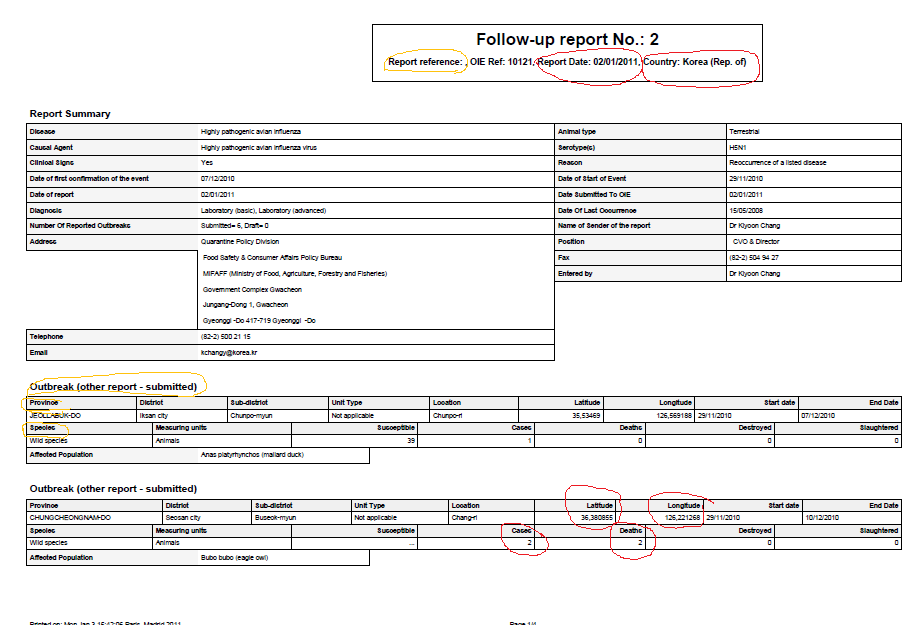
# Parsing H5N1 Outbreak Data

# Using Python And Pdf Text Converter.

The source data to be extracted for visualization is contained with in PDF reports hosted on the website of the World Organization for Animal Health <http://www.oie.int/animal-health-in-the-world/update-on-avian-influenza/2011/> . The reports are a compilation of submitted outbreak reports.

Example:



The task is to extract values for the following fields:

Report Date, Country, Latitude, Longitude, Cases and Deaths.

## 3 phases to getting the data.

1. We used Beautiful Soup API to crawl on each of the page of the website. Scraping out the link for each PDF publication then down loading the file to local disc.
2. The Highly complex internal format of PDF is to formidable directly code for. So we chose to find a tool that would convert the PDF into a more digestible format. After evaluating many tools that converted to different types such as HTML, XML, excel, plain text we finally found the tool *intraPDF* “PDF to TEXT converter “ <http://www.intrapdf.com/convert_pdf_to_text.htm>.

The tool converted the PDF to text preserving spacing relations give us a format amenable to parsing . As of now the tool is run manually on the pdf files, no documentation so far could be found on how to run from a command line, where we could automate this from python.

1. Parsing is done in Python, we loop on file line reads looking for keys word such as the table header s { OutBreak, Province, Species}. Once an line of interest is found it is parsed with a RegEx expression to find the date, numeric value or string value. Once all that data for an Event is collect it is flushed out to TSV file and the next Out Break section is searched for.

Future Work:

The years for 2011 to 2008 have been so far collected and processed. Older repots have a slightly different table formatting so the parser needs minor refactoring to handle those reports.