CD-Ark

A Tool for Cooperative Processing of Optical Discs

Zdeněk Hruška
The Moravian Library
Czech Republic
zdenek.hruska@mzk.cz
0000-0001-7606-5693

Abstract – A valuable and unique part of library collections, certain types of optical discs are becoming obsolete and face a variety of risks ranging from physical damage to loss or theft. In order to protect the content of optical data discs in Czech libraries, a collaborative software tool was developed. The CD-Ark software is designed to store data from the discs, creating data packages that include an ISO disc image alongside with a checksums, as well as technical and bibliographic metadata; the whole package is uploaded onto a joint server, ensuring bit-level preservation.

Keywords – Optical discs, Cooperation, Iso image, Bit-level preservation, Libraries.

Conference Topics – Collaboration; Building Capacity, Capability and Community

I. Introduction

Though in the past they seemed an attractive option to store information, nowadays it is obvious that CD and DVD discs do not age well. They are prone to physical damage due to poor storage conditions and easily lost or stolen. It is mainly the data stored on the disks that needs to be preserved for future generations, but the carrier itself has a certain cultural value and its image, too, should be conserved. However, up until very recently little attention has been paid to the issue of their protection and long-term preservation in spite of the large number of optical disks in the collections of libraries across Czech Republic. The number is not accurate due to imprecise catalogue records, but out of tens of thousands optical discs throughout Czech Republic, the Moravian library alone stores around 25,000. Therefore a project CD-Ark was launched which developed a software that would process their content and ensure its long-term protection.

II. CD-Ark

CD-Ark is a set of tools which enables cooperative processing of optical data discs (CD-ROMs and DVD-ROMs in particular) and the storage of the bit copy on the central server. It is designed mainly for large libraries that can collectively process optical data disks they hold in their collections. Duplicity is avoided at the moment by the system checking the bibliographic metadata and the disc image checksum.

As far as legal issues are concerned, the project was carried out in accordance with the Copyright Act as well as the Library Act. According to Article 18 of the Library act, libraries are under obligation to ensure the protection of the library collection, even by the conversion of library documents to another type of carrier, if necessary. Therefore, as the law stands, libraries can make copies of carries for archival use.

CD-Ark consists of two software applications: the CDArk-client and the CDArk-server which are described below in greater detail.

III. Workflow

In collaborative effort, libraries can achieve better results and the processing of optical disks is made more efficient. This is the very premise on which CD-Ark is built. With more libraries involved, the data producer is identified by a unique identification number that all Czech libraries have.

How does processing optical discs work with this set of tools? The CD-Ark-client (being a desktop application) is installed on a workstation in institutions

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.

that take part in the project who also have disk space allocated on the central server. First, a metadata record is downloaded from the electronic catalogue, followed by the CD-Ark-client application software creating a data image of the optical disc (the so-called ISO image based on ISO 9660). As the carrier itself is a cultural artefact, the top of the disk alongside with the booklet and cover are scanned in order to preserve the optical disc's visual appearance. Once all is in one place, a package of data along with the checksums is sent to the central server for further processing.

Here, on the CD-Ark server, output from DROID (a file format identification tool) is added into the data package. An OCR (in TXT and ALTO) is created from the scans and finally everything is wrapped in BagIt (the standard data package format created by the Library of Congress for data storage and transfer). Subsequently, all this data is stored on the server and the checksums enable regular automatic checks for data corruption.

IV. Conclusion

Thus at least bit-level data protection is solved. That being said, the problem of logical long-term preservation still needs to be addressed. Now, however, it is crucial that we transfer the data to a more reliable medium; not only are the optical discs at risk of physical damage, loss or theft, but some are inevitably approaching the end of their lifespan. For future long term preservation, an open source LTP system ARCLib, is currently being developed in collaboration of various Czech libraries. Hopefully having tackled problems arising from the logical preservation of certain formats (such as software and games), the system is planned to ensure logical data preservation.

A long and arduous journey lies in front of us; extracting the data from the discs was no more than the first step leading towards the long-term protection of optical disc contents.

References

[1] Act No. 121/2000 Coll., on Copyright and Rights Related to Copyright and on Amendment to Certain Acts. 2006 [Accessed: 17-Sep.-2019]. Available: https://wipolex.wipo.int/en/text/137175.

[2] ARCLib: Popis projektu, c2019 [Accessed: 17-Sep.-2019]. Available: https://arclib.cz/popis-projektu/.

[3] DROID: (Digital Record and Object Identification). GitHub [online]. 2018 [Accessed: 17-Sep.-2019]. Available: https://digital-preservation.github.io/droid/.

[4] CDArk-client. GitHub [online]. c2019 [Accessed: 17-Sep.-2019]. Available: https://github.com/moravianlibrary/CDArcha-client.

[5] CDArk-server. GitHub [online]. c2019 [Accessed: 17-Sep.-2019]. Available: https://github.com/moravianlibrary/CDArcha-server.

[6] ISO 9660. The File Format Wiki [online]. 2019 [Accessed: 17-Sep.-2019]. Available: http://fileformats.archiveteam.org/wiki/ISO_9660.

[7] J. Kunze, J. Littman, L. Madden, E. Summers, A. Boyko, and B. Vargas, The BagIt File Packaging Format (V0.97) [online]. 2016 [Accessed: 17-Sep.-2019]. Available: https://tools.ietf.org/id/draft-kunze-bagit-14.txt.

[8] Law No. 257/2001 Coll. of 29 June 2001 on Libraries and Terms of Operating Public Library and Information Services (Library Act) [online]. 2001 [Accessed: 17-Sep.-2019]. Available: https://ipk.nkp.cz/ipk-en/legislation/02_LibAct.htm.

16th International Conference on Digital Preservation
iPRES 2019, Amsterdam, The Netherlands.