔다는 뜻이 된다. 1,200건을 모두 보는 것은 시간이 걸리지만 20건쯤이야 1-2분 정도면 한눈에 확인할 수 있다.

드디어 제목이 [이순신 장군과 거북선]이라는 제목의 연구논문이 검색이 되었고 영문스펠이 Yi Sun-Shin이라고 쓰여져 있는걸 확인하였다. 그리고 그내용을 친구에게 보였다.

“이름의 철자를 모르면 검색 불가능이 아니라 주변상황을 나타내는 키워드를 잘 골라내면 정확하게 희망하는 정보를 검색할 수가 있지요.”라면서 어깨를 으쓱거리면서 설명한 기억이 난다.

정보화, 세계화와 더불어 우리의 문화가 점차 로마자로 표기해야 할 일들이 급증하였다. 그러나 이것이 표준이 없이 소리나는대로 적어버리는 습관이 일반화 된 현실에서 데이터베이스화가 순조로울 수가 없다.

지난호에서도 보고했지만 국제표준화기구에서 다루고 있는 [한글의 로마자 옮겨적기]가 아직도 남북한의 팽팽한 대립으로 금년에도 아무런 결론이 나오지 않았다. 아무래도 내년까지 다시 논의를

미루어야 할 모양이다. 국제표준화에 앞서서 국내 표준만이라도 조속히 이루어져야 할터인데 이나마 순조롭게 나아가고 있는 것 같지가 않다.

이러한 상황이면 앞으로 만들어지는 모든 데이터베이스에서 대구를 찾으려면 TAEGU인지 DAEGU인지 분명하지 않고 부산을 찾으려면 PUSAN인지 BUSAN인지 분명하지 않는 상태에서 정보검색에 임해야 한다.

7-8년전, 치안본부에 방문하여 본인의 이름으로 주민등록번호를 시험삼아 찾은 일이 기억난다. [유경희]란 키워드를 넣으니 수십만명의 유경희가 검색되어 나왔지만 본인 것이 없다. 서글퍼서 [나는 대한민국 국민이 아닌가?]하고 한탄을 하는데 컴퓨터를 조작하던 경위 한분이 물었다. [혹시 버들류씨의 유씨가 아닙니까?] [그렇습니다만...]

“그러면 류경희로 찾아야 합니다. 차류종친회에서 공문이 왔는데 모든 버들류씨의 표기를 류로 표기하기로 하였습니다.” “그렇습니까? 바뀌어진 연유야 나중에 알기로 하더라도 당장은 대한민국 국민이란 것이 증명이 되어서 매우 행복합니다.”

유씨를 류씨로, 이씨를 리씨로, 이렇게 고치기 시작하면 한글철자도 만만치 않을 터인데 그래도 되는지 잘 모르겠다. 좌우간 바뀌어지면 정보검색이 잘 안되면 데이터베이스 제작이전에 만들어져야 할 표준이 덜 되어 있다는 것을 의미하는 것이다.

이러한 상태대로 데이터베이스를 계속 만들어 나간다는 것은 결국은 Garbage In Garbage Out(GIGO) 즉 쓰레기를 넣으면 쓰레기밖에 나오지 않는다는 속담처럼 데이터베이스가 아닌 쓰레기공장이 되어버리지 않을까 무척 염려스럽다.

〈한국정보유통센터 회장〉