

# ASCII Table 읽고 쓰기



## ASCII 테이블 파일

과학자들이 netCDF, HDF, FITS, DICOM 등의 여러 가지 폼 나쁜 전문 데이터 포맷을 많이 사용하고 있지만, 아무래도 결국 ASCII 파일을 가장 많이 쓰는 것 같습니다. 특히나 열 방향으로 데이터가 나열되는 테이블 형 아스키 파일은 매우 많이 쓰이는 형태이며 Fortran 프로그래밍 기본 교재에도 가장 많이 등장합니다. 이런 형태의 파일은 Excel과 같은 스프레드시트에서도 읽을 수 있지요. 오른쪽 테이블 파일을 보세요. 첫 번째 열은 별의 식별번호, 이어서 적경의 시,분,초, 적위의 도,분,초, V, I, V-I, R-I, eV, eI, eVI, eRI의 순서로 정리되어 있습니다. 이 문서의 목적은 이런 스타일의 데이터 파일을 매우 간단히 읽고 쓰는 방법에 대한 소개입니다.

ID	RA	DEC	U	I	V-I	R-I	e(V)	e(I)	e(VI)	e(RI)	Nabs	REMARK
1294	11 15 8.20	-61 15 47.1	13.305	11.684	1.621	0.000	0.005	0.009	0.010	0.000	2 2 2 0	X
10001	11 15 4.87	-61 15 46.5	21.727	18.976	2.751	1.540	0.166	0.024	0.168	0.073	1 1 1 1	X
10002	11 15 4.88	-61 15 51.8	99.999	21.007	99.999	1.725	0.000	0.000	0.000	0.225	0 1 0 1	
10003	11 15 4.91	-61 15 45.0	19.119	17.179	1.940	1.099	0.029	0.013	0.032	0.024	1 1 1 1	h
10004	11 15 4.95	-61 15 50.3	23.004	21.451	1.553	99.999	0.316	0.006	0.327	0.000	1 1 1 0	
10005	11 15 4.99	-61 15 45.5	99.999	20.681	99.999	1.300	0.000	0.078	0.000	0.121	0 1 0 1	
10006	11 15 5.45	-61 15 47.0	99.999	20.686	99.999	1.692	0.000	0.077	0.000	0.161	0 1 0 1	
10007	11 15 5.08	-61 15 39.2	99.999	21.629	99.999	1.489	0.000	0.087	0.000	0.263	0 1 0 1	
10008	11 15 5.09	-61 15 42.9	21.497	18.348	3.149	1.558	0.099	0.012	0.091	0.033	1 1 1 1	X
10009	11 15 5.11	-61 15 46.5	18.344	16.650	1.604	0.937	0.019	0.007	0.020	0.012	1 1 1 1	X SB56372A
10010	11 15 5.13	-61 15 44.3	99.999	20.888	99.999	1.540	0.000	0.069	0.000	0.133	0 1 0 1	
10011	11 15 5.13	-61 15 46.1	22.626	20.618	2.008	1.193	0.272	0.097	0.289	0.132	1 1 1 1	SB56372C
10012	11 15 5.13	-61 15 49.2	99.999	20.798	99.999	1.169	0.000	0.060	0.000	0.098	0 1 0 1	
10013	11 15 5.14	-61 15 44.1	99.999	21.474	99.999	1.467	0.000	0.056	0.000	0.205	0 1 0 1	
10014	11 15 5.14	-61 15 44.7	23.302	21.268	2.042	0.971	0.320	0.078	0.329	0.118	1 1 1 1	
10015	11 15 5.15	-61 15 34.5	99.999	20.549	99.999	1.780	0.000	0.065	0.000	0.145	0 1 0 1	
10016	11 15 5.15	-61 15 41.7	22.599	20.213	2.386	1.133	0.188	0.056	0.196	0.105	1 1 1 1	
10017	11 15 5.16	-61 15 37.2	22.789	19.617	3.172	1.535	0.189	0.027	0.191	0.075	1 1 1 1	h
10018	11 15 5.17	-61 15 29.6	28.038	17.382	2.648	1.643	0.073	0.021	0.076	0.200	1 1 1 1	+ Ha_From_HST
10019	11 15 5.17	-61 15 32.1	99.999	21.136	99.999	1.618	0.000	0.081	0.000	0.186	0 1 0 1	
10020	11 15 5.17	-61 15 42.5	18.525	16.042	1.683	0.944	0.028	0.009	0.022	0.014	1 1 1 1	X

여러 데이터 필드가 열방향을 정렬된 테이블(NGC 3603의 측광자료). 자료를 받을 수 있는 곳 : <http://arcsec.sejong.ac.kr/~sungh/paper.html>.  
성환경 등, 2004,AJ,127,1014 The Initial Mass Function and Stellar Content of NGC 3603

## IDL Astronomy User's Library

이 문서에서 소개할, READCOL, FORPRINT와 같은 루틴들은 유명한 IDL Astronomy User's Library에 포함되어 있습니다. <http://idlastro.gsfc.nasa.gov>에서 약 1Mb 정도의 작은 파일을 다운로드 받아 설치할 수 있으며 천문학자가 아니더라도 유용한 루틴들이 많아 일단 설치해 두는 것이 좋습니다.

## 간단히 파일을 읽어 봅시다.

IDL 예제 파일 중에 plot.txt 파일은 오른쪽과 같이 두 줄의 헤더가 있고 세 개의 컬럼으로 되어 있습니다.

IDL> file=filepath('plot.txt', subdir=['examples', 'data'])

IDL> readcol, file, x, y, z ;세 컬럼을 읽어 각각 x, y, z 변수에 담습니다.

```
% READCOL: Format keyword not supplied - All columns assumed floating point
% READCOL: Skipping Line 1
% READCOL: Skipping Line 2
% READCOL: 9 valid lines read
```

```
;This file ...
11 2.9 1.9
12 3.2 3.9
13 6.0 7.1
14 8.5 7.3
15 9.2 10.1
16 9.8 8.9
17 12.7 13.9
18 8.2 7.2
19 5.8 6.9
```

plot.txt

READCOL은 많은 부분을 알아서 처리하기 때문에 위와 같은 메시지(에러 메시지가 아닙니다)를 종종 보게 됩니다. 위 메시지는 "1)아무 지정이 없으므로 모든 컬럼은 Float형이라고 생각하겠다. 2)1행과 2행은 데이터 행이 아니므로 건너 뛴다. 3)총 9행을 읽었다." 는 의미의 보고입니다. 이러한 메시지가 보기 싫으면 /SILENT 키워드를 사용하면 되겠습니다. 끝입니다. 간단하지요?

## READCOL의 문법

앞의 예에서 우리는 파일 이름과 변수 세 개(세 개의 컬럼)만을 지정해 주었지만, 현재 버전에서는 40개 컬럼까지 자유롭게 쓸 수 있습니다. 옵션으로 설정할 수 있는 키워드는 다음과 같은 것이 있습니다.

- COMMENT : 하나의 문자로 지정하며 이 문자로 시작하는 행은 건너 줍니다.
- FORMAT : 각 컬럼의 데이터 타입을 지정합니다. 이를 지정하지 않으면 모든 컬럼은 Float 형으로 간주됩니다. A(문자열), B(Byte), D(Double), F(Float), I(Integer), L(Long Integer), Z(16진수), X(해당 컬럼 건너 뛴)을 의미합니다.
- /DEBUG : 이 키워드를 설정하면 파일 이름, 파일의 줄 수와 같은 조금 더 자세한 정보를 화면에 보여 줍니다.
- /SILENT : /DEBUG와 반대입니다. 화면상에는 조용하게 일을 처리합니다.
- SKIPLINE : 지정 행 수만큼 건너뛰고 읽기 시작합니다. READCOL이 헤더를 알아서 생략하므로 사실 자주 쓰이지는 않습니다.
- NUMLINE : 읽고자 하는 행 수를 설정합니다. 생략하면 끝까지 읽습니다.
- DELIMITER : 각 컬럼을 구분하는 기준이 되는 구분자를 지정합니다. 일반적으로 하나의 문자(예를 들어 기상청 자료 포맷의

경우에는 '#'를 쓰지만 여러개의 문자일 수도 있습니다(예를 들어 HTML의 테이블인 경우 '<TD>'). 이 옵션을 생각하면 콤마(',')나 띄어쓰기(' ')로 컬럼이 구분된다고 가정합니다. 그래서 많은 경우 DELIMITER 키워드가 생략될 수 있는 것입니다.

- /NAN : 이 키워드가 설정되면 비어 있는 필드를 NaN(수 아님)으로 처리합니다. 디폴트는 빈 필드를 0.0으로 처리합니다.
- /PRESERVE\_NULL : Null 필드를 포함한 행을 유효 행으로 처리합니다. 빠진 컬럼이 있어도 다른 컬럼을 읽고자 할 때 유용합니다. 이 키워드가 설정되지 않으면 Null 필드가 있을 경우 데이터 형이 맞지 않은 것으로 간주하여 건너 뜁니다.
- STRINGSKIP : COMMENT와 같지만 한 글자가 아닌 문자열 사용할 수 있습니다.
- COUNT = 변수 : 읽은 유효 행 수를 변수에 저장합니다.
- NLINES = 변수 : 파일의 총 행 수를 변수에 저장합니다. FILE\_LINES() 함수와 같은 역할을 합니다.

## READCOL 예제(1)

```
47185#D0913347#23:16:14#000:00#33.29420#126.16270#1007.8#18.2#88.0#8.2#50.0#69.0#69.2#16.2#0.0#33.29420#126.16270# =
47185#D0913347#23:16:14#000:05#33.29420#126.16270#1005.6#18.2#87.1#9.6#27.8#88.2#88.4#16.0#3.8#33.29404#126.16255# =
47185#D0913347#23:16:14#000:10#33.29420#126.16270#1002.6#17.9#85.2#10.7#21.1#113.6#113.9#15.4#5.1#33.29383#126.16244# =
... 이하 생략 ...
```

위 테이블은 레원존데 데이터 파일인데, '#'으로 컬럼을 구분하고 있습니다. 각 컬럼은 "지점번호#고유ID#시작시각#경과시간#관측소위도#관측소경도#기압#온도#습도#풍속#풍향#고도#지오폠펀셀고도#노점온도#상승속도#위도#경도#="을 의미합니다. 이 중에서 기압,온도,습도,고도의 4개 컬럼만 읽고 싶은 경우 다음과 같이 처리합니다.

IDL> file='MyData\UPP\_RAW\_2010011000.txt' ;파일 이름

IDL> readcol, file, pres, temp, hum, hgt, delimiter='#', format='X,X,X,X,X,X,F,F,X,X,X,F'

건너 뛴 컬럼은 X로 처리합니다. 읽고자 하는 마지막 컬럼 이후 필드에 대해서는 아무런 설정이 필요하지 않습니다.

## READCOL 예제(2)

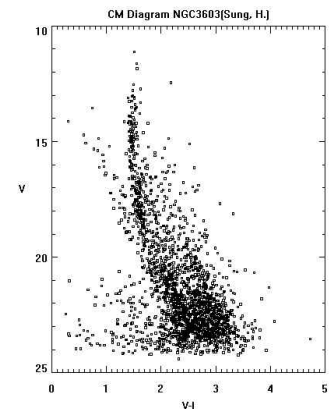
앞에 소개된 NGC3603의 측광자료, n3603hst.dat를 읽어 봅시다. ID와 V, I, V-I, R-I 등급만 필요한 경우 다음과 같이 읽을 수 있습니다.

IDL> file='MyData\n3603hst.dat'

IDL> readcol, file, id, v, i, vi, ri, format='I,X,X,X,X,X,X,F,F,F,F'

% READCOL: 3457 valid lines read

IDL> plot, vi, v, xrange=[0, 4.5], yrange=[23, 10], psym=3 ;오른쪽 CM도를 그림니다.



## FORPRINT로 파일 쓰기

위 측광테이블에는 V, I, V-I, R-I 등급이 기록되어 있습니다. 우리가 만일 ID, V, R, I 등급으로 구성된 테이블 파일을 새로 쓰려면 어떻게 해야 할까요? FORPRINT는 이런 일을 간단하게 처리합니다. ((R-I)+I = R 등급)

IDL> forprint, id, v, ri+i, i, TEXTOUT='D:\hst3603vri.dat' ;파일로 씁니다.

IDL> forprint, id, v, ri+i, i ;TEXTOUT 키워드가 없으면 화면(콘솔)에 출력합니다.

## FORPRINT의 문법

READCOL과 유사한 키워드를 사용합니다. 현재 버전에서는 18개의 컬럼까지 허용합니다.

- COMMENT : 파일의 첫줄에 쓸 내용(헤더)입니다. 생략하면 파일을 쓰는 시각이 첫 줄에 기록됩니다. 어떤 헤더도 쓰지 않으려면 /NOCOMMENT 키워드를 사용합니다.
- TEXTOUT : FORPRINT의 소스코드를 보면 몇가지 사용 옵션이 있습니다만 실제로는 "TEXTOUT=파일이름"의 형태로 대부분 씁니다. 생략하면 화면(콘솔)에 출력합니다.
- FORMAT : 'F12.8, I4, A6' 과 같이 자릿수를 명확히 지정해 줄 수 있습니다. 'F, I, A'와 같이 자릿수를 생략할 수 있고 이 경우 자릿수는 디폴트 설정대로 잡힙니다. FORMAT 키워드를 생략한 경우 출력 변수의 데이터 타입을 따라 자동 설정됩니다.
- /SILENT : 화면(콘솔)에 보고 내용을 출력하지 않습니다.
- STARTLINE : 배열의 몇 번째부터 출력할지 설정할 수 있습니다. 물론 디폴트는 0번부터(STARTLINE=1)입니다.
- SUBSET : 출력 대상이 되는 배열의 인덱스를 지정할 수 있습니다. 예를 들어, V 등급이 20보다 밝은(작은) 별만 출력하고자 할 경우에는, ok=where(v lt 20) 으로 해당 인덱스를 찾아, SUBSET=ok 와 같이 키워드를 설정합니다. 유용하게 씁니다.