1. **Các trường thông tin cần extract từ file text:**

* **12 biến** trong 12 cột theo bảng. Trong đó quan trọng nhất là :

Name of issuer

Title of Class

CUSIP Number

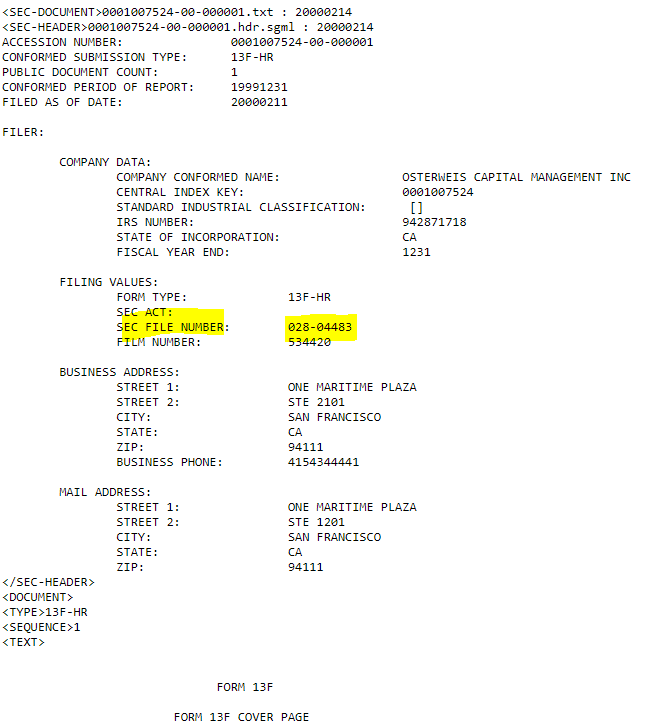
Value(x1000)

Shares/Principl amt

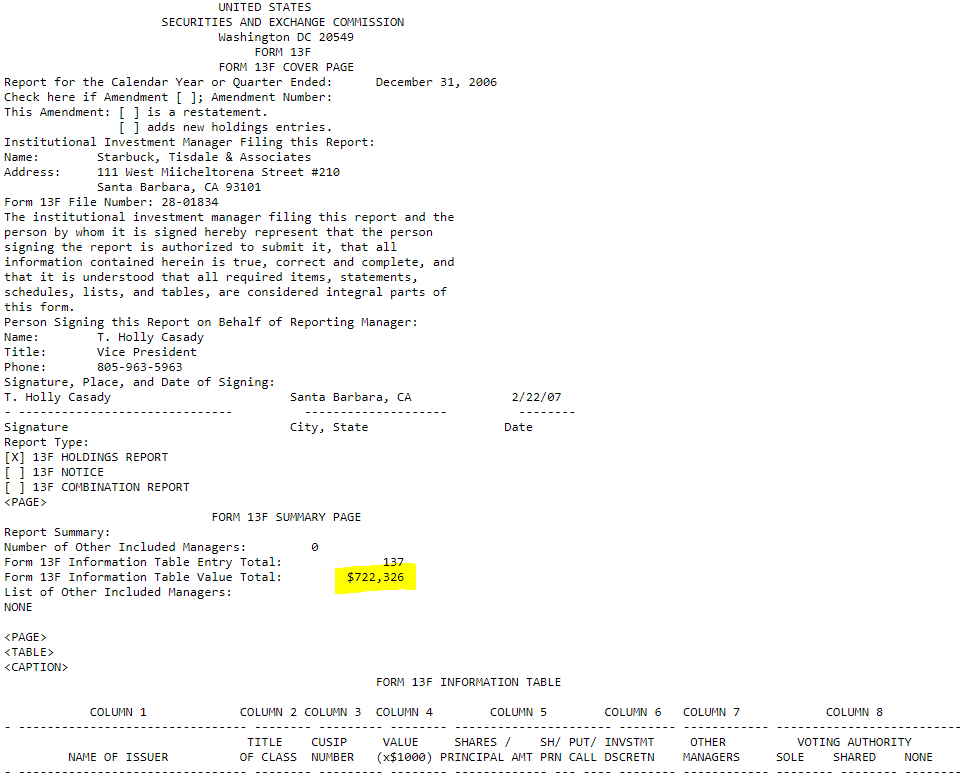
SH/PRN (đơn vị của biến trên)

PUT/ CALL

* Biến **Form13F\_File\_Number**: ở vị trí sau:



* **Tổng giá trị** (Form 13F Information Table Value Total): có thể tìm thấy ở dòng sau  
  Form 13F information Table Value Total: $xxx,xxx . Thực ra giá trị này là theo đơn vị thousands (x1000)



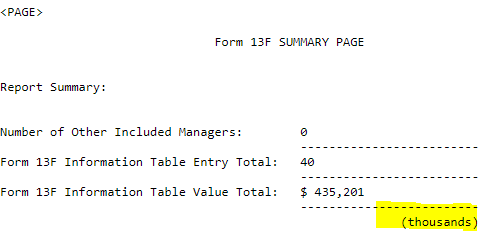
* 1 số biến thể: Form 13F Information Table Value Total (x$1000): $xxx,xxx.
* 1 biến thể khác:



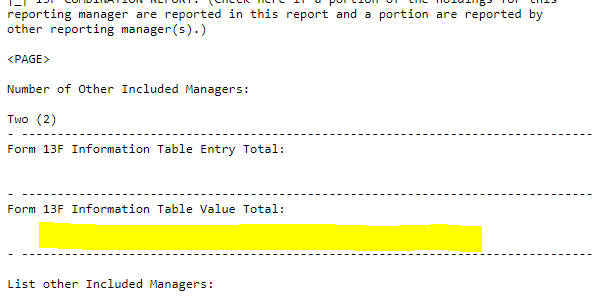
Hoặc cùng 1 dòng:

Form 13F Information Table Value Total: $ 2,694,264 (in thousands)

Hay lại như này:

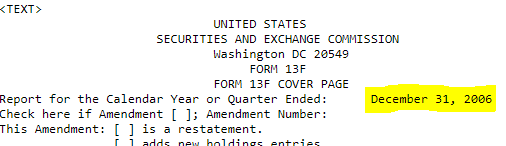


Hoặc thậm chí là bị **để trống:**

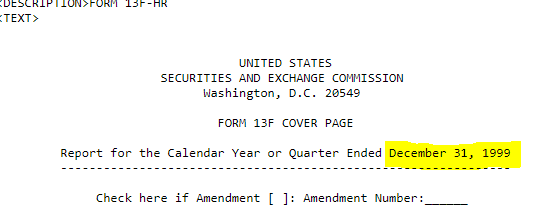


🡺 **1 cách khác để lấy Total Value cho mỗi 1 file đó là** cộng tổng của cột Value(x1000) rồi lưu vào thành biến số 13: Total\_Value, trước khi add dữ liệu của file txt đó vào database.

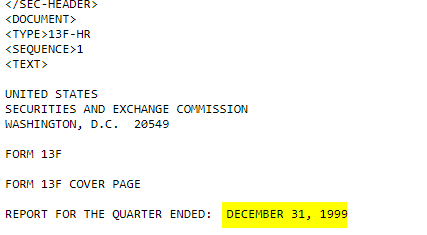
* 1 biến quan trọng khác đó là ngày file: Filing date: thường ở ngay dưới dòng “FORM 13F COVER PAGE”: sẽ có 1 dòng ghi dạng: “Report for the Calendar Year or Quarter Ended: Date



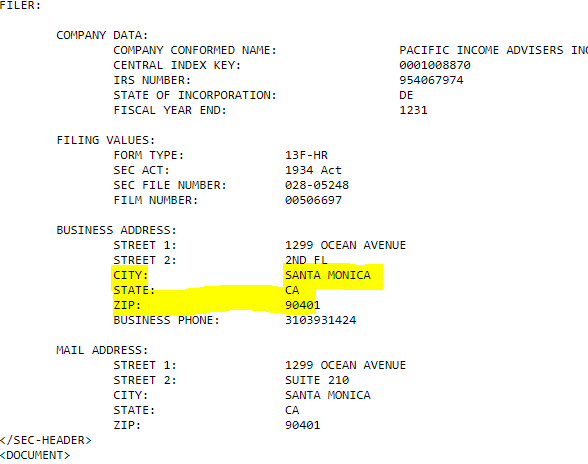
Hoặc có thể không có “:” sau chuỗi “quarter ended”:



Hoặc chuỗi chỉ là “Report for the quarter ended”



* 3 biến nữa lấy ở đầu file, đó là biến CITY, ZIP, STATE của phần BUSINESS ADDRESS (3 biến này nếu lấy đc thì tốt không thì thôi):



2. **Các dạng file text có thể gặp:**

1. FORMAT chuẩn:
   * Có 12 cột data
   * Cột 1 (tên) bắt đầu bằng <S> ở dòng đầu, 11 cột kia có <C>. Tuy nhiên có thể có biến thể: <S> và <C> ở phía trên header hoặc dưới header:
   * Toàn bộ bảng nằm gọn trong 1 cặp tag <TABLE> </TABLE>
   * Toàn bộ bảng nằm gọn trong 1 <PAGE> </PAGE>
   * Tên bảng “FORM 13F INFORMATION TABLE” Có thể nằm bên trong tag <TABLE> hoặc nằm ngoài, phía trên.
   * Có thể có hoặc ko có tag <CAPTION> bên trong <TABLE>
   * Có thể có hoặc ko có dòng COLUMN 1 COLUMN2
   * Dòng header (tên biến) có thể nằm gọn 1 dòng hoặc 2 dòng. Nhưng thứ tự các biến từ trái sang phải thì luôn giống nhau. Cho nên có thể nhận dạng header và bỏ qua ko đọc header.

