University of Stuttgart
University Library

Longterm Archiving of Data & Software

Sibylle Hermann

Interface between library and researcher

What I do

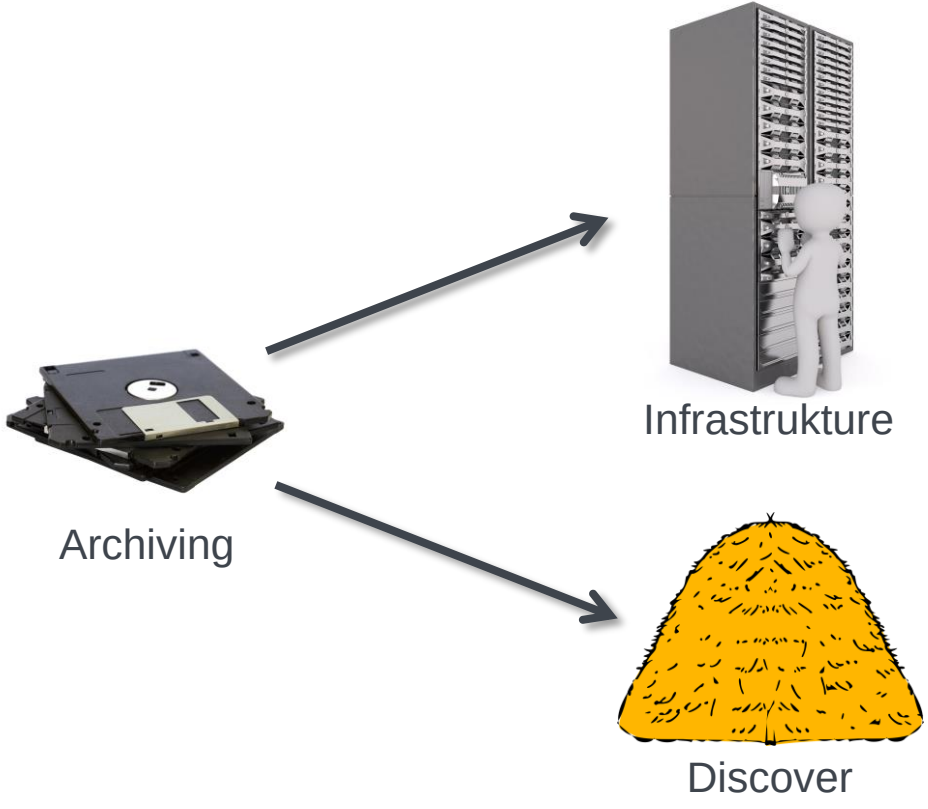
bwDataBib

Build an infrastructure for research data and software at the University of Stuttgart

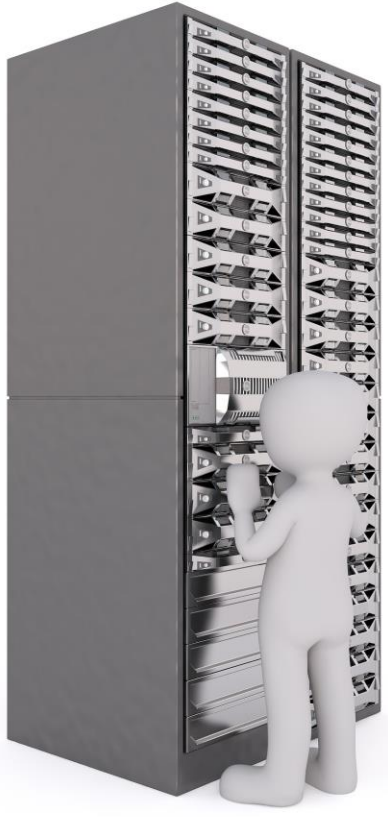
ReSUS?

SusI

Outline



Infrastructure



What would you like to have available in 10 years time?

Manage Digital Preservation

- Usability: intellectual content of the item must remain usable
- Discoverability: content must have logical bibliographic metadata
- Authenticity: provenance of the content must be proven
- Accessibility: content must be available for use

Byte Replication

LOTS OF COPIES KEEP STUFF SAFE

Decentralized and distributed preservation

Participating libraries acquire copies of important “stuff”

collect and **preserve** software in source code form

curate and make collected software accessible

Emulation as a Service

- establish a shareable infrastructure that provides on-demand access to old software
- recreating the original software environment on a current-day device
- enable access to at least 3,000 applications, including operating systems, scientific software, office and email applications, design and engineering software, and software for creative pursuits like video editing or music composition
- $1 million grants from The Andrew W. Mellon Foundation and the Alfred P. Sloan Foundation
- 2018-2020

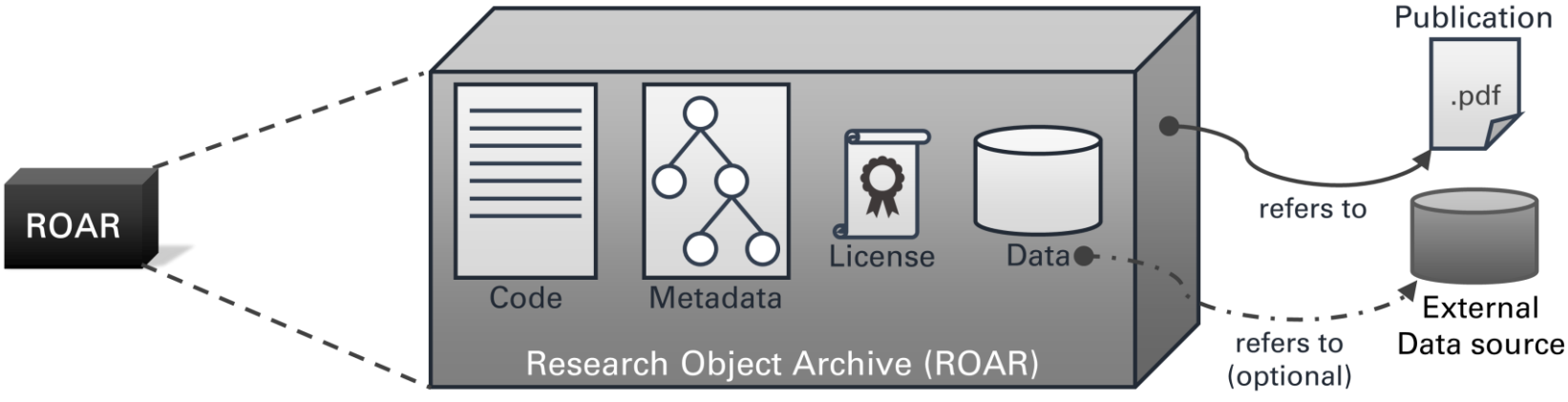
- http://eaas.uni-freiburg.de/
- http://www.softwarepreservationnetwork.org/eaasi/

Encapsulation

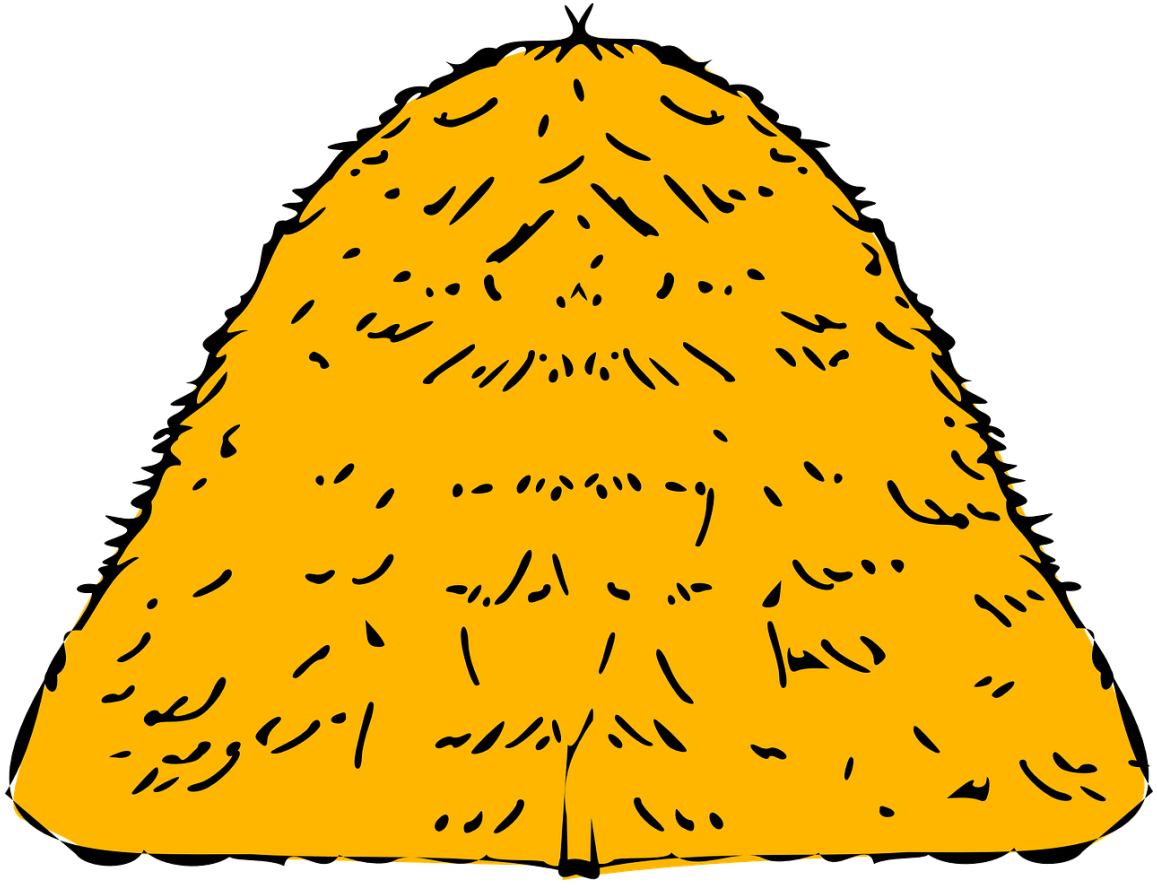
- grouping together a digital object and metadata necessary to provide access to that object
- encapsulate metadata with a digital object include reference, representation, provenance, fixity and context information
- considered as a key element of emulation

Research Objects

- Publish everything togehter so that



Discover



Publish

- Just what is published can be archived
- Credit for
  - code, software and scripts
  - data
  - Software Sustainability Institute
- Cite and reference
  - CITATION files
  - good metadata help others to find, use and credit the work

- Repositories

- Software Papers and Data Papers
- Publish with a DOI

Data Cite

- enabling easier access to research data
- strengthening the acceptance of research data as a relevant, citable component of the scientific track record
- support data archiving so that research results can be verified and reused
- development of standards, workflows and best practices

- DOI-Registration for Research Data
- global consortium supported by local institutions

DataCite
FIND, ACCESS, AND REUSE DATA

Data Cite: Metadata

Mandatory	Recommended	Optional
Identifier	Subject	Language
Creator	Contributor	Altrenate ID
Title	Date	Size
Publisher	Related Identifier	Format
Publication year	Description	Version
Resource Type (since 2016)	GeoLocation	Rights

Software mentions in publications

Mention Type	Count (n=286)	Percentage
Cite to publication	105	37%
In-text name only	90	31%
Instrument-like	53	19%
Cite to name or website	15	5%
URL in text	13	5%
Cite to users manual	6	2%
Not even name	4	1%

J. Howison and J. Bullard. Software in the scientific literature: Problems with seeing, finding, and using software mentioned in the biology literature. Journal of the Association for Information Science and Technology, 2015. In press. http://dx.doi.org/10.1002/asi.23538 .

Citing Software

Software Citation Principles

- Making software a citable entity in the scholarly ecosystem
- Metadata depends on the use case
- Unique Identifier, Softwarename, Author(s), Contributor(s), Version number, Location/repository, Release date, Indexed citations, Software license, Description, Keywords

Smith AM et al., FORCE11 Software Citation Working Group.(2016) Software Citation Principles. *PeerJ Computer Science* 2:e86. DOI: 10.7717/peerj-cs.86

CodeMeta Project

- Minimal metadata schema for science software and code, in JSON-LD
- Crosswalk between repositories

Jones, M. et al. 2017. **CodeMeta: an exchange schema for software metadata. Version 2.0.** KNB Data Repository. doi:10.5063/schema/codemeta-2.0

The benefit must be higher than the effort!

RePlay - DH

- Use Git for version control in the local workspace folder
- Assist user in elicitation of process metadata during/after every workflow step
- Use Git commit messages to store collected metadata in JSON format
- Allow user to navigate the workflow graph and/or export workflow information
- Allow user to export results to local repository

Metadata

Process metadata

- Used to describe interrelation between items
- Developed together with the researcher
- Can be adapted for different communities

Object metadata

- For publishing: DataCite
- Describing each item

https://GitHub.com/RePlay-DH/

Software Licence

- Unlicensed code is closed code, so any open license is better than none
- Licences can be used to facilitate access as well as to restrict it
- Who is the legal rights owner: researcher or institution
- Which License to choose?
- Permissive vs. copyleft

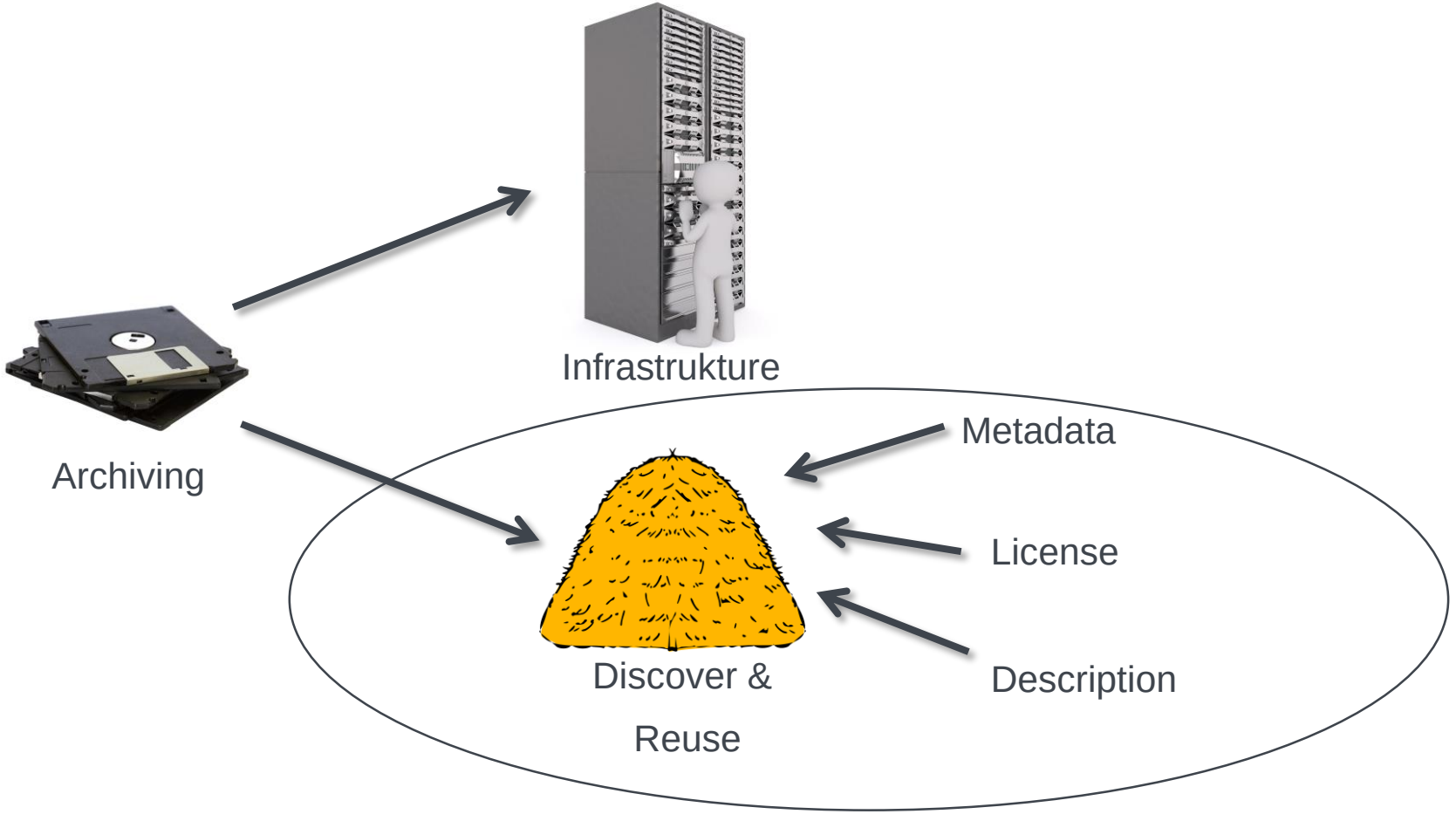
- https://choosealicense.com/

Research Data Licence

- DataJus – legal framework for research data (Germany)
- Who is the owner of the data?
- Copyright protected? → individuality
- Qualitative vs. quantitative data
- Case-by-case examination without legal certainty
- Data protection
- Patent law
- https://ufal.github.io/public-license-selector/

- How to licence research data

Summary


Infrastrukture
Archiving
Metadata
License
Description
Discover &
Reuse

Conclusion

- Archiving is more than infrastructure
- Discover and reuse need metadata and describing
- There are ideas how long-term archiving can work
- Reuse needs Licensing
- Important to know: legal rights owner

Librarians are the secret masters of the world. They control information. Don’t ever piss one off.

Spider Robinson