TrapperKeeper:
The Case for Using Virtualization to Add Type Awareness to File Systems
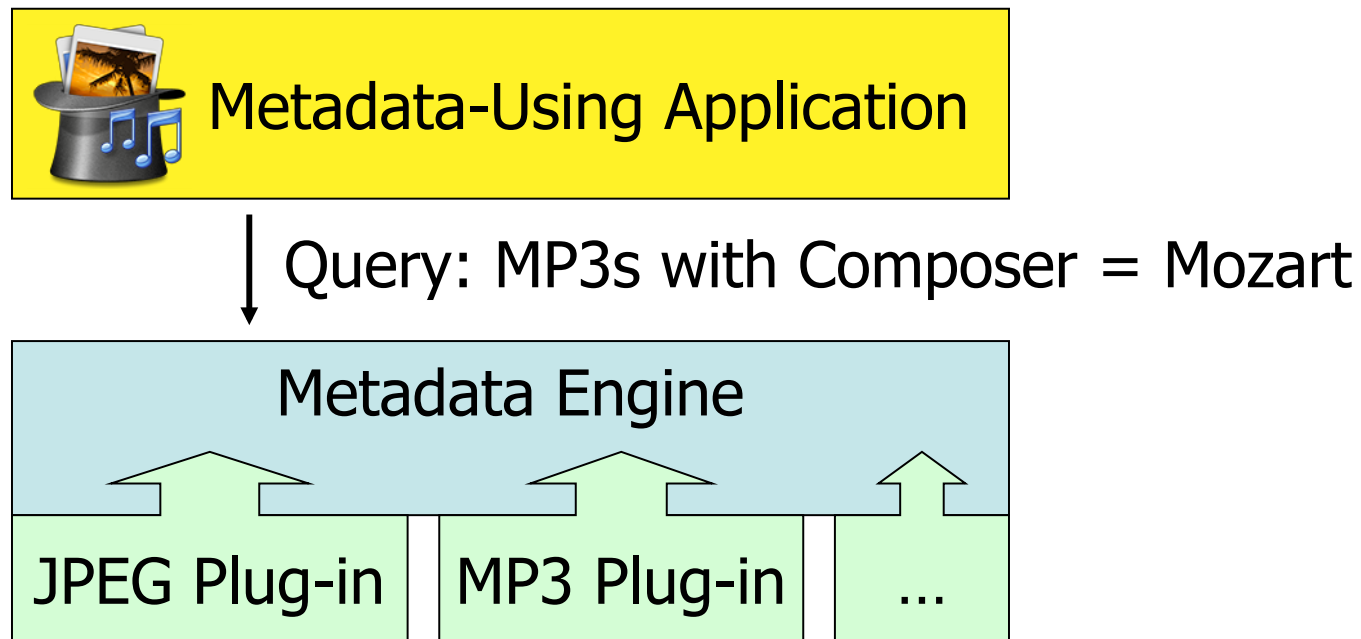
Daniel Peek

Facebook

Jason Flinn

University of Michgan

Trapper Keeper

- Need access to type-specific metadata
  - Searching, Organization, Presentation

- Extracting metadata is hard
  - Lots of file types out there
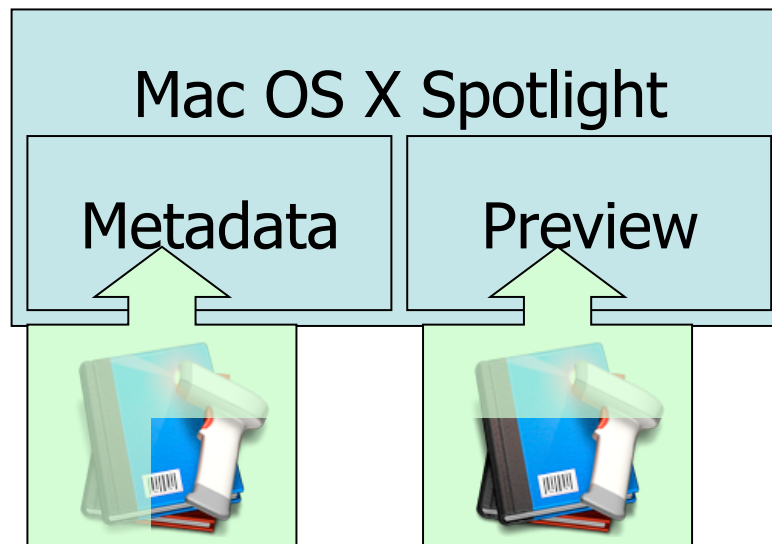 - Custom code required for each type

- A better way to get metadata

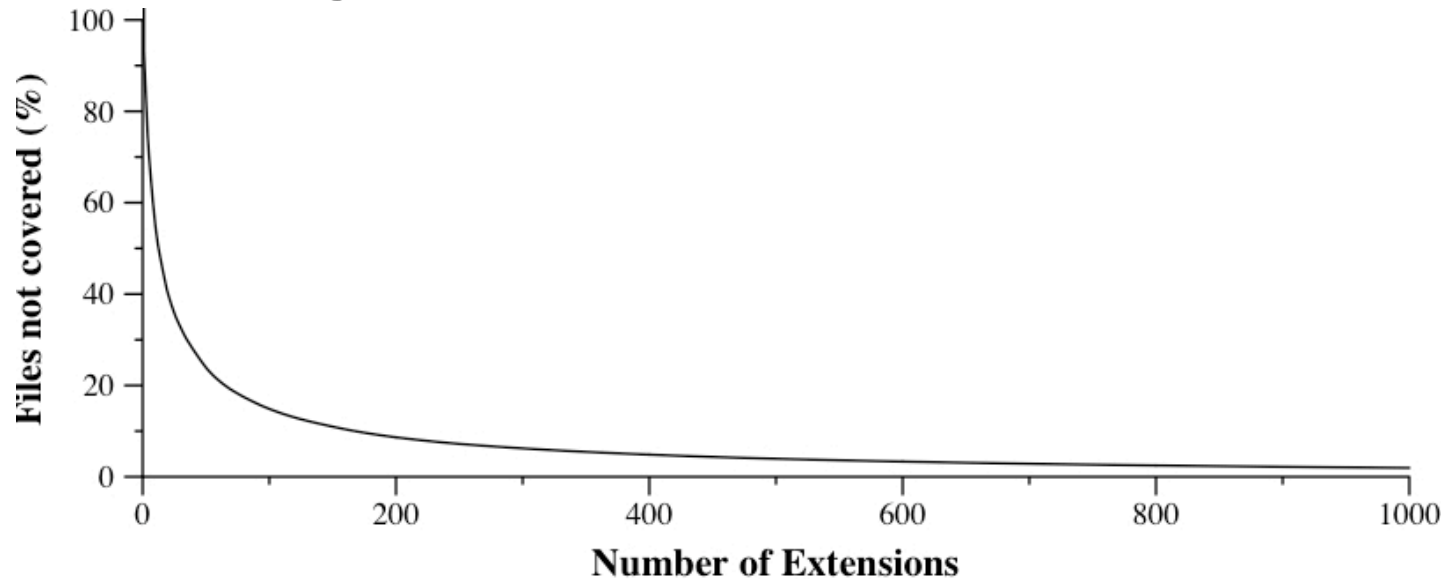# The Plug-in Solution

- Developers make plug-ins for each type

The Plug-in Solution

- Lots of work for developers

Mac OS X Spotlight

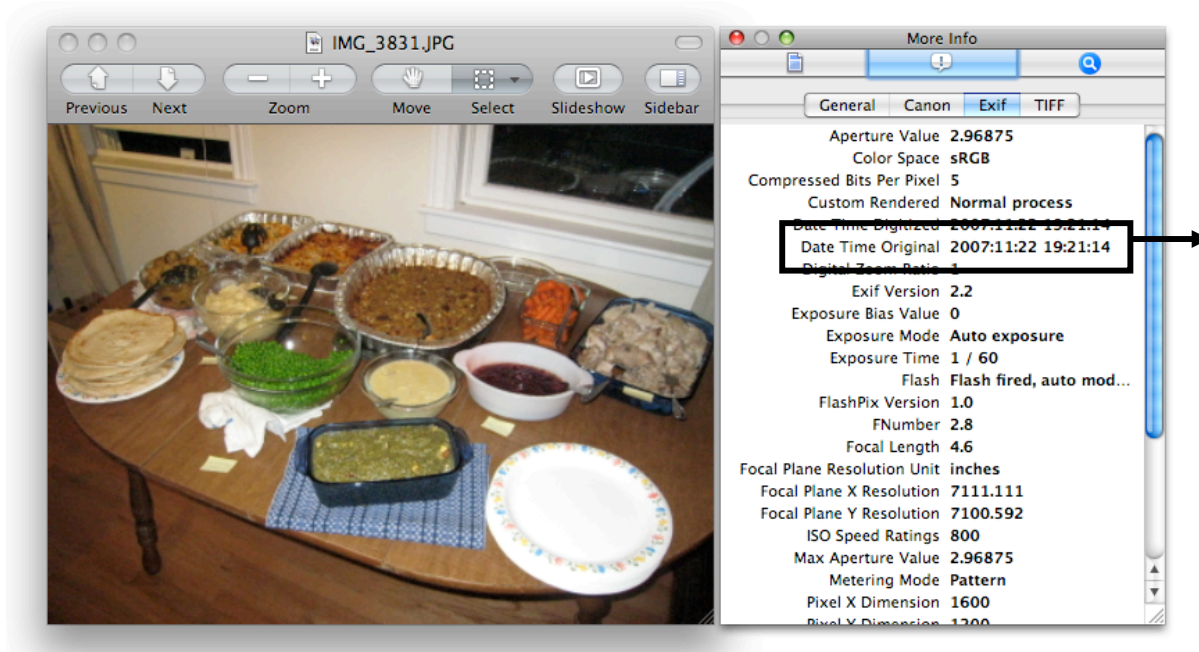Metadata   Preview

The Long Tail

- How big is this problem?

- Big (Agrawal et al.)
- Uneconomical to support all types
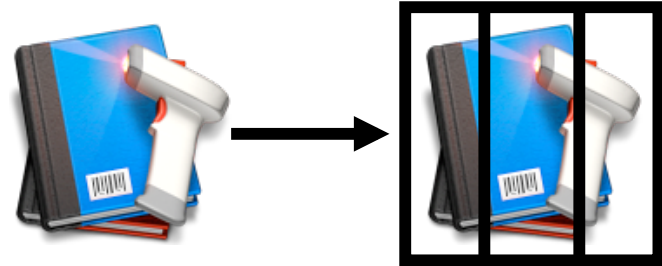
The TrapperKeeper Solution

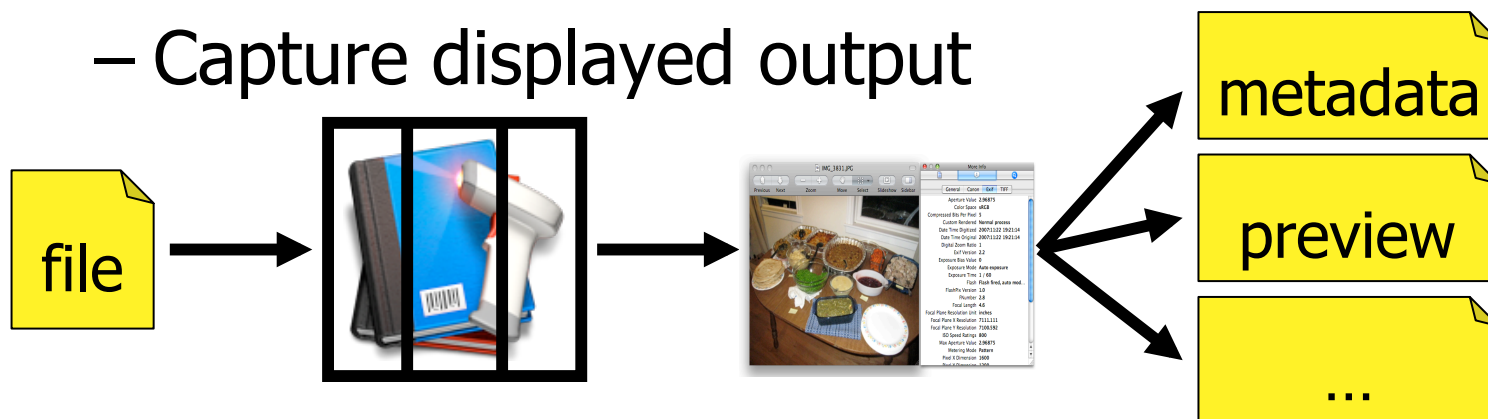- Already have apps that parse these files
  - Apps expose information through GUI

Date Time Original = 2007:11:22 19:21:14

The TrapperKeeper Process
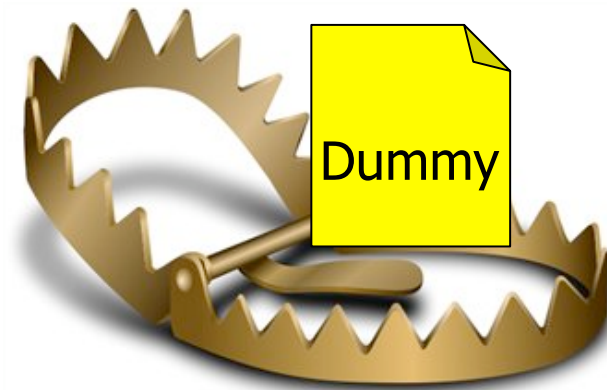
- ## Once Per Application
  - – Trap the application

- ## Once Per File
  - – Use trapped application to parse the file
  - – Capture displayed output
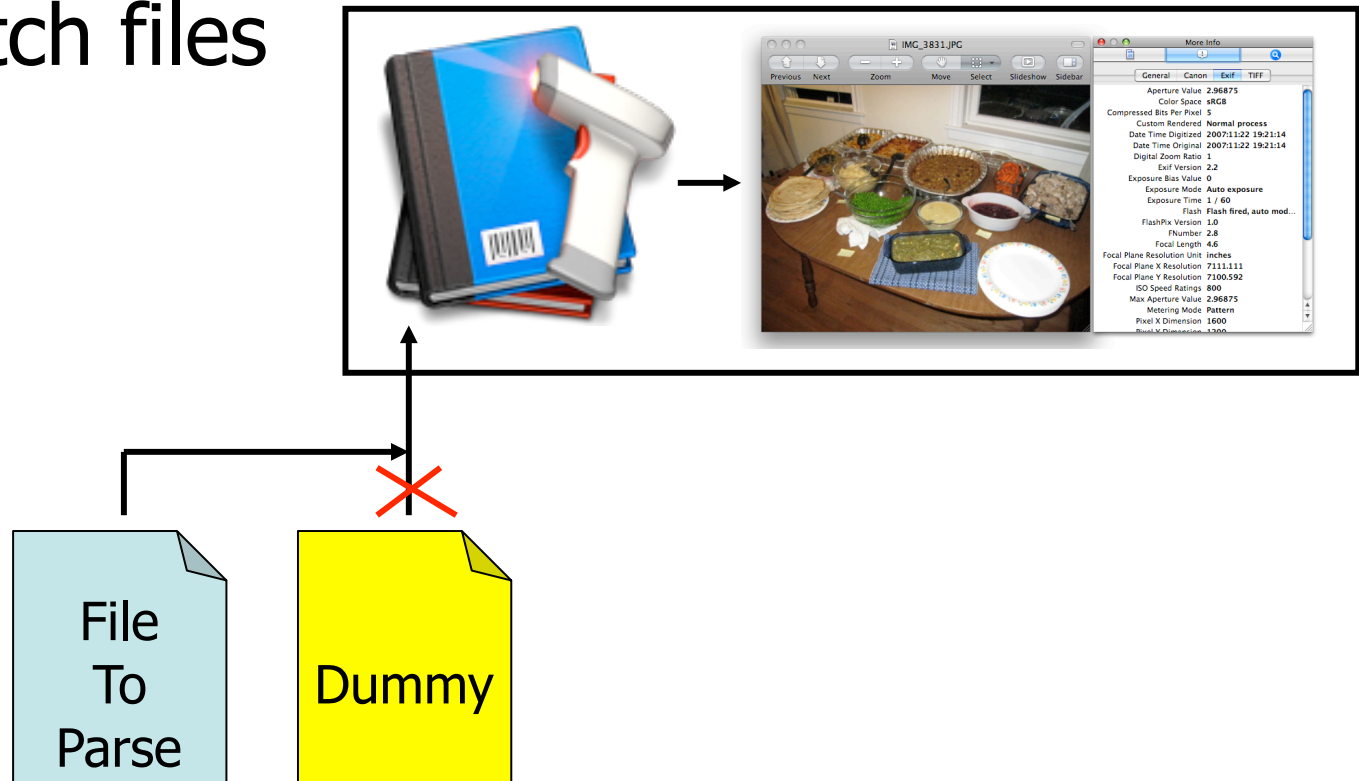
file → metadata, preview, ...

Trapping Applications

- Run app inside a VM
  - Contains app effects
- Make app open a dummy file
- Snapshot at moment of open()
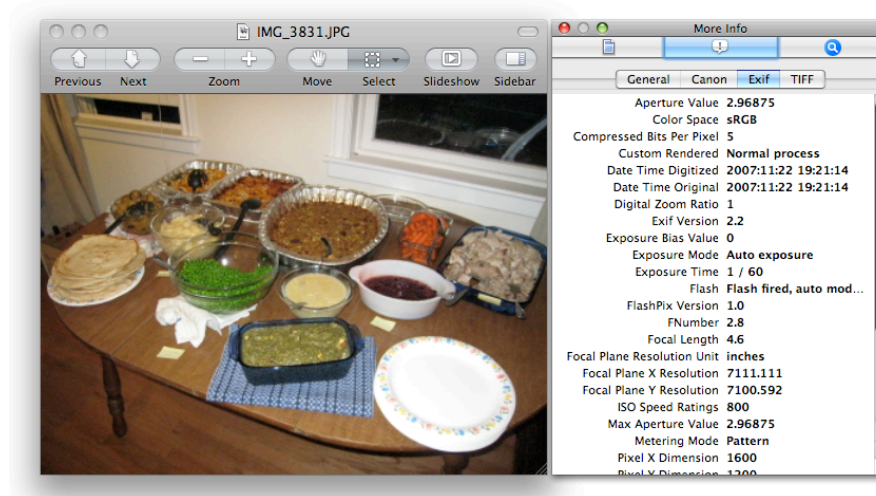 - About to execute file parsing behavior

Parsing with Trapped Apps

- Restart VM
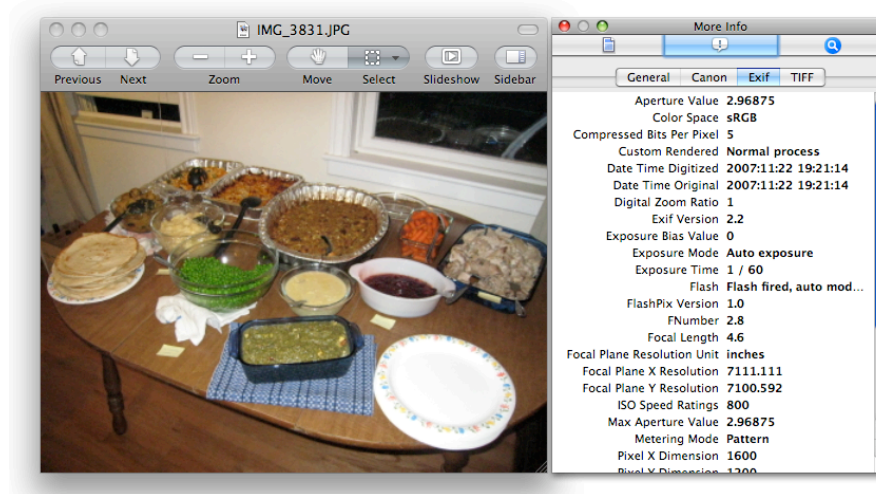- Switch files

Accessibility

Window
└── Tab Pane
    ├── ...
    ├── Text Label: "Date Time Original"
    │   Text: "2007:11:22 19:21:14"
    └── ...

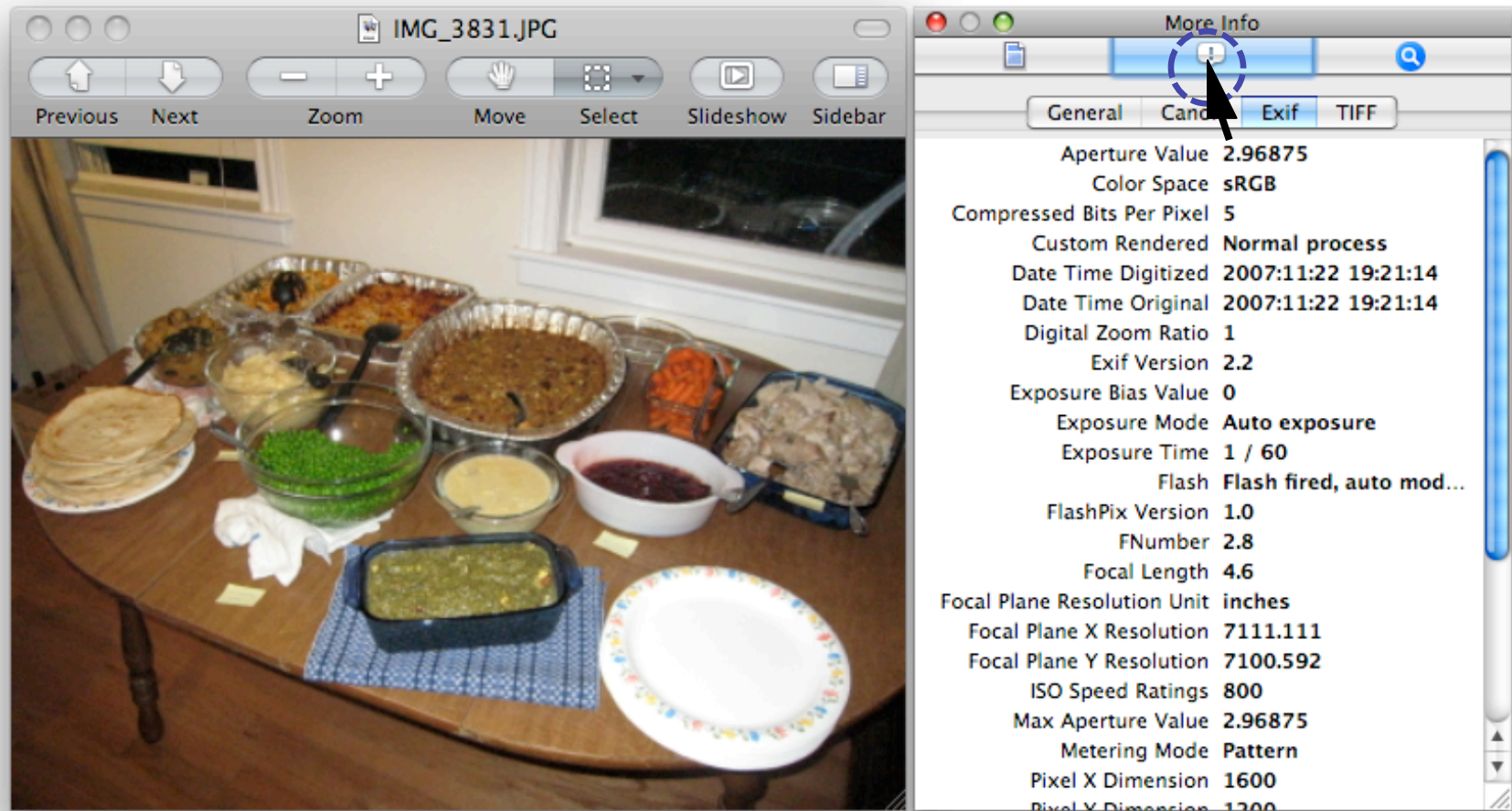# Accessibility

Window
└─ Tab Pane
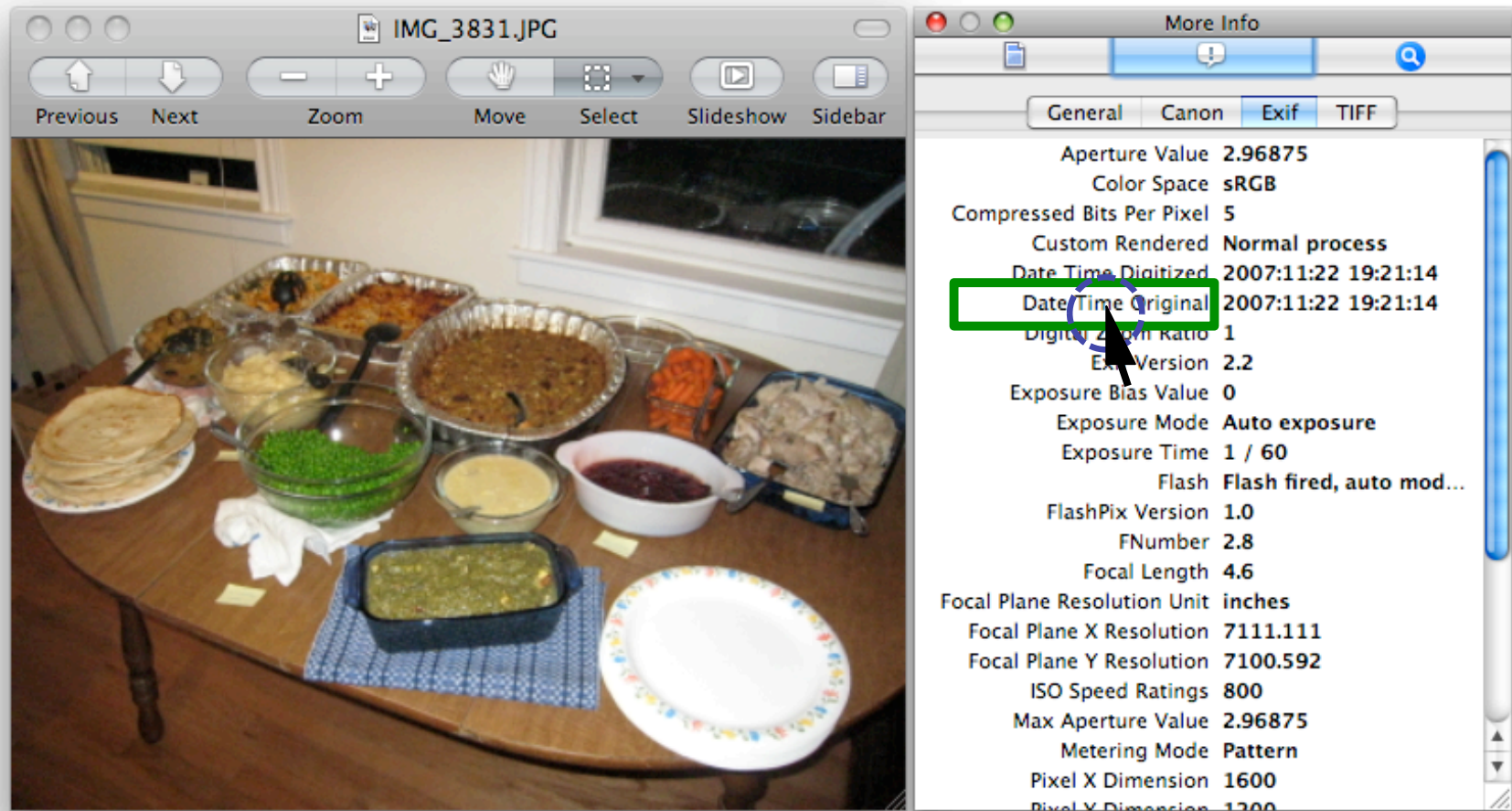   ├─ ...
   ├─ Text Label: "Date Time Original"
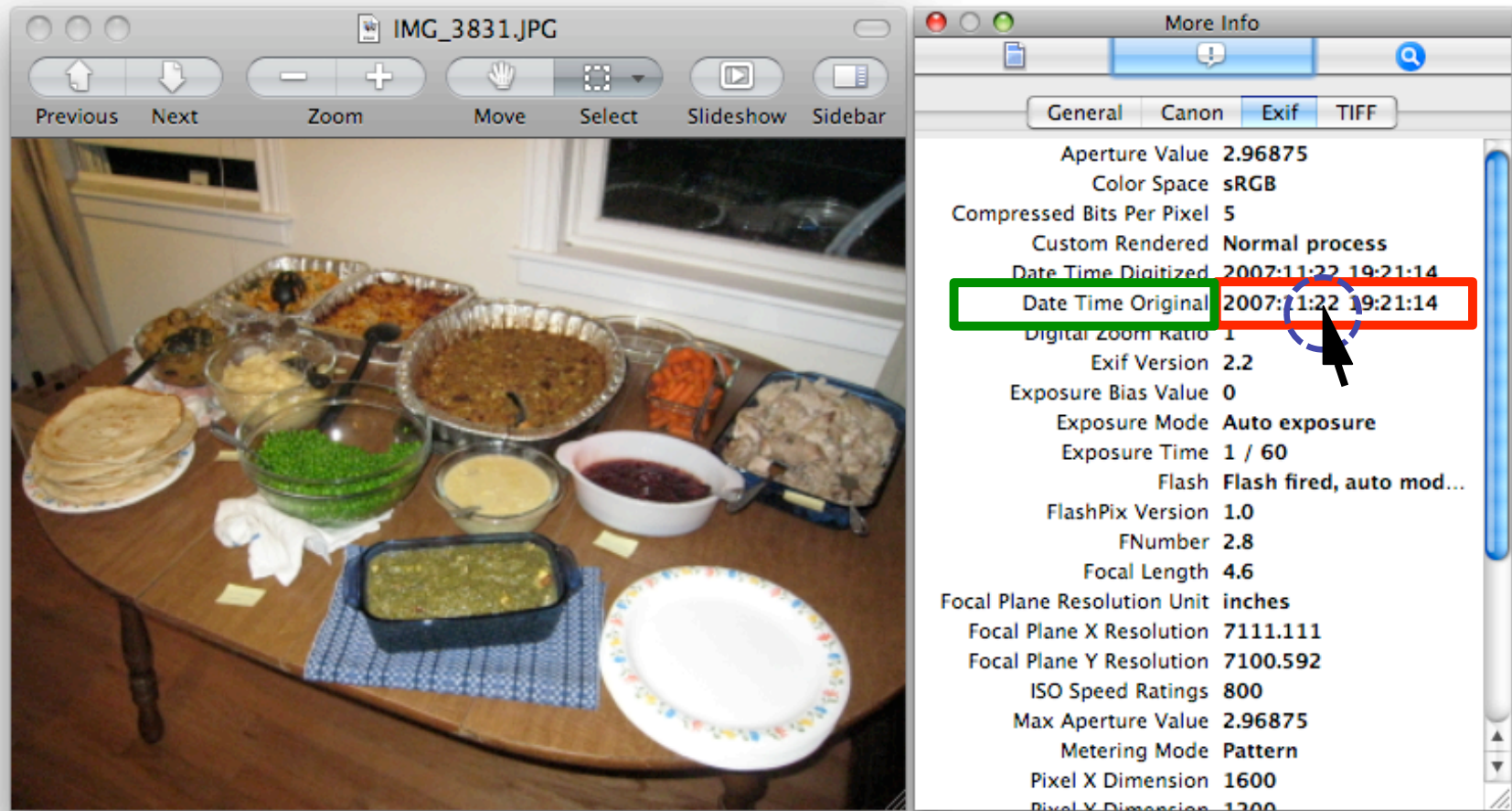   ├─ Text: "2007:11:22 19:21:14"
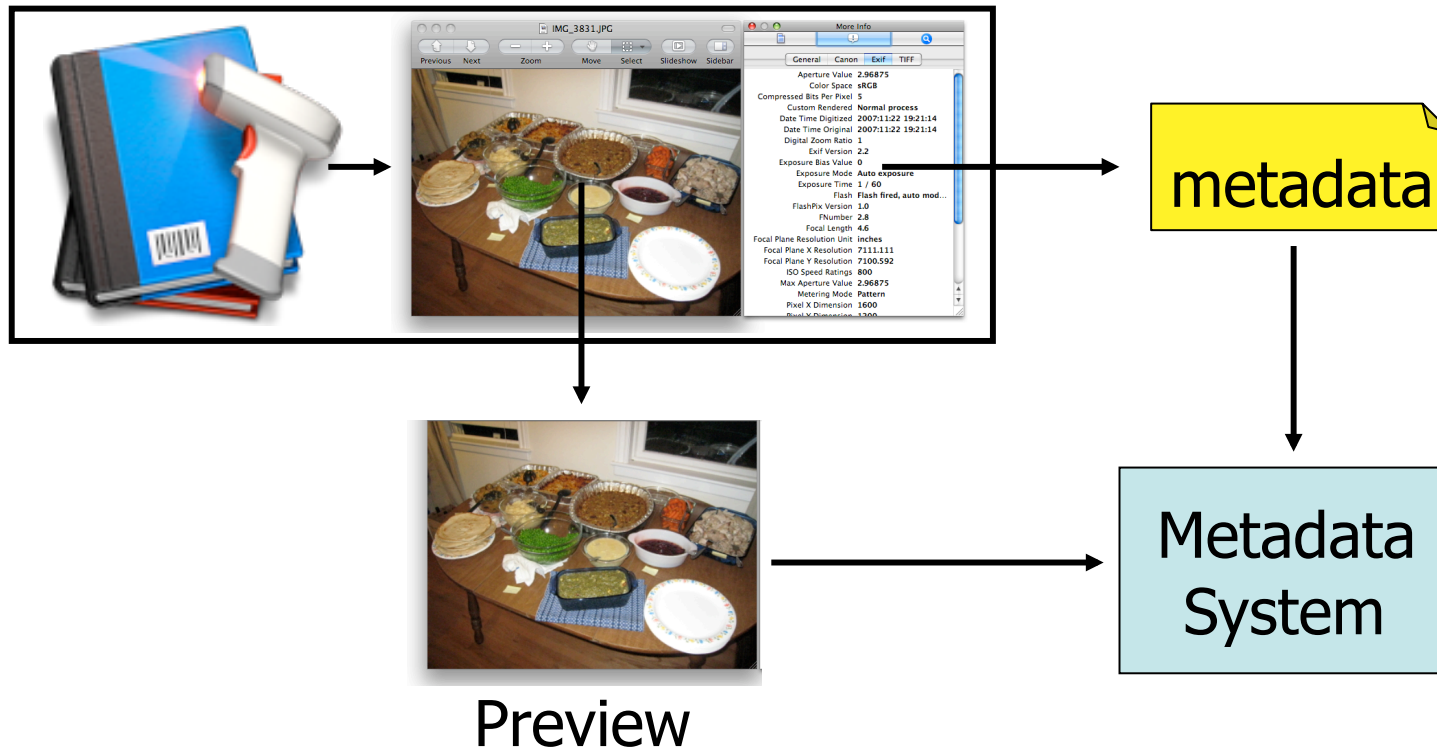   └─ ...

Guided Extraction

Guided Extraction

Guided Extraction

Execute Features

- Snapshot window in VM

Preview

metadata

Metadata System

TrapperKeeper Results

- Makes it easy to extract metadata
  - No development skill
  - No source code
  - Just be able to use the application
- Successful use
  - All GUI apps in Ubuntu 7.10 in a day
  - Parses over 100 file types
  - Rate of 318 files/hour

Tricky Situations

- Application has no accessibility support
- Application does not expose metadata
- Application needs external info to parse
  - Configuration files
  - License servers
  - Internet connections

Tricky Situations

- Performance: a sudden influx of files
  - Fresh installation
  - Download from digital camera
- Which metadata is the right metadata?

Questions