# DATA MANAGEMENT PLAN

of the

# Center for Interdisciplinary Research and Documentation of Inner and South Asian Cultural History at the University of Vienna

Document Version and Date: V2.0, 22.07.2016

Loosely based on the Data Management Plan for projects at the University of Vienna . Version 2.0 (Phaidra o:407974)

# Table of Contents

# 1 | Preamble

The Data Management Plan of the Center for Interdisciplinary Research and Documentation of Inner and South Asian Cultural History at the University of Vienna (CIRDIS) aims at codifying strategies and concepts for the preparation, archiving and publication of original research data from the greater Himalayan area in CIRDIS' online archives. It is furthermore conceived as a general guideline for researchers and projects affiliated and/or cooperating with the research center now or in the future (cf. "How to cooperate with us").

CIRDIS has been gaining respectable experience and know-how in creating and working with digital archives, the needed infrastructure, connected technical and legal issues and common pit-falls (a significant amount of it has been gained by starting early and making a lot of mistakes).

As openness and preservation of research data and metadata has become more and more important for digital societies, CIRDIS declares itself a strong supporter of these emerging ideals, concepts and standards and strives to adopt them wherever feasible.

# 2 | Archives, Data Collections and Metadata

CIRDIS' Western Himalaya Archive Vienna (WHAV) is home to a collection of around 130.000 analogue and visual assets – mainly mounted slides – largely resulting from research travel made by scholars since the 1990s. The homonymous online image database[1] was one of the first large image databases of its kind and, early on, made large parts of the WHAV's physical collection available digitally. It has since been continuously maintained as a research database for images, more and more of them originally digital since the early 2000s.

The Himalaya Archive Vienna (HAV)[2], modern multimedia counterpart and part-successor to the WHAV, is currently under development and will provide modern functionality and flexibility in the structuring and visualising of all our multimedia assets.

While ingesting and representing the largest part of the materials from the WHAV as one of its future collections, the WHAV will remain as a platform to maintain citeability and exclusively house some legacy data not suitable for the HAV. To ensure seamless interoperability with the HAV, sensible integration with PHAIDRA and the largest possible openness of our research data, we have initiated a large-scale, multi-step process of raising the metadata quality and clarifying the licensing situation in the WHAV.

The materials forming the initial data-sets to be integrated into the HAV have been collected in the course of field trips to Inner and South Asia and consist of images, video and audio recordings, annotations and field notes. The materials collected in the last decade is almost entirely original digital data and often provided with media-inherent metadata such as time-codes, GPS data etc.

[1] https://whav.aussereurop.univie.ac.at
[2] http://www.univie.ac.at/cirdis/documentation/himalaya-archive-vienna-hav

## a) TYPE AND AMOUNT OF DATA

The HAV is expecting a first series of new data-sets in the order of 1.3 TB. Additionally a subset of the visual material already archived at the WHAV (at the time of writing the WHAV is hosting ~130.000 images, about 6TB of image data) will also be integrated into the HAV. The initial new data sets consist of heterogeneous corpora of largely ethnographic research data:

- a collection of original documentation from field research trips of Prof. Martin Gaenszle from the 1980s onward (150+ audio tapes, photographs and field notes) equal to an estimated total of 300GB digitized data
- the field research documentation of the FWF Project "Ritual, Space, Mimesis: Performative Traditions and Ethnic Identity among the Rai of Eastern Nepal" (P23204-G15); digital data totalling ~500 GB
- a subsection of the field research documentation of the FWF Project "Text, Art and Performance in Bon Ritual" (P24701-G21), in particular large parts of the audiovisual documentation comprised of original digital data totalling ~500GB

## b) FILE FORMATS AND TYPICAL SIZES

Lossless and/or open and file formats are to be preferred for the creation of new data. Since we are unfortunately not always confronted with suitable file formats from the start (e.g. legacy digital data, formats dictated by the recording devices etc.) we aim at archiving either the original file and/or a lossless derivative of it.

We recommend (preferences marked) the following file formats:

Images: JPEG, SVG, PNG, GIF, TIFF, DNG;
Audio: AIFF, FLAC, MP3, OGG, WAV;
Video: AVI, MPEG2, MPEG4, VP8, VP9, WMV;
Text: DjVu, DOC, ODF, PDF/A, TXT;
Structured Text: HTML, JSON, Markdown, TEI, TEX, XML;
Tables: CSV, ODS, TSV, XLS;
Databases: MySQL, MariaDB, PostgreSQL

## c) METADATA

The more extensive the data is described, the easier will be its retrieval, analysis, reuse and meaningful presentation. The implementation of standardized vocabularies and classifications (like ÖFOS, Eurovoc, ACM or Getty) further help in making the data more visible and reusable.

Metadata Structure and Standards

CIRDIS is currently in the process of analysing its initial data corpora and defining a set of metadata suitable for representing said data and its complex multidimensional inter-relations. Basic Dublin Core information, guaranteeing the interoperability with Phaidra[3] and other

[3] Phaidra, short for "Permanent Hosting, Archiving and Indexing of Digital Resources and Assets" is the University of

national and international virtual libraries such as Europeana[4] and BASE[5], will be extended by a domain-specific metadata model designed to fit the needs of our material and the multiple scientific disciplines involved in CIRDIS' research. Special attention will be paid to:

- ensure the largest possible intersection between our metadata model and existing metadata standards in order to ease translation of our metadata and thus interoperability with other systems;
- incorporate means to enable easy extendability of the metadata information the HAV will be able to handle in order to provide flexibility for collections to be integrated in the future.

All initial data-sets have largely been assessed and – despite being provided with different levels of detail in their respective metadata – satisfy a common core of information needed to be archived and meaningfully represented in the HAV and Phaidra.

# 3 | Archiving, Storage and Preservation

CIRDIS pays great attention to ensure that no data is lost, that everything is secured long-term and that only authorized users have access to modify its metadata.

## a) Hardware and Software

CIRDIS promotes and uses free and open software solutions wherever possible, which, thanks to the flourishing the open source movement experienced over the last decades, now practically spans the entire range of server and end user software. Whenever the need for the usage of proprietary or commercial software arises, CIRDIS pays particular attention to ensuring all data and metadata produced or manipulated with such software can be saved or exported to standardized, lessless and possibly open formats (cf. 2.a point 1) suitable for long-term preservation to ensure re-usability of data and metadata in the future.

The online multi-media archive HAV will – like the WHAV – be based on a LAPP-stack consisting of the open source software Linux, Apache, PostgreSQL and Python and is itself aimed at being released under an open software license.
CIRDIS' online archives are currently running on a dedicated web and storage server owned and administered by CIRDIS and professionally housed at the University Vienna. The server is currently equipped with a quad-core Intel Xeon (Nehalem), 8GB of DDR3 memory, 150GB SSD Raid1 + 9TB HDD RAID5 + hotspare and redundant power supplies.

CIRDIS can support ongoing and future field-work endeavours with a small stock of high grade semi-professional gear such as photo equipment a professional hd video camera, basic sound recording equipment and gps trackers.

CIRIDS at time of writing runs its own scanning workstation equipped with a high grade slide scanner for 35mm slides, 35mm film strips and APS film (Nikon Coolscan LS5000ED + slide feeder) and a

Vienna's digital asset management system with long-term archiving functions: https://Phaidra.univie.ac.at/.

[4] Europeana – think culture (http://www.europeana.eu)

[5] Bielefeld Academic Search Engine (https://www.base-search.net/)

professional monitor including colorimeter. Additionally flatbed scanners as well as a duplex document scanner are available.

## b) STORAGE AND BACKUPS

All the WHAV's and HAV's data together with the archives' database dumps containing - among other web application specific information - the accompanying metadata, is secured by daily differential backups. Here CIRDIS is making use of one of the many professional IT-infrastructure service provided by the Vienna University Computer Center (ZID[6]). The implementation, operation and monitoring of and, in case of an incident, recovery from the backups are covered by CIRDIS' IT-administrator. To prevent unneeded downtime or a lockout in case CIRDIS' IT-administrator should be unavailable a secondary administrator's access is available to CIRDIS' long-term collaborator and lead developer.

At the moment our server has a free storage capacity of around 3TB remaining. We are currently in the process of preparing the integration of a large part of legacy data from the WHAV into Phaidra, which will free up a significant amount of our storage capacities as the high-resolution/high-definition digital source material together with its accompanying metadata will be hosted by Phaidra. Of these files, CIRDIS' web and mass-storage server will only store a number of lossy derivatives and extended domain-specific metadata to needed for its specific archiving functionality and presentation on the web (e.g. JPEG for images; MP3 for audio; MP4, WEBM for video).

At the same time, with our current transition to a multi-modal, multimedia-archive, the HAV, we are faced with heterogeneous sets of research data and a much more diverse body of source file formats. A significant amount of these will have to be transcoded into a suitable format for long-term preservation in Phaidra (e.g. proprietary RAW-images, AVCHD streams,..). Wherever this is the case, we are striving to retain and archive the original source formats on our server in order to maintain maximum flexibility and avoid information loss in the transcoding process (as is the case in the transition of RAW-image formats to high quality TIFs suitable for long term preservation). In order to support this effort an extension of our storage capabilities will be of necessity. We are to combine this upgrade with the needed exchange of hard disk drives latest by the end of 2017, as the first of them will have reached their End of Life at this point.

## c) ACCESS AND SECURITY

It is CIRDIS' goal to have most of its data published fulfilling the legal and technical prerequisites of Open Research Data, thus being openly accessible by definition. Nevertheless, certain subsets of data will need to – be it temporarily or permanently – remain subject to strict access restrictions (due to restrictive copyright situations, moral or ethical considerations, ongoing embargo periods). All but a small subset of legacy data with either unclear or very restrictive copyright situations will be permanently archived in Phaidra.

CIRDIS' is taking a number of precautions in securing the WHAV and HAV server and the data and metadata it is hosting. Our server is physically securely located through the server housing facilities of the University of Vienna's Computer Center. CIRDIS' provides its own, dedicated IT administrator responsible for regular software and hardware maintenance, ensuring timely security updates, regular

[6] Zentraler Informatikdienst: https://zid.univie.ac.at/en/home/

backups, user access management, monitoring and additional security measurements. The server is secured by a two-tier firewall (institutional and local), intrusion detection and monitoring systems, and strong and secure cryptographic authentication and authorization systems.

All web-traffic is encrypted through a securely implemented SSL-connection based on modern, well documented, well tested and open software solutions. The same holds true for the authentication and authorization mechanisms implemented for the internal user area of the WHAV's and HAV's web interfaces.

Different levels of access to both the front- and backend of our online archives are realised via user-roles, this ensures that our collaborators can access their data even in the cases it can not be openly published (yet).

CIRDIS' IT staff upholds similar measures for the desktop computers and laptops used at the center and makes sure security relevant updates are being deployed on time and daily backups are in place.

### d) WHAT IS THE LONG-TERM PRESERVATION PLAN FOR THE DATASET?

With Phaidra as the host for our archival data, our infrastructural assets can be used for providing additional services such as preserving original source data not suitable for long-term storage (cf. “Formats” and “Storage and Backup”). Furthermore CIRDIS will focus on the individual, domain-specific needs and complexities our research data entails. This encompasses providing ease of use in the archiving and subsequent retrieval processes, meaningful presentation, developing research tools, increasing metadata quality as well as open data and linked open data efforts.

## 4 | Open Access and Data Sharing

In accordance with the reusability considerations of FAIR[7] both our data and metadata will be supplied under clearly specified and accessible (both to humans and machines) licensing in order to guarantee its open availability and reusability.

CIRDIS is currently in the process of making its metadata in the WHAV and HAV available under Creative Commons Attribution 4.0 International (CC BY 4.0). The larger part of the data is being prepared to be made available under Creative Commons Attribution-ShareAlike 4.0 International (CC BY-SA 4.0).

The WHAV and HAV will additionally support more open licences such as CC BY and CC0. On the other end of the spectrum – to address copyright and complex ethical and/or moral issues – a number of non-open licenses such as CC BY-NC-SA 4.0[8], CC BY-NC-ND 4.0 and “all rights reserved” are available in order to accommodate for various needs. In addition the HAV will provide the possibility of archiving materials with a predefinable embargo period. It so allows for the secure archiving of research data prior to its open release (e.g. while still being the object of ongoing research within a running project).

A small subset of legacy data and metadata in the WHAV will unfortunately have to remain “internal use only” for the time being due to either restricted copyright situations and/or ethical concerns. This

[7] The “FAIR Data Principles” as formulated by FORCE11 (https://www.force11.org/group/fairgroup/fairprinciples).

[8] Due to strong reservations to “commercial use” on the side of some of our local communities and informants (cf. ethical considerations) certain data will have to be released as Non Commercial (NC).

data will therefore, despite the possibility of locking down data sets in Phaidra, not be included. We believe misusing Phaidra as mere storage solution - an error we made ourselves early on - should be avoided.

## 5 | Ethical and Gender Considerations

Since projects associated with CIRDIS have to deal with the sensitive ethical issues of conducting research involving partners from different cultures and societies, care has to be taken to protect the personal rights of local consultants, collaborators, interpreters, institutions in the documentation process. All partners should be informed about the context and intentions of the research. For all materials that will be made accessible through our digital archives "informed consent" is mandatory. In accordance with the 'Ethics for Researchers'[9] we will only process data necessary for meeting our aims. We likewise adhere to the principles and rules as stated in 'The European Code of Conduct for Research Integrity'[10]. CIRDIS is paying close attention to any concerns about a possible commercialisation by third parties as well as any partner's wish to be protected by anonymity.

Gender issues are an integrate part of any research and documentation process and CIRDIS follows the principles expressed in the "EU Toolkit Gender in EU-funded Research".[11]

How to Cooperate with Us

CIRDIS welcomes all research initiatives working in the extended Himalayan area and wishing to preserve and share their data according to the policies of the open science community. If you want to profit from and contribute to our interdisciplinary approach and multi-modal archiving, CIRDIS can provide its expertise and infrastructure. We encourage you to contact us - we are always looking forward to collaborations strengthening the field of Himalayan studies through open and innovative ideas.

[9] http://ec.europa.eu/research/participants/data/ref/fp7/89888/ethics-for-researchers_en.pdf

[10] http://www.esf.org/fileadmin/Public_documents/Publications/Code_Conduct_ResearchIntegrity.pdf

[11] http://www.yellowwindow.com/genderinresearch/

# 6 | Imprint

CIRDIS' own Data Management Plan will be implemented by its authors i.e. CIRDIS' core team. Our Data Management Plan has been created in cooperation with internal and external collaborators. It will be maintained, revised and extended by CIRDIS core team.

Authors of this document:

| | |
|---|---|
| Jürgen Schörflinger, BA | e-mail: juergen.schoerflinger@univie.ac.at |
| Mag. Jan Seifert | e-mail: jan.seifert@univie.ac.at |
| Dr. Verena Widorn | e-mail: verena.widorn@univie.ac.at |

Contact us:

CIRDIS

Department for South Asian,
Tibetan and Buddhist Studies
Spitalgasse 2-4, Hof 2.1
A-1090 Vienna, Austria
Tel.: +43 1 4722 41474

| | |
|---|---|
| Univ.-Prof. Dr. Martin Gaenszle | e-mail: martin.gaenszle@univie.ac.at |
| Dr. Verena Widorn | e-mail: verena.widorn@univie.ac.at |