The FROGBEAR project is supported in part by funding from the Social Sciences and Humanities Research Council of Canada.

(Meta)Data Collection for FROGBEAR: Looking Back and Looking Ahead

Bruce Rusk
Department of Asian Studies, University of British Columbia
June 11, 2024

Introduction

The FROGBEAR repository, hosted at the Library of the University of British Columbia (UBC), contains (as of April, 2024) over 1,450 records, each with one or more digital objects (photographs, videos, audio, and other files) related to East Asian religions (University of British Columbia, 2024a). Most of this material was collected by field visits of the research clusters of From the Ground Up: Buddhism and East Asian Religions (FROGBEAR), a Partnership Grant project sponsored by the Social Science and Humanities Research Council of Canada and running from 2016 to 2024, while some was contributed by other partners. What makes this collection useful—beyond the fact that it includes media documenting some little-known sites and objects—is that it each record includes, in addition to the digital objects themselves, a standardized set of metadata that makes the material interpretable, accessible, and searchable for end users. The data and metadata are fully open access under a Creative Commons license and readable through an open API as well as the Library's web interface. The metadata is accessible via a REST API (JSON format) and the images can be displayed through IIIF (International Image Interoperability Framework).

This openness, for instance, allowed the FROGBEAR team to build a web front-end through which users can search and display the repository's contents, something that any developer could do, with no special permissions or access.

Making the material in the collection work with this system was not always straightforward. Creating appropriate digital objects involved learning techniques for identifying what to document, how to capture suitable images and videos, how to take useful notes, how to organize the files collected, and how to do all of this while working as a team with fellow researchers with different sets of skills and knowledge. Once data had been collected, team members needed to produce structured metadata to accompany it, according to standards based on those of the UBC Library.

Many of these aspects of data collection and metadata production were unfamiliar to the majority of the participants, who were principally graduate students and researchers in the humanities. This working paper seeks to describe some of the experiences in the creation of this collection and highlight some lessons learned from the process that could be helpful to future projects of a similar nature, whether in a related domain or in another field entirely.

The occurrence of the COVID-19 pandemic in the middle of the project's scheduled run inevitably made many of the planned activities impossible, and forced clusters to find alternative arrangements, largely through online activities. Some of these also involved the creation of database records, a fortuitous outcome that was not part of the original plan but suggests possible future avenues and ways of working.

This paper draws on one participant's experience as a cluster member, as a member of the project's Data Committee, and as a producer of training material, coordinator of metadata production and quality control, and in assisting with the creation of the web interface. It summarizes the project's practices and their outcomes in standards creation, preparation/training, execution in the field, and creating new tools such as the front-end interface. Finally, it suggests some lessons learned that could be of value for future projects.

Standards and Training

The FROGBEAR repository was planned with a set of objectives, and the field research operated under a number of constraints, all of which called for the use of clear set of standards. First, the data was, from the outset, intended to be open access, in accordance with the Tri-Council principles under which the SSHRC operates (Government of Canada, 2021). Second, the collection was intended to be sustainable, accessible well into the future and maintainable, to the extent possible, even if technical standards and practices change. Third, the data and metadata were designed to be accessible through as wide a range of devices and for as wide a range of users as possible, making them readable to both humans with a range of capacities and to digital systems that could analyze and manipulate the data and metadata.

To serve these objectives, the project set a number of standards for the data to be collected. First, all digital objects should, to the extent possible, be in non-proprietary formats that are well-documented and free from constraints such as patent and licensing encumbrances, as well as widely interoperable with current software. This meant that all images should be in an open format such as JPEG, TIFF, or PNG, even at the cost of some image quality when raw camera images were available, since these are generally either in camera manufacturers' proprietary formats or in Adobe's freely-usable but proprietary Digital Negative (DNG) format. Raw camera data is generally not suitable for display, and although some of the information it contains is lost when the image converted to a viewable format this loss is necessary to ensure that the images would be viewable. Even if the current formats are superseded in the future, it should be trivial to convert them to new ones. The same applies to video, where the output of many cameras is in proprietary formats that can be converted to open formats such as MP4. However, in the case of tabular data, the advantages of maintaining formatting and other features in Excel spreadsheets led to a different solution: keeping two copies of the data in each record, once in Excel (a proprietary if well-documented and widely-used format) and once in plaintext CSV (comma-separated values), easily read by most software. Setting these standards generally did not create much difficulty for participants, though some did not initially pay attention to the standards and submitted material in formats that could not be accepted and had to be reprocessed. Moreover, some kinds of data are more difficult to find an appropriate standard for, such as more complex databases and 360° video (which could also not be displayed with the UBC Library's web interface).

The more challenging standards to set and to maintain were those for metadata. The UBC Library hosted the material in its Open Collections, a large institutional repository based on the DSpace content management system (University of British Columbia Library, 2024). Because this is a large system that contains a wide range of different types of objects from a diverse array of sources, its cataloguing standards are somewhat general-purpose, and there is a limited scope to adjust them to the needs of a particular project. For FROGBEAR, this customization consisted mainly of selecting which fields to include (some are optional and did not need to be included in our metadata process), how to map these to what the participants collected, and how to format these to suit the needs and goals of the project.

The standards for this metadata were largely inherited from the technical constraints of the underlying content management system, of the metadata standards used (based in Dublin Core, with elements from Library of Congress cataloguing standards), in addition to the particular format of the Open Collection system (University of British Columbia Library, 2024).

Sometimes these formats do not closely align with the expectations of users. For example, each record is tagged with a Creator field, which is used to store the name(s) of the person(s) who produced the data—in this case, the maker of the digital objects such as photos or videos. There are also separate fields to record the maker's institutional affiliation, academic status (i.e., student, faculty, etc.). This structure works best with records that have a single maker, since each field contains a separate list and there is no way to link an individual name to a status or affiliation. Moreover, there is no affordance for linking individual digital objects to their respective creators. Because many FROGBEAR records contain objects created by multiple individuals, the accurate attribution of data can become impossible: a record would contain a list of names, of statuses, and of affiliations, with no way to connect them or to know which individual images or videos they are tied to. This is a negative side effect of an ill-fitting standard, albeit a minor one.

A greater and more systematic challenge was the handling of multilingual metadata. By its nature, the material collected by the project was multilingual in nature. Although the basic language for all metadata is English, most records would also include, at the very least, material in one East Asian language (for example, place names, the names of sites, and the names of religious figures or texts depicted). In many cases multiple languages were involved: for example, a copy of a Chinese Buddhist text in Japan might have a Japanese title that is a transliteration of a Classical Chinese title that could, in turn, be a translation of a Sanskrit name. Hence it would be appropriate to include all three languages (Japanese, Chinese, and Sanskrit) in addition to English, to ensure that a researcher or student interested in the text would find the record regardless of which language they used to search. Moreover, for some languages (notably Korean) there are multiple Romanization systems in common use, so it is important to set standards in this area as well.

After consultation with the UBC Library, the data collection team decided on best practices for metadata. In titles, English terms normally comes first, with Romanization and (if appropriate) Sinographic characters following in parentheses. However, some variation in this order is common in free text fields such as description because scholarly practice varies even within a given field, with some writers preferring to give an English term followed by an original language term in parentheses, others following the original language term with the original. And some metadata fields, such as Geographic Location, had existing standards (in this case from the Library of Congress) which mandate the form to use, in this case requiring the Romanized form of the name of administrative units (e.g., "Sichuan sheng, China" rather than "Sichuan province, China").

Another area where best practices had to be devised was in establishing references and cross-references within and between records. All the metadata is in plain text, with no external links and no internal links such as ones between the description of a series of images in a record and the descriptions of those images. As a result, conventions had to be established for maintaining consistency across a set of records, such as those produced on a single visit or at a particular site. In the repository, each record is a self-contained entity, and there is no internal means of defining groups of records other than through consistent naming practices. To address this, general standards for file naming and for referencing specific images (using leading numbers in the filenames) were established. This solution worked, but being entirely manual and requiring manual renaming of files it was labour-intensive to carry out consistently.

Establishing relationships across records was even more challenging, and led to even more *ad hoc* solutions such as sequential numbering of series of objects (e.g., for a sequence of niches in a cave complex) and/or instructions in the Description field to "See…" a record identified by title (which could, then, be found by manually searching the database). This approach is inefficient for the user, but the open nature of the database would allow a developer to create a more structured external presentation by drawing on the relevant subset of data.

A final set of standards issues relates to rights and permission: UBC Library policy permits only the inclusion of material that is free from copyright encumbrances in its Open Collections, so all the data collectors had to agree to release the material they collected under a Creative Commons License. Likewise, no recognizable representations of living people could be included (viz., images/videos showing their faces or recordings of their voices) without the permission of the individual.

To enable the participants in field visits to produce appropriate, high-quality, and well-documented records, the FROGBEAR team produced a set of documentation and offered training on data collection and metadata authoring. This included a wiki (University of British Columbia, 2024b) and a YouTube channel (UBC Frogbear Project, 2018–2024) introducing the basics of photography, of working in a team to document a site, and of metadata authoring, as well as a template spreadsheet for metadata, which included sample records. Creative Commons licensing and image release documents were also included. Much of this material was available in multiple languages (English and Chinese, at minimum). These materials were updated from time to time based on experience with the field visits.

Field Visit Execution

The fifteen clusters (in two phases, 2016–19 and 2020–23) planned a range of field activities, mainly scheduled for the summer, and a number involving the collection of data at religious sites in East Asia, plus a short trial in which a group traveling to Japan collected materials and field-tested draft processes. The clusters were led by faculty affiliated with the project and field visits included both researchers and graduate students from a variety of institutions. Student training was an important objective of the project, and that included training in data collection and metadata production.

The Data Committee recommended that each team running a field visit plan its data collection beforehand and incorporate training on data collection and metadata production. Recommended planning included identifying key contents to document to the extent this could be known in advance, listing tasks that would need to be performed and the roles that team members would need to take on, and inventorying and acquiring equipment such as computers, cameras, lighting, and measurement tools. It was also recommended that time be allocated to each step of the process: training before on-site work began, collection of data, and production and editing of metadata (ideally, interspersed with data collection).

In practice, many clusters found it challenging to balance the time needed to prepare and to train their participants during the short time in the field with the other demands of the project: familiarity with the scholarship on the topics and sites being studied and time studying (rather than documenting) the material. Many clusters found it difficult to integrate work on organizing the data collected and drafting metadata into the field visits' daily schedule unless a significant amount of time was set aside for these purposes. It was also a challenge to maintain consistency of standards among the teams into which each cluster was divided, for example in the formatting and wording of metadata.

In general, clusters found it most effective to divide into smaller teams (roughly 3 to 6 people). Depending on the scale of the sites being investigated and the nature of the research topic, some spread out over a single site such as a temple complex while in other cases each team would go to individual locations such as widely-dispersed small shrines. In either case, it was helpful to assign roles such as photography, lighting, note-taking, transcription, measurement, and so on. Although the metadata needed to ultimately be entered into a database, no more efficient method for taking field notes was found than paper notebooks. For recording geographic locations, the GPS functions on mobile phones proved sufficiently accurate, so long as consistent notes were taken. To track complex sites or sequences of objects (such as a large set of inscriptions or sequence of icons), a handheld whiteboard and dry-erase marker were useful: the whiteboard could be photographed with a note about the object before the object itself. The photo of the whiteboard would not become part of the repository but assisted in sorting and classifying the images.

The process of organizing and annotating the data collected by the field visits to make it usable for the repository proved to be more difficult than many of the clusters had expected. The time required was significantly greater than most groups had budgeted for—the person-hours required to sort through and produce metadata for the data collected were, we estimated, more than twice what it took to collect the material while in the field. Doing this work in the field ideally could have involved a shorter time each day spent on data collection and a longer time on metadata production, but this would have extended the length of trips, and thus their cost, very significantly. Alternatively, some clusters focused on creating fewer records but documenting them more carefully, and this approach was often effective.

For clusters whose metadata production was not completed during the field visits, much of the metadata work was done afterward, when the participants had scattered and could online communicate online and mainly asynchronously, as they had to communicate with FROGBEAR staff, one another, and the UBC Library, often in a chain back and forth; some participants became unreachable and as a result some of the material could not be used. This was generally slow and inefficient, compared to work on site immediately after data collection.

Alternative Field Visit Models

Some of the data collection did not follow the main model for research clusters. For example, in 2017 Cluster 1.2: Religion and Technology did not focus on a single site of historical importance but sent small teams to visit a range of small temples and altars in northern Taiwan, collecting photographs and precise GPS coordinates. This created a different kind of dataset from many of the others, but also helped train the participants—mainly graduate students—in the collection of geographic data and the production of useful notes even on sites where they had limited information about each site. The focus on training rather than fundamental research questions did not detract from the quality of the data collected, but rather led to the efficient production of a large set of high-quality records.

The COVID-19 pandemic made field visits for most of the Phase 2 period impossible or impractical, and forced the cancellation or rethinking of many planned activities. However, it also created the opportunity to trial other approaches, such as "virtual field visits" employing previously gathered material in student training and metadata. Cluster 3.4: Typologies of Text-Image Relations organized an online training workshop in which students attended seminars on Buddhist sites in the Sichuan area then worked with a large collection of thousands of photographs taken mainly by Professor Christoph Anderl (University of Ghent) to organize and annotate them. This depended also on Professor Anderl's kind agreement to release his photographs under an open access license. This produced hundreds of records meticulously analyzing the iconography of the many stone carvings of these important sites. The participants joined a training session on FROGBEAR metadata standards and consulted with FROGBEAR personnel and subject matter experts during the process; as a result the material they produced was of high quality and consistency. In addition to creating the opportunity for research work when travel is not feasible, such "virtual visits" are a useful approach for lower-cost training and can also serve to "crowdsource" the analysis and publication of field data collected by an individual researcher. The contrast between the relatively short time it took to collect the photographs (one person spending a single day at each site, for the most part) and how long it took to produce detailed metadata (multiple people working for hours on each record) is likewise a reminder of the imbalance between the different stages of the process.

Another sizable body of data was contributed by a third party, a graduate student (Hannibal Taubes, University of California, Berkeley) who had over the course of several years documented a number of temple sites in North China, with a focus on their interior murals. Many of these sites are previously unstudied or not well documented, and many are under threat of destruction—indeed some were destroyed within a few years of being photographed, for example by thieves stealing the murals by peeling them from the temple walls. To preserve this valuable material, FROGBEAR worked with the student to organize and create metadata for dozens of these sites, ensuring that a photographic record would be preserved, along with detailed iconographic and art historical analyses. This was a welcome and unexpected addition to the project's repertoire, but it fit the aims and framework of the project, as well as coming at a time when there was additional capacity to ingest such material because the number of regular field visits was greatly reduced.

Lessons Learned

The experience of data collection in the FROGBEAR project suggests a number of lessons for similar undertakings, most of which will be familiar to those working in related fields. First, any project involving data collection with humanists should devote time in advance to the creation of standards for data and metadata, and best practices for their implementation, while being prepared to iterate these as participants work in the field. Second, training material should be abundant and, ideally accessible—videos are generally more effective than text, and should ideally be available in multiple relevant languages. Third, it is important to balance the time spent on different aspects of data collection and to recognize that the compilation of even basic metadata is time-consuming and most effective when done when a team is together on site, rather than in retrospect. This could also be accomplished by hiring additional personnel, such as research assistants. Fourth, it is important to identify participants with skills in various areas (such as photography, video production, and especially metadata production, in addition to subject matter expertise) and make sure that they play an important role in sharing their knowledge with fellow cluster members. This both helps data collection go more smoothly and ensures that more participants learn such transferable skills, which is an important goal that can be valuable for students in particular. Finally, the experience with "virtual field visits" suggests that working remotely with existing data can be a fruitful alternative to on-site field collection that is less expensive, less carbon-intensive, and more accessible than international travel would be.

Finally, the future accessibility and findability of FROGBEAR data is limited by the nature of the library Open Collections format. A lack of human resources made it impossible to produce sufficient metadata for all of the material collected by field visits, so some of it was never added to the collection, but could potentially be useful to future researchers (this includes data produced by field visits that does not fit into the formats housed by the library, such as 360° images/video). And it is not straightforward to bulk download the data or a subset of it. Hence it would be useful to collect all of this material, both what is currently in the library repository and what is not, into another repository such as the Internet Archive or Zenodo, with minimal metadata to make the collection as a whole findable and interpretable. This would require some additional labour, but much less than the production of individual repository records would.

The major outputs of the data collection side of the project are the repository of thousands of images and videos available for the use of researchers and teachers, as well as the training that scores of students received. Hopefully both will contribute to the field of East Asian religions and beyond for a generation to come.

References

Government of Canada. (2021, January 20). *Tri-Agency Statement of Principles on Digital Data Management*. https://science.gc.ca/site/science/en/interagency-research-funding/policies-and-guidelines/research-data-management/tri-agency-statement-principles-digital-data-management

UBC Frogbear Project. (2018–2024). *Training Videos*. https://www.youtube.com/playlist?list=PLUkA_ZXM8fg3dTzW_9BewWdxoa3Mc3hjY

University of British Columbia. (2024a). *Database of Religious Sites in East Asia*. From the Ground Up: Buddhism and East Asian Religions. https://frogbear.org/app/#/list

University of British Columbia. (2024b). *FROGBEAR Data Collection*. UIBC Wiki. https://wiki.ubc.ca/Documentation:Library:Circle/FROGBEAR_Data_Collection

University of British Columbia Library. (2024). *Open Collections*. https://open.library.ubc.ca/