EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Seminar für Sprachwissenschaft

The BagIt Format

Claus Zinn

CLARIN 2021 Center Meeting, 10.06.2021

CLARIN

Motivation (from the researcher‘s perspective)

- How to transfer data *safely* between researcher and archive:
  - Complete (all files should make the transfer)
  - Valid (each file should preserve its integrity/nature)

- Need for something better than “traditional” methods:
  - E.g., all data is on a CD, a USB drive, a cloud space
  - E.g., all data is attached to this email
- Ideally, have a tool to construct such “bags”
  - Bagman, see https://weblicht.sfs.uni-tuebingen.de/bagman/
    - (also generates CMDI)

Motivation (from the archive‘s perspective)

- How to best receive data *from* researchers:
  - Get a single (digital) parcel with a description what’s in it
  - Open the parcel and see whether it is complete and not corrupted
    - Use tool support for bag validation
  - Once bag is validated, start ingestion process with repository system
  - Ideally, have a good CMDI to come along as well
    - (Bagman)

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

BagIt Specification

- BagIt File Packaging Format (v1.0)
  - https://datatracker.ietf.org/doc/html/rfc8493
- A set of hierarchical file layout conventions for storage and transfer of arbitrary digital content.
- A “bag” has just enough structure to enclose descriptive metadata “tags” and a file “payload” but does not require knowledge of the payload’s internal semantics.
- Suitable for reliable storage and transfer
- Developed and used by Library of Congress
- Has strong integrity assurances as it uses cryptographic-quality hash algorithms (checksums) – in constrast to ZIP, TAR
- “Bag it and tag it”

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

BagIt Layout Conventions

```
<base directory>/
|
+-- bagit.txt
|
+-- manifest-<algorithm>.txt
|
+-- [additional tag files]
|
+-- data/
|     |
|     +-- [payload files]
|
+-- [tag directories]/
      |
      +-- [tag files]
```

optional: *bag-info.txt*

optional: tagmanifest-sha256-txt

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

BagIt

3. Complete and Valid Bags

A _complete_ bag MUST meet the following requirements:

1. Every required element MUST be present (see Section 2.1).

2. Every file listed in every tag manifest MUST be present.

3. Every file listed in every payload manifest MUST be present.

4. For BagIt 1.0, every payload file MUST be listed in every payload manifest. Note that older versions of BagIt allowed payload files to be listed in just one of the manifests.

5. Every element present MUST conform to BagIt 1.0.

A _valid_ bag MUST meet the following requirements:

1. The bag MUST be _complete_.

2. Every checksum in every payload manifest and tag manifest has been successfully verified against the contents of the corresponding file.

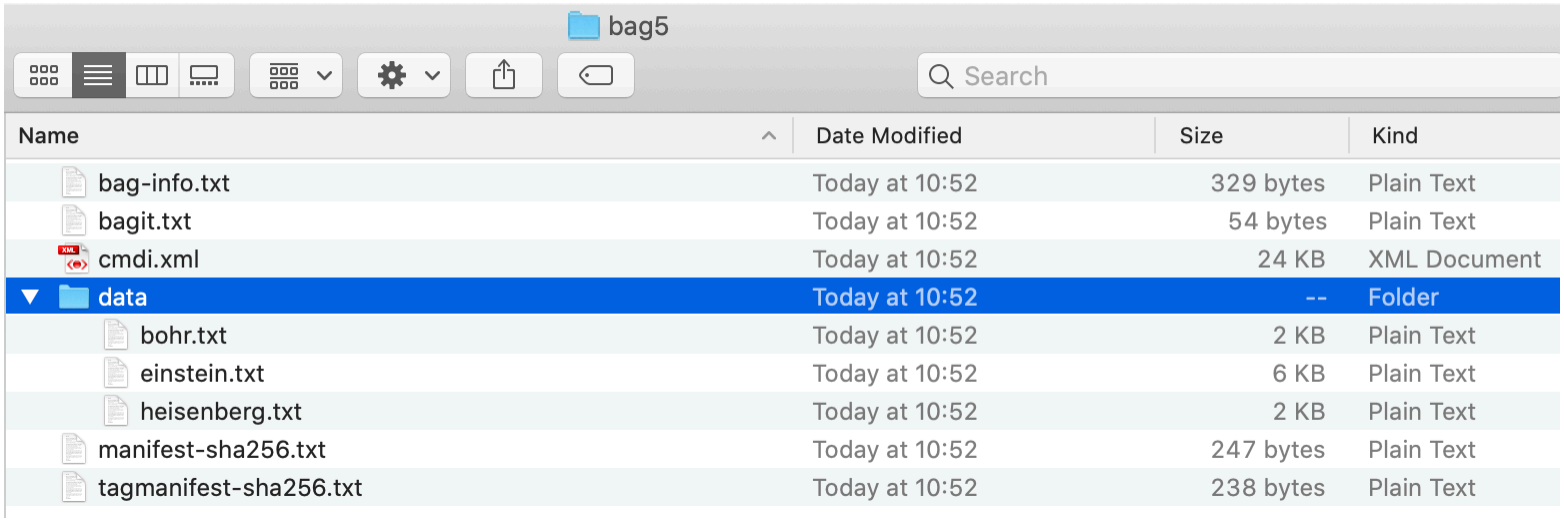
Bagit tools

- https://github.com/LibraryOfCongress/bagit-python
  - Command line tool, or
  - Programmatically inside **Python** programs
- https://github.com/LibraryOfCongress/bagger
  - **Java-based** GUI to BagIt specification
- https://github.com/LibraryOfCongress/bagit-java
  - **Java-based** library to support the creation, manipulation and validation of bags.
- https://github.com/LibraryOfCongress/bagger-js
  - Pure **Javascript** implementation of the BagIt specification, simple web app for bagging and transfering

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Bagging tool (python)

- *python3 -m bagit scientists*
- Bags the directory scientists:
  1. Move the contents of scientists to data directory (payload)
  2. Then creates the bag-specific .txt files
  3. That's it. No additional container



EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

BagIt – files generated (by Bagman)

```
Source-Organization: Eberhard Karls Universität Tübingen
Contact-Name: Ada Mustermann
Contact-Phone: +49 (0) 7071-29 73968
Contact-Email: claus.zinn@uni-tuebingen.de
Description: Second Language Acquisition in Parrots
Bagging-Date: 2021-06-09
BagIt-Version: 0.97
Tag-File-Character-Encoding: UTF-8
Bag-Count: 3
Bag-Size: 10 KB
```

```
BagIt-Version: 0.97
Tag-File-Character-Encoding: UTF-8
```

```
802736b1ad298a2bafcef9c678f2c27095d29097a32c723ba68a8be09c98bb13 data/bohr.txt
9aa10330eda9b69f27c887b4344946c4e5a3b387b3d6d46b5a1706b5674b7cda data/heisenberg.txt
7344b01bd8e7a907b63ef6708bd9703137c1c5d71ca559a7e1ed4d30388906e8 data/einstein.txt
```

```
ea4e4b66e0a092d5184b86f75699a26a7f772d8dd81e5f83f7658376acd3bf70 manifest-sha256.txt
1cc6c37de491b806aa40455d47f095c24f9796c304b29565ed9d0693accb26f5 bagit.txt
fff2d9e645dc50d9ee232eb7c510e1232263cc740743632458b22de1df5aada9 bag-info.txt
```

EBERHARD KARLS
UNIVERSITÄT
TÜBINGEN

Validation tool (python)

```
Clauss-MBP:Downloads zinn$ python3 -m bagit --validate  bag5
2021-06-09 14:48:12,893 - INFO - Verifying checksum for file /Users/zinn/Downloads/bag5/data/bohr.txt
2021-06-09 14:48:12,893 - INFO - Verifying checksum for file /Users/zinn/Downloads/bag5/data/heisenberg.txt
2021-06-09 14:48:12,894 - INFO - Verifying checksum for file /Users/zinn/Downloads/bag5/data/einstein.txt
2021-06-09 14:48:12,894 - INFO - Verifying checksum for file /Users/zinn/Downloads/bag5/manifest-sha256.txt
2021-06-09 14:48:12,894 - INFO - Verifying checksum for file /Users/zinn/Downloads/bag5/bagit.txt
2021-06-09 14:48:12,895 - INFO - Verifying checksum for file /Users/zinn/Downloads/bag5/bag-info.txt
2021-06-09 14:48:12,895 - INFO - bag5 is valid
```