Born digital: a symposium exploring digital architectural and built environment records

Monday 18 + Tuesday 19 April 2016

Extracting metadata from the RMIT Storey Hall and the Ridgway Building digital corpus

Georg Grossmann

Born digital: a symposium exploring digital architectural and built environment records

18 April 2016

University of
South Australia

Starting Point

- **Ridgway Apartment Building**
- **RMIT Storey Hall**
- Collected **4,312** digital files

Ridgway Building

RMIT Storey Hall

Requirements

- Extract information from folders and create a database with file record information.
- The record should include the following information:
  - File name
  - File location
  - Last Date Modified
  - Author
  - File Type
  - File Size

Challenges

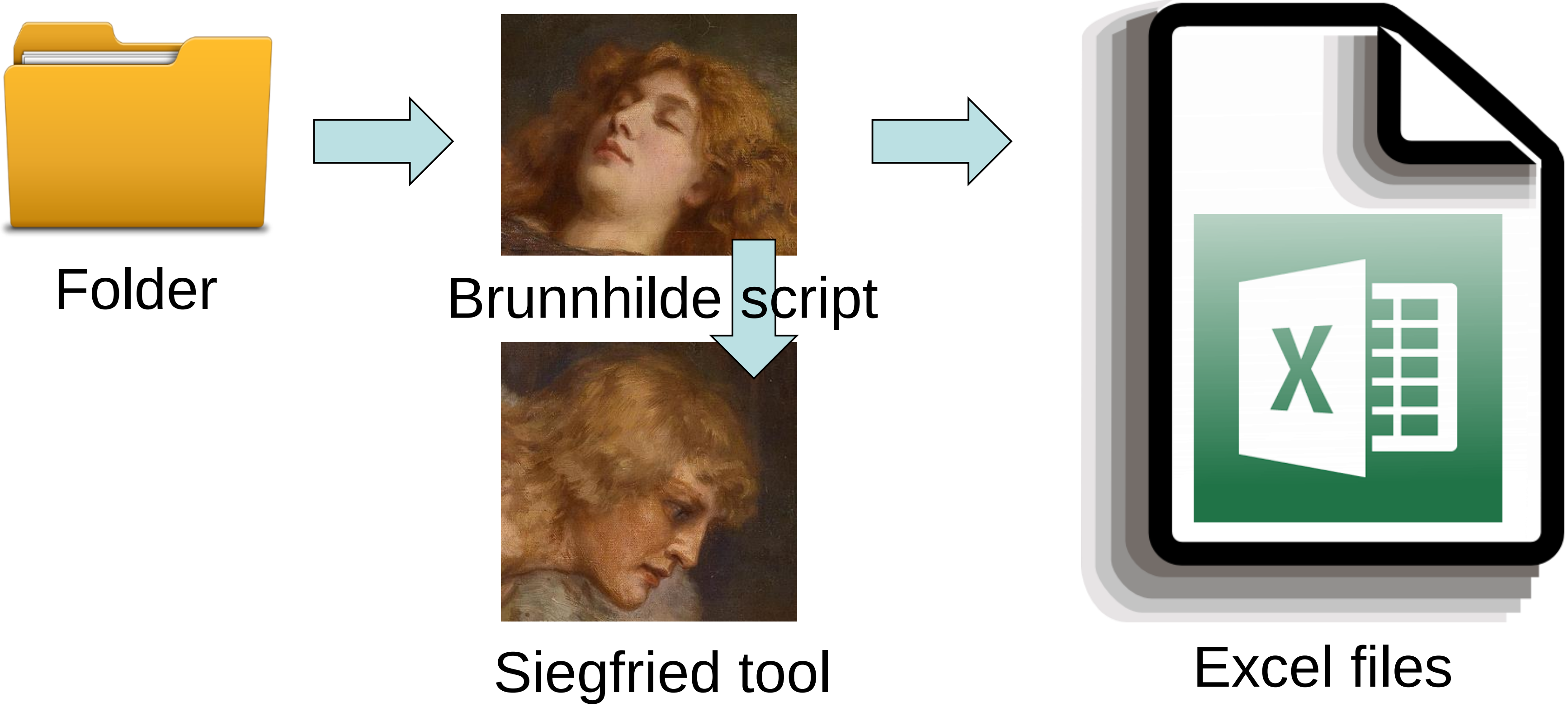
- **Medium:**
  - Complex folder structure
- **Files:**
  - Duplicate files
  - Missing/wrong file types
  - Old version of software (Word, Excel)
  - Different version of same file
- **Metadata:**
  - Identify how often files were copied
  - When printed last?
  - Author?
  - Identify links between files
  - Chronological order of when files were created and edited
- **Content:**
  - Scanned documents
  - Drawings

Tools

- **Existing tools**
  - **+** free, proven technique, efficient
  - **–** costs (commercial), may be difficult to use and embed within other tools
  - Examples: Brunnhilde, Siegfried
- **Write your own**
  - **+** specific to your needs and requirements
  - **–** costs, maintenance

Application of Brunnhilde

- Python script
- Uses **Siegfried** (a signature-based file format identification tool)



Folder

Brunnhilde script

Siegfried tool

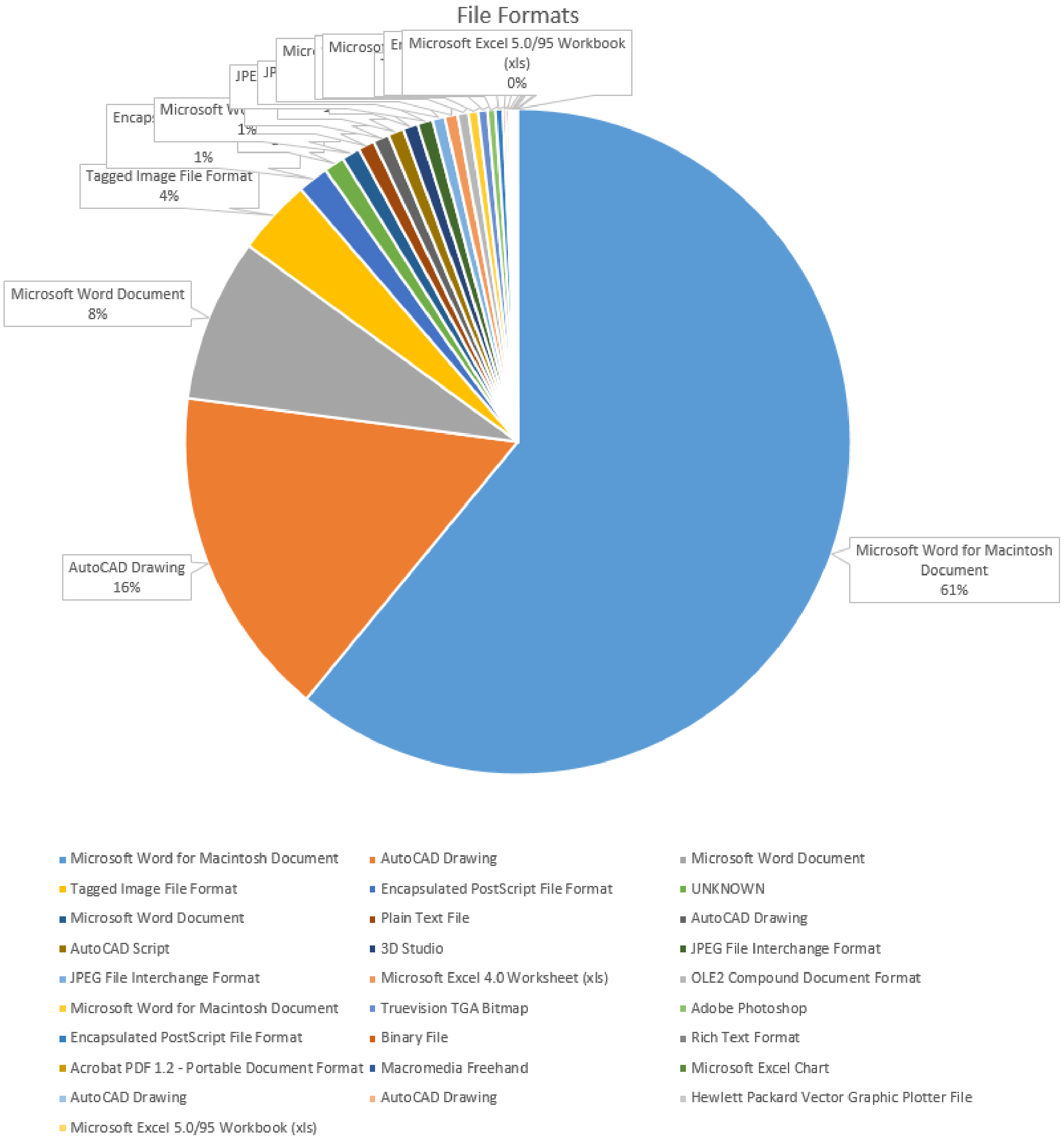
Excel files

Outcomes – RMIT Storey Hall

- **2,611 files**
- **Errors/Warnings:**
  - 4 errors indicating 4 empty files
  - 1,856 warnings with the majority being file types missing (but identified)
  - 411 files are duplicates
- **Many different file formats:**
  - Majority are Word for Macintosh (61%)
  - AutoCAD (16%)
  - Microsoft Word (8%), and
  - Tagged Image File Formats (4%).
- **Dates:** 1980 – 2016

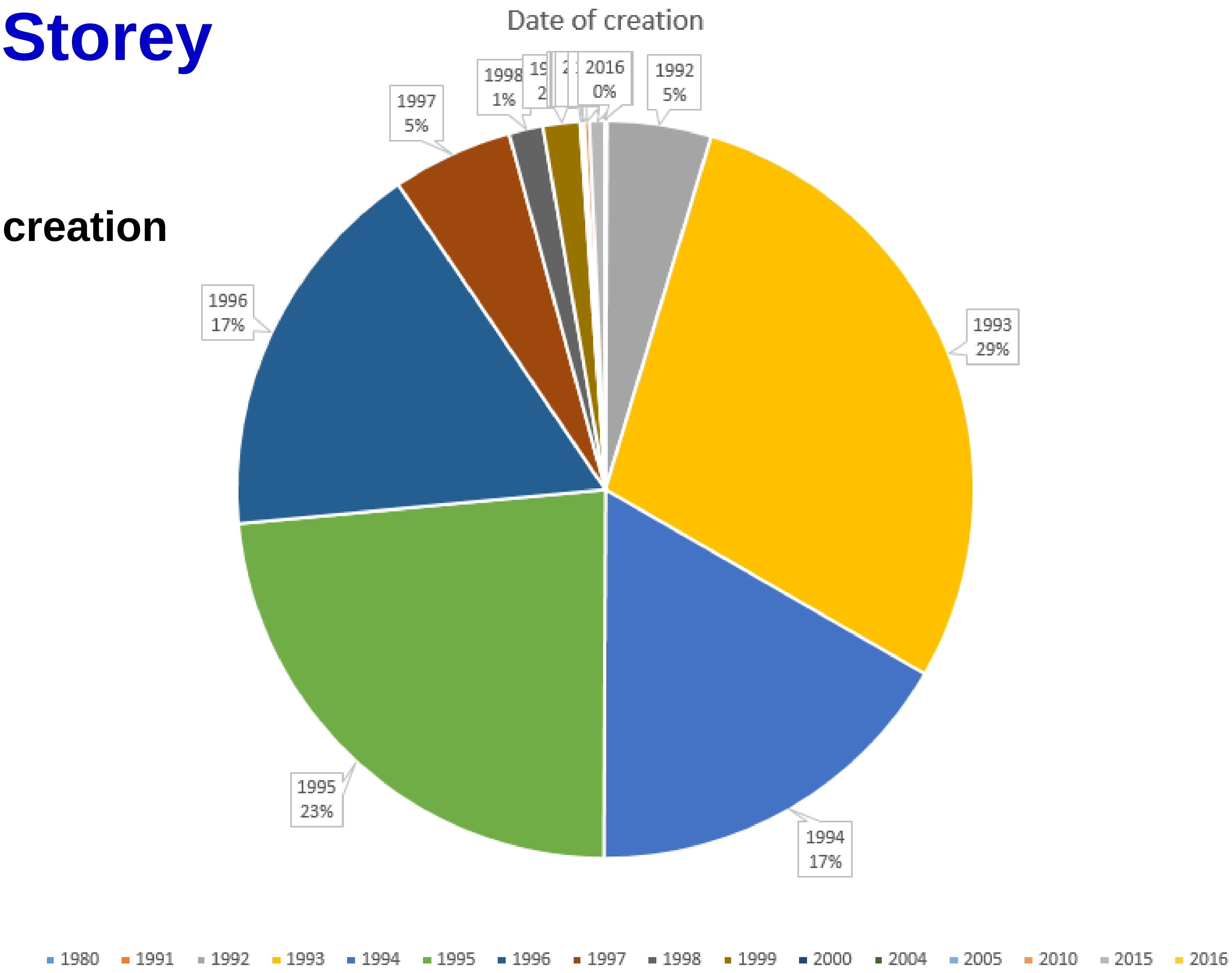
RMIT Storey Hall

- **File formats**



RMIT Storey Hall

- **Date of creation**



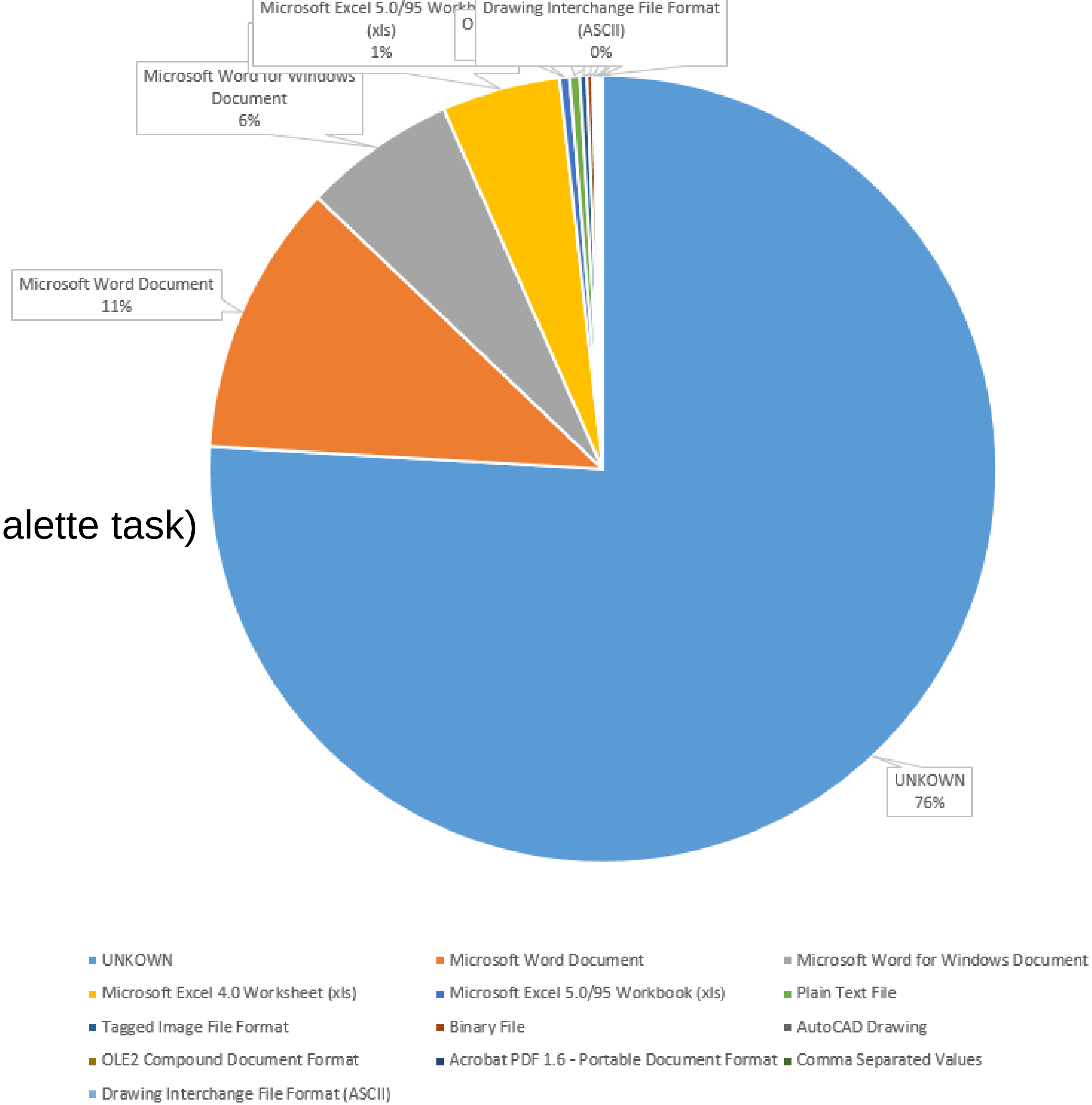
Outcomes – Ridgway Building

- **1,707 files**
- **Errors/Warnings:**
  - There was only one error indicating the file is corrupt
  - 1,311 warnings because of unknown file type
  - Only one duplicate
- **Different file formats:**
  - Majority is unknown (76%),
  - Microsoft Word 6.0/95 (11%),
  - Microsoft Word 2.0 (6%), and
  - Microsoft Excel 4.0 (5%).
- **Dates:** 1992 – 2016

Ridgeway Building

File formats

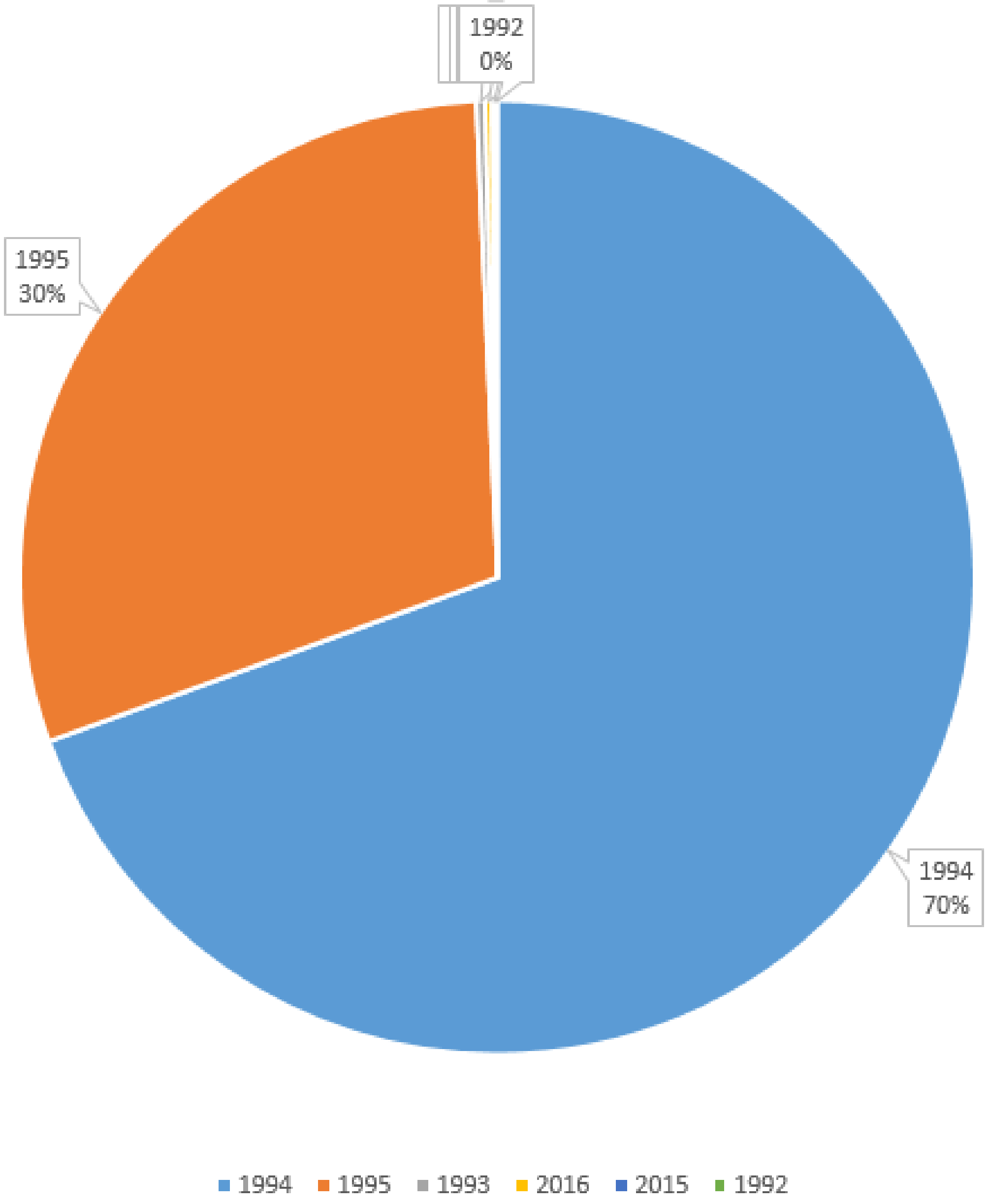
- Unknown:
  - ***.WND**
    (WinG Hidden Palette task)
  - ***.DAT**
    (Backup files)



Ridgway Building

- **Date last modified:**

1992	1
1993	5
1994	1185
1995	510
2015	2
2016	4



Open Challenges

- **Open files:**
  - How to open Backup files?
  - How to open old files?
- **Metadata extraction:**
  - Microsoft metadata
  - AutoCAD drawings
- **New functionality:**
  - Text search across all files
  - Extract dictionary from content
  - Link documents
  - Verification of extracted information (e.g. dates)