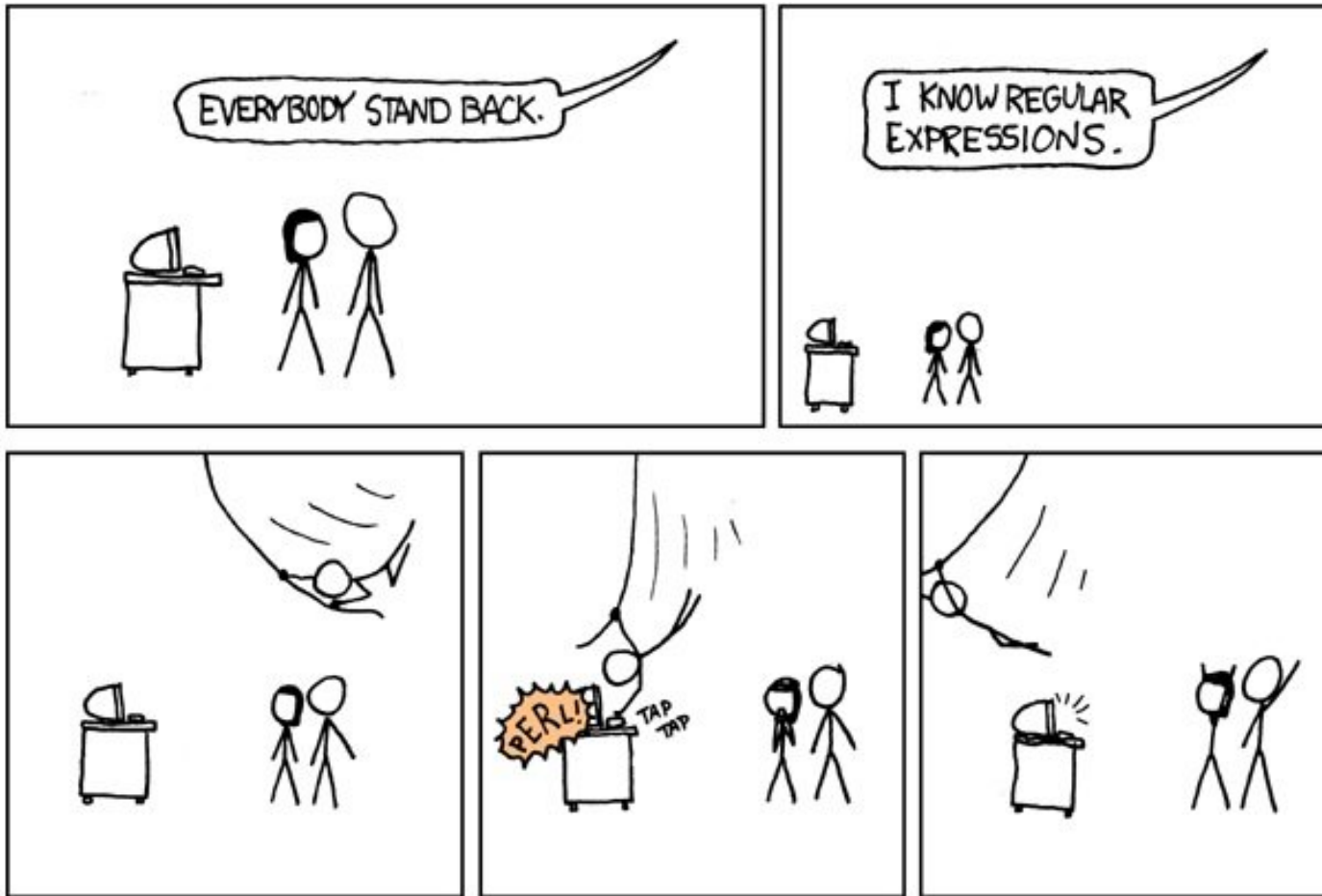# EAD Automation with Regular Expressions

Megan Mummey

Collections Management Archivist

University of Kentucky Special Collections Research Center

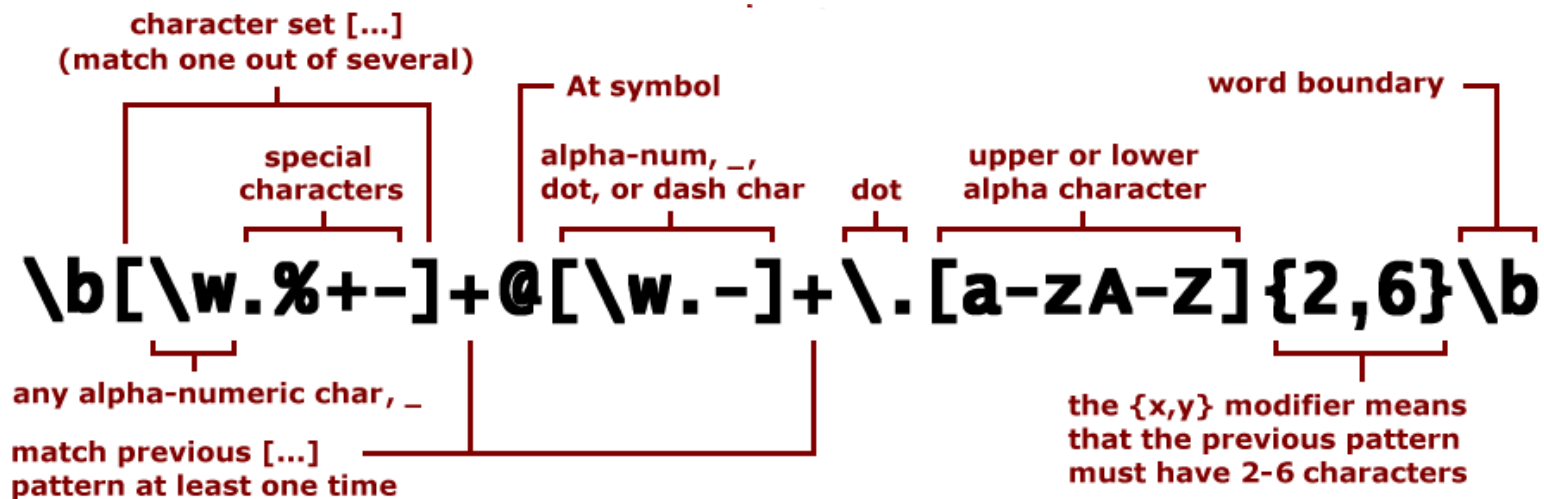11/7/2014

Regular expressions

What are they?

- Sequence of characters forming a search pattern
- Used for find and replace operations
- Option in most computing languages and software

What do they look like?

- Made up of metacharacters and regular characters

character set [...]
(match one out of several)

special
characters

At symbol

alpha-num, _,
dot, or dash char   dot

upper or lower
alpha character

word boundary

\b[\w.%+-]+@[\w.-]+\.[a-zA-Z]{2,6}\b

any alpha-numeric char, _

match previous [...]
pattern at least one time

the {x,y} modifier means
that the previous pattern
must have 2-6 characters

Parse: username@domain.TLD (top level domain)

Metacharacters

- . = any single character
- * = matches the preceding element 0 or more times
- ( ) = defines a subexpression, the string in the parentheses can be recalled later
- \n = (where n=digit from 1-9) matches the marked subexpression matched

What I use

(.*) = ignore all

```
<c02>
    <container >(.*)</container>
    <container >(.*)</container>
    <unitdate>(.*)</unitdate>
</c02>
```
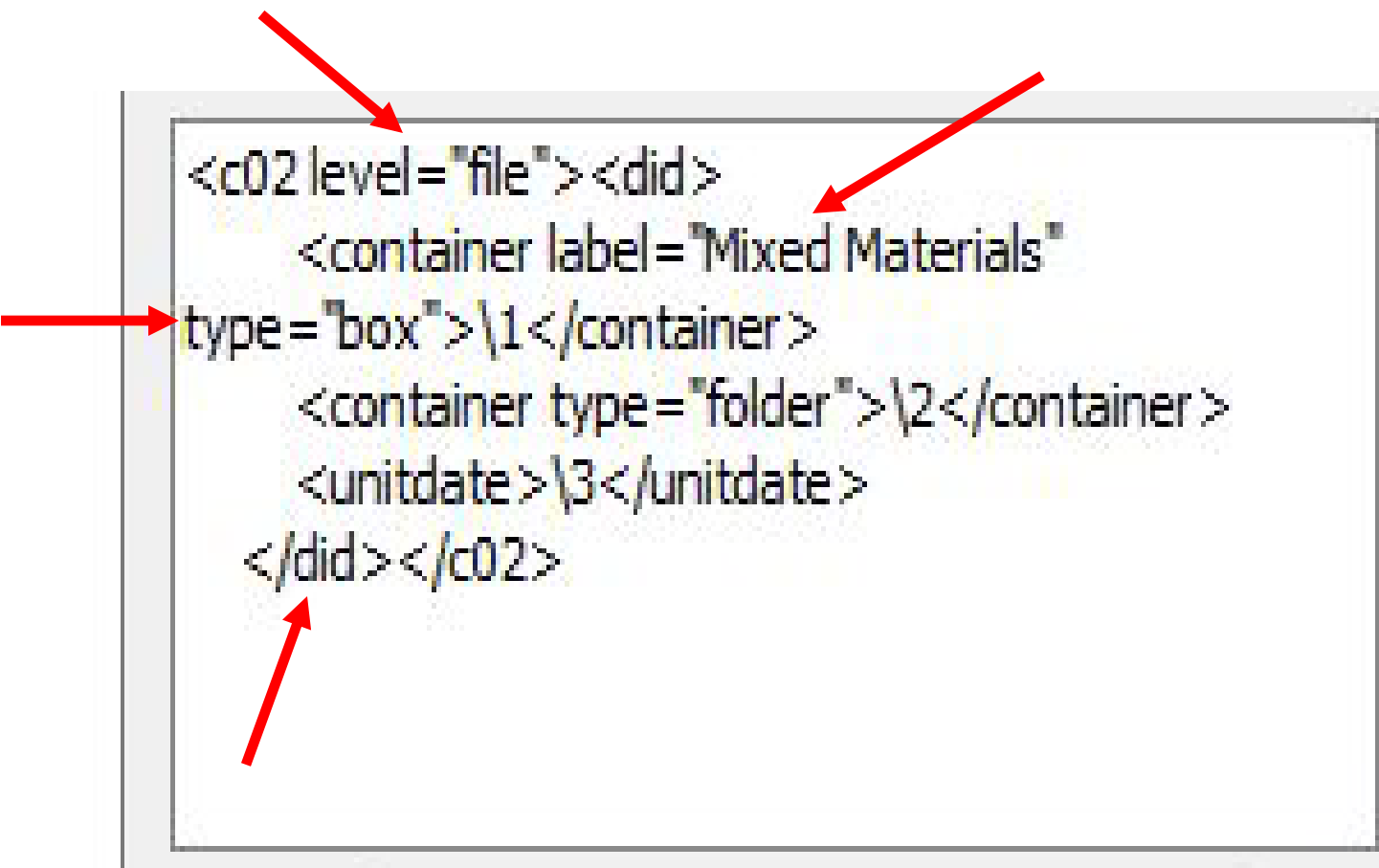
Replace with:

```
<c02>
    <container>\1</container>
    <container>\2</container>
    <unitdate>\3</unitdate>
</c02>
```

```
<c02 level="file"><did>
     <container label="Mixed Materials"
type="box">\1</container>
     <container type="folder">\2</container>
     <unitdate>\3</unitdate>
</did></c02>
```

How I use them

- Globally edit EAD

- Rearrange tags in EAD

- Import flat finding aids from Excel

Questions?

Exercises

- Exercise 1 = rearranging tags
- Exercise 2 = changing tag content
- Exercise 3 = importing flat finding aids from excel

Exercise 1

- **Scenario** – Finding aid encoded with <unitdate> outside of <unittitle>. Move all of the <unitdate> tags inside <unittitle>.
- Use xml file **exercise 1**.

Exercise 2

- **Scenario** – Finding aid that starts over the box numbers from 1 in the photograph series. Due to the way Archivist's Toolkit works this makes it impossible to assign locations. We need to go in and add a "P" before each box number.
- Use xml files **Exercise 2** and **Exercise 2 photo series**.

Exercise 3

- **Scenario** – Legacy finding aid with a flat structure in excel. Need to encode but several thousand entries.

- Use excel file **exercise 3** or if using Notetab use xml file **exercise 3.**

Regular expressions tutorials

- http://www.regular-expressions.info/quickstart.html
- http://www.zytrax.com/tech/web/regex.htm
- http://regexone.com/ (simple interactive tutorial)