



Excel 2 XML

Extract data from the most used database in the world and transform it in something useful



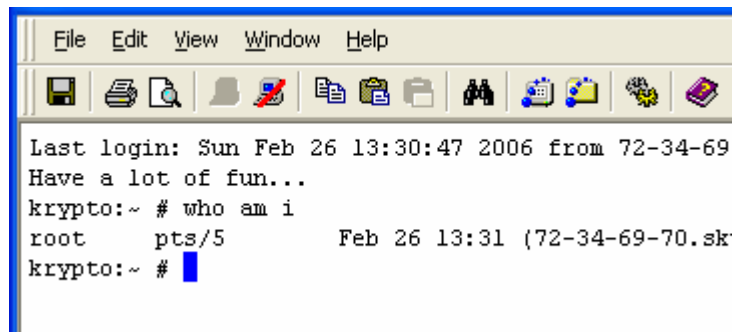
Agenda

- Who?
- Why?
- How?
- What?
- When?
- Resources



Who am I

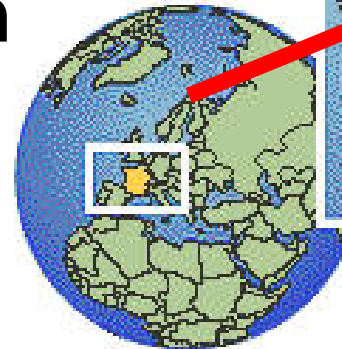
- o My Unix box usually answers...



```
File Edit View Window Help
Last login: Sun Feb 26 13:30:47 2006 from 72-34-69
Have a lot of fun...
krypto:~ # who am i
root pts/5 Feb 26 13:31 (72-34-69-70.sk)
krypto:~ #
```

● ● ● | Ok, no joke now...

- Jean Georges Perrin
- Dev. Tools (xGL, Java EE, PHP)
- Web since 1994
- IIUG BoD since 2002
- Focus on **Software Engineering**
- Proudly lives in



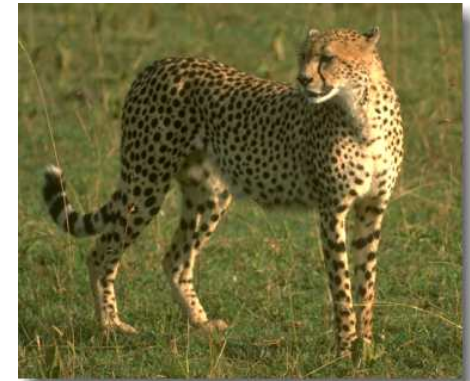


And you?

- Java gurus?
- PHP?
- XML gurus?
- Excel gurus?

Why?

- Need to publish Excel data on the Web
- Need to store Excel data in XML-oriented databases
 - Informix Cheetah, Oracle, DB2 “pureXML”
- Part of an SOA chain





How?

- Go to **File, Save As...**
- Why not?
 - Microsoft centric
 - Microsoft XML dependent
 - No (easy) automation



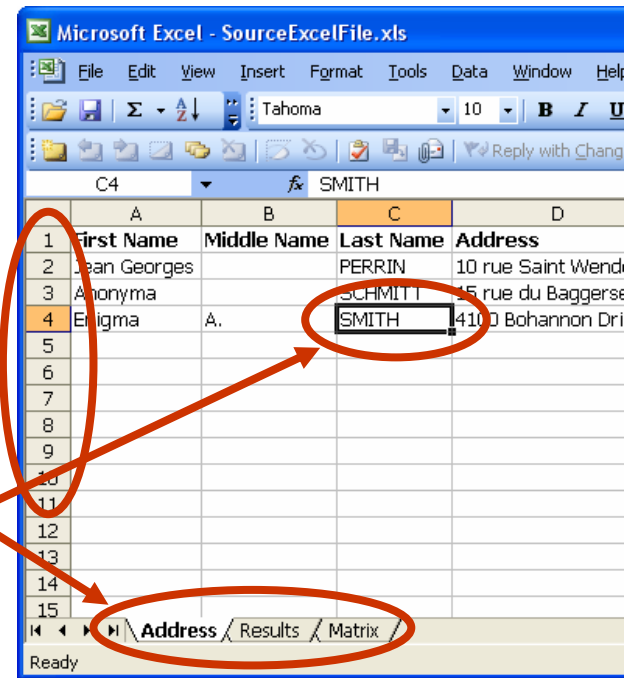
How?

- o Parse Excel files with POI
- o Build XML document with Xerces
- o Using Java 5.0, running on **any platform**

Excel Anatomy

o Workbook

- Contains **sheets**
- Contains **rows**
- Contains **cells**



o Similar to something you know?



Database Anatomy

○ Database ↔ ○ Workbook

● Contains **tables** ↔ ● **sheets**

● Contains **rows** ↔ ● **rows**

● Contains **fields** ↔ ● **cells**

○ Excel is the most popular database and nobody knows it!

Understanding the “table”

- o 3 type of tables in Excel
 - Vertical
 - Horizontal
 - Raw

Microsoft Excel - SourceExcelFile.xls

1	First Name	Middle Name	Last Name	Address	Zip
2	Jean Georges		PERRIN	10 rue Saint Wendelin	
3	Anonyma		SCHMITT	15 rue du Baggersee	
4	Enigma	A.	SMITH	4100 Bohannon Drive	94
5					
6					
7					
8					
9					
10					
11					
12					

Microsoft Excel - SourceExcelFile.xls

1	First name	Jean Georges	Anonyma	Enigma
2	Last name	PERRIN	SCHMITT	SMITH
3	Math Q1	70.00	34.00	78.00
4	Math Q2	80.00	23.00	12.00
5	Math Q3	92.46	64.79	23.00
6	Math Q4	93.27	89.00	34.46
7	Math Y	83.93	52.70	36.86
8	French Q1	60.00	76.00	56.00
9	French Q2	64.00	75.00	23.78
10	French Q3	62.00	74.24	23.00
11	French Q4	64.00	77.00	23.00
12	French Y	62.50	75.56	31.45

Microsoft Excel - SourceExcelFile.xls

1	A	B	C	D	E	F
2	Ni	Nickel	28	58.71	1453	2732
3	H	Hydrogen	1	1.0079	-259.14	-252.87
4	Rb	Rubidium	37	85.467	38.89	688
5						
6						
7						
8						
9						
10						
11						
12						



XML

- o a `<database/>`
 - contains one or more `<table/>`
 - contains one or more `<row/>`
 - contains one or more `<field/>`

```
<?xml version="1.0" encoding="ISO-8859-1"?>
<database>
  <table name="Address" orientation="vertical">
    <row id="1">
      <field dataType="text" name="First Name">Jean Georges</field>
      <field dataType="unknown" name="Middle Name"></field>
      <field dataType="text" name="Last Name">PERRIN</field>
      <field dataType="text" name="City">Haguenau</field>
      <field dataType="unknown" name="State"></field>
      <field dataType="text" name="Country">France</field>
      <field dataType="date" name="Birth Date">Tue Oct 05 00:00:00 CET 1971</field>
      <field dataType="text" name="e-mail">jgp@jgp.net</field>
    </row>
    <row id="2">
      <field dataType="text">Anonyma</field>

```

...



What?

Have a look at what it looks like...



When?

- A bit of code to write
 - Vertical tables
 - Raw data
 - Consistent field type for the column
- A few challenges to solve
 - Formulas
 - Storage of meta data
- Better analysis and factoring
 - Better understanding of Excel
 - Better code design
- Test
- Documentation
- Any volunteers?



Resources

- JGP_{net}

- Source code & project
- <http://www.jgp.net>

-  Apache POI



- Part of the Jakarta Apache project
- <http://jakarta.apache.org/poi>

- Xerces

- Apache project
- <http://xerces.apache.org>

- Eclipse, JDeveloper, NetBeans...



Thanks

Feel free:

jgp@jgp.net

jgp@iiug.org