Technology Fundamentals for Digital Preservation

Developed in partnership with the DigitalPreservationCoalition

This POWRR Institute is generously funded by the INSTITUTE of Museum and Library SERVICES

What We'll Be Learning

Common Computer Systems & File Formats

We'll become familiar with the main aspects of many computer systems we may encounter, including operating systems, file systems, and file formats

Open Source Software, Packages, & Metadata

We'll learn how to begin choosing and deploying open source software at our institutions

OAIS Standard

We'll gain familiarity with the main concepts of OAIS, particularly with regards to the Information Model

Common Computer Systems
&
File Formats

Developed in partnership with the  DigitalPreservationCoalition

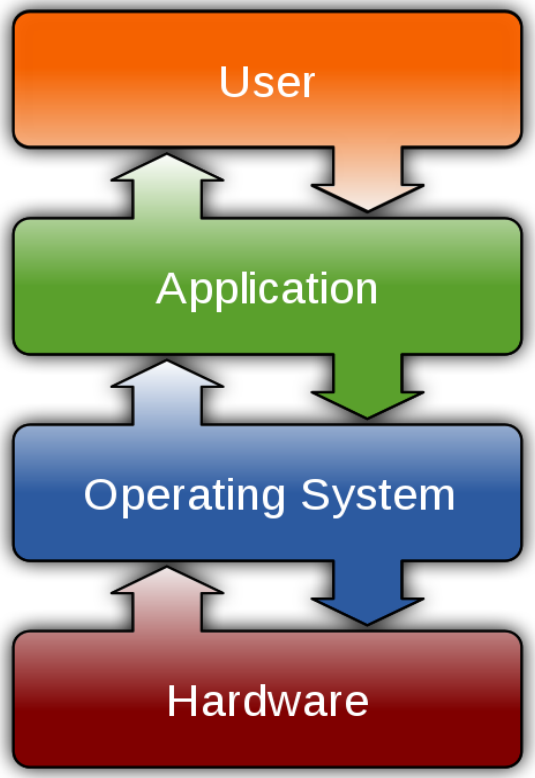
Expected Outcomes

Common Computer Systems & File Formats

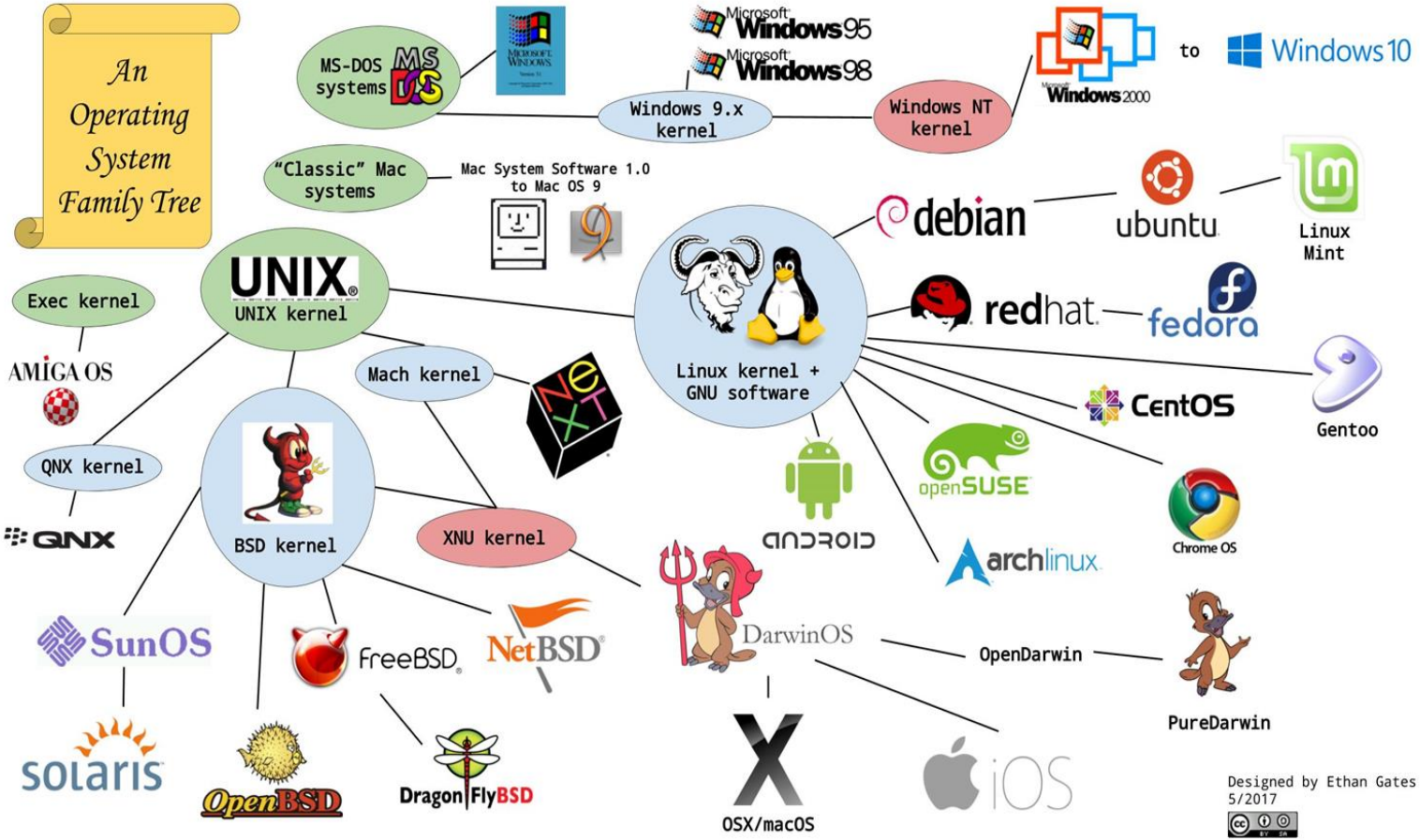
- ✓ Identify Windows and Unix Operating Systems and the key similarities and differences
- ✓ Navigate the standard file systems for these OS and use basic functions
- ✓ Describe the main issues relating to the preservation of common file formats

What is an Operating System?

- System software
- Manages hardware and software programs
- Schedules tasks
- Exist on all platforms
  - PCs/Laptops
  - Smart phones/tablets
  - Servers



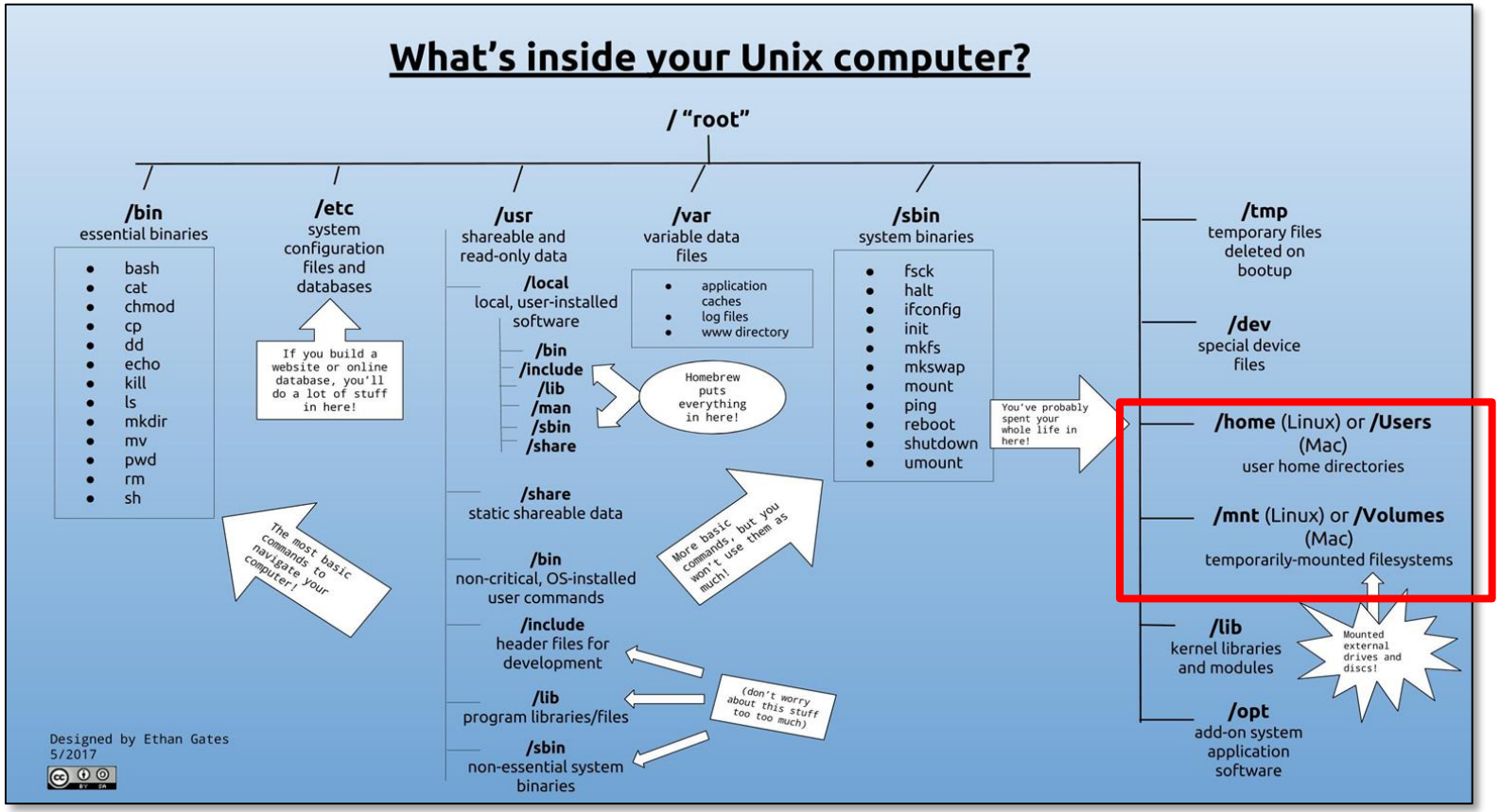
Many Flavors: The OS Family Tree



Some Important Differences

- Cost
- Licenses
- Customization
- Command Line and GUIs
- Storage

Getting to Know What's Inside



Don’t Fear the Command Line

Before GUIs, this was the primary way to interact with computers

Benefits:

- Fewer system resources used
- More control, power and precision
- Can automate common processes

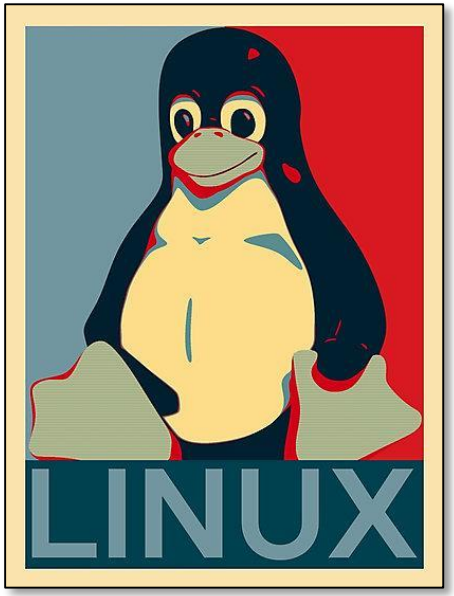
Used to run many digital preservation tools

File Formats: Just Keep the Bits…

What's In a File?

```
10101010101111010001010101
00100101001011101001010101
10010010010101010010000101
01010101111101000101010100
10010100101110100101010101
01001001010101001000010101
01010111110100010101010010
01010010111010010101010100
00100101010100100000101010
01011010101011110100010101
10001001010010111010010101
10100100100101010100100000
01010101011111010001010101
00100101001011101001010101
10010010010101010010000101
01010101111101000101010100
10010100101110100101000100…
```

```
SOI
APP0 JFIF
1.2
APP13 IPTC
APP2 ICC
DQT
SOF0
200x392
DRI
DHT
SOS
ECS0
RST0
ECS1
RST1
ECS2…
```



What Are the Risks?

- Media obsolescence
- Media failure or decay (such as “bit rot”)
- Natural / human-made disaster
- File format obsolescence

Images by Aldric Rodríguez Iborra, Erin Standley, Marie Van den Broeck, Edward Boatman and Dilon Choudhury from the Noun Project

What Is the Result?

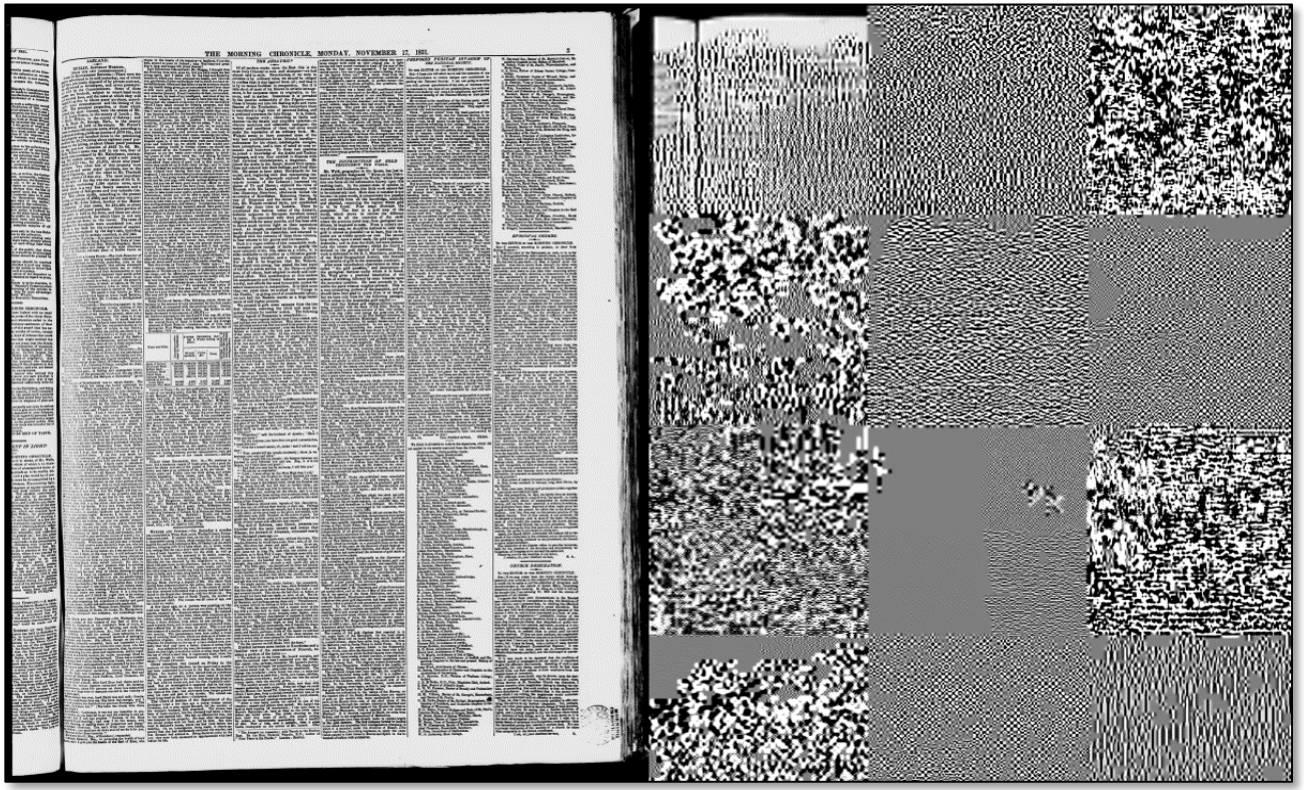


Image courtesy of the British Library

Stuff Happens

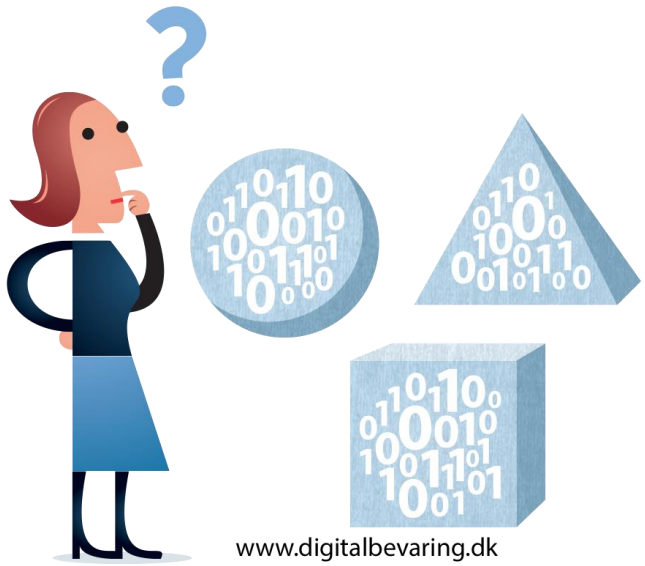
Whenever a digital collection is moved, processed, curated or altered in any way.... things can go wrong!

- Network dropouts at critical times
- Disks get full, subsequent data copied there is lost
- Software bugs lead to unexpected results
- Human error leads to all sorts of issues

Stuff happens a lot more at scale!

How Do We Solve These Problems?

- Keep more than one copy
- Refresh storage media
- Know what you have
- Integrity check your data (also called “Fixity”)
- Use ‘open’ formats
- Carry out preservation actions



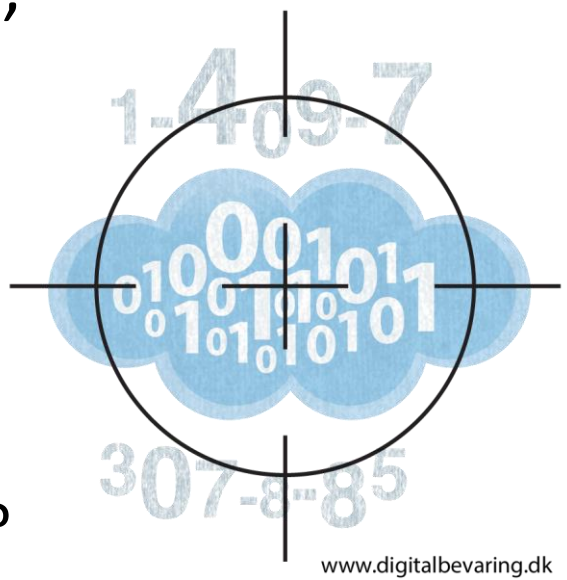
Making Sense of a Collection

Understand the data, then assess risks, plan, take action to preserve

Characterization:

- How many files?
- How big are the files?
- What file formats?
- Is the data dynamic or interactive?
- Does it contain personal information?
- Is it encrypted?

Scale = automation = software tools



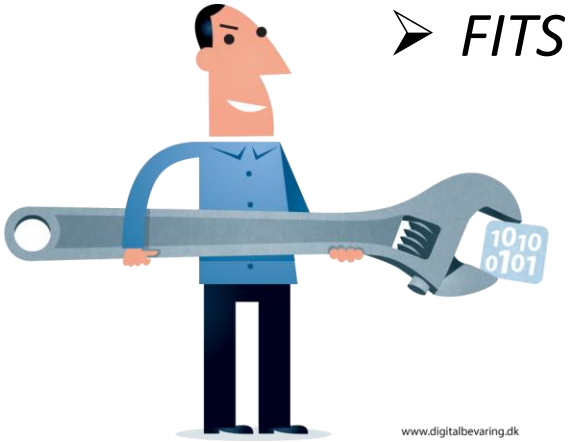
Characterization Tools

Pronom: a register of file formats and their behaviors (probably the world's most boring database)

DROID: a tool that analyses the files on a system (using the most boring database in the world)

Also in this space:

- *C3PO*
- *JHOVE*
- *TIKKA*
- *FITS*




TRUST NO ONE
Assume nothing,
validate everything

What is a “checksum” or “hash value”?

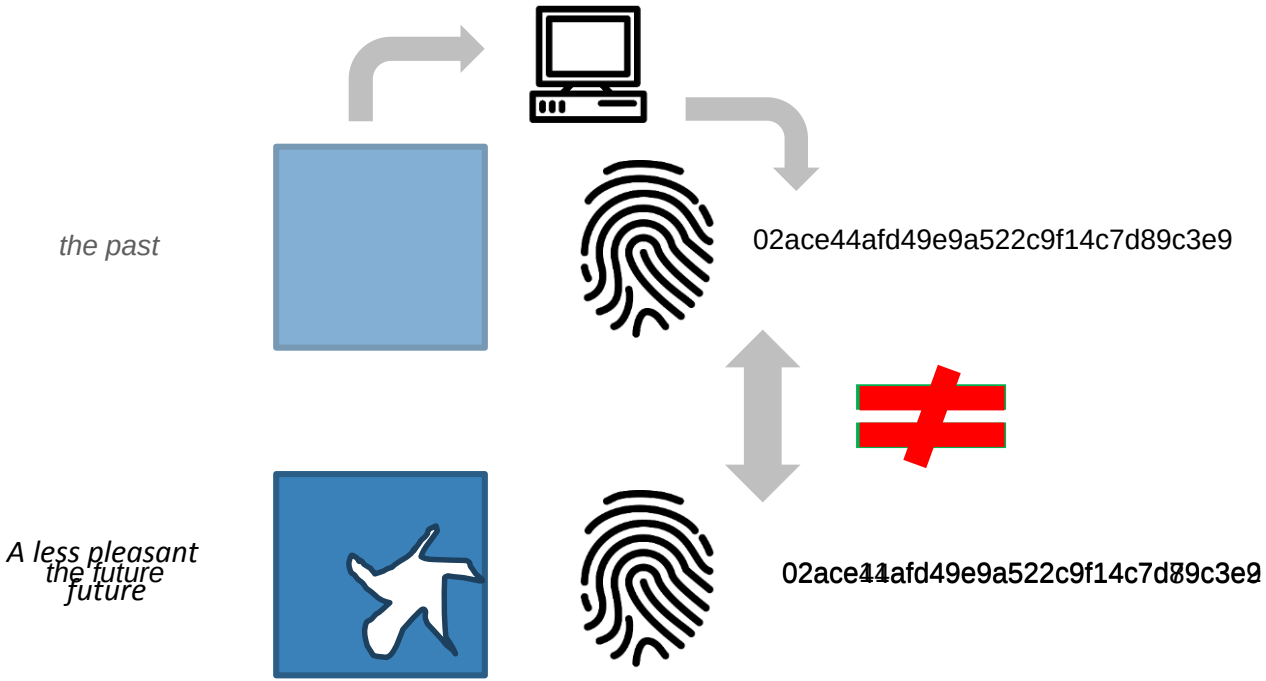
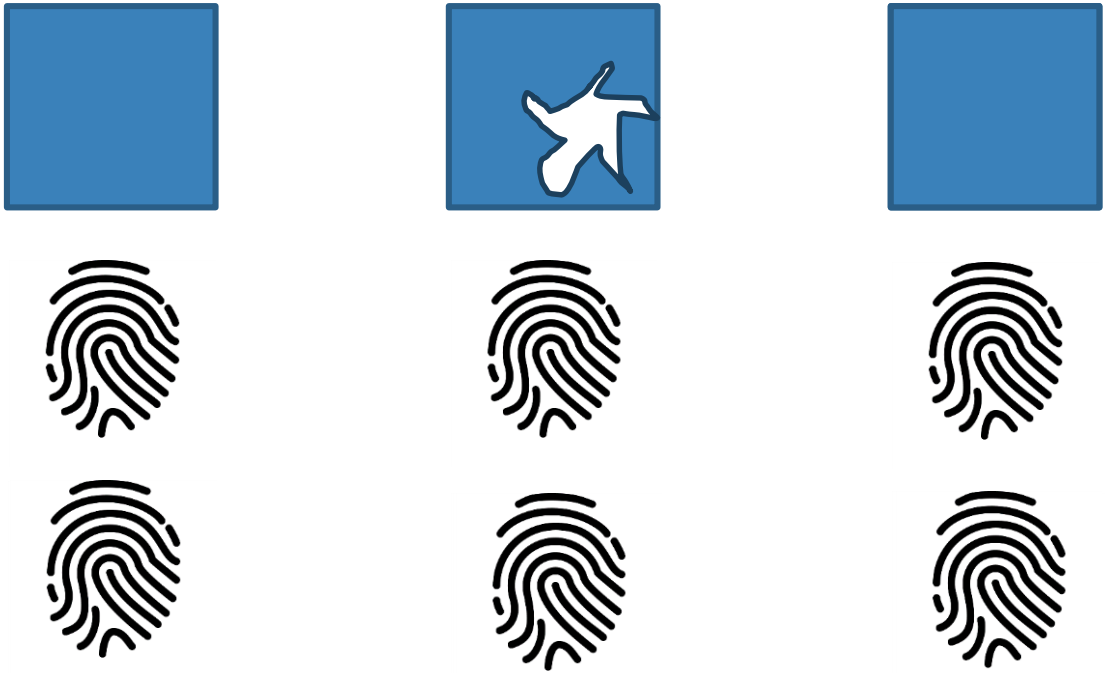


Image by Arthur Shlain from the Noun Project

Combined Strategies:

Keep 3 Copies & Perform Integrity Checks



Integrity Checking – Tools

Fixity

https://www.avpreserve.com/tools/fixity/

Auditing Control Environment (ACE)

https://wiki.umiacs.umd.edu/adapt/index.php/Ace

For alternatives – see COPTR

http://coptr.digipres.org/Category:Fixity

Approaches to Preservation

- Bit-Level
- Migration
- Emulation
- Hardware Preservation
- Digital Archaeology

etc.....

Illustration by Jørgen Stamp
digitalbevaring.dk CC BY 2.5 Denmark

Migration

Normalization

To New Versions

Emulation


Start
PL ER 1
PLAYER 2
RAN
RANK
99
99
SECTOR 1
SCORE
FUEL
TEMP
SCORE
00000
© 1981 SEGA
00000
Footer

Common Computer Systems
&
File Formats

QUESTIONS?

Open Source Software

Developed in partnership with the 

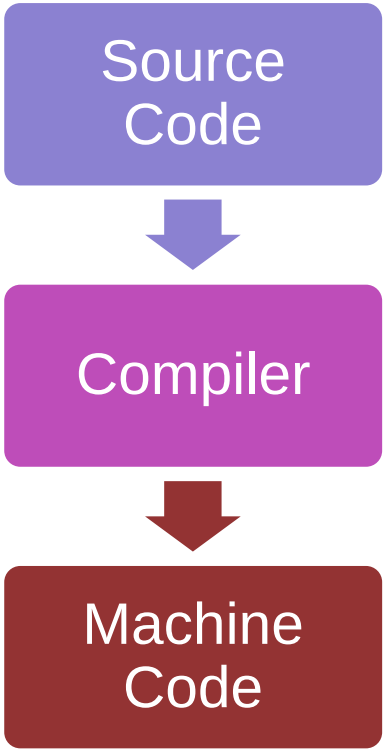
Expected Outcomes

Open Source Software

- ✓ Explain the ethos of the open source software movement and the main benefits and constraints of using this type of software product
- ✓ List the main digital preservation open source software tools for libraries and archives
- ✓ Describe the differences between using open source software and products offered by a vendor.

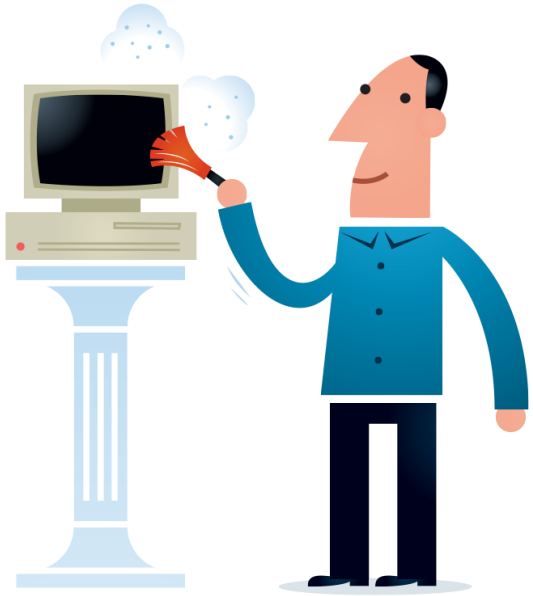
Software 101

- Written in a human-readable programming language
- Most often ‘Compiled’ using an intermediary program into computer-readable form
- Proprietary software provides only compiled version
  - Can’t make modifications beyond program’s inbuilt functionality



History of OSS

- First conceived in late 1990s
- Adopt best practices from Free and Commercial Software
- Open development = better software
- First program released as OSS: Netscape browser
- Server/software infrastructure early priorities



www.digitalbevaring.dk

Ethos of OSS

"Software should be made universally available in its entirety, with everyone afforded the opportunity to understand, change and re-distribute it."

Andrew McHugh, DCC Manual, 2005

Key Elements of OSS:

- Transparency
- Openness
- Community

Ten Criteria for OSS

1. Free Redistribution
2. Include Source Code
3. Allow Derived Works
4. Integrity of Author's Source Code
5. No Discrimination Against Persons or Groups
6. No Discrimination Against Fields of Endeavor
7. Inherited Distribution of License
8. License Must Not Be Specific to a Product
9. License Must Not Restrict Other Software
10. License Must Be Technology-Neutral

A Free Beer, A Free Cat, or Free Speech?

A Free Beer

- OSS is not necessarily free as in ‘gratis’

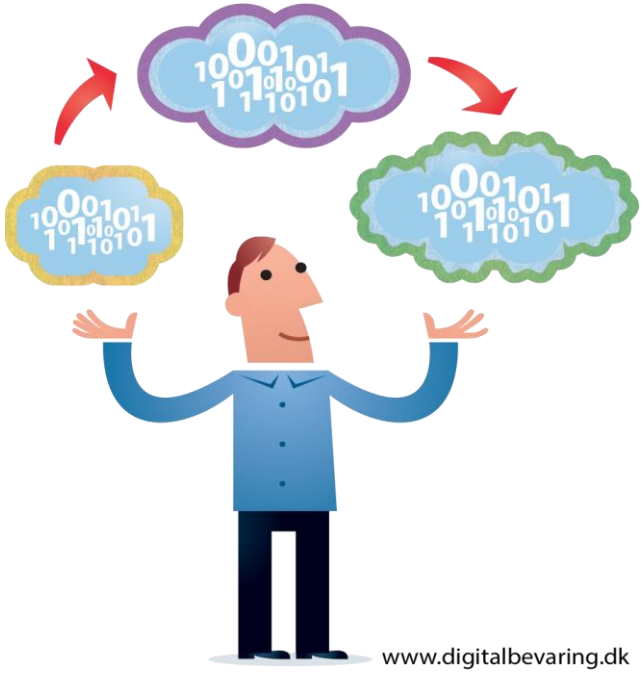
A Free Cat

- Costs relating to implementation, upkeep, training, support, etc.

Free Speech

- Access to source code
- Ability to adapt to own needs
- Can redistribute

Development Model



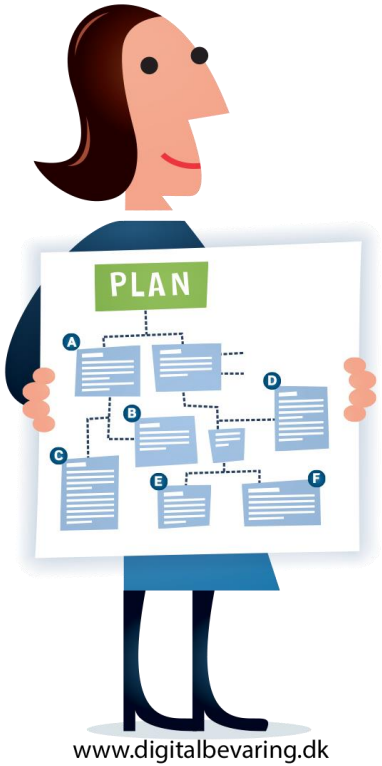
- Users as co-developers
- Early releases
- Frequent integration
- Different versions: beta vs stable
- High modularization
- Dynamic decision-making

Different Types of Contributions

"Give as you can"

Help with:

- Scoping developments
- Identifying requirements
- Writing code
- Providing feedback
- Identifying Bugs



SPRUCE Project

Community orientated approach to digital preservation

Collaboration on tools and resources

Held 3 Mashups and 1 Hackathon

SPRUCE Mashup Manifesto

- ➢ Be agile
- ➢ Re-use, don’t reinvent the wheel
- ➢ Keep it small, keep it simple
- ➢ Make it easy to use, build on, re-purpose and ultimately, maintain
- ➢ Share outputs, exchange knowledge, learn from each other

Some Major OSS Organizations

- Open Source Initiative
- Apache Foundation
- Mozilla
- Linux Foundation
- Free Software Foundation
- WordPress

Benefits/Opportunities

- Likely to be lower cost
- More freedom
- Influence new tools/functionality
- Fewer license restrictions
- Improved debugging
- Builds communities
- Easier to emulate
- Can share tools with data creators

Risks/Constraints

- Tech resources/skills needed
- Lack of clear leadership and governance
- Requires community engagement
- Variable documentation
- Misconception about costs
- Securing institutional buy-in
- Potentially less diversity
- Too much customization
- Funding/sustainability

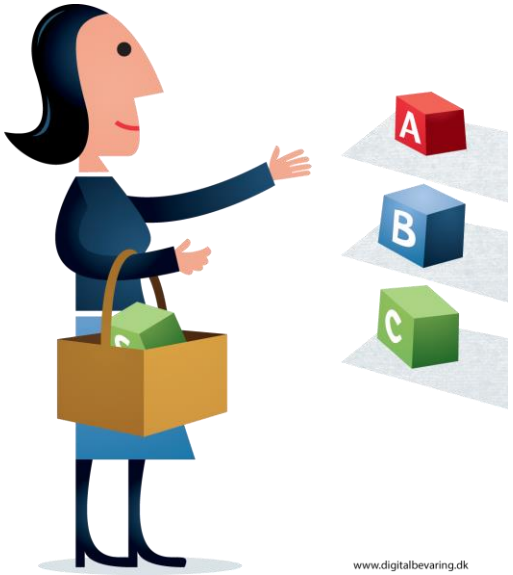
OSS Licenses

- ‘Copyleft’ licenses
- Approved by OSI
- Emphasis on collaboration, openness and reuse
- Derived works must have same license
- Popular licenses include:
  - Apache License 2.0
  - GNU General Public or Library General Public Licenses
  - BSD 3-Clause or 2-Clause Licenses
  - Mozilla Public License

Comparison with Vendor Solutions

Issue	OSS	Vendor
Initial Cost		
Installation		
Source Code		
Customization		
Licenses		
Bugs		
Support		
Documentation		
Training		
Motivation for Developments		
Succession		

Things to Consider When Selecting OSS



- Longevity
- Stability
- Costs
- Ubiquity
- Skills required
- Documentation/training
- Compatibility

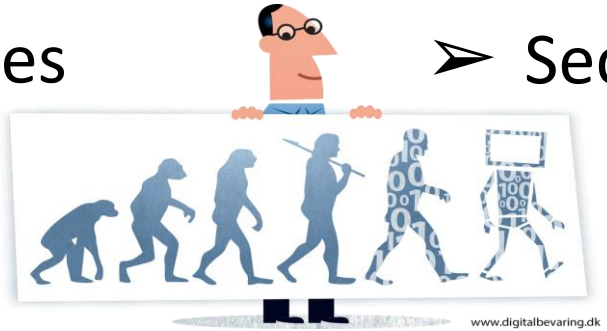
Beta vs. Stable

Beta

- Version for community testing
- More bugs
- Latest features
- More updates

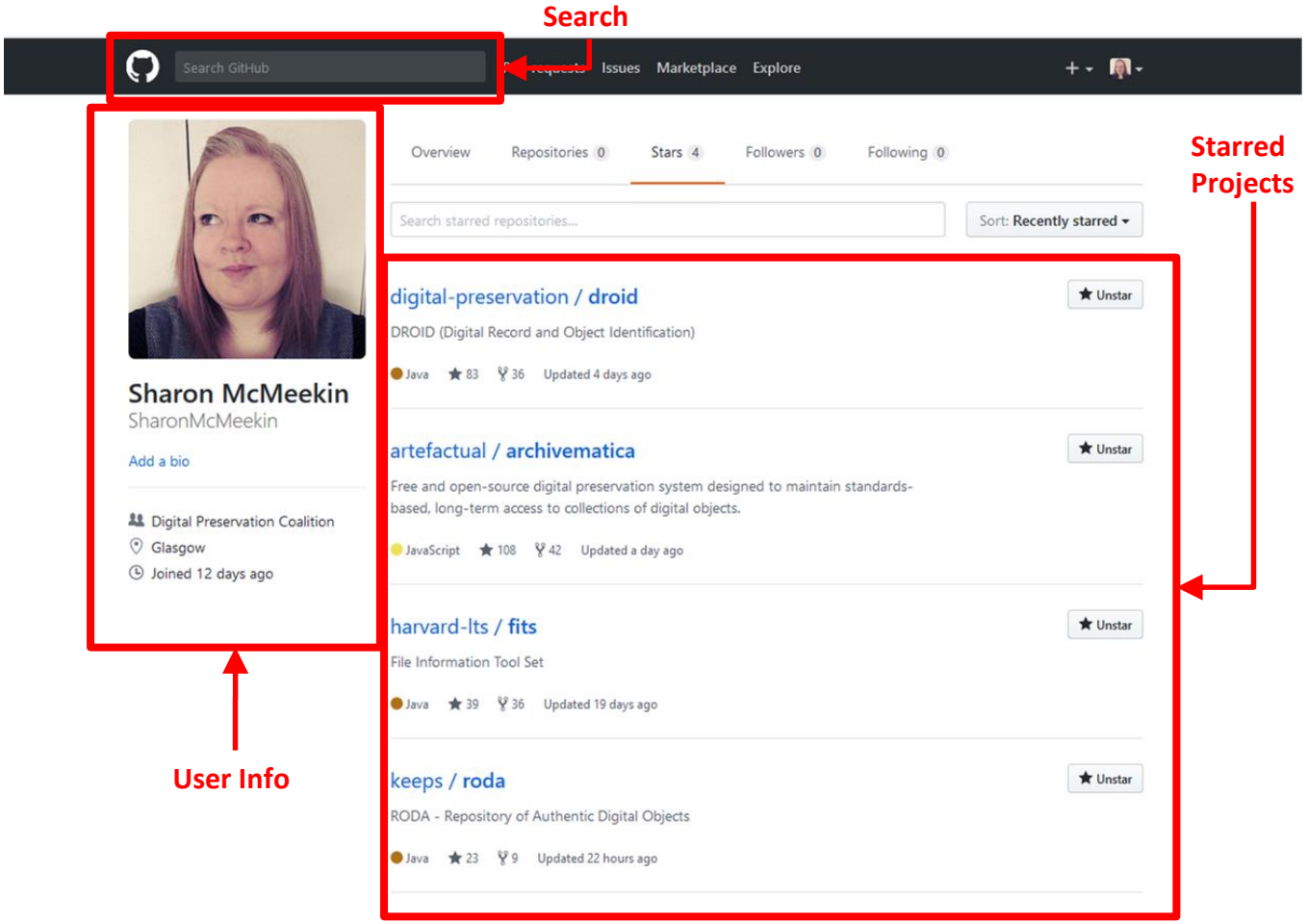
Stable

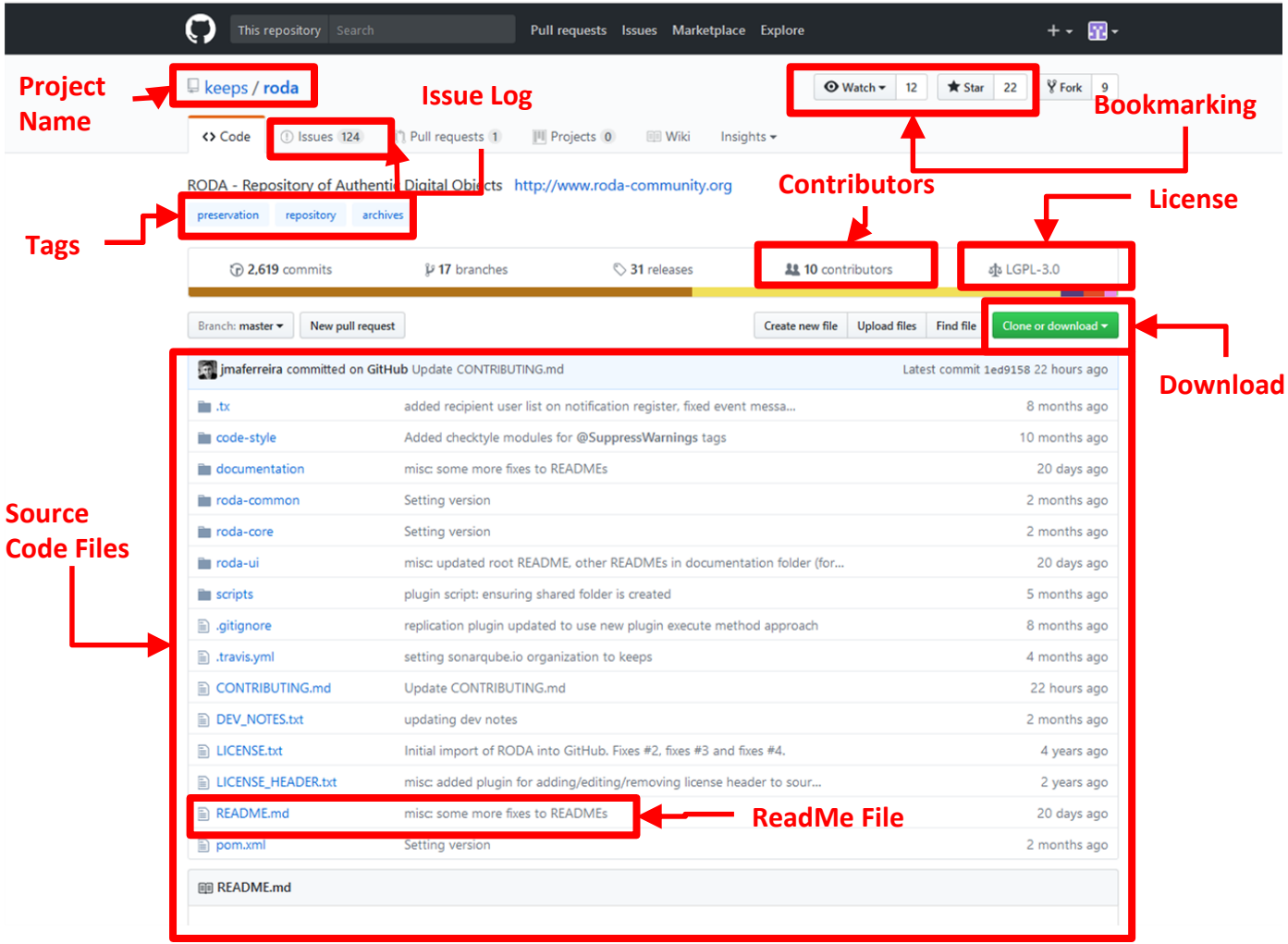
- Thoroughly tested
- Less buggy
- May lack new features
- Security updates

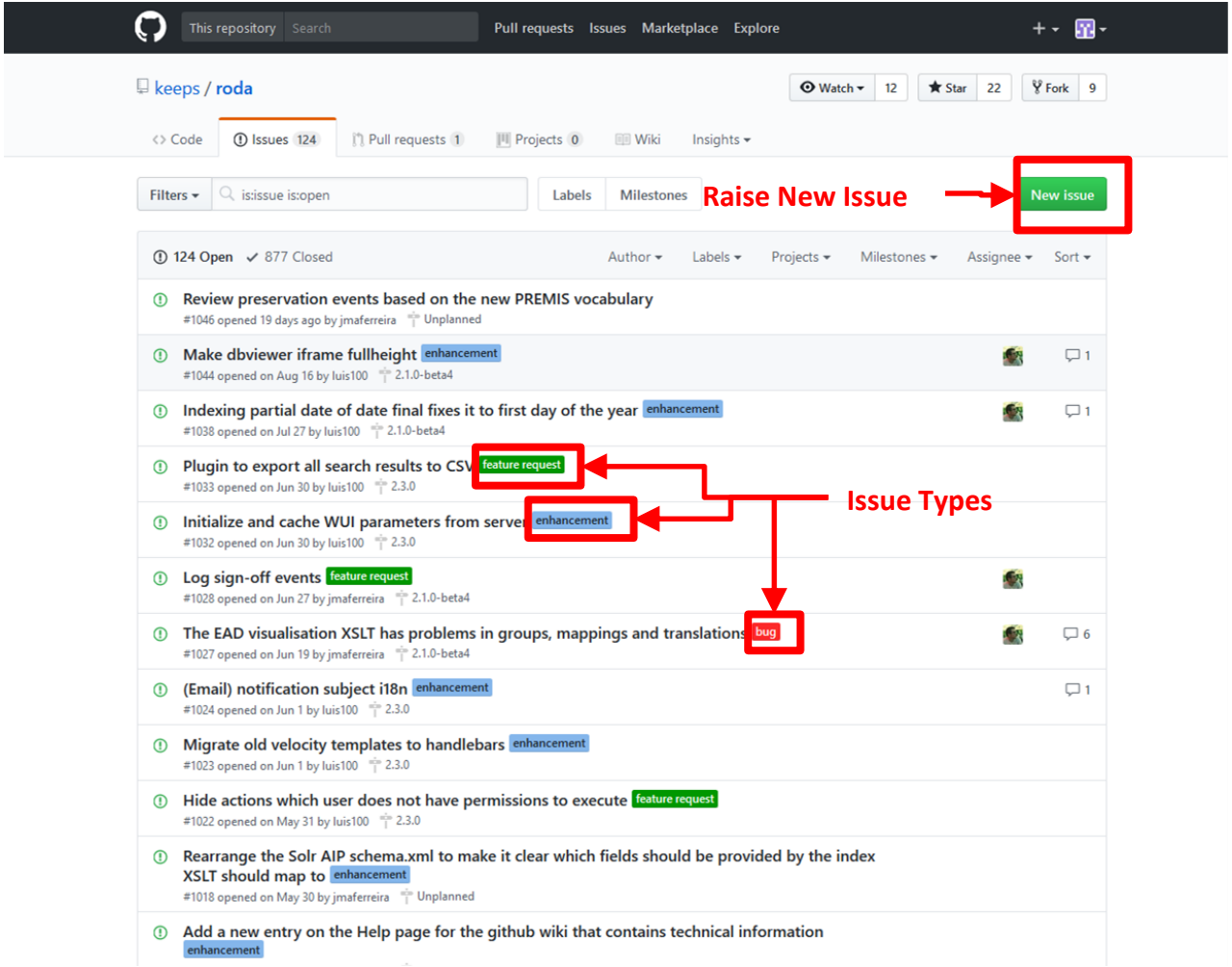


GitHub

- A code hosting platform
  - Collaboration
  - Version Control (Git)
- Used by developers of the majority of OSS digital preservation tools and solutions
- Public and private development spaces
  - Basic account = free
- Access to full source code
- Best way to contribute to software development


Search
Search GitHub
Issues
Marketplace
Explore
Overview
Repositories 0
Stars 4
Followers 0
Following 0
Search starred repositories...
Sort: Recently starred
Starred Projects
digital-preservation / droid
Unstar
DROID (Digital Record and Object Identification)
Java 83 36 Updated 4 days ago
artefactual / archivematica
Unstar
Free and open-source digital preservation system designed to maintain standards-based, long-term access to collections of digital objects.
JavaScript 108 42 Updated a day ago
harvard-lts / fits
Unstar
File Information Tool Set
Java 39 36 Updated 19 days ago
keeps / roda
Unstar
RODA - Repository of Authentic Digital Objects
Java 23 9 Updated 22 hours ago
Sharon McMeekin
SharonMcMeekin
Add a bio
Digital Preservation Coalition
Glasgow
Joined 12 days ago
User Info


This repository
Search
Pull requests
Issues
Marketplace
Explore
Project Name
keeps / roda
Issue Log
Watch 12
Star 22
Fork 9
Bookmarking
Code
Issues 124
Pull requests 1
Projects 0
Wiki
Insights
RODA - Repository of Authentic Digital Objects
http://www.roda-community.org
Contributors
License
Tags
preservation
repository
archives
2,619 commits
17 branches
31 releases
10 contributors
LGPL-3.0
Branch: master
New pull request
Create new file
Upload files
Find file
Clone or download
Download
jmaferreira committed on GitHub Update CONTRIBUTING.md
Latest commit 1ed9158 22 hours ago
.tx added recipient user list on notification register, fixed event messa... 8 months ago
code-style Added checktyle modules for @SuppressWarnings tags 10 months ago
documentation misc: some more fixes to READMEs 20 days ago
roda-common Setting version 2 months ago
roda-core Setting version 2 months ago
roda-ui misc: updated root README, other READMEs in documentation folder (for... 20 days ago
scripts plugin script: ensuring shared folder is created 5 months ago
.gitignore replication plugin updated to use new plugin execute method approach 8 months ago
.travis.yml setting sonarqube.io organization to keeps 4 months ago
CONTRIBUTING.md Update CONTRIBUTING.md 22 hours ago
DEV_NOTES.txt updating dev notes 2 months ago
LICENSE.txt Initial import of RODA into GitHub. Fixes #2, fixes #3 and fixes #4. 4 years ago
LICENSE_HEADER.txt misc: added plugin for adding/editing/removing license header to sour... 2 years ago
README.md misc: some more fixes to READMEs 20 days ago
pom.xml Setting version 2 months ago
README.md
Source Code Files
ReadMe File


This repository
Search
Pull requests
Issues
Marketplace
Explore
keeps / roda
Watch 12
Star 22
Fork 9
Code
Issues 124
Pull requests 1
Projects 0
Wiki
Insights
Filters
is:issue is:open
Labels
Milestones
Raise New Issue
New issue
124 Open
877 Closed
Author
Labels
Projects
Milestones
Assignee
Sort
Review preservation events based on the new PREMIS vocabulary
#1046 opened 19 days ago by jmaferreira
Unplanned
Make dbviewer iframe fullheight
enhancement
#1044 opened on Aug 16 by luis100
2.1.0-beta4
Indexing partial date of date final fixes it to first day of the year
enhancement
#1038 opened on Jul 27 by luis100
2.1.0-beta4
Plugin to export all search results to CSV
feature request
#1033 opened on Jun 30 by luis100
2.3.0
Initialize and cache WUI parameters from serve
enhancement
#1032 opened on Jun 30 by luis100
2.3.0
Issue Types
Log sign-off events
feature request
#1028 opened on Jun 27 by jmaferreira
2.1.0-beta4
The EAD visualisation XSLT has problems in groups, mappings and translations
bug
#1027 opened on Jun 19 by jmaferreira
2.1.0-beta4
(Email) notification subject i18n
enhancement
#1024 opened on Jun 1 by luis100
2.3.0
Migrate old velocity templates to handlebars
enhancement
#1023 opened on Jun 1 by luis100
2.3.0
Hide actions which user does not have permissions to execute
feature request
#1022 opened on May 31 by luis100
2.3.0
Rearrange the Solr AIP schema.xml to make it clear which fields should be provided by the index XSLT should map to
enhancement
#1018 opened on May 30 by jmaferreira
Unplanned
Add a new entry on the Help page for the github wiki that contains technical information
enhancement

Types of OSS for Digital Preservation

Two main types of open source for digital preservation

Large-scale applications

- Repository systems
- Storage
- Workflow

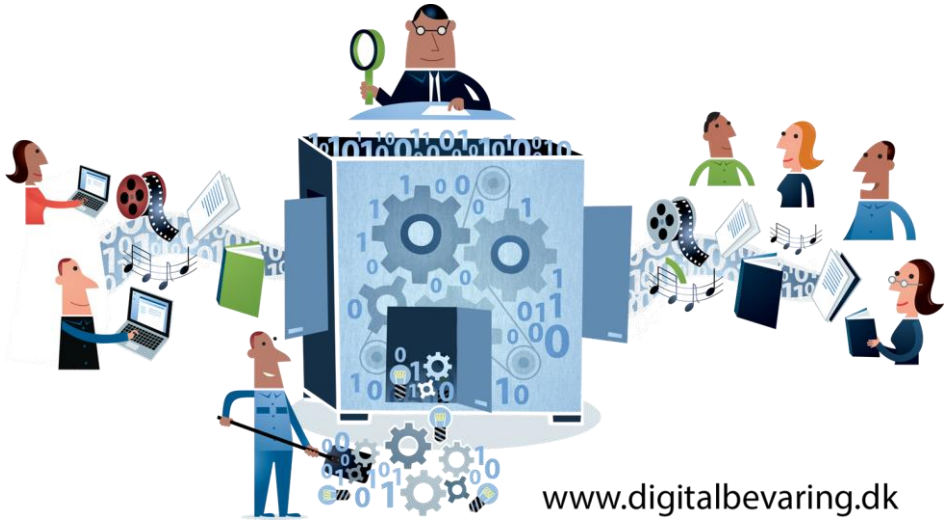
Tools for particular functions

- Characterization
- Migration
- De-duplication
- …

Example Repository Systems

OSS repository systems include:

- Archivematica
- RODA
- DSpace
- Fedora
- Islandora
- Eprints
- Samvera (Hyku)



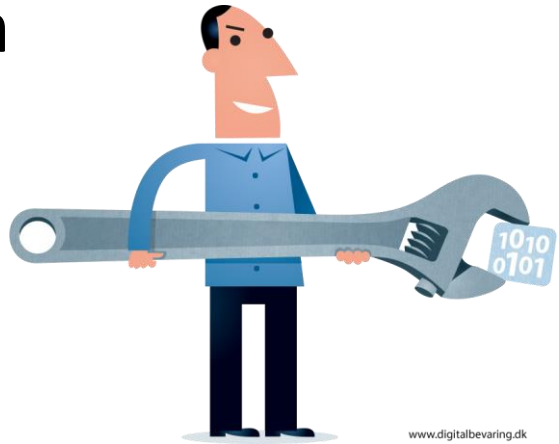
Example Tools: Characterization

Various tools with different functionality:

- DROID
- Apache Tika
- C3PO
- FIDO
- JHOVE
- FITS

Other Types of Tools

- De-duplication
- Forensics
- Decryption
- Fixity
- Planning



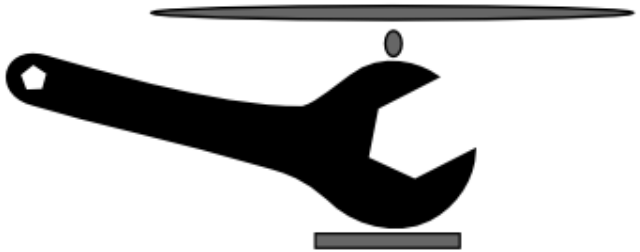
- Migration
- Emulation
- Validation
- Policy
- etc……

COPTR

- Tools registry for digital preservation
- Includes OSS and Vendor solutions
- Part of DigiPres Commons
- Hosted by the Open Preservation Foundation

Browse by:

- Name
- Function
- Type of content



POWRR Tool Grid

	Access, Use and Reuse	Create or Receive (Acquire)	Cross-Lifecycle Functions	Dispose	Ingest	Preservation Action	Preservation Planning	Store
Audio	2	5	3		15	11	1	
Binary Data			4					
Container						5		
Database	1	1	3		3	14		
Disk Image		7	4		3	1		1
Document	3	1	4		35	15		
EBook					5	2		
Email			5		2	4		1
Geospatial					1			
Image	2	2	3		23	23		
Project Management Data	1					1		
Research Data	2	8	13		4		16	17
Software		1	1		2	2		1
Spreadsheet					6	3		
Video	1	3	1		10	8	1	
Web	3	21	2		7	3	1	1
~Not Content Type Specific~	22	38	83	9	69	61	31	51

Open Source Software

QUESTIONS?

The OAIS Reference Model, Packages, & Metadata

Developed in partnership with the  DigitalPreservationCoalition

Expected Outcomes

OAIS, Packages, & Metadata

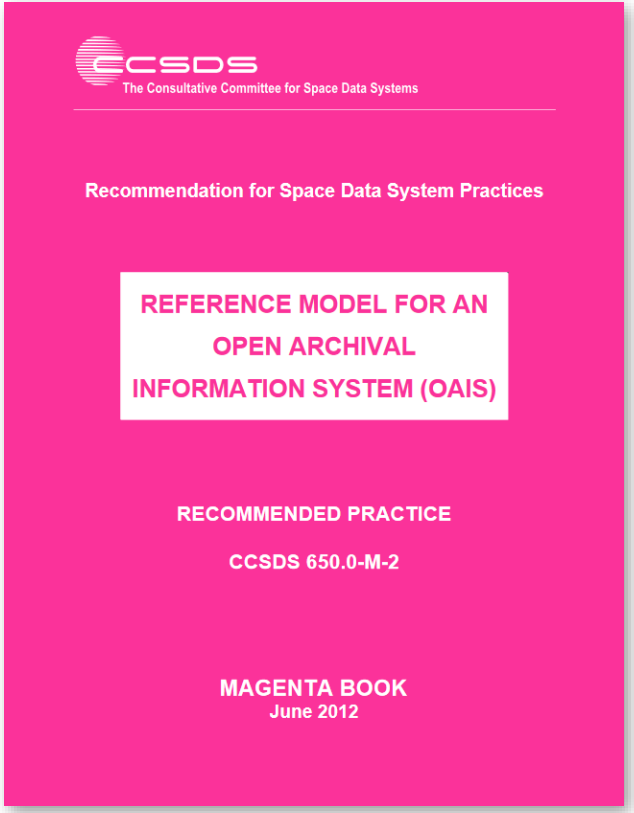
- ✓ Explain at a high level the main components of the OAIS standard; including the mandatory responsibilities, functional model, information model and key terms
- ✓ Describe the elements of the information model and their relevance to the preservation lifecycle
- ✓ Design a basic information package and select relevant metadata standards

Why Do We Need Models?

- High-level conceptual map for activities
- Can help set requirements
- Supports identification and development of standards
- Framework for comparing and assessing approaches

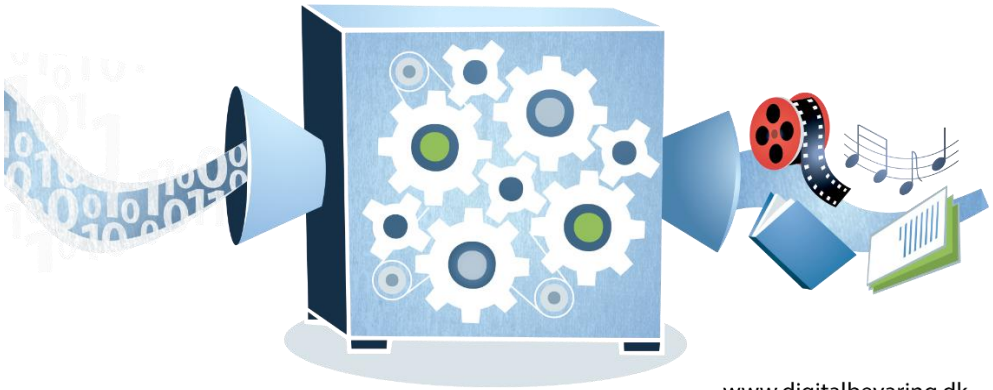
What is OAIS?

- Open Archival Information System Reference Model
- Originally developed by Consultative Committee for Space Data Systems
- An international standard
  ISO 14721:2012
- Vocabulary and basic framework for much digital preservation work



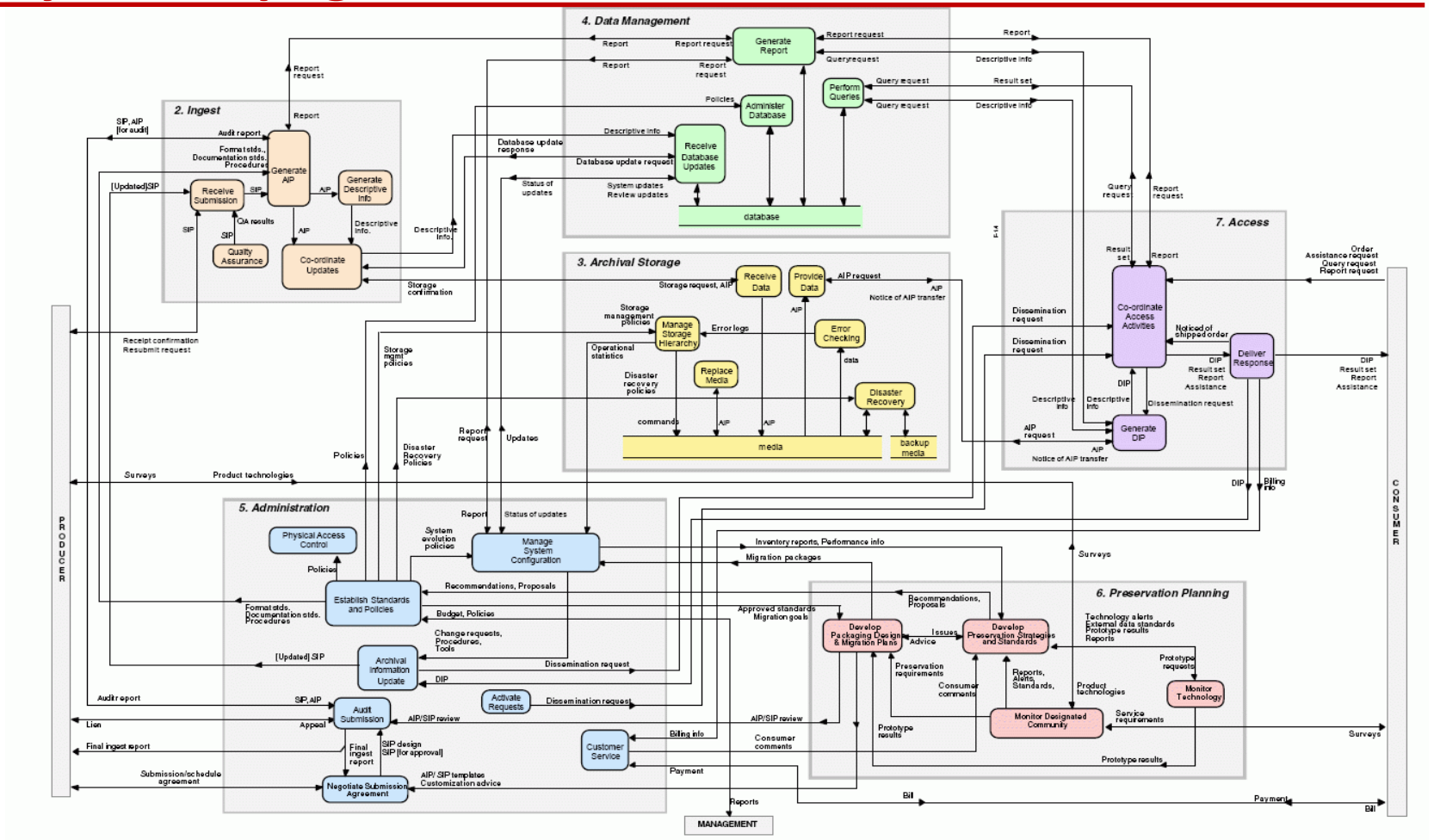
Basic Definition of an OAIS

a reference model … to establish a **system** for archiving information, both digitalized and physical, with an **organizational scheme** composed of people who accept the **responsibility** to **preserve information** and make it available to **a designated community**



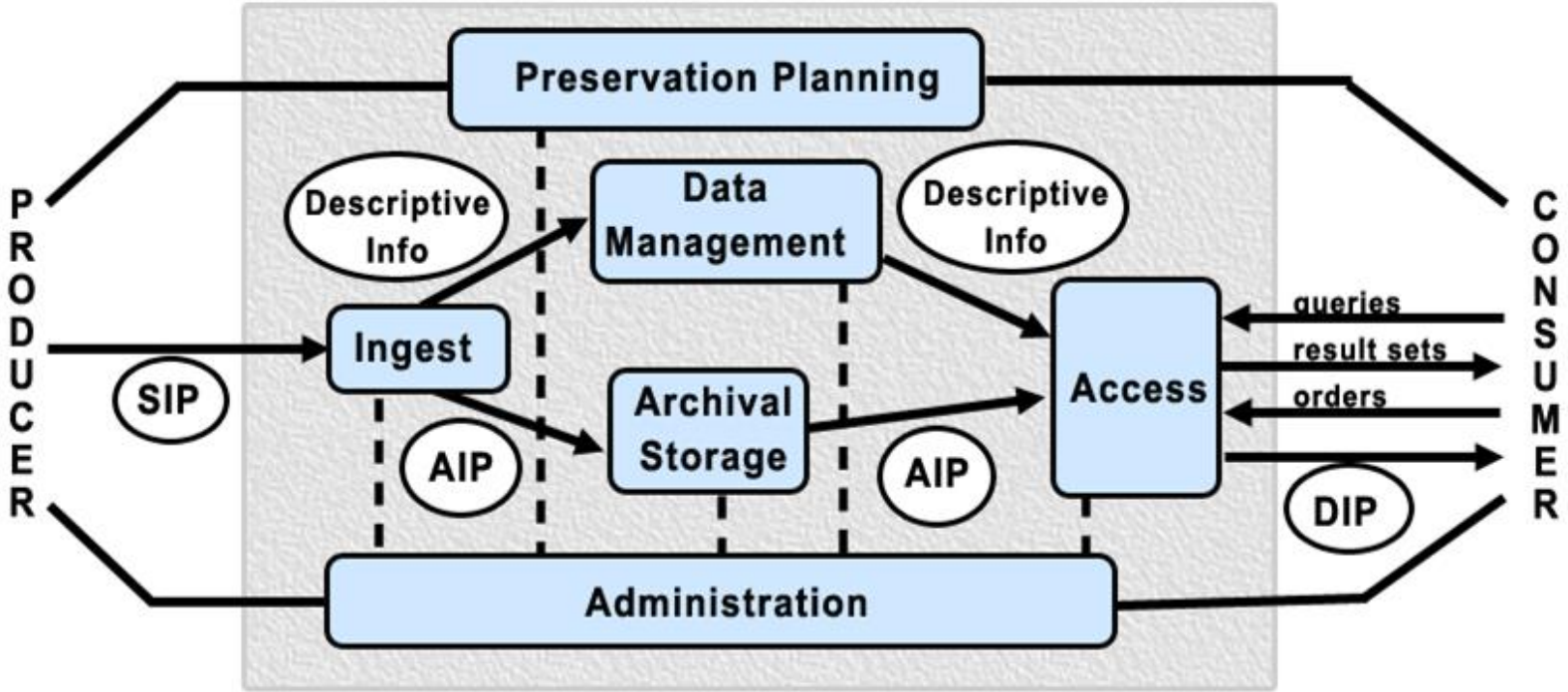
www.digitalbevaring.dk

Scary OAIS Spaghetti Monster



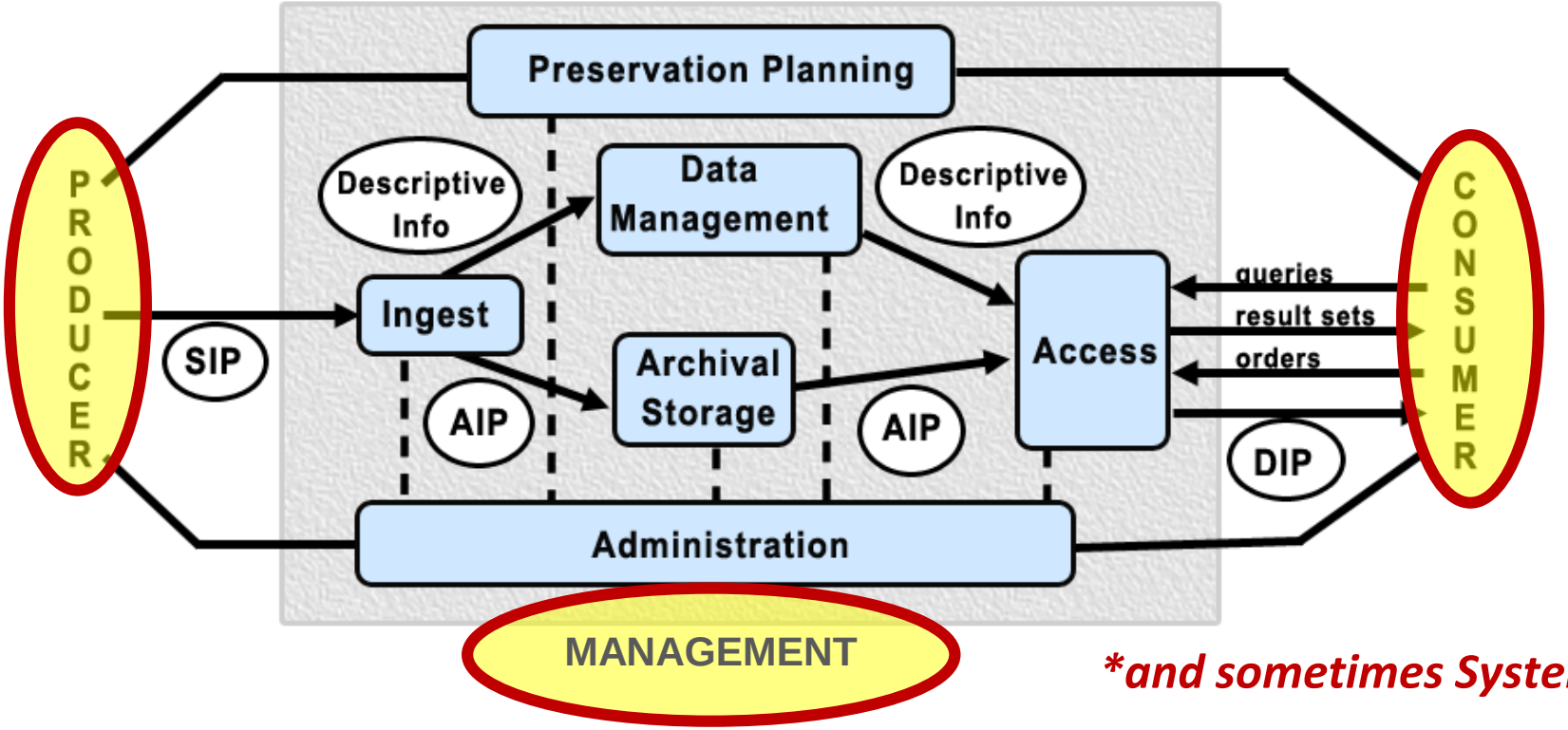
Functional Model….

…still scary but let's give it a chance.



Actors….

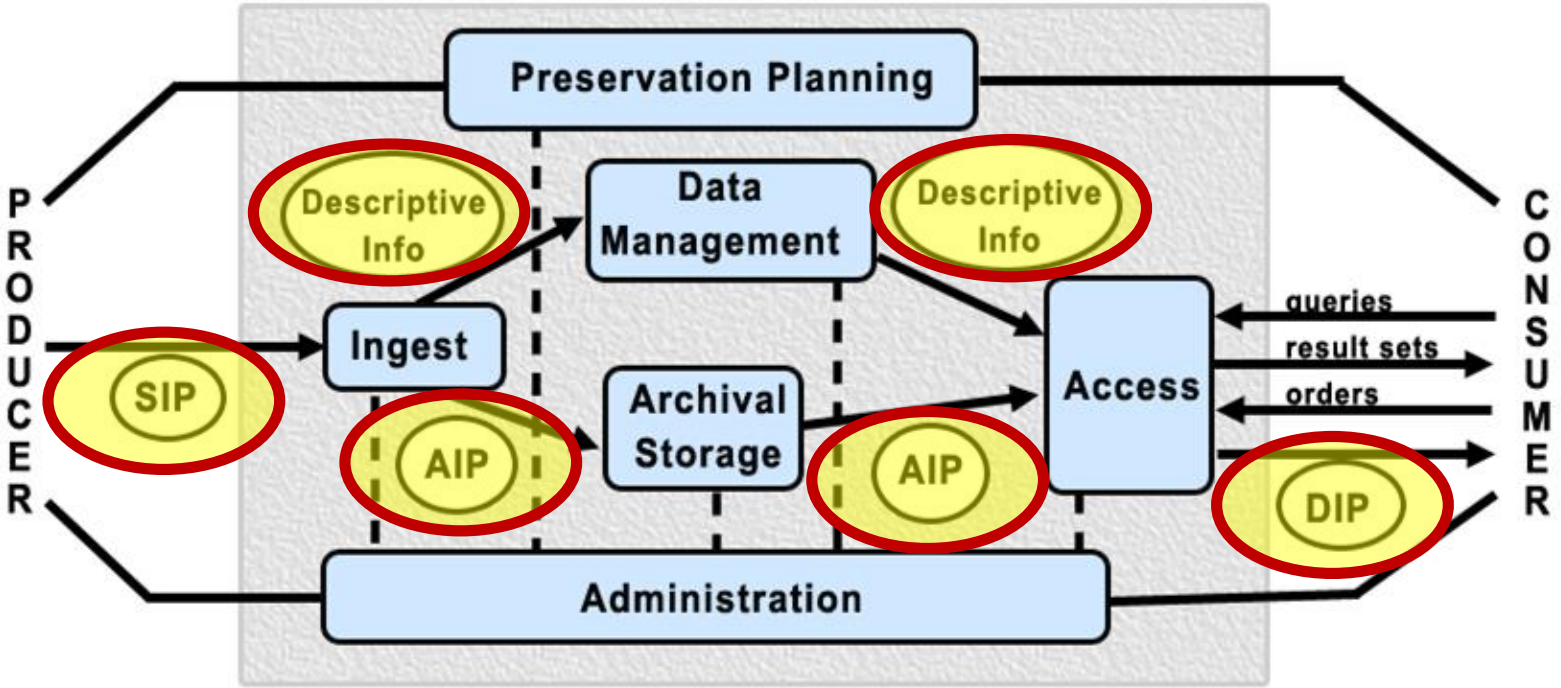
…are just the folks* in your normal professional encounters.



**and sometimes Systems*

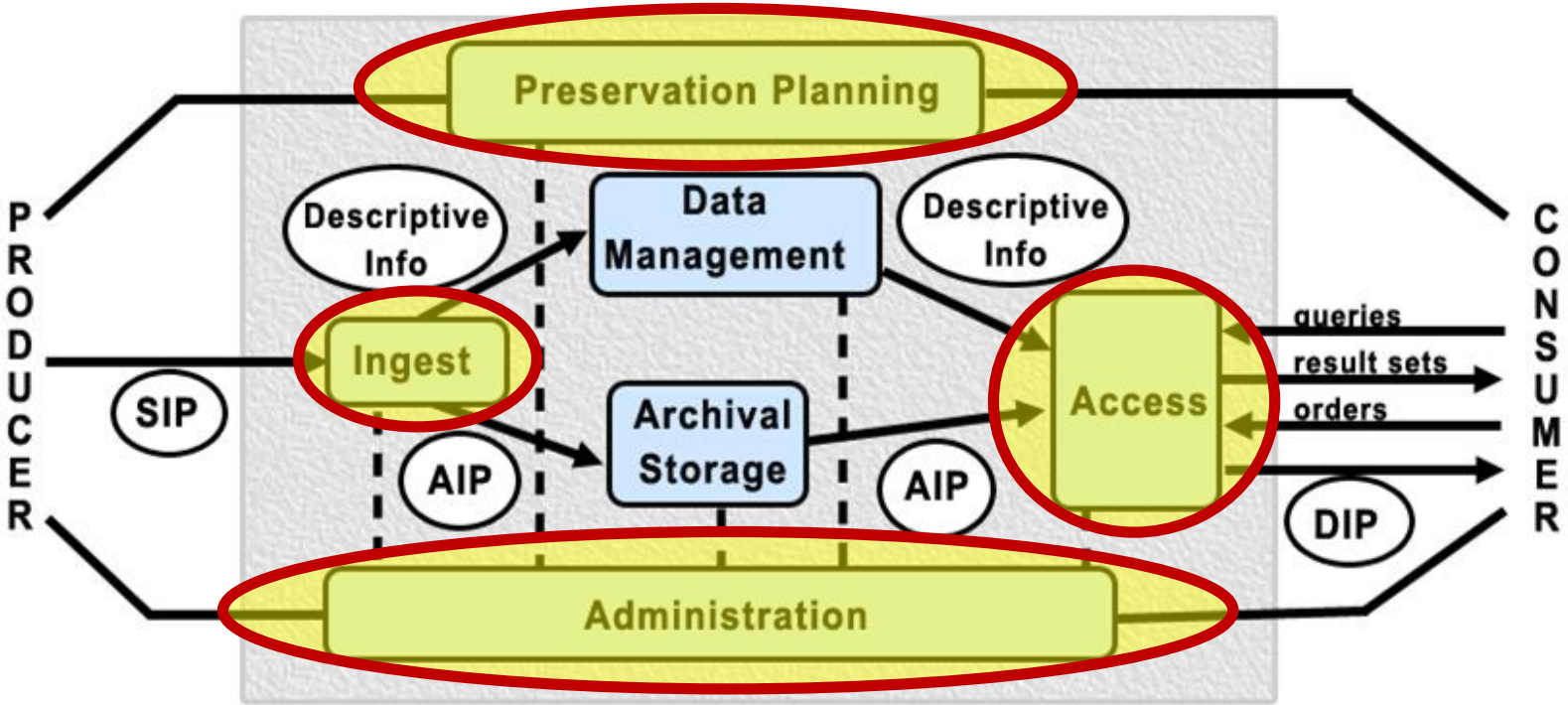
Objects….

…are just the materials and the information about them that bounce around your world.



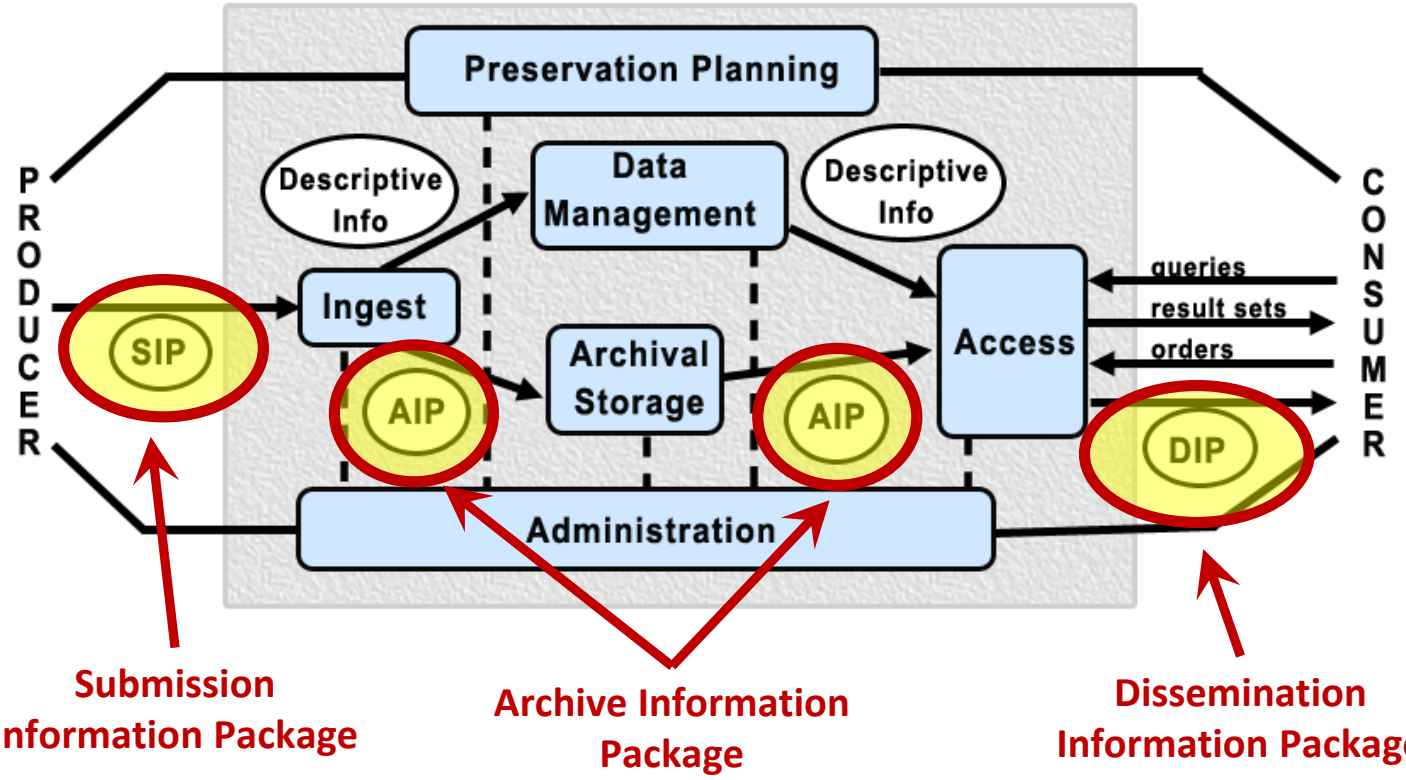
Functional Entities….

…are just the activities that *someone* needs to do in your world.



Information Packages….

…are just a way to keep the materials and the necessary information about them together.



Information Package Structures

May be influenced by:

- Designated community needs
- Existing systems
- Resources available
- Preservation plans

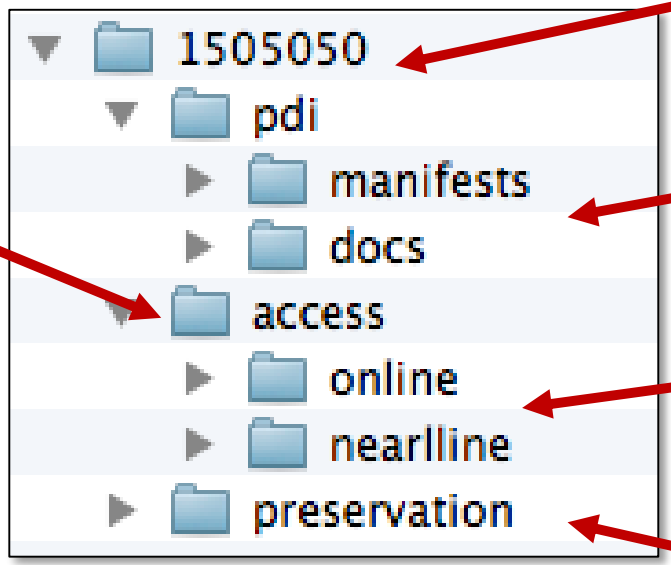
Options from simple to complex

- Standard folder system
- Databases
- XML wrappers

Tools available to help with creation of IPs

What's in the AIP? An Example

AIP Example from Chris Prom at UIUC Archives



Unique ID
Accession #
System ID

Descriptive info

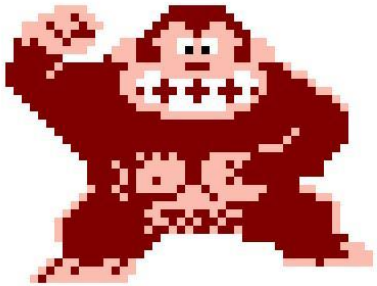
Access copies of original digital files, maybe migrated to new formats

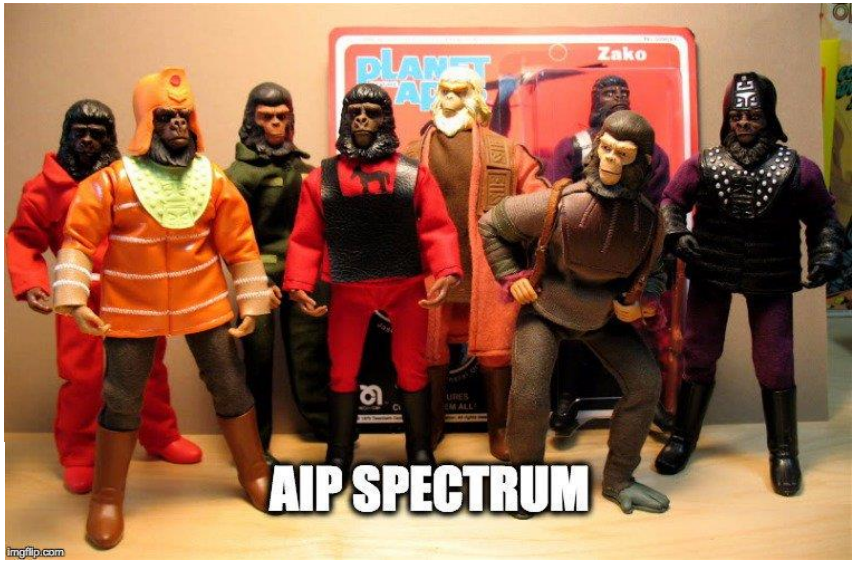
Online = online version
Nearline = made available in person

Original submitted digital files, pre-preservation actions

http://e-records.chrisprom.com/archival-information-packet-structure/

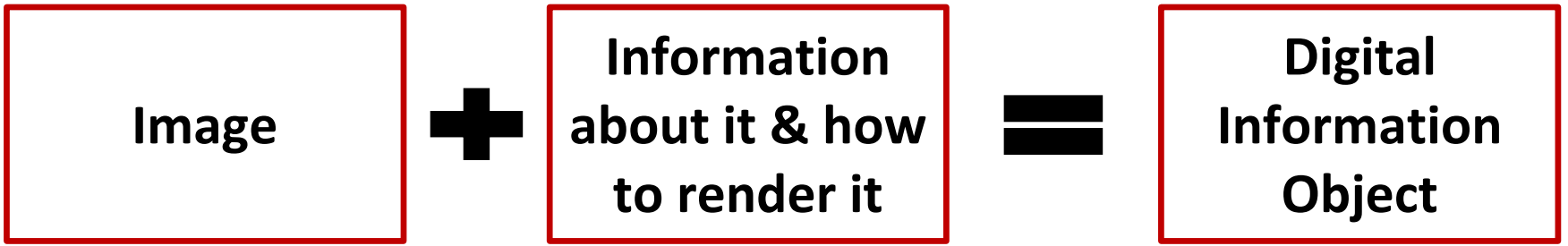
Planet of the AIPs





Getting From Objects To Information

DIGITAL



PHYSICAL

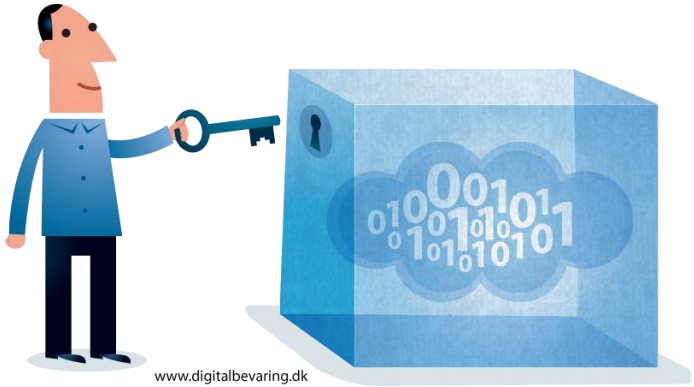
Representation Information

- Two types:
  - Structure Information
    - File Format, Software….**how to render it…the projector!**
  - Semantic Information
    - User Documentation, Data Dictionary….**the information about it!**

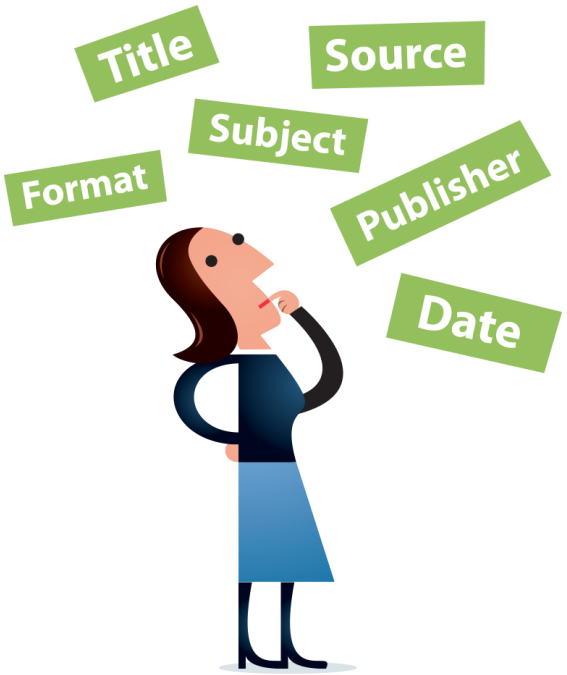
- Can be simple through to very complex

- Determined by needs of your Designated Community

- Tends to become more complex over time



Preservation Description Information (PDI)



Supports preservation, authenticity and dissemination

Describes‘the past and present states of the data

Consists of 5 components:

- Reference information
- Context information
- Provenance information
- Fixity information
- Access Rights information

A Little Bit On PREMIS

- Widely adopted preservation metadata standard.
- Covers elements of representation information and preservation description information.
- Output is NOT created by hand; depends upon the output of tools who perform actions on your files.
- Record can grow over time, as preservation actions occur.
- Steep learning curve. ☹
- But various repository platforms; Archivematica, DataAccessioner and other tools/systems will create PREMIS records for you.

What does PREMIS capture?

PREMIS can capture:

- The program on which the file was created.
- The version of that program.
- The operating system on which that program ran.
- Who created the file.
- The rights associated with the file.
- When the file was ingested into the preservation system.
- Dates the file was validated.
- And more….

METS

Standard for packaging.

Wrapper for XML metadata – you put PREMIS, Dublin Core, MODS, etc **INSIDE** it.

Contains seven sections:

- Header
- Descriptive Metadata
- Administrative Metadata
- File Section
- Structural Map
- Structural Links
- Behavior

METS and PREMIS cover *most* of the metadata requirements of OAIS.

```
<mets>
  <metsHdr/>
  <dmdSec/>
  <amdSec/>
  <fileSec/>
  <structMap/>
  <structLink/>
  <behaviorSec/>
</mets>
```

The OAIS Reference Model, Packages, & Metadata

QUESTIONS?

OAIS in the Wild

A quick case study

An Introduction to RCAHMS

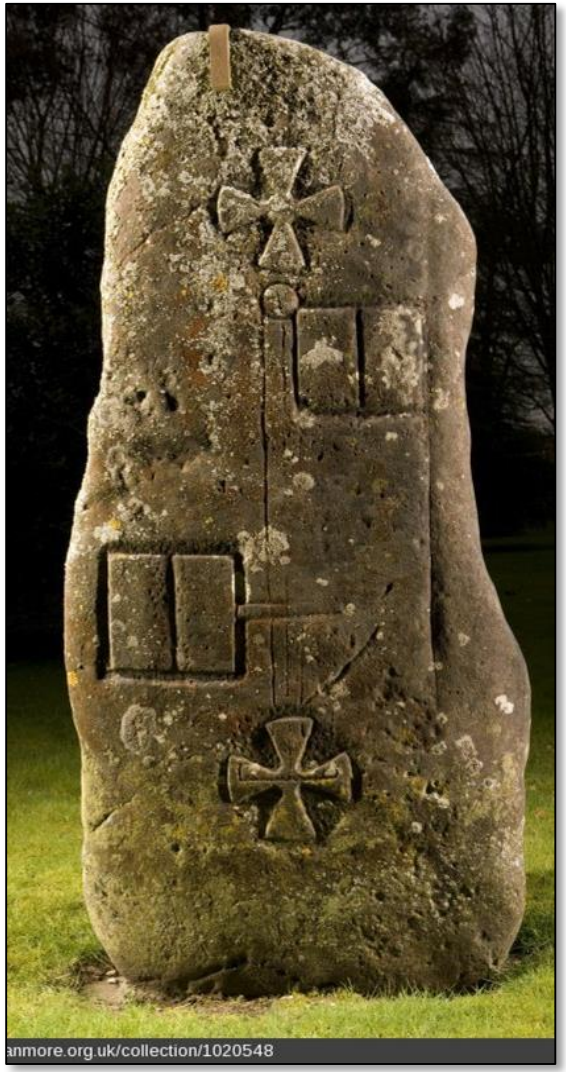
A medium-sized archive and survey institution based in Edinburgh

Mission to record Scotland's built heritage

Archive built from:

- Outputs of RCAHMS' own survey work
- Material collected from external depositors

https://canmore.org.uk/collection/1167305

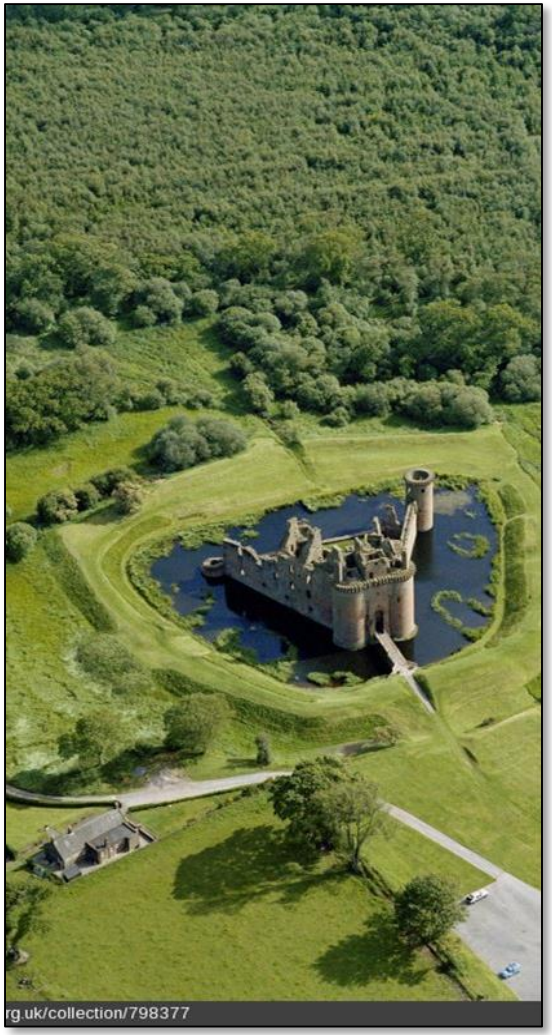


First Digital Archive

- First received digital data in 1992
- Report detailing preservation needs
- Contract to develop systems in 2003
- Limited standards and tools available
- Systems:
  - Area in database to record metadata
  - Dedicated storage area
  - Batch processing for digital images
- No preservation or dissemination systems

Motivations for Redevelopment

- New Digital Archivist hired
- Exponential growth of digital deposits
- Emergence of new standards and tools
- A more strategic approach to management and development required



A Plan of Action

Questions considered:

- What might success look like?
- How could buy-in be secured from stakeholders?
- How to identify useful standards and tools?

OAIS core to the process:

- Helped set aims
- Provided a framework to guide choices
- Used to carry-out a gap analysis

Gap Analysis: Ingest